



UNIVERSITAT<sub>DE</sub>  
BARCELONA

## Unravelling genetic predisposition to familial breast and ovarian cancer: new susceptibility genes and variant interpretation by *in silico* approaches

Alejandro Moles Fernández



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution 4.0. Spain License.**

**Unravelling genetic predisposition to familial breast and ovarian cancer: new susceptibility genes and variant interpretation by *in silico* approaches**

Alejandro Moles Fernández







# Unravelling genetic predisposition to familial breast and ovarian cancer: new susceptibility genes and variant interpretation by *in silico* approaches

Doctoral thesis submitted by

**Alejandro Moles Fernández**

for the doctoral degree by University of Barcelona (UB)

This work was developed under the supervision of

**Dra. Sara Gutiérrez-Enríquez** and **Dr. Orland Diez Gibert**

at the Vall d'Hebron Institute of Oncology (VHIO)

Biomedicine Doctoral Program

Faculty of Biology, University of Barcelona (UB)

Tutor: **Dr. Alex Sánchez Pla**

**Dra. Sara Gutiérrez-Enríquez**



**Dr. Orland Diez Gibert**



**Dr. Alex Sánchez Pla**



**Alejandro Moles Fernández**



Barcelona, 2021



*A las pacientes*



## ACKNOWLEDGMENTS

Bueno, pues ya está, ea, ya estamos aquí. Con el trabajito que ha costado! He disfrutado mucho este camino durante los últimos años y quiero agradecerse a algunas personas.

En primer lugar, a mis padres, Silvia y Antonio, sois un ejemplo de nobleza y esfuerzo, gracias por vuestro apoyo infinito y amor incondicional. No os imagináis cuanto os quiero.

A mi familia, a mi abuelo Óscar, y a mi abuela Pepi, la alegría hecha persona. A mis abuelos Chelo y Antonio, y a mis tías Toñi y Mari siempre tan generosas. A mis tías Rosa, David, Sandra y Chelo, a Antonio e Isabel y a todas mis primas. En especial a Julia, Sofía y al golfo de mi primo Javi.

He tenido mucha suerte de aprender y disfrutar de la ciencia en el grupo de Oncogenética. Qué difícil ha sido... Pero, al final hemos llegado! Sara y Orland, he aprendido mucho con vosotros y me habéis hecho sentirme muy valorado. Gracias por vuestro esfuerzo y paciencia, creo que poca gente ama la ciencia tanto como vosotros, con tanta curiosidad, y me lo habéis transmitido.

Hacer la tesis ha sido mucho más sencillo con tan buena gente cerca... Gemma Montalban, Laura y Sandra, que pechá de reír nos hemos dado, quina alegría treballar amb vosaltres, y mira que me costó enseñaros malagueño, cullons. Y Joanna y Ester, las últimas en llegar, tan generosas, es un gustazo estar en el laboratorio con vosotras. También a Judith y al laboratorio de la UARPC, esta tesis no habría sido posible sin vosotras.

A la gente del VHIO, con la que he compartido cientos de cervezas y cafés en estos años, Gemma, Marc, Queralt, Carmen, Estefanía, Ana Belén, Olga, Cate, Sara Arce... En especial, quiero agradecer a Marc y Gemma, por el cachondeo y las conversaciones, no me he podido reír más con vosotros, gracias por estar ahí.

A la gente de Barcelona, David, Leire, Aida, Patri, María, Leire R, Xenia, David Mex. Aunque David, colega...no sé si darte las gracias o preguntarte cómo hemos sido capaces de terminar la tesis...que bien nos lo hemos pasado!

A Lucía, y a mis amigos de toda la vida, mi otra familia, Alfonso, Álvaro, Manu, Juanca y Diego, tengo una suerte inmensa de teneros cerca.

A ti, Carmen, por hacer la vida fácil y divertida, solamente siendo tú... esta tesis es también tuya.



## **LIST OF ABBREVIATIONS**

ACMG: American College of Medical Genetics and Genomics

AMP: Association for Molecular Pathology

BC: Breast cancer

BRRM: Bilateral risk reduction mastectomy

CAGI: Critical Assessment of Genome Interpretation

CNV: Copy number variations

ESEs: Exonic splicing enhancers

ESRs: Exonic splicing regulators

ESSs: Exonic splicing silencers

FAP: familial adenomatous polyposis

HBOC: Hereditary breast and ovarian cancer

HSF: Human Splice Finder

IARC: International Agency for Research on Cancer

ISEs: Intronic splicing enhancers

ISRs: Intronic splicing regulators

ISSs: Intronic splicing silencers

LoF: Loss of function

MCC: Matthews correlation coefficient

MES: MaxEntScan

ML: Machine learning

OC: Ovarian cancer

OR: Odds ratio

PJS: Peutz Jeghers syndrome

PRS: Polygenic risk score

PWM: Position weight matrix



RRSO: risk reduction salpingo-oophorectomy

SEOM: Spanish society of medical oncology

SNP: Single nucleotide polymorphism

SnRNPs: Small nuclear ribonucleoproteins

SNVs: Single nucleotide variants

SR proteins: Serine rich proteins

SREs: Splicing regulatory elements

SSF: Splicing Site Finder

VUS: Variants of uncertain significance

WES: whole-exome sequencing

WGS: whole-genome sequencing

**INDEX**



LIST OF ABBREVIATIONS.....	XX
SUMMARY .....	19
INTRODUCTION .....	23
1. Hereditary cancer .....	25
1.1 Epidemiology.....	25
1.2 Genetic predisposition to hereditary breast and ovarian cancer .26	
1.2.1. Clinical criteria for germline genetic analysis.....	30
1.2.2. Gene panel analysis.....	32
1.2.3. Clinical management in carriers of pathogenic variants in HBOC related genes.....	33
1.3. Identification of new susceptibility genes.....	35
2. Variant interpretation .....	37
2.1 Variant classification system .....	38
2.1.1 ACMG/AMP classification guidelines .....	40
2.2 The challenge of interpretation of variants of uncertain significance (VUS).....	42
3. <i>In silico</i> tools usage in RNA splicing and hereditary breast/ovarian cancer .....	44
3.1. Elements of Splicing .....	45
3.2. Spliceogenic variants in genetic diseases.....	48
3.3 Splicing <i>in silico</i> tools.....	50
3.3.1 Splicing <i>in silico</i> tools for specific sequence regions .....	53
HYPOTHESIS.....	53
OBJECTIVES.....	57
RESULTS.....	61
CONTENTS .....	67
SUMMARY OF RESULTS.....	68
REPORT OF IMPACT FACTOR OF THE ARTICLES INCLUDED IN THE THESIS... 70	
Article 1 .....	73
Article 2 .....	87
Article 3 .....	105
Article 4 .....	123

DISCUSSION .....	155
1. <i>In silico</i> tools for spliceogenic variants identification in HBOC genes.	157
1.1 Identification of variants altering donor and acceptor splice sites in HBOC using <i>in silico</i> tools .....	158
1.2 Deep intronic variant identification using <i>in silico</i> tools .....	163
1.3 Importance of SREs balance in the pseudoexon inclusion caused by deep intronic variants.....	165
1.4 Future of splicing variants identification: through a unified <i>in silico</i> pipeline and <i>in vitro</i> RNA sequencing .....	169
2. Adaptation of ACMG guidelines to the <i>ATM</i> gene.....	172
3. Candidate genes identification and validation in a case-control analysis	175
CONCLUSIONS .....	181
BIBLIOGRAPHY.....	185
References.....	187
URL of online resources and tools.....	202
APPENDIX .....	203
Additional publications .....	205
Article 5 .....	205
Article 6 .....	215
Article 7 .....	229
Article 8 .....	241
Article 9 .....	261

## **LIST OF FIGURES AND TABLES**



Figure 1. Absolute risk of breast cancer in carriers of truncating variants, .....	29
Table 1. Risk of breast cancer associated with protein-truncating variants .....	29
Figure 2. Relative risk and allele frequencies of high-, moderate-, and low-risk genetic variants associated to breast and ovarian cancer.....	30
Table 2. SEOM clinical criteria for germline genetic analysis in hereditary breast and ovarian cancer patients.....	31
Figure 3. Variant classification levels .....	39
Table 3. ACMG/AMP evidence codes by category .....	40
Table 4. Number of submitted variants per significance .....	43
Figure 5. Schematic representation of cis (DNA splicing sequences) and trans (protein binding splicing elements). .....	46
Figure 6. Types of splicing variants. ....	50
Table 5. Summary of commonly used in silico splicing tools .....	54
Figure 7. SRE mapping of exonic and intronic regions.....	166
Table 6: In silico splicing analysis proposed pipeline .....	170





## **SUMMARY**



Patients with hereditary breast and ovarian cancer (HBOC) in whom a causative pathogenic variant is not identified after genetic analysis may not benefit from prevention, early detection, or precision treatment measures. This negative or inconclusive results are due, among other causes, to the detection of variants of uncertain significance (VUS). The main objective of this thesis is to increase the capacity of genetic diagnosis of patients with HBOC, by focusing on i) the optimisation in the interpretation of exonic and intronic variants that might affect RNA quality or quantity but remain as variants of uncertain significance (VUS) and ii) the identification of new susceptibility genes for HBOC.

The article included in this thesis, Moles-Fernández et al., 2018 (DOI: 10.3389/fgene.2018.00366) explains an optimization in the identification of potentially spliceogenic variants located near to splicing sites, and provides recommendations to use for analysing donor and acceptor sites. Moreover, the creation or activation of cryptic sites along deep intronic regions could alter splicing causing the inclusion of intronic sequences in RNA. In the article, Moles-Fernández et al., 2021 (DOI: 10.3390/cancers13133341), a framework for the identification of deep intronic spliceogenic is provided, after the performance analysis of SpliceAI *in silico* tool in a dataset of spliceogenic and non-spliceogenic deep intronic variants. In addition, the importance of the splicing regulatory elements balance in the pseudoexon creation is described.

The American College of Medical Genetics (ACMG) variant interpretation guidelines provide general recommendations to classify variants. In the included article Feliubadalò et al., 2021 (DOI: 10.1093/clinchem/hvaa250), ACMG guidelines were adapted to *ATM* gene. We focused on *in silico* splicing evidence (PP3/BP4). After reclassification of variants following the adapted guidelines, a reduction of VUS was obtained.

On the other hand, in patients without pathogenic variants identified in HBOC related genes, the phenotype could be due to deleterious variants in genes still not known associated with the disease. For this reason, in Moles-Fernández et al., (article *in preparation*), the aim was to identify candidate genes through exomes and extended panel analysis and validate their risk association by performing a case-control study. The significant identification of loss-of-function variants in *ALKBH3*, *BLM*, *CAMKK1*, *FANCD2*, *FANCM*, *NEIL3*, *PER1*, *RBL1*, *RECQL4*, *WRN* and *XRCC4* genes in patients with HBOC suggests that they might be breast/ovarian cancer susceptibility genes.

## **INTRODUCTION**



# 1. Hereditary cancer

## 1.1 Epidemiology

According to estimates from the World Health Organization (WHO) in 2019, cancer is the first or second leading cause of death before the age of 70 years in 112 of 183 countries and ranks third or fourth in a further 23 countries.

Cancer disease occurring in an individual with a family history of the disease is known as “familial” cancer. The term “hereditary” is used to describe families in which there is a higher-than-normal occurrence of certain types of cancer, caused by pathogenic variants in certain genes passed from parents to children with a known genetic pattern of inheritance. When cancer occurs in an individual without a family history is often referred to as “sporadic”.

Several cancer hereditary syndromes have been described among last decades (Ngeow and Eng, 2016) related to affected tissues. Hereditary breast and ovarian cancer syndrome (HBOC), hereditary colorectal cancer (Lynch Syndrome), familial adenomatous polyposis (FAP), or hereditary prostate cancer among others, are examples of the most common hereditary cancer syndromes. In addition, a few hundreds of rare and extremely rare syndromes have been described, such as Li-Fraumeni (Malkin, 2011), Cowden syndrome (Pilarski et al., 2013), Fanconi Anaemia (Tischkowitz and Hodgson, 2003) and xeroderma pigmentosum (Berneburg and Lehmann, 2001).

Overall, hereditary syndromes could collectively explain approximately 5-10% of all cancer cases (Ngeow and Eng, 2016). They are usually characterized by earlier ages of diagnosis, multiple incidences of cancer in an individual (or bilaterality), and family history. Hereditary breast and ovarian cancer (HBOC) is a syndrome that involves increased predisposition primarily to breast cancer (BC) and/or to ovarian cancer (OC). Current estimates indicate that



approximately 10-15% of breast and ovarian cancer cases can be explained by inherited deleterious genetic variants in high and moderate penetrance genes (Samadder et al., 2019; González-Santiago et al., 2020). This percentage comprises a high number of patients and families, since female breast cancer is the leading cause of global cancer incidence in 2020, with an estimated 2.3 million new cases, namely 11.7% of all cancer cases. Similarly, ovarian cancer accounts for 3.4% of female cancer incidence and 4.7% of female cancer deaths (Sung et al., 2021).

### **1.2 Genetic predisposition to hereditary breast and ovarian cancer**

Hereditary cancers are characterized by the occurrence of germline pathogenic variants in specific genes, such as *BRCA1*, *MLH1* or *TP53* (associated with breast cancer, colon cancer, and Li-Fraumeni syndrome, respectively). Carriers of a germline alteration in a cancer predisposition gene have a higher risk of developing certain tumours throughout life and usually at a younger age than in the general population (Sociedad Española de Oncología Médica, 2019).

Approximately a hundred genes for hereditary predisposition to cancer associated with a large number of neoplasms have been described in the literature (Sociedad Española de Oncología Médica, 2019). Moreover, pathogenic variants in specific genes have been associated with susceptibility to more than one cancer hereditary disease (Bonadona et al., 2011; Nyberg et al., 2020).

Focusing on the HBOC syndrome, multiple genes have been associated, being the most important *BRCA1* and *BRCA2* (*BRCA1/2*) (Hall et al., 1990; Miki et al., 1994; Wooster et al., 1994). Pathogenic variants in these genes, involved in homologous recombination repair, explain approximately 10-25% of all hereditary breast/ovarian cancer cases (Petrucci et al., 2016). More than

65,000 unique variants have been described in these genes (BRCAExchange, accessed July 2021) in multiple populations and different ethnicities. However, variants with increased incidence, known as recurrent and founder variants, have been described in specific populations (Díez et al., 2003; Ferla et al., 2007; Fachal et al., 2014). It has been estimated that women carrying pathogenic *BRCA1* and *BRCA2* variants have a cumulative risk by 80 years of age of developing breast cancer of 55-72% and 69% and that of developing ovarian cancer of 44% and 17%, respectively (Kuchenbaecker et al., 2017; Dorling et al., 2021). Deleterious *BRCA1/2* variants also increase the risk of prostate and pancreatic cancer, primarily in individuals with a *BRCA2* pathogenic variant (Mersch et al., 2015).

Other high penetrance genes associated with rare hereditary syndromes also cause breast cancers, among others. The *TP53* gene has been implicated in hereditary breast cancer as part of the Li-Fraumeni syndrome (Malkin, 1993), and accounts for a small proportion of breast cancer patients diagnosed before 30 years of age (Woodward et al., 2021). The *PTEN* gene has been identified as the causal gene in Cowden syndrome, in which early-onset breast cancer is associated with a variety of other features, including hamartomas of the skin and mucous membranes, thyroid adenomas, colonic polyps (including juvenile polyps), and craneomegaly (Liaw et al., 1997). *STK11* is associated with the dominantly inherited condition Peutz-Jeghers syndrome, characterized by benign polyps throughout the gastrointestinal tract and mucocutaneous pigmentation (particularly on the lips) and confer a cumulative risk of breast and gynaecological cancers (van Lier et al., 2011). Additionally, pathogenic variants in the *CDH1* gene cause the hereditary diffuse gastric and lobular breast cancer, a dominantly inherited condition that confers to a carrier women a 40–60% lifetime risk of breast cancer (Fitzgerald et al., 2010).

The spectrum of HBOC associated genes has been expanded during the last 20 years including a large number of genes with a critical role in homologous recombination repair (Hoang and Gilks, 2018). Among these genes, *PALB2*

## INTRODUCTION

(BRCA2 interactor), considered a breast cancer high-risk susceptibility gene (Rahman et al., 2007), shows 3834 unique variants entry in ClinVar database (accessed July 2021). Various studies indicate that odds ratio (OR) of *PALB2* deleterious mutations for breast cancer was comparable to that of *BRCA2* pathogenic variants (Dorling et al., 2021; Hu et al., 2021). Genes as *CHEK2*, *ATM*, *BRIP1*, *RAD51C*, and *RAD51D* are related to moderate risks of developing breast/ovarian cancer (Meijers-Heijboer et al., 2002; Renwick et al., 2006; Seal et al., 2006; Meindl et al., 2010; Loveday et al., 2011) explaining 4.6% of cases (Tung et al., 2016). Women with pathogenic variants in *ATM* or *CHEK2* face a breast cancer cumulative risk at 80 years of 20 to 30% (Dorling et al., 2021) while *RAD51C* and *RAD51D* generate an ovarian cancer risk of 11% and 13% (Yang et al., 2020). Also, some evidence of *BARD1* as a breast cancer susceptibility gene have been collected during the last decade (Alenezi et al., 2020; Dorling et al., 2021).

Recently, population and family-based studies using a high number of cases and healthy controls have been performed to determine the risk of the associated genes to breast cancer (Dorling et al., 2021; Hu et al., 2021) (Figure 1). Protein-truncating variants in five genes (*ATM*, *BRCA1*, *BRCA2*, *CHEK2*, and *PALB2*) were associated with a significant risk of breast cancer overall ( $P < 0.0001$ ), and in four other genes (*BARD1*, *RAD51C*, *RAD51D*, and *TP53*), albeit with a  $p$  of less than 0.05 in most of these genes, the odds ratio differed according to breast cancer subtype (Dorling et al., 2021).

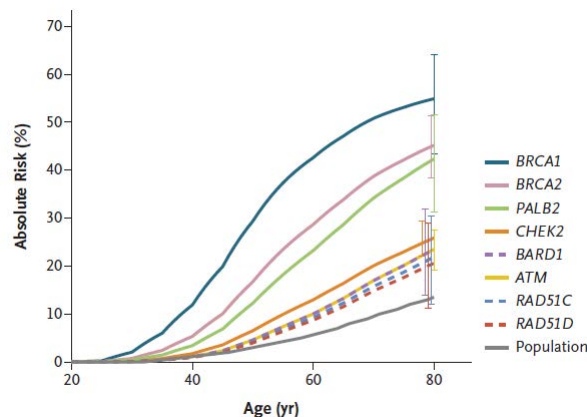


Figure 1. Absolute risk of breast cancer in carriers of truncating variants of *BRCA1*, *BRCA2*, *PALB2*, *CHEK2*, *BARD1*, *ATM*, *RAD51C*, and *RAD51D* compared with the general population. Taken from Dorling et al., 2021.

These large studies (Dorling et al., 2021; Hu et al., 2021), defined the risk of the susceptibility for breast cancer and evidence the stratification between high penetrance and moderate penetrance risk genes, and discard other candidates that did not show association with the disease (e.g., *NBN*). Nevertheless, despite the size of these studies, the evidence of an association with breast cancer risk for several of the genes that were analysed (e.g., *FANCM*, *MSH6*, and *NF1*) remains equivocal (Dorling et al., 2021) (Table 1).

Table 1. Risk of breast cancer associated with protein-truncating variants in the population-based study reported in Dorling et al., 2021. Adapted from Dorling et al., 2021.

Gene	Breast Cancer Patients (n=48,826)	Controls (n=50,703)	Odds Ratio (95% CI)	<i>p</i>
<i>ATM</i>	294	150	2.10 (1.71–2.57)	9.2×10 <sup>-13</sup>
<i>BARD1</i>	62	32	2.09 (1.35–3.23)	0.00098
<i>BRCA1</i>	515	58	10.57 (8.02–13.93)	1.1×10 <sup>-62</sup>
<i>BRCA2</i>	754	135	5.85 (4.85–7.06)	2.2×10 <sup>-75</sup>
<i>CHEK2</i>	704	315	2.54 (2.21–2.91)	3.1×10 <sup>-39</sup>
<i>FANCM*</i>	302	300	1.06 (0.90–1.26)	0.48
<i>MSH6</i>	39	23	1.96 (1.15–3.33)	0.013
<i>NF1</i>	31	17	1.76 (0.96–3.21)	0.068
<i>PALB2</i>	274	55	5.02 (3.73–6.76)	1.6×10 <sup>-26</sup>
<i>RAD51C</i>	54	26	1.93 (1.20–3.11)	0.007
<i>RAD51D</i>	51	25	1.80 (1.11–2.93)	0.018

\*Over-represented in ER-negative breast cancer (Dorling et al., 2021).

On the other hand, large genome-wide association studies have been performed for longer than a decade to identify associations between common variants and disease. These approaches have led to the robust identification of

## INTRODUCTION

more than 300 single nucleotide polymorphisms (SNPs). These common low-risk alleles only confer a small risk by themselves, but when combined in a polygenic risk score (PRS) they provide a more significant risk estimate. It has been calculated that these SNPs currently explain up to 28% of the familial risk of breast cancer (Woodward et al., 2021). Similarly, it has been proposed that these SNPs can modify the risk of pathogenic variant carriers (Barnes et al., 2020; Gallagher et al., 2020; Yanes et al., 2020).

In summary, the genetic landscape of hereditary breast and ovarian cancer is heterogeneously composed of various high, moderate, and low susceptibility genes (Figure 2). However, approximately half of hereditary breast-ovarian cancer patients remain genetically undiagnosed (Couch et al., 2014).

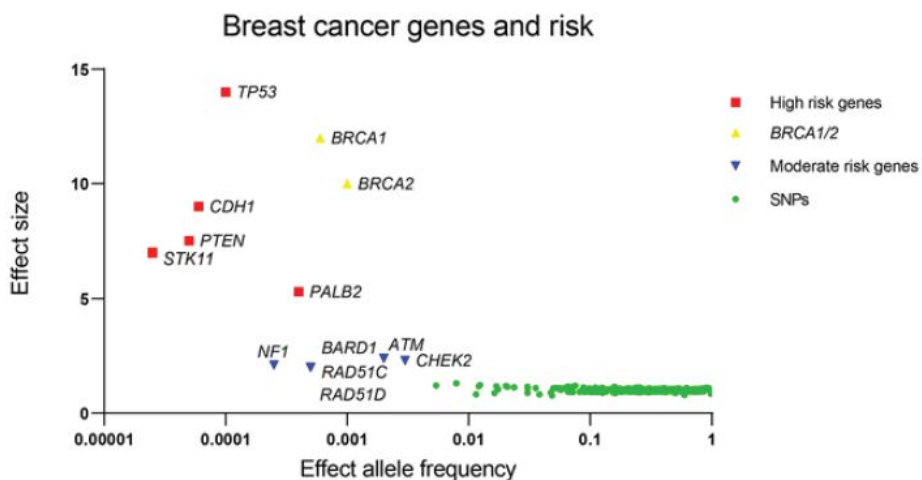


Figure 2. Relative risk and allele frequencies of high-, moderate-, and low-risk genetic variants associated to breast and ovarian cancer. Taken from Woodward et al., 2021.

### 1.2.1. Clinical criteria for germline genetic analysis

Different European guidelines containing recommendations for *BRCA1/2* testing have been published in the last years. These criteria are associated with a probability of  $\geq 10\%$  pathogenic variant detection. Clinical criteria for genetic testing differ between guidelines, but all of them are based on clinical risk

factors such as age, hormone receptor status, ancestry with founder mutations, and personal or family history of cancer (Table 2) (González-Santiago et al., 2020).

*Table 2. SEOM clinical criteria for germline genetic analysis in hereditary breast and ovarian cancer patients. Adapted from Gonzalez-Santiago et al., 2020.*

<b>Selection criteria for germline testing</b>
<b>Regardless of family history:</b>
Women with synchronous or metachronous breast and ovarian cancer
Breast cancer ≤ 40 years
Bilateral breast cancer (the first diagnosed ≤ 50 years)
Triple-negative breast cancer ≤ 60 years
High-grade epithelial non-mucinous ovarian cancer (or fallopian tube or primary peritoneal cancer)
Ancestry with founder mutations
BRCA somatic mutation detected in any tumour type with a tumour allele frequency > 30% (if it is known)
Metastatic HER2-negative breast cancer patients eligible to consider PARP inhibitor therapy
<b>2 or more first degree relatives with any combination of the following high-risk features:</b>
Bilateral breast cancer + another breast cancer < 60 years
Breast cancer < 50 years and prostate or pancreatic cancer < 60 years
Male breast cancer
Breast and ovarian cancer
Two cases of breast cancer diagnosed before age 50 years
<b>3 or more direct relatives with breast cancer (at least one premenopausal) and/or ovarian cancer and/or, pancreatic cancer or high Gleason (≥ 7) prostate cancer</b>

Most of these guidelines are based predominantly on the probability of carrying pathogenic variants in *BRCA1* or *BRCA2*. Thus, the sensitivity of these criteria to

## INTRODUCTION

identify other pathogenic alterations in different high or moderate-risk genes is limited. Recent research supports *BRCA1* and *BRCA2* testing in a broader range of individuals, if not in every breast cancer patient. This recommendation is based on the findings of studies that conclude that the traditional approach may miss an elevated number of pathogenic variant carriers (Beitsch et al., 2019; González-Santiago et al., 2020).

New criteria for germline testing, regardless of family history, are arising thanks to improvements in massive tumour sequencing techniques, as well as in predicting response to new therapeutic agents. Following detection of a somatic mutation in a cancer predisposition gene with high tumour allele frequency, it is advisable to rule out or verify the existence of a germline pathogenic variant considering possible implications in genetic counselling.

The Spanish Society of Medical Oncology (SEOM) recommends genetic risk evaluation and genetic counselling (before and after germline testing) for patients who are at high risk of harbouring a pathogenic variant in one of the breast/ovarian cancer predisposition genes. Genetic counselling is a process that guarantees a discussion about the benefits and limitations of genetic testing, including information about cancer risk, recommendations for early detection and risk reduction interventions, as well as advice regarding reproductive options, and support for psychological well-being (González-Santiago et al., 2020).

### **1.2.2. Gene panel analysis**

To identify pathogenic variants in susceptibility genes, customized massively parallel sequencing panels are widely used in the clinical practice, allowing the analysis of high and moderate penetrance associated risk genes in germline DNA of affected patients (Bonache et al., 2018; Hauke et al., 2018; Tsaousis et al., 2019).

However, there are differences between the genes included in the panels across hospital laboratories or commercial companies. In fact, some centres and companies offer wide panels including genes that have never been linked convincingly to breast/ovarian cancer and other hereditary cancers. Although multi-gene panels differ among testing laboratories, they commonly include the high penetrance genes *BRCA1*, *BRCA2* and *PALB2*, other high-penetrance genes associated with rare genetic syndromes (*TP53*, *PTEN*, *CDH1*, and *STK11*), and genes associated with moderate risks such as *CHEK2*, *ATM*, *RAD51C* and *RAD51D* (Woodward et al., 2021). The mismatch repair genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*), may also be included in an opportunistic detection approach (Feliubadaló et al., 2019).

In Spain, SEOM recommended to include *BRCA1* and *BRCA2*, which are the most common mutated susceptibility genes in breast/ovarian tumours, followed by *PALB2* (in BC) and genes with pathogenic variants that confer moderate penetrance cancer risk, such as *ATM* and *CHEK2* (in BC) and *BRIP1*, *RAD51C*, *RAD51D*, *MLH1*, *MSH2*, and *MSH6* (in OC) (González-Santiago et al., 2020). Clinical validity for *BRCA1*, *BRCA2* and *PALB2* (in BC/OC), and *BRIP1*, *RAD51C*, *RAD51D*, *MLH1*, *MSH2*, and *MSH6* (in OC) have been established with subsequent surveillance and preventive clinical options (Domchek and Robson, 2019; González-Santiago et al., 2020).

### **1.2.3. Clinical management in carriers of pathogenic variants in HBOC related genes**

The identification of pathogenic variants in genes associated with susceptibility to the disease allows the genetic accurate risk assessment and medical management of patients and families.



## INTRODUCTION

Early detection, risk-reducing surgery or personalized therapy are clinical strategies available to carriers of deleterious variants in *BRCA1* or *BRCA2* (Pashayan et al., 2020; Pujol et al., 2021):

- **Early detection strategies.** Annual mammography and annual breast magnetic resonance imaging in women with deleterious variants in *BRCA1/2* are recommended (Warner, 2018).
- **Risk-reducing surgery.** Bilateral risk reduction mastectomy (BRRM) decreases the occurrence of breast cancer in women with a moderate-high risk by 90% without a decrease in all-cause mortality (Li et al., 2016). Also, bilateral risk reduction salpingo-oophorectomy (RRSO) demonstrated a risk reduction of OC, fallopian tube cancer, and primary peritoneal cancer of ~80% in women with *BRCA1/2* loss-of-function variants. RRSO in carriers of moderately penetrant pathogenic alleles of *BRIP1*, *RAD51C* or *RAD51D* should be contemplated on a case-by-case basis, and is also an option to be considered in carriers of pathogenic variants in Lynch syndrome genes (González-Santiago et al., 2020).
- **Chemoprevention.** Preventive treatments such as tamoxifen are an option for female *BRCA1/2* carriers who do not want to undergo BRRM (González-Santiago et al., 2020). Oral contraceptives in *BRCA1/2* carriers can reduce the risk of OC; however, there are discrepancies in its recommendations of use (Huber et al., 2020).
- **Treatment strategies.** Along last years, poly (ADP-ribose) polymerase (PARP) inhibitors agents have shown their benefits in ovarian cancer patients carrying *BRCA1/2* deleterious variants (Ledermann, 2016; Cook and Tinker, 2019; Mirza et al., 2020). In addition, these molecules are prime candidates for the treatment of breast cancers associated with germline or acquired mutations of *BRCA1/2* and potentially also *PALB2* given their functional roles in homologous recombination pathways (Woodward 2021). The identification of a germline *BRCA1/2*

pathogenic variant in the oncology setting has important therapeutic implications, with PARP inhibition of advanced breast cancers being associated with increased progression-free survival over standard care (Robson et al., 2017; Litton et al., 2018). Given their use now in a maintenance setting in ovarian cancer, they would be employed after primary treatment even for earlier-stage breast cancer (Woodward et al., 2021). The use of this therapy in patients has an enormous potential and interest and more than 350 clinical trials involving this molecule are active (Clinicaltrials.gov, accessed July 2021). To date, PARP inhibitors have been approved for the treatment of germline *BRCA1/2*-mutated ovarian, breast, prostate, and pancreatic cancers (Lynparza | European Medicines Agency. <https://www.ema.europa.eu/en/medicines/human/EPAR/lynparza>. Accessed July 2021).

### **1.3. Identification of new susceptibility genes**

Susceptibility gene identification strategies have been evolved since 90s decade. Initially, multiple-case families were studied to identify high-risk susceptibility genes by linkage analysis, detecting markers that co-segregated with the disease. This approach led to the identification of *BRCA1* and *BRCA2* genes (*BRCA1/2*) (Hall et al., 1990; Miki et al., 1994; Wooster et al., 1994). In addition, multiple-case families were also used to know the role of high penetrance genes in rare hereditary syndromes causing breast or ovarian cancers, among others, such as *TP53* (Malkin, 1993), *PTEN* (Liaw et al., 1997), *STK11* (van Lier et al., 2011) and *CDH1* (Fitzgerald et al., 2010).

In the last two decades, following a case-control strategy and segregation analysis strategies, *PALB2* high-risk susceptibility gene was identified (Rahman et al., 2007). Moreover, a higher proportion of truncating variants in affected

## INTRODUCTION

individuals compared to healthy controls were identified in *CHEK2*, *ATM*, *BRIP1*, *RAD51C*, *RAD51D*, and *BARD1* genes, associating these genes with low to moderate risks of developing breast/ovarian cancer (Meijers-Heijboer et al., 2002; Renwick et al., 2006; Seal et al., 2006; Meindl et al., 2010; Loveday et al., 2011; Dorling et al., 2021).

However, a part of the missing heritability in this disease may be due to new genes related to the susceptibility to HBOC still unknown. The advent of massively parallel sequencing has led to timely testing of multiple genes using panels, whole-exome sequencing (WES), or whole-genome sequencing (WGS). Interestingly, WES has become a common approach to identify rare variants by performing a staged study starting with sequencing of cohorts of small number of cases with strong familial aggregation of hereditary cancer, highlighting potential candidate genes (Rotunno et al., 2020). In addition, functional analysis or mutational tumour signature indicating the relevance of candidate genes in developing the disease are also valuable approaches to identify potential risk genes (Polak et al., 2017; Hernández et al., 2018). Following this approach, a few genes have been identified so far as potential candidates in colorectal cancer and hereditary breast and ovarian cancer (Te Paske et al., 2020); Rotunno et al., 2020; Subramanian et al., 2020). However, these genes have to be validated in large case-control studies to clarify if they are associated with the disease.

Published studies identifying candidate genes comprehend both the identification and the validation using a different cohort of patients, and sequencing healthy controls or using public control databases. For example, germline *RBBP8* variants have been recently associated with early-onset breast cancer, firstly identifying potentially deleterious variants in a small cohort of patients, and secondly validating its association by sequencing a large cohort of patients and by functional assays (Zarrizi et al., 2020). In addition, *RECQL5* gene was highlighted as a potentially related gene by identifying a deleterious variant

by WES in an HBOC family and observing an enrichment of deleterious variants in affected patients after a comparison with healthy controls.

These genes are an example of how the identification of new candidate genes is possible using massively parallel approaches in HBOC patients. In this thesis, the author describes a study consisting of the identification of candidate genes and their subsequent validation with an analysis of cases and controls. First results are presented since the phase of validation and calculation of associated risk is in process.

## **2. Variant interpretation**

Most of the diagnostic panels performed in breast/ovarian cancer patients focus on the sequencing of coding regions and exon-intron boundaries. Variants identified in DNA could be single nucleotide variants (SNVs), insertions, deletions, duplications, indels, or inversions. These types of variants can also affect RNA processing. Moreover, copy number variations (CNV) or even Alu insertions can be identified using massively parallel sequencing panels (Kerkhof et al., 2017; Qian et al., 2017). In affected patients, the genetic test can result in the identification of novel or already known benign or pathogenic variants or variants that we can still not clearly define and classify, termed variants of uncertain significance (VUS), or no identification of variants. The growing accumulation of genetic data generates a high amount of percentages of VUS (Lumish et al., 2017), and this is especially true in oncological diseases, for which gene-panel sequencing is often required in large groups of cancer patients (Federici and Soddu, 2020).

## 2.1 Variant classification system

The variant classification has evolved since Cotton et al. (Cotton and Scriver, 1998), more than 20 years ago, delineated several types of evidence that could prove helpful in understanding a missense variant role in disease causation, including segregation analysis, nature of the amino acid substitution and functional assays (Mester and Pesaran, 2019; Harrison et al., 2021).

Major steps forward in harmonizing the approach to variant classification occurred with publications from the American College of Medical Genetics and Genomics (ACMG) in 2000 (Kazazian and Boehm, 2000) and 2008 (Richards et al., 2008). However, these guidelines did not define what degree of certainty was required to classify a variant as disease-causing or harmless, how much weight should be assigned to different types of evidence, or how to combine different pieces of evidence to reach a classification. The first effort toward combining evidence types was published in 2004 by Goldgar et al., whose multifactorial likelihood based-model incorporated several clinical data points together with conservation and functional data to arrive at likelihood ratios that the authors deemed high enough to support (1000:1) or refute (100:1) causality for the *BRCA1* and *BRCA2* hereditary breast and ovarian cancer susceptibility genes (Goldgar et al., 2004; Mester and Pesaran, 2019; Harrison et al., 2021).

A step forward in unifying variant classification terminology occurred in 2008 with the publication of a 5-tier classification system developed by the International Agency for Research on Cancer (IARC) Unclassified Genetic Variants Working Group (Plon et al., 2008). The IARC stated that it is essential to discriminate between variants with scarce information (class 3) and variants with strong but not undeniable evidence of association or not with the disease (classes 4, “likely pathogenic” and 2, “likely not pathogenic”, respectively) (Figure 3) (Plon et al., 2008; Moghadasi et al., 2016).



Figure 3. Variant classification levels. Taken from Mester and Pesaran, 2019.

At that point, genome diagnostic laboratories and researchers had broadly accepted the use of a standard, 5-tier scheme for classifying variants: benign, likely benign, variant of uncertain significance (VUS), likely pathogenic, and pathogenic also described as class 1 through 5, respectively. Although this system standardized the naming of the different variant classifications, it did not cover what evidence would be required to get a variant classified in each category. This issue was tackled by the ACMG/AMP classification guidelines, published in 2015 (Richards et al., 2015), giving detailed recommendations on how to build up the evidence to classify a variant into one of these five categories. Several improvements to the ACMG/AMP guidelines have been published since, as some of the original specifications were open to different interpretations. Additional modifications were sometimes required to apply the guidelines for certain genes and/or diseases (Harrison et al., 2021).

In addition, other classification guidelines exist following a 5-tier scheme, such as ENIGMA (Evidence-based Network for the Interpretation of Germline Mutant Alleles), an international research consortium focused on developing and applying methods to determine the clinical significance in breast-ovarian cancer predisposition genes (Spurdle, et al., 2012). ENIGMA has developed variant classification criteria that utilize both quantitative (multifactorial based-model) and qualitative (rules-based) methods to assess the clinical significance of variants in *BRCA1* and *BRCA2* (<http://enigmaconsortium.org/>) (Parsons et al., 2019). InSiGHT (International Society for Gastrointestinal Hereditary Tumors) also applied quantitative strategies to classify variants in the mismatch repair genes associated with Lynch syndrome along with a qualitative system able to

be utilized in the absence of data supporting a quantitative, multifactorial analysis (Plazzer et al., 2013; Thompson et al., 2014).

### 2.1.1 ACMG/AMP classification guidelines

To standardize variant classification, the 2015 ACMG/AMP guideline outlines 28 types of evidence, or criteria, that are encountered during variant assessment (Richards et al., 2015). Each criterion is assigned a direction, either pathogenic (P) or benign (B), and a relative strength: stand-alone (A), very strong (VS), strong (S), moderate (M), or supporting (P). The combination of the direction and relative strength creates an evidence code that refers to a specific evidence criterion. For example, pathogenicity (P) of moderate (M) strength and a specific number assigned to the evidence. Each evidence code corresponds to a single criterion (Harrison et al., 2021) (Table 3).

To assess and subsequently classify a variant, the variant curator must determine which criteria are applicable based on all available evidence. All the criteria are grouped into categories of evidence to aid in their assessment when they are similar or use the same source of data (Table 3). In addition to considering gene-level evidence, the 2015 guideline notes that with professional judgment, some criteria listed at a certain strength can be moved to a stronger or weaker level of evidence (Harrison et al., 2021).

Table 3. ACMG/AMP evidence codes by category (adapted from Harrison et al., 2021).

Evidence	Benign	Pathogenic
<b>Population data</b>	Variant is too common for disorder (BA1, BS1)	Case control studies, multiple affected probands (PS4)

	Variant identified in unaffected individuals (BS2)	Variant is absent in population databases (PM2)
<b>Variant type, location, and predictive data</b>	Missense variant in a gene where LOF is mechanism (BP1)	LOF or missense variant in gene with relevant mechanism (PVS1, PP2)
	In-frame indel in repeat region (BP3)	Different nucleotide changes in same codon (PS1/PM5)
	Benign <i>in silico</i> prediction (BP4)	Pathogenic <i>in silico</i> prediction (PP3)
	Synonymous variant (BP7)	Variant located in functional domain (PM1)
		Variant changes protein length (PM4)
<b>Functional data</b>	Functional studies demonstrating no effect on gene product (BS3)	Functional studies supporting pathogenicity (PS3)
<b>Case-level data</b>	Lack of segregation with disease (BS4)	Co-segregation with disease (PP1)
	Variant observed with another pathogenic variant in the same or different gene (BP2, BP5)	Variant in <i>trans</i> with pathogenic variant in recessive disorder (PM3)
		De novo observation (PS2, PM6)
		Phenotype consistent with disease (PP4)

These guidelines are intended to be generic, and thus some evidence codes will not be relevant for variant curation for a specific gene. The ClinGen consortium (Rehm et al., 2015), <https://www.clinicalgenome.org/> has engaged with expert



groups to develop adaptations of the guidelines to specify which rule codes and strengths are appropriate for a specific gene-disease relationship and to provide guidance on the phenotypic features that are most predictive of variant pathogenicity (Rivera-Muñoz et al., 2018). To date, adaptations of the ACMG/AMP criteria have been completed for hereditary cancer or rare diseases related-genes such as *PTEN* (Mester et al., 2018), *CDH1* (Lee et al., 2018), *TP53* (Fortuno et al., 2021), *LDLR* (<https://www.medrxiv.org/content/10.1101/2021.03.17.21252755v1>) or Rett and Angelman-like disorders related genes ([https://clinicalgenome.org/site/assets/files/6050/clingen\\_rettas\\_acmg\\_specifications\\_v1.pdf](https://clinicalgenome.org/site/assets/files/6050/clingen_rettas_acmg_specifications_v1.pdf)) while other gene-adaptations are in development.

Other hereditary cancer-related genes, such as *ATM*, among others, remained with no specific classification guidelines. The adaptation of ACMG guidelines to specific hereditary cancer or rare diseases related genes will provide a comprehensive framework to diminish the VUS number and optimize the variant classification.

The thesis includes the collaborative work, with the participation of the author, of the Spanish *ATM* hereditary cancer variant interpretation Working Group, to develop a specific guide for *ATM* gene.

## **2.2 The challenge of interpretation of variants of uncertain significance (VUS)**

The type of VUS identified during routine genetic testing can include synonymous, missense, small in-frame insertions/deletions, and intronic variants, for which sequence information alone is not sufficient to infer a cancer-associated risk. These variants have unknown functional effects on

proteins and cannot be clinically classified as either “Pathogenic” or “Benign”. As a result, VUS carriers and their family members cannot benefit from risk assessment measures and personalized cancer screening programs.

In ClinVar database (<https://www.ncbi.nlm.nih.gov/clinvar>), the genes with the highest number of submitted variants include the most relevant tumour risk genes, such as *BRCA1* and *BRCA2*, *APC*, mismatch repair genes and *ATM*. Moreover, many ClinVar submitted variants have conflicting interpretations, and a large number of total variants in the database are VUS (Table 4) (Federici and Soddu, 2020). Concerning HBOC-related genes, up to 35% of *BRCA1/2* variants submitted in ClinVar are classified as VUS or have conflicting interpretations (ClinVar database, accessed July 2021).

*Table 4. Number of submitted variants per significance (of all variants included in ClinVar) (adapted from Federici et al., 2020).*

Submission significance	Variants	Genes
<b>Uncertain significance</b>	266,759	13,346
<b>Likely benign</b>	203,141	9,515
<b>Benign</b>	128,364	14,810
<b>Pathogenic</b>	91,322	9,998
<b>Likely pathogenic</b>	41,404	4,198

VUS are difficult to classify for several reasons (Federici and Soddu, 2020) such as: i) lack of sufficient population-based statistical evidence, ii) scarcity of functional evidence, and iii) different evaluations by clinicians and researchers. In the first case, VUS might be not so rare, but found in many different pathological conditions and population subgroups, impeding appropriate statistical evaluations and classifications. The second reason is mainly due to the nature of the variant itself: VUS are mainly missense or synonymous substitutions, substitutions of biochemically similar residues, intronic SNVs, or in-frame insertions/deletions. They may be found in non-coding regions, at less

conserved residues, at splicing boundaries, or in less functionally relevant domains compared to true pathogenic variants. Thus, the impact of such VUS on the proteins and their functions is more difficult to uncover than nonsense variants. The third reason is due to different approaches taken by scientists and clinicians for classifying variants. Medical genetic counsellors mainly consider pathogenic variants with documented involvement in the disease. Furthermore, different laboratories do not necessarily adopt the same standardized reporting format. These divergent approaches create a gap in knowledge and make VUS challenging to use, overlooking potentially relevant information for the disease (Federici and Soddu, 2020).

Moreover, it has been noted that hereditary cancer genes are highly susceptible to splicing variants (Rhine et al., 2018). Exonic and intronic VUS in HBOC associated risk genes can affect splicing, altering the RNA and leading to protein defects, similar to truncating variants (Wai et al., 2020).

### **3. *In silico* tools usage in RNA splicing and hereditary breast/ovarian cancer**

In eukaryotic organisms, genes are organized in coding regions (exons) separated by non-coding DNA (introns). The process by which introns are excised from the pre-mRNA is named “splicing.” This process is dependent on the presence and interaction between the called *cis* and *trans* elements. The *cis*-elements are the conserved DNA sequences that define exons, introns, and other regulatory sequences necessary for proper splicing. Spliceosome proteins together with small nuclear RNAs (snRNAs), splicing repressors, and activators

recognize the conserved DNA *cis*-splicing elements and are called *trans*-acting elements (Anna and Monika, 2018; Ule and Blencowe, 2019).

The splicing process is performed in two steps. The first step is the recognition of the splicing sites at intron/exon junctions, and the second one is the intron removal and exon ends joining. During the splicing process, different complexes between the pre-mRNA and spliceosome are formed (Anna and Monika, 2018).

### **3.1. Elements of Splicing**

#### **i) *Cis*-elements of splicing**

The *cis*-acting core consensus sequences include: (i) the splice sites evolutionarily conserved defined by GU at 1 and 2 of the 5' donor splice site (DS) and AG at 1 and 2 of the 3' acceptor splice site (AS); (ii) intronic and exonic nucleotides adjacent to these invariable nucleotides also highly conserved: CAG/GUAAGU in donor sites and NYAG/G in acceptor sites; (iii) the polypyrimidine tract preceding the 3' splicing site, and (iv) the branch point, located anywhere from 18 to 40 nucleotides upstream from the 3' end of an intron (Ohno et al., 2018). Other *cis*-splicing regulatory elements (SREs) modelling splicing are classified as exonic splicing enhancers (ESEs) or silencers (ESSs) when they serve as promoters or inhibitors of exon inclusion, and as intronic splicing enhancers (ISEs) or silencers (ISSs) when they enhance or inhibit the use of adjacent splice sites or exons from an intronic location (Wang and Burge, 2008), helping to the correct exon inclusion and intron exclusion. All these elements shape a recognizable landscape to include a specific region in mature mRNA. The high fidelity of splicing is critically dependent on the recognition of the *cis*-acting pre-mRNA sequences (Fig. 5) (Dufner-Almeida et al., 2019).

## INTRODUCTION

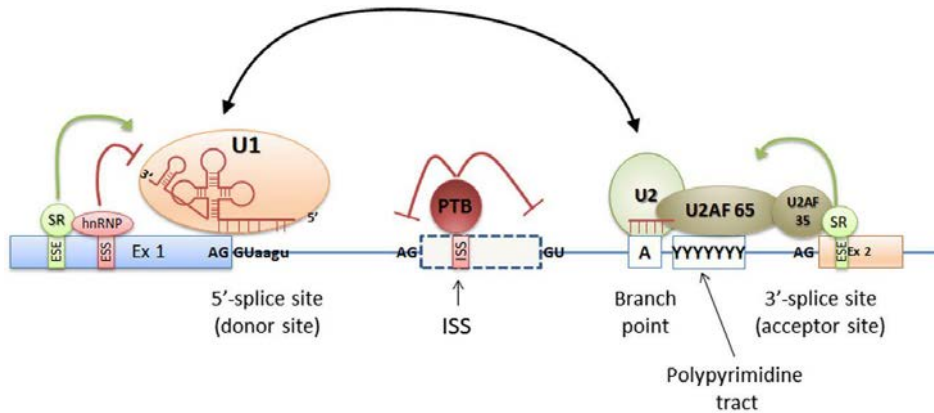


Figure 5. Schematic representation of *cis* (DNA splicing sequences) and *trans* (protein binding splicing elements). ESE (exonic splicing enhancer); ESS (exonic splicing silencer); ISS (intronic splicing silencer); SR (Serine rich protein). Taken from Anna and Monika, 2018.

### - Splicing sites

These sites are recognized multiple times during spliceosome assembly, and introns are removed from the primary transcripts by cleavage at the splice sites. In most cases (98.7%), the exon-intron boundary sequences contain GT and AG motifs at the 5' and 3' ends of the intron, respectively. Non-canonical GC-AG and AT-AC sequences at the splice sites occur in 0.56 and 0.09% of the splice site pairs, respectively (Anna and Monika, 2018). Although the classic splicing motifs are typically essential for splicing, it is not generally sufficient for accurate splice-junction definition.

### - Branch point and polypyrimidine tract

Branch point and polypyrimidine tract intronic sequences bind specific proteins involved in the formation of splicing complexes (Figure 5). The branch point motif, that might be localized between -9 and -400 bp upstream from the acceptor site with the consensus sequence yUnAy in humans, is essential for early spliceosome complex formation through the formation of a lariat RNA intermediate which is debranched and subsequently degraded after exon junction (Gao et al., 2008; Corvelo et al., 2010). As the sequences of the branch

point are highly degenerated, their exact localization is difficult to determine, but it is present mainly between -18 to -44 intronic nucleotides (Leman et al., 2020).

The polypyrimidine tract with sequence enriched in pyrimidine nucleotides  $/(Y)_{12-17}/$  is located between -5 and -40 bp from the acceptor splice site. This sequence binds the U2AF65 spliceosome subunit and polypyrimidine tract-binding protein. Any mutation in these sequences could lead to splicing alterations (Ohno et al., 2018).

#### - **Splicing regulatory elements (SREs)**

SREs are *cis*-regulatory RNA motifs, often 6–8 nt long, degenerated, and sometimes overlapping, that modulate RNA splicing by interacting with *trans*-acting factors that either activate or repress the splice site recognition and intron removal (Cartegni et al., 2002; Chasin, 2007). They include exonic splicing enhancers and silencers (ESEs and ESSs, also called ESRs for exonic splicing regulators), as well as their intronic counterparts (ISEs and ISSs, collectively known as ISRs). Importantly, it has been estimated that ESEs cover about half of all exonic nucleotides in the human genome (Chasin, 2007). These elements serve as binding sites that recruit *trans*-acting factors (e.g., SR and hnRNP proteins) that activate or suppress splice site recognition of spliceosome assembly by various mechanisms (Fu and Ares, 2014).

#### ii) ***Trans*-elements of splicing**

The *trans*-elements include proteins and ribonucleoproteins required for the splicing machinery (spliceosome) and its regulation. The spliceosome is a highly dynamic and supramolecular ribonucleoprotein complex, composed of five small nuclear ribonucleoproteins (snRNPs) and more than 100 proteins, including kinases, phosphatases and helicases, many of which are required for spliceosomal function, as well as associated proteins such as mRNA-export

factors and transcription factors (reviewed in Wang and Cooper, 2007; Wilkinson et al., 2020). The spliceosome assembly is further coordinated by the interaction of auxiliary splicing *cis*-elements and enhancer or repressor protein complexes, as exon splicing enhancers that bind serine/arginine (SR)-rich related (SR) proteins and recruit and stabilize binding of spliceosome components such as U2AF (reviewed in Wang and Cooper, 2007). Interestingly, several studies have shown that mutations in components of the splicing machinery can contribute to a dysregulated RNA splicing and tumorigenesis (Wang and Aifantis, 2020). However, alterations in these trans-elements are out of the scope of this thesis.

### **3.2. Spliceogenic variants in genetic diseases**

A DNA variant disrupting any of the *cis*-acting core or regulating elements may lead to incorrect splicing, resulting in partial or complete exon loss/intron gain in the mature mRNA, thus generating aberrant non-functional transcripts (truncating or in-frame) proteins which could cause disease (Baralle and Buratti, 2017) (Figure 6). According to HGMD, nearly 9% of all variants (ascertained September 2020) leading to human genetic diseases affect pre-mRNA splicing. However, the rate of miss-RNA splicing variants causing disease is thought to be much higher as synonym, missense, and deep intronic variants outside the donor and acceptor splicing site motifs are rarely considered and evaluated as splicing disruptors (Polak et al., 2017; Canson et al., 2020).

In general, variants in the canonical acceptor and donor sites affect strongly conserved sequences that define exon-intron boundaries. The most known affect + 1 and + 2 residues at the 5' donor splice site and -1 and - 2 residues at the 3' acceptor splice site (Anna and Monika, 2018). However, exonic/intronic variants located within the proximity of canonical splice sites have also been

shown to alter splicing (Montalban et al., 2018b, 2019; Duran-Lozano et al., 2019).

Some variants create similar sequences to splicing sites, known as new splicing sites. Intronic or exonic *cis*-acting variants also have the potential to disrupt the use of alternative (cryptic) splice sites, changing the proportions of naturally occurring mRNA transcripts which may also lead to disease (Wang and Cooper, 2007).

Moreover, variants affecting splicing regulatory elements have also been described as causing non-functional proteins in HBOC syndrome and other genetic diseases (Tubeuf et al., 2020). Variants disrupting polypyrimidine tract or branch point sequences have been less reported, but recent studies have identified many variants altering these elements (Anna and Monika, 2018; Leman et al., 2020).

Along DNA intronic regions there are numerous sequences similar to authentic splicing elements. The activation or disruption of these elements (such as cryptic splice sites activation or SREs alteration) could lead to the formation of pseudoexons (Vaz-Drago et al., 2017) (Figure 6).

On the other hand, due to the lack of sequencing of deep intronic regions in the diagnostic panels, a reduced number of spliceogenic variants have been identified in these regions in genes related to hereditary cancer or other rare genetic diseases. Interestingly, in previous research conducted in our laboratory, the first *BRCA1* deep intronic pathogenic variant was identified (Montalban et al., 2018a).



## INTRODUCTION

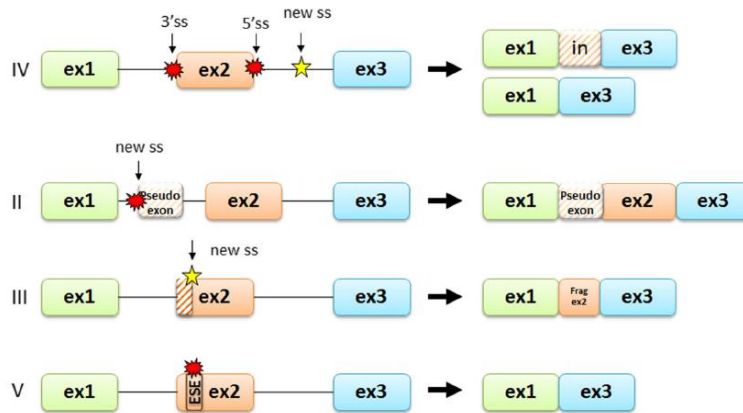


Figure 6. Types of splicing variants. Consequences in mRNA caused by variants disrupting DNA conserved elements. Taken from Anna and Monika, 2018.

### 3.3 Splicing *in silico* tools

Given that all types of genetic variation in any *cis*-splicing element may result in RNA alteration, the potential effect on splicing of all identified genetic variants should be evaluated (Canson et al., 2020). To detect splice alterations, *in vitro* splicing assays with patient's RNA or minigenes are widely used. However, testing all detected potentially spliceogenic variants is time-consuming and expensive. To prioritize variants to be experimentally evaluated, multiple computational prediction tools have been developed to assess the effect of DNA variants on splicing, based on scoring the functionality of the affected *cis*-element (Ohno et al., 2018). Also, *in silico* algorithms are included as one of the evidence criteria utilised for variant interpretation by the American College of Medical Genetics and Genomics (Richards et al., 2015).

Current bioinformatic filtering strategies and clinical interpretation guidelines tend to focus on amino acid level effects. This can lead to synonymous, intronic, or missense variants being filtered out at an early stage of analysis, even though such variants may affect splicing. Similarly, although deep intronic variant data

are increasingly available via massively parallel sequencing in approaches like exome or genome sequencing, such non-coding variants are rarely considered owing to a lack of evidence on which to base interpretations. When bioinformatic predictions suggest that a variant affects splicing, there can be scope for additional RNA-based investigations (Wai et al., 2020).

*In silico* tools that analyse the splicing strength have evolved from approaches based on position weight matrix (PWM) to machine learning and deep learning methods (Rowlands et al., 2019). In general, bioinformatic tools predicting donor and acceptor splicing sites alteration are more reliable than those applied to more loosely conserved elements like SREs (Ohno et al., 2018; Canson et al., 2020).

Historically, computational tools are based on different premises, the most common being used based on position weight matrix (PWM), in which each nucleotide on the splice site sequence is scored and ranked according to its frequency from its aligned consensus sequence. Splicing Site Finder (SSF) and Human Splice Finder (HSF) tools are based on PWM (Shapiro and Senapathy, 1987; Desmet et al., 2009).

Neural network programs, such as NNSplice (NNS), are previously trained on examples with consensus splice sites. Based on the result of the network, the exact location of the splice site is predicted without prior knowledge of the donor or acceptor splice sites in the analysed sequence (Reese et al., 1997; Johansen et al., 2009). Tools based on maximum entropy distribution models such as MaxEntScan, take into account the dependencies between nucleotide positions given a set of constraints defined as low-order marginal distributions, and generates two models based on a set of real and decoy splice sites. It then compares the probability that a given nucleotide sequence belongs to each of the two distributions and returns how much more likely the sequence is to be

## INTRODUCTION

a real site, rather than a decoy site (Yeo and Burge, 2004; Rowlands et al., 2019). Also, the combination of different algorithms such as SpliceSiteFinder and MaxEntScan (SPiCE tool) has been proved to be efficient to identify splicing sites altering variants (Leman et al., 2018).

Interestingly, some *in silico* tools have been developed using experimentally observed evidence, e.g. ESRseq (Ke et al., 2011) and HEXPLORER (Erkelenz et al., 2014). Using *in vitro* RNA approaches, scores for each hexamer as splicing regulatory element sequence were obtained. Then, after calculating the difference between wild type and variant sequence,  $\Delta$ ESRseq and  $\Delta$ HZEI scores are retrieved (Ke et al., 2011; Erkelenz et al., 2014).

Some tools adopt a meta-analytical approach by incorporating output from other tools as features, such as the use of different tool scores integrated into S-CAP. This tool combines sequenced-based features, evolutionary conservation, and existing metrics, including SPIDEX, CADD, and LINSIGHT, and comprises six different splicing prediction models, each designed to predict the pathogenicity of rare single nucleotide variants in a different splicing region (Jagadeesh et al., 2019).

In recent years, splice prediction tools have incorporated a wide range of machine learning-based models requiring training and testing. A key element of machine learning is the use of features or variables that are incorporated into the models and from which inferences are ultimately made. Many of these features are often sequence-based, showing the frequency or position of particular nucleotide sequences over a given region. Biochemical features, such as GC content and thermodynamic properties, are often also employed (Rowlands et al., 2019).

One major contributing factor to the rapid upgrowth in the number of machine learning-based splice prediction tools is the increased availability of publicly available data. Particularly valuable are experimentally-derived RNA-seq datasets. Many tools use raw sequence data as input. In such cases, these sequences are taken from a reputed transcript model, as in the cases of MMSplice (Cheng et al., 2019b) and SpliceAI (Jaganathan et al., 2019). SpliceAI is one of the most interesting deep learning-based tools. It analyses each position in a pre-mRNA transcript and evaluates whether it is likely to be a donor and acceptor splice site or neither. To train the model, the authors selected over 20,287 principal protein-coding transcripts from the GENCODE v24 annotation and used those from a selection of particular chromosomes as a training set. This tool is able to check any position in the genome and analyse up to 10,000 nt from the location of the variant. The authors demonstrate the ability of SpliceAI to faithfully identify authentic splice sites from nucleotide sequence alone, allowing the recreation of entire gene transcripts. SpliceAI-10k exhibits 95% top-k accuracy and an area under precision-recall curve (PR-AUC) of 0.98 (Jaganathan et al., 2019). The tool has been built to identify the variant impact on acceptor/donor loss or gain. However, the performances for the identification of variants altering other elements like SREs remain unknown.

### **3.3.1 Splicing *in silico* tools for specific sequence regions**

Some of the previously mentioned tools are dedicated to specific splicing elements or regions, and others can analyse all types of splicing elements. For example, SSF, HSF, MES, or SPiCE are *in silico* tools limited to identifying variants altering natural splicing sites or creating similar sequences to them (Table 5). These tools have been the most commonly used in spliceogenic variant detection due to the more historical knowledge about these conserved splicing sequences.

## INTRODUCTION

Variants affecting the polypyrimidine tract located near the acceptor splicing site can be optimally identified using MaxEntScan (Yeo and Burge, 2004). On the other hand, the collection of experimental information of branch points allowed the development of tools like Branch Point Prediction (BPP) and Branchpointer (Zhang et al., 2017; Signal et al., 2018) (Table 5).

In addition, HAL (Rosenberg et al., 2015), a combination of machine learning and synthetic biology trained with mini-gene sequences, ESRseq (Ke et al., 2011), and HEXPLORER (Erkelenz et al., 2014) tools are designed to identify variants creating or disrupting SREs (ESE/ESS or ISE/ISS) (Table 5).

Finally, some *in silico* tools like SpliceAI, SPANR, MMSplice, S-Cap or SPiP can potentially identify alterations in various sequences or splicing elements (Table 5).

*Table 5. Summary of commonly used in silico splicing tools, their targeted analysed regions and elements and their model of analysis.*

<b>Tool</b>	<b>Region/element of analysis</b>	<b>Model base</b>	<b>Citation</b>
Splice Site Finder (SSF)	Spl sites and new/cryptic sites	PWM	(Shapiro and Senapathy, 1987)
MaxEntScan (MES)	Spl sites, new/cryptic sites, and PPT	Maximum Entropy Principle	(Yeo and Burge, 2004)
Human Splice Finder (HSF)	Spl sites and new/cryptic sites	PWM	(Desmet et al., 2009)
dbscSNV	Spl sites	Adaptive boosting and random forests	(Jian et al., 2014)
SPiCE	Spl sites and new/cryptic sites	PWM + Maximum Entropy Principle	(Leman et al., 2018)

BranchPointer	Branch Points	ML	(Signal et al., 2018)
BPP	Branch Points / PPT	Mixture Model	(Zhang et al., 2017)
ESRseq	Splicing Regulatory Elements	Experimentally inferred	(Ke et al., 2011)
HAL	Splicing Regulatory Elements	Machine learning and synthetic biology	(Colombo et al., 2021)
HEXPLOER	Splicing Regulatory Elements	Experimentally inferred	(Erkelenz et al., 2014)
SpliceAI	All sequences	ML (Deep neural network)	(Jaganathan et al., 2019)
SPANR	Exons, plus +/-300 intronic nts	ML (Neural network)	(Xiong et al., 2015)
S-CAP	Exons, plus 50 bp flanking intronic sequence	ML (Gradient boosting tree)	(Jagadeesh et al., 2019)
SPiP	All sequences	Meta-predictor	<a href="https://github.com/raphaelleman/SPiP">https://github.com/raphaelleman/SPiP</a>

PWM, Position Weight Matrix; ML, Machine Learning; PTT, Polypyrimidine Tract; Spl, Splicing.

In summary, tools based on different approaches have been launched, some of them dedicated to specific elements, and others able to compute various types of splicing elements. Performance analysis comparing tools using different datasets have been carried out (Tang et al., 2016; Tubeuf et al., 2020). However, there is no consensus of which tool has to be used, in which conditions, or in what type of variants has to be applied. Moreover, their inclusion in the ACMG guideline is not clearly defined and do not cover the whole *cis*-splicing elements landscape. For this reason, independent performance studies with a large set of variants, comparing tools and optimizing their use, are needed.



**HYPOTHESIS**





Patients with HBOC in whom a causative pathogenic variant is not identified after genetic analysis may not benefit from prevention, early detection, or precision treatment measures. This negative or inconclusive results are due, among other causes, to the detection of variants of uncertain significance (VUS), some of them potentially spliceogenic, and the existence of still unknown susceptibility genes.

The hypotheses of this work consist of:

-Variants detected using massive parallel sequencing could disrupt or create splicing elements and consequently alter RNA leading to non-functional or partially functional proteins. The development of *in silico* tools to identify this type of variants (including deep intronic variants), could help to select with high grades of sensitivity and specificity those variants that should be prioritized for subsequent RNA analysis, increasing the possibility of reaching a diagnosis.

-Improving the classification of variants of uncertain significance (VUS) by adapting variant classification guidelines to specific genes, such as *ATM*, can reduce the number of such variants and the uncertainty in patients and clinicians.

-The analysis of exomes or extended panels in patients with HBOC negative for known risk genes, could identify new candidate genes, that have to be validated in later case-control studies. These new susceptibility genes will enlarge the clinical benefit for patients who have not obtained a prior genetic diagnosis.



**OBJECTIVES**



The main objective of this thesis is to increase the capacity of genetic diagnosis of patients with HBOC, by focusing on i) the optimisation in the interpretation of exonic and intronic variants that might affect RNA quality or quantity but remain as variants of uncertain significance (VUS) and ii) the identification of new susceptibility genes for HBOC. According to this, the specific aims of this project are:

- 1) To evaluate the performance of native splicing site alteration predictions made by commonly used *in silico* tools, comparing their outputs with the experimental evidence obtained by *in vitro* RNA analysis of variants detected in HBOC genes, for their implementation in the clinical variant interpretation guidelines.
- 2) To provide an *in silico* framework to prioritize deep intronic variants for their experimental RNA analysis and to elucidate the importance of the landscape of splicing elements and SRE balance in the inclusion of sequences in mature RNA.
- 3) To adapt the ACMG variant interpretation guideline to *ATM* gene for patients with cancer predisposition syndromes, especially focusing on splicing *in silico*.
- 4) To analyse exomes and a panel of candidate genes in patients with early-onset BC/OC and no pathogenic variant in known genes, to select new genes potentially associated with risk, and validate their relationship with susceptibility to HBOC in patients and healthy Spanish controls.



## **RESULTS**





**CONTENTS**

The present thesis comprises four articles that follow the order of objective section:

**Article 1** “Computational tools for splicing defect prediction in breast/ovarian cancer genes: How efficient are they at predicting RNA alterations?” Moles-Fernández A, Duran-Lozano L, Montalban G, Bonache S, López-Perolio I, Menéndez M, Santamariña M, Behar R, Blanco A, Carrasco E, López-Fernández A, Stjepanovic N, et al. 2018. *Front Genet* 9:366.

**Article 2** “Role of Splicing Regulatory Elements and *In Silico* Tools Usage in the Identification of Deep Intronic Splicing Variants in Hereditary Breast/Ovarian Cancer Genes” Moles-Fernández, A.; Domènech-Vivó, J.; Tenés, A.; Balmaña, J.; Díez, O.; Gutiérrez-Enríquez, S. 2021. *Cancers*, 13, 3341. <https://doi.org/10.3390/cancers13133341>

**Article 3** “A Collaborative Effort to Define Classification Criteria for ATM Variants in Hereditary Cancer Patients”. Feliubadaló L, Moles-Fernández A, Santamariña-Pena M, Sánchez AT, López-Novo A, Porras L-M, Blanco A, Capellá G, la Hoya M de, Molina IJ, Osorio A, Pineda M, et al. 2021. *Clin Chem* 67:518–533.

**Article 4** “Unravelling genetic predisposition to familial breast and ovarian cancer: identification of new susceptibility genes by case-control study” A. Moles-Fernández, E. Aguado-Flor, C. Zamarreño-Pastor, Tu Nguyen-Dumont, Melissa C Southey, M. Antolín, S. Bonache, A. López-Fernández, L. Feliubadaló, J. Fernández-Navarro, C. Lázaro, J. Balmaña, O. Díez, S. Gutiérrez-Enríquez. Article *in preparation*.

## SUMMARY OF RESULTS

This thesis aimed to provide tools for the classification of VUS identified in HBOC risk-associated genes, to reduce the rate of undiagnosed patients. Variants of uncertain significance could affect RNA due to disruption or creation of *cis*-splicing elements, and *in silico* tools can help to identify these potential alterations.

Splicing sites are one of the most important splicing conserved elements. **Article 1** explains an optimization in the identification of potentially spliceogenic variants located near to these sequences, comparing different *in silico* algorithms, alone or in combination, and providing recommendations to use HSF+SSF-like or HSF+SSF-like+MES for analysing donor sites and SSF-like for acceptor sites, after their validation in a large set of data.

Moreover, the creation or activation of cryptic sites along deep intronic regions could alter splicing causing the inclusion of intronic sequences in RNA, potentially leading to non-functional proteins. In **Article 2**, a framework for the identification of deep intronic spliceogenic was provided, after the performance analysis of SpliceAI *in silico* tool in a dataset of spliceogenic and no-spliceogenic deep intronic variants. In addition, the importance of the splicing regulatory elements balance in the pseudoexon creation was described.

The ACMG variant interpretation guidelines provide general recommendations to classify variants. This approach causes a non-totally accurate classification because guidelines are not adjusted to specific gene characteristics, resulting in a VUS over-classification. In **Article 3**, ACMG guidelines were adapted to *ATM* gene. We focused on *in silico* splicing evidence (PP3/BP4). After reclassification of variants following the adapted guidelines, a reduction of VUS from 58% to 42% was obtained.

On the other hand, in patients without pathogenic variants identified in HBOC related genes, the phenotype could be due to deleterious variants in genes still not known associated with the disease. For this reason, in **Article 4**, the aim was to identify candidate genes through exomes and extended panel analysis and

validate their risk association by performing a case-control study. After the analysis of affected patients and healthy controls, a set of genes were associated to susceptibility to hereditary breast/ovarian cancer.

**REPORT OF IMPACT FACTOR OF THE ARTICLES INCLUDED IN THE THESIS**

Sara Gutiérrez Enríquez, PhD and Orland Díez Gibert, PhD, with DNI 23820736G and 36509106X, as directors of the doctoral thesis carried out by Mr./a Alejandro Moles Fernández entitled “Unravelling genetic predisposition to familial breast and ovarian cancer: new susceptibility genes and variant interpretation by *in silico* approaches” enrolled in the Biomedicine doctoral programme at the University of Barcelona, notify the contribution of Alejandro Moles Fernández in the articles included in this thesis:

**Article 1.** Computational Tools for Splicing Defect Prediction in Breast / Ovarian Cancer Genes : How Efficient Are They at Predicting RNA Alterations ? **Moles-fernández, A.**; Duran-lozano, L.; Montalban, G.; Bonache, S.; López-perolio, I.; Menéndez, M.; Santamariña, M.; Behar, R.; Blanco, A.; Carrasco, E.; et al. *Front. Genet.* 2018, 9, doi:10.3389/fgene.2018.00366. **Impact factor: 4.599**  
**Contribution: First author.** Conception and design, acquisition of data, data analysis and interpretation, drafting the workcritical revision of the article and final approval of the version to be published.

**Article 2.** Role of Splicing Regulatory Elements and In Silico Tools Usage in the Identification of Deep Intronic Splicing Variants in Hereditary Breast/Ovarian Cancer Genes. **Moles-Fernández, A.**; Domènech-Vivó, J.; Tenés, A.; Balmaña, J.; Díez, O.; Gutiérrez-Enríquez, S. *Cancers* 2021, 13. **Impact factor: 6.639**  
**Contribution: First author in co-authory.** Conceptualization, methodology, software, validation, formal analysis, research, data curation, writing - preparing the original draft, writing - review and editing, visualization.

**Article 3.** A Collaborative Effort to Define Classification Criteria for ATM Variants in Hereditary Cancer Patients. Feliubadaló, L.; **Moles-Fernández, A.**; Santamariña-Pena, M.; Sánchez, A.T.; López-Novo, A.; Porras, L.-M.; Blanco, A.; Capellá, G.; de la Hoya, M.; Molina, I.J.; et al. *Clin. Chem.* 2021, 67, 518–533, doi:10.1093/clinchem/hvaa250. **Impact factor: 8.327** **Contribution: Second author in co-authory.** Conceptualization and design, data acquisition and analysis, interpretation of data, writing and review of the article for its intellectual content. Special involvement in the sections of *in silico* evidence and hot-spot regions.

**Article 4.** Unravelling genetic predisposition to familial breast and ovarian cancer: identification of new susceptibility genes by case-control study. **A. Moles-Fernández,** E. Aguado-Flor, C. Zamarreño-Pastor, Tu Nguyen-Dumont, Melissa C Southey, M. Antolín, S. Bonache, A. López-Fernández, L. Feliubadaló, J. Fernández-Navarro, C. Lázaro, J. Balmaña, O. Díez, S. Gutiérrez-Enríquez. *In*

*preparation. Contribution: First author.* Conception and design, acquisition of data, data analysis and interpretation, drafting the workcritical revision of the article and final approval of the version to be published.

Barcelona, 13 September of 2021

Sara Gutiérrez-Enríquez

Orland Díez

A handwritten signature in blue ink that reads "Sara G. Gutiérrez-Enríquez". The signature is written in a cursive style and is underlined with a single horizontal line.A handwritten signature in blue ink that reads "Orland Díez". The signature is written in a cursive style and is underlined with a single horizontal line.



## Article 1

### Computational tools for splicing defect prediction in breast/ovarian cancer genes: How efficient are they at predicting RNA alterations?

Moles-Fernández A, Duran-Lozano L, Montalban G, Bonache S, López-Perolio I, Menéndez M, Santamariña M, Behar R, Blanco A, Carrasco E, López-Fernández A, Stjepanovic N, Balmaña J, Capellá G, Pineda M, Vega A, Lázaro C, de la Hoya M, Díez O, Gutiérrez-Enríquez S.

**Frontiers in Genetics.** 2018 Sep 5;9:366  
doi:10.3389/fgene.2018.00366.





# Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations?

Alejandro Moles-Fernández<sup>1</sup>, Laura Duran-Lozano<sup>1</sup>, Gemma Montalban<sup>1</sup>, Sandra Bonache<sup>1</sup>, Irene López-Perolio<sup>2</sup>, Mireia Menéndez<sup>3,4,5</sup>, Marta Santamariña<sup>6</sup>, Raquel Behar<sup>2</sup>, Ana Blanco<sup>6</sup>, Estela Carrasco<sup>7</sup>, Adrià López-Fernández<sup>7</sup>, Neda Stjepanovic<sup>7,8</sup>, Judith Balmaña<sup>7,8</sup>, Gabriel Capellá<sup>3,4,5</sup>, Marta Pineda<sup>3,4,5</sup>, Ana Vega<sup>6</sup>, Conxi Lázaro<sup>3,4,5</sup>, Miguel de la Hoya<sup>2</sup>, Orland Diez<sup>1,9\*</sup> and Sara Gutiérrez-Enríquez<sup>1,†\*</sup>

## OPEN ACCESS

### Edited by:

Paolo Peterlongo,  
IFOM - The FIRCC Institute  
of Molecular Oncology, Italy

### Reviewed by:

Rachid Karam,  
Ambry Genetics, United States  
Logan Walker,  
University of Otago, New Zealand

### \*Correspondence:

Orland Diez  
odiez@vhio.net  
orcid.org/0000-0001-7339-0570  
Sara Gutiérrez-Enríquez  
sgutierrez@vhio.net  
orcid.org/0000-0002-1711-6101

† These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Cancer Genetics,  
a section of the journal  
Frontiers in Genetics

Received: 23 May 2018

Accepted: 22 August 2018

Published: 05 September 2018

### Citation:

Moles-Fernández A, Duran-Lozano L,  
Montalban G, Bonache S,  
López-Perolio I, Menéndez M,  
Santamariña M, Behar R, Blanco A,  
Carrasco E, López-Fernández A,  
Stjepanovic N, Balmaña J, Capellá G,  
Pineda M, Vega A, Lázaro C,  
de la Hoya M, Diez O and  
Gutiérrez-Enríquez S (2018)  
Computational Tools for Splicing  
Defect Prediction in Breast/Ovarian  
Cancer Genes: How Efficient Are  
They at Predicting RNA Alterations?  
Front. Genet. 9:366.  
doi: 10.3389/fgene.2018.00366

<sup>1</sup> Oncogenetics Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain, <sup>2</sup> Laboratorio de Oncología Molecular – Centro de Investigación Biomédica en Red de Cáncer, Instituto de Investigación Sanitaria San Carlos, Hospital Clínico San Carlos, Madrid, Spain, <sup>3</sup> Hereditary Cancer Program, Catalan Institute of Oncology, Institut d'Investigació Biomèdica de Bellvitge, Hospitalet de Llobregat, Barcelona, Spain, <sup>4</sup> Program in Molecular Mechanisms and Experimental Therapy in Oncology (Oncobell), Institut d'Investigació Biomèdica de Bellvitge, Hospitalet de Llobregat, Barcelona, Spain, <sup>5</sup> Centro de Investigación Biomédica en Red de Cáncer, Madrid, Spain, <sup>6</sup> Grupo de Medicina Xenómica-USC, Fundación Pública Galega de Medicina Xenómica-SERGAS, CIBER de Enfermedades Raras, Instituto de Investigación Sanitaria, Santiago de Compostela, Spain, <sup>7</sup> High Risk and Cancer Prevention Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain, <sup>8</sup> Medical Oncology Department, University Hospital Vall d'Hebron, Barcelona, Spain, <sup>9</sup> Area of Clinical and Molecular Genetics, University Hospital Vall d'Hebron, Barcelona, Spain

*In silico* tools for splicing defect prediction have a key role to assess the impact of variants of uncertain significance. Our aim was to evaluate the performance of a set of commonly used splicing *in silico* tools comparing the predictions against RNA *in vitro* results. This was done for natural splice sites of clinically relevant genes in hereditary breast/ovarian cancer (HBOC) and Lynch syndrome. A study divided into two stages was used to evaluate SSF-like, MaxEntScan, NNSplice, HSF, SPANR, and dbSCSNV tools. A discovery dataset of 99 variants with unequivocal results of RNA *in vitro* studies, located in the 10 exonic and 20 intronic nucleotides adjacent to exon–intron boundaries of *BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *ATM*, *BRIPI1*, *CDH1*, *PALB2*, *PTEN*, *RAD51D*, *STK11*, and *TP53*, was collected from four Spanish cancer genetic laboratories. The best stand-alone predictors or combinations were validated with a set of 346 variants in the same genes with clear splicing outcomes reported in the literature. Sensitivity, specificity, accuracy, negative predictive value (NPV) and Mathews Coefficient Correlation (MCC) scores were used to measure the performance. The discovery stage showed that HSF and SSF-like were the most accurate for variants at the donor and acceptor region, respectively. The further combination analysis revealed that HSF, HSF+SSF-like or HSF+SSF-like+MES achieved a high performance for predicting the disruption of donor sites, and SSF-like or a sequential combination of MES and SSF-like for predicting disruption of acceptor sites. The performance confirmation of these last results with the validation dataset, indicated that the highest sensitivity, accuracy, and NPV (99.44%, 99.44%, and 96.88, respectively) were attained with HSF+SSF-like or HSF+SSF-like+MES for donor sites and SSF-like (92.63%, 92.65%, and 84.44, respectively) for acceptor sites.

We provide recommendations for combining algorithms to conduct *in silico* splicing analysis that achieved a high performance. The high NPV obtained allows to select the variants in which the study by *in vitro* RNA analysis is mandatory against those with a negligible probability of being spliceogenic. Our study also shows that the performance of each specific predictor varies depending on whether the natural splicing sites are donors or acceptors.

**Keywords:** hereditary cancer genes, NGS of gene-panel, VUS classification, *in silico* tools, splicing, RNA alteration

## INTRODUCTION

The increasing use of massive parallel sequencing of customized multi-gene panels, for germline clinical testing of hereditary breast and ovarian cancer (HBOC) and Lynch syndrome, is leading to higher detection of genetic variants of unknown significance (VUS).

All exonic or intronic VUS can be potentially spliceogenic by disrupting the *cis* DNA sequences that define exons, introns, and regulatory sequences necessary for a correct RNA splicing process. Specifically, the *cis* DNA elements include: (i) exon–intron boundary core consensus nucleotides (GT at +1 and +2 of the 5′ donor site and AG at -1 and -2 of the 3′ acceptor site); (ii) intronic and exonic nucleotides adjacent to these invariable nucleotides that are also highly conserved and have been found to be critical for splice site selection: CAG/GUAAGU in donor sites and NYAG/G in acceptor sites; (iii) branch point and polypyrimidine tract sequence motifs, essential for the spliceosome complex formation; (iv) intronic and exonic sequences that act as splicing enhancers (ISE and ESE) or silencers (ISS and ESS), regulatory motifs that are usually bound by serine/arginine (SR)-rich proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs), respectively (Cartegni et al., 2002; Soukari et al., 2016; Abramowicz and Gos, 2018). A nucleotide change in any of these elements could lead to incorrect splice site recognition, creating new ones or activating the cryptic ones, resulting in aberrant transcripts and in non-functional proteins associated with disease such as hereditary cancer.

Interestingly, it has recently been described that hereditary cancer genes (including some HBOC and Lynch genes) are enriched for spliceogenic variants (Rhine et al., 2018). This finding highlights the importance of both the identification and the functional interpretation of variants causing RNA alterations in hereditary cancer genes. In HBOC syndrome and Lynch Syndrome, the clinical classification of VUS is essential since carriers of pathogenic variants may benefit from cancer prevention and risk-reducing strategies, make informed decisions about prophylactic surgery, and benefit from targeted treatments (Moreno et al., 2016). Conversely, carriers of non-pathogenic variants can be excluded from intensive follow-ups and avoid unnecessary risk-reducing surgery (Eccles et al., 2015).

To detect splice site alterations, *in vitro* splicing assays with patient's RNA or minigenes are widely used. However, testing all variants detected in the vicinity of exon–intron boundaries can be time consuming and expensive. In consequence, to select variants to be experimentally evaluated, a large number of prediction

programs have been developed. These splicing computational tools are based on different premises. The most commonly used are based on Position Weight Matrix (PWM), in which each nucleotide on the splice site sequence is scored and ranked based on its frequency from its aligned consensus sequence (Shapiro and Senapathy, 1987; Desmet et al., 2009). Neural network programs use sets of sequences from databases to identify splicing sites (Reese et al., 1997). Tools based on Maximum Entropy Distribution models take into account the dependencies between nucleotide positions (Yeo and Burge, 2004). Approaches like SPANR (Xiong et al., 2015) use DNA and RNA sequence information and a machine learning method, to predict splicing alterations, enabling the identification of variants affecting *cis* and *trans* splicing factors. Another type of splicing tool has been developed using ensemble learning methods (adaptive boosting and random forest) taking advantage of individual computational tools (Jian et al., 2014a).

Several studies have analyzed the performance of these tools for genes related to cancer and other diseases and report discordant results without a consensus guideline recommending which programs should be used (Houdayer et al., 2008, 2012; Holla et al., 2009; Vreeswijk et al., 2009; Desmet et al., 2010; Théry et al., 2011; Colombo et al., 2013; Jian et al., 2014a; Tang et al., 2016) (Table 1). Here, we present an evaluation of the performance of commonly used splicing *in silico* tools, comparing their output with the experimental evidences obtained by RNA *in vitro* analysis of variants detected in HBOC and Lynch syndrome genes. In the first phase of the study, we assessed the accuracy of the splicing *in silico* tools with a dataset of RNA *in vitro* outcomes collected from four Spanish cancer genetic units. Subsequently, we validated the best algorithms obtained in the discovery phase, with findings obtained after RNA analysis extracted from different curated databases and reported literature.

## MATERIALS AND METHODS

### Variant Selection

#### Discovery Set

We restricted the study to variants located within the last 10 exonic and 20 first intronic nucleotides from the 5′ splice donor site, and the last 20 intronic and the first 10 exonic nucleotides from the 3′ splice acceptor site (−10 to +20 and −20 to +10, respectively). *BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6*, and *PMS2* variants were selected from HBOC and Lynch

TABLE 1 | Publications evaluating *in silico* splicing site tools.

Reference	Number of variants	Source of the variants and <i>in vitro</i> data	Gene(s)	Region analyzed	Experimental design	Prediction tools evaluated	Accuracy of recommended tools	Consensus guideline
Houdayer et al., 2008	39	*Experimental evidence	<i>RBT1</i>	±60 nucleotides from an AG/GT site	One evaluation stage	NNSplice, PWM, MES, ASSA, ESEfinder, RESCUE-ESE	NA	Not specifically provided
Holla et al., 2009	18	Experimental evidence	<i>LDLR</i>	Intronic: 5' until +5, 3' until -16	One evaluation stage	MES, NNSplice, NetGene2	NA	Not specifically provided
Vreeswijk et al., 2009	29	Experimental evidence	<i>BRCA1/BRCA2</i>	Intronic: 5' until +60, 3' until -20	One evaluation stage	NNSplice, NetGene2, PWM, ASSA, MES, HSF	NA	Not specifically provided
Desmet et al., 2010	623	UMD locus-specific databases, HGMD, and datasets from previous studies	Multiple	Not specifically stated	One evaluation stage	GENSCAN, GeneSplicer, HSF, MES, NNSplice, SplicePort, SplicePredictor, SpliceView, SROOGLE	Invariable position: HSF, MES, SpliceView and SROOGLE 100%. Intronic SS +3, +5 and last exonic position: MES. Other SS intronic positions: MES 100%. Other SS intronic position: MES and SplicePort 5' 76/68% and 3' 77.21/77.27%	Invariable position: HSF, MES, SpliceView and SROOGLE. Intronic SS +3, +5 and last exonic position: MES. Other SS intronic positions: MES and SplicePort
Théry et al., 2011	53	Experimental evidence	<i>BRCA1/BRCA2</i>	Not specifically stated	One evaluation stage	PWM, GeneSplicer, NNSplice, MES, HSF	NA	Not specifically provided
Houdayer et al., 2012	272	Experimental evidence	<i>BRCA1/BRCA2</i>	Not specifically stated	One evaluation stage	NNSplice, SSF, MES, ESEfinder, RESCUE-ESE, HSF	Accuracy as AUC: MES: 0.956, SSF-like: 0.914	Sequential MES and SSF
Colombo et al., 2013	24	Experimental evidence	<i>BRCA1/BRCA2</i>	Not specifically stated	One evaluation stage	PWM, MES, NNSplice, GeneSplicer, HSF, NetGene2, SpliceView, SplicePredictor, ASSA	NA	HSF and ASSA
Jian et al., 2014b	2,959	HGMD, SpliceDisease database and DBASS. Negative variants from 1000 Genomes Phase 1	Multiple	5': from -3 to +8, 3': from -12 to +2	Evaluation of individual tools + new model construction + validation stage	SSF-like, MES, NNSplice, GeneSplicer, HSF, NetGene2, GENSCAN, SplicePredictor, **obscSNV	SSF-like: 91.1%; MES: 89.5%; obscSNV: 93.3%	SSF-like, MES/dbscSNV
Tiang et al., 2016	272	HGMD (damaging variants) and negative variants from 1000 Genomes Phase 1	Multiple	Intronic: 5' from +3 to +7, 3' from -3 to -9	One evaluation stage	HSF, MES, NNSplice, ASSP	Accuracy as AUC: MES: 0.878 ASSP: 0.881 HSF: 0.834	MES, ASSP, and HSF combination
Leman et al., 2018	395	Experimental evidence	Multiple	5': from -3 to +8, 3': from -12 to +2	Training + evaluation stage	HSF, MES, SSF-like, NNSplice, GS, SPOCE (MES and SSF combination)	SPOCE 95.6%	SPOCE (Th <sub>95</sub> threshold with MES and SSF combination)

\*Experimental evidence, experimental *in vitro* RNA FMO results collected specifically for the study, derived from either patient blood cells or mitogene assay. \*\*obscSNV: database containing the adaptive boosting and random forests scores. UMD, Universal Mutation Database; DBASS, Aberrant Splice Database; Splice NNSplice, Site Prediction by Neural Network; PWM, Position Weight Matrix; MES, MaxEntScan; ASSA, Automated Splice-Sites Analyses; HSF, Human Splice Finder; GS, GeneSplicer; SROOGLE, splicing regulation online graphical engine; ASSP, Alternative Splice Site Predictor; NA, information not available in the paper; SS, splicing site; AUC, area under the curve; Th<sub>95</sub>, optimal sensitivity threshold.



syndrome patients routinely analyzed for diagnostic purposes. We also included *ATM*, *BRIP1*, *CDH1*, *PALB2*, *PTEN*, *RAD51D*, *STK11*, and *TP53* variants obtained in a research series of *BRCA1* and *BRCA2* negative HBOC patients. Genetic variants with unequivocal experimental evidences showing presence or absence of alterations in the mRNA, were collected from four different Spanish centers: Hospital Universitari Vall d'Hebron (HUVH), Barcelona; Hospital Clínico San Carlos (HCSC) Madrid; Fundación Pública Galega de Medicina Xenómica (FPGMX), Santiago de Compostela; Institut Català d'Oncologia (ICO), Hospital Duran i Reynals, Barcelona.

The variants included in the discovery set were analyzed *in vitro* in carriers and controls. RNA was isolated from whole blood leukocytes or short-term lymphocyte cultures, phytohaemagglutinin stimulated, and treated with and without puromycin. The contributing laboratories used diverse isolation protocols and/or cDNA synthesis strategies following ENIGMA recommendations (Colombo et al., 2014; Whiley et al., 2014). Briefly, the splicing products generated by reverse transcription-polymerase chain reaction (RT-PCR) assays were characterized using agarose gel or capillary electrophoresis in a QIAxcel instrument with QIAxcel DNA High Resolution Kit (QIAGEN) or an Agilent 2100 Bioanalyzer (Agilent), and Sanger sequencing. PCR primers were designed to amplify at least one whole exon 5' and 3' flanking the exon harboring the variant of interest. Primer sequences are available upon request.

The study was approved by the Institutional Review Board of each participating center. Patients received genetic counseling and written informed consent was obtained for further genetic and research studies.

### Validation Set

At this stage, the predictors that presented the best performance alone or in combination, were applied to compare their predictions with the *in vitro* RNA results from the dataset obtained through literature and databases. We chose a collection of variants reported in INSIGHT, ClinVar and published works that were (i) located within the regions defined for the discovery set; (ii) identified in the set of cancer risk genes included above; (iii) experimentally confirmed as spliceogenic and non-spliceogenic in blood samples or with minigene assay at least by RT-PCR, agarose gel and Sanger Sequencing analysis; and (iv) not located at exonic splicing enhancer (ESE) regions with specific experimental evidence of causing splicing alteration.

### In silico Splice Tools

A total of six splice-site prediction software programs were selected for this study. Two ensemble prediction scores constructed by Jian et al. (2014a) using adaptive boosting and random forests ensemble learning methods, were extracted from dbSCSNV database<sup>1</sup>. Splicing-based Analysis of Variants (SPANR), a computational model of splicing derived from the application of “deep learning” computer algorithms (Xiong

et al., 2015) was ascertained by its own web site<sup>2</sup>. Splice Site Finder (SSF-like) (based on Shapiro and Senapathy, 1987), MaxEntScan (MES) (Yeo and Burge, 2004), Splice Site Prediction by Neural Network (NNPLICE) (Reese et al., 1997), and Human Splicing Finder (HSF) (Desmet et al., 2009) accessed through Alamut Visual 2.10 (Interactive Biosoftware). The GeneSplicer program is also included in the splicing module of Alamut, but it was excluded from the study since we noticed it had an exceedingly high missing scores (no estimation was obtained for 30% of the variants analyzed; data not shown), which had also been reported by Jian et al. (2014a). SPANR and dbSCSNV do not analyze insertions and deletions and dbSCSNV gives estimations for variants only located from -3 to +8 at 5' and -12 to +2 at 3' (Supplementary Table 1).

To interrogate the splicing prediction tools, we calculated the score variation caused by the variant in the donor site or acceptor site. To do that, we compared the score computed in the wild-type sequence (WT) to the score computed in the variant sequence (VAR) as:

$$\%scorevariation = (VARscore - WTScore)/WTScore * 100$$

We calculated the % score variation for four out of the six tools (SSF-like, HSF, MES, and NNPLICE), since dbSCSNV and SPANR already provide a score change.

To consider a % score change as a positive prediction of a splicing motif disruption caused by the variant, which would lead to aberrant splicing, we adopted thresholds pre-established in the literature (Supplementary Table 1). When two programs were combined, a correct prediction of splicing alteration was considered if at least one of them scored above the threshold. When three, four, five, or six programs were combined, all tools but one had to score above the threshold to indicate splicing alteration.

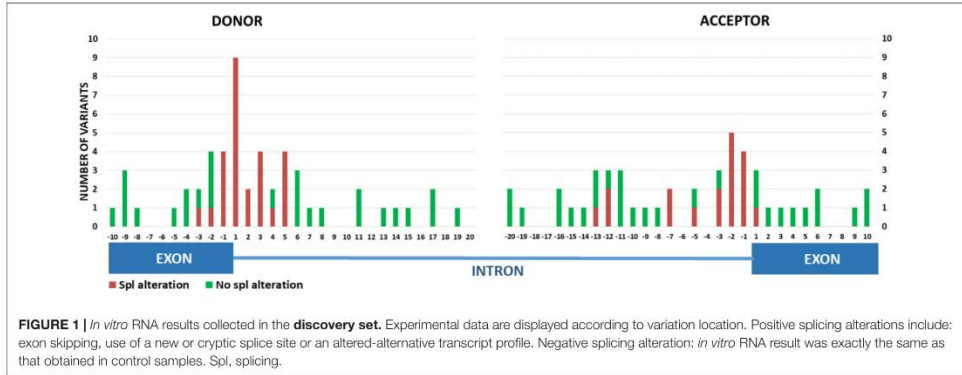
### Performance Assessment

In the discovery and validation phases, the experimental RNA results for each collected variant were annotated as positive splicing alteration when they unequivocally, verified by gel electrophoresis and Sanger sequencing, lead to: exon skipping, use of a new or cryptic splice site or altered alternative transcript profile. In contrast, a negative splicing alteration was annotated when the *in vitro* RNA result was exactly the same as that obtained in control samples.

For both stages, we calculated the overall accuracy (ratio of overall correct predictions to the total number of predictions), specificity (correct identification of non-spliceogenic variants; true negative rate), and sensitivity (correct identification of deleterious variants; true positive rate). The positive predictive values (PPV, proportion of positive predictions that were true positives), negative predictive values (NPV, proportion of negative predictions that were true negatives), false negative rates (FNR, proportion of false negative detection), and false positive rates (FPR, proportion of false positive detection) were also

<sup>1</sup><https://sites.google.com/site/jpoggen/dbNSFP>

<sup>2</sup><http://tools.genes.toronto.edu/>



calculated. Matthews correlation coefficient (MCC) was used to provide a balanced comparison between *in silico* tools.

**RESULTS**

**Discovery Set**

A total of 99 variants with unequivocal RNA *in vitro* results were studied, located within positions -10 to +20 from the 5' donor site, and within -20 to +10 from the 3' acceptor site (Supplementary Table 2). Forty-four of the 99 variants generated a splice defect, with 11 and 9 disrupting the canonical GT or AG dinucleotides, respectively. The 24 remaining variants with aberrant splicing were located outside invariable GT or AG positions, with 15 variants altering the 5' splice site and nine altering the 3' splice site. Fifty-five variants did not yield an aberrant splicing, all located outside invariant dinucleotides. Figure 1 displays the number of positive and negative splicing results relative to variant location.

Six *in silico* tools were used to interrogate the 99 variants, and their corresponding % score variation was obtained. These outputs were compared to the experimental RNA results. The respective thresholds pre-established in the literature were adopted for each program (Supplementary Table 1).

Supplementary Table 2 lists the % score variation obtained from each splicing tool used to assess the 99 variants, highlighting which scores were in agreement with the RNA analysis outcome. Of note, seven insertions or deletions were not computed by SPANR and dbSCSNV, while estimations for 33 substitutions were not provided by dbSCSNV.

Table 2 shows separately, for 5' (52 variants), 3' (47 variants), and both splice sites (global, 99 variants), the results of performance analysis for each one of the tools. The six predictors detected wild type (WT) splice sites in reference sequences for all the genes of interest.

On average, predictions for variants located in 5' regions have higher accuracy (90.98%), sensitivity (90.44%) and specificity (91.28%) compared to those located in 3' regions (83.74%,

84.52%, and 82.30%, respectively) (Table 2). The predictions computed by HSF (with a score change threshold of -2%) were the most accurate and sensitive for variants at donor site, while for variants at acceptor sites or affecting either acceptor or donor sites (global), SSF-like were the most accurate (with a score change threshold of -5%). MES program (with a score change threshold of -15%) showed 100% of sensitivity on all predictions, but its specificity did not reach 87% in any case. In contrast, SPANR program showed the highest values of specificity for predictions of variants at donor site or all variants affecting either at acceptor or donor splice sites, but the lowest values of sensitivity (Table 2).

Accordingly, the lowest false negative rates for 5' splice site were reached by the HSF and MES predictors, while at 3' splice sites, the SSF-like and MES predictors obtained the lowest false negative rates (Table 2 and Figure 2). In contrast, SPANR predictor had the highest false negative and the lowest false positive rates in almost all cases (Table 2 and Figure 2). Regarding the estimation of the proportion of negative predictions that were true negatives (NPV), HSF or MES and SSF-like or MES achieved the highest values (100%) for donor and acceptor sites, respectively (Table 2).

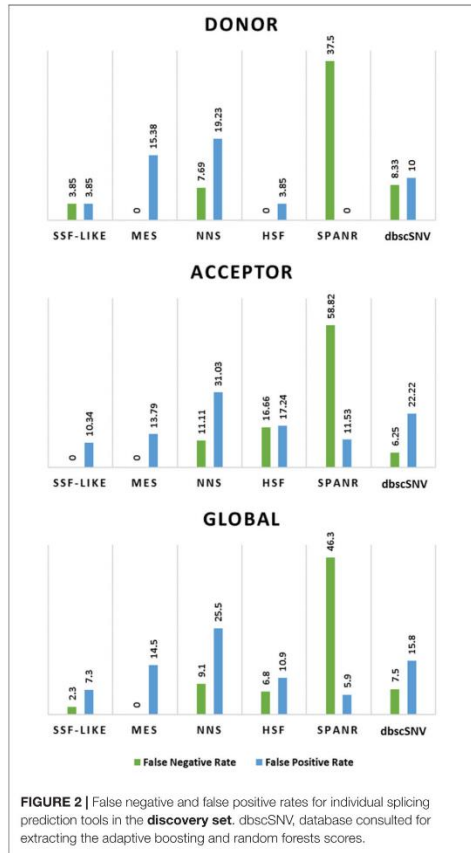
The accuracy of all possible predictor combinations was further assessed. For 5' donor splice sites, predictions of HSF alone or HSF together with seven different combinations, SSF-like+SPANR and SSF-like+MES+SPANR reached a 98.08% of accuracy with the highest sensitivity for all the models (100%), obtaining 96.15% of specificity, 0.96 MCC and 100% of NPV (Supplementary Table 3). For 3' splice sites, a sequential combination recommended by Houdayer et al. (2012) using MES as first-line analysis with a cut-off of 15% followed by SSF-like with a 5% threshold achieved the best performance, with a 100% of sensitivity, 96.55% of specificity, 97.87 % of accuracy, 0.96 MCC, and 100% of NPV (Supplementary Table 4). However, SSF-like alone and two more combinations including it also showed a 100% of NPV together with 100% sensitivity and high values of accuracy

TABLE 2 | Performance of the individual *in silico* tools in the discovery dataset.

	Sensitivity	Specificity	Accuracy	MCC	Positive Predictive Value	Negative Predictive Value	False Negative Rate	False Positive Rate	False Discovery Rate	False Omission Rate
<b>Donor (5')</b>										
HSF	<b>100.000</b>	96.154	<b>98.077</b>	<b>0.962</b>	96.296	<b>100.000</b>	<b>0.000</b>	3.846	3.704	<b>0.000</b>
SSF-like	96.154	96.154	96.154	0.923	96.154	96.154	3.846	3.846	3.846	3.846
MES	<b>100.000</b>	84.615	92.308	0.856	86.687	<b>100.000</b>	<b>0.000</b>	15.385	13.333	<b>0.000</b>
dbcsnV	91.667	90.000	91.176	0.795	95.652	81.818	8.333	10.000	4.348	18.182
NNS	92.308	80.769	86.538	0.735	82.759	91.304	7.692	19.231	17.241	8.696
SPANR	62.500	<b>100.000</b>	81.633	0.677	<b>100.000</b>	73.529	37.500	<b>0.000</b>	<b>0.000</b>	26.471
<b>Acceptor (3')</b>										
SSF-like	<b>100.000</b>	<b>89.655</b>	<b>93.617</b>	<b>0.877</b>	85.714	<b>100.000</b>	<b>0.000</b>	<b>10.345</b>	14.286	<b>0.000</b>
MES	<b>100.000</b>	86.207	91.489	0.839	81.818	<b>100.000</b>	<b>0.000</b>	13.793	18.182	<b>0.000</b>
dbcsnV	93.750	77.778	88.000	0.736	<b>88.235</b>	87.500	6.250	22.222	<b>11.765</b>	12.500
HSF	83.333	82.769	82.979	0.649	75.000	88.889	16.667	17.241	25.000	11.111
NNS	88.889	68.966	76.596	0.563	64.000	90.909	11.111	31.034	36.000	9.091
SPANR	41.176	88.460	69.760	0.343	70.000	69.697	58.824	11.538	30.000	30.303
<b>Global (5' and 3')</b>										
SSF-like	97.727	92.727	<b>94.949</b>	<b>0.900</b>	91.489	96.077	2.273	7.273	8.511	1.923
MES	<b>100.000</b>	85.455	91.919	0.850	84.615	<b>100.000</b>	<b>0.000</b>	14.545	15.385	<b>0.000</b>
HSF	93.182	89.091	90.909	0.818	87.234	94.231	6.818	10.909	12.766	5.769
dbcsnV	92.500	84.211	89.831	0.767	<b>92.500</b>	84.211	7.500	15.789	<b>7.500</b>	15.789
NNS	90.909	74.545	81.818	0.653	74.074	91.111	9.091	25.455	25.926	8.869
SPANR	53.659	<b>94.118</b>	76.087	0.533	88.000	71.642	46.341	<b>5.882</b>	12.000	28.358

Results of the performance evaluation is grouped by donor, acceptor or both splice sites. The best performance scores are highlighted in bold. False Discovery Rate represents the rate of false positives of the total of variants positively predicted and False Omission Rate represents the rate of false negatives of the total negative predicted variants. dbcsnV, database consultant for extracting the adaptive boosting and random forests scores.





(for predictions at acceptor site, **Supplementary Table 4**). Considering the tool combinations for predicting disruption caused by variants located in any of the two splice sites (global), MES and SSF-like sequential combination achieved the best accuracy with a 96.97% and 0.94 of MCC, followed for two combinations, including SSF-like and MES, which showed 100% sensitivity and 100% of NPV (**Supplementary Table 5**).

**Validation Set**

In order to validate the predictors with the best performance obtained in the discovery set, we analyzed a dataset of 346 variants with RNA *in vitro* results published or detailed in free available databases. At donor region, 210 variants were included, 177 showing *in vitro* splicing alterations (65 at intronic GT positions) and 33 showing no splicing effects (all outside intronic

GT) (**Figure 3** and **Supplementary Table 6**). One hundred thirty-six variants were located at the acceptor region, 95 showing splicing alterations (67 of them at intronic AG positions), and 41 with absence of alterations (40 of them outside intronic AG) (**Figure 3** and **Supplementary Table 7**). Only SSF-like and SPANR were able to identify all WT splice sites in reference sequences for all the genes of interest.

We selected for validation, the HSF stand-alone and the combinations HSF+SSF-like and HSF+SSF-like+MES for 5’ donor sites (**Supplementary Table 3**), and the SSF-like alone and the sequential MES and SSF combination for 3’ acceptor sites (**Supplementary Table 4**), considering sensitivity, accuracy, MCC and NPV scores. We excluded the combinations including SPANR or dbcsSNV since they do not provide predictions on insertions and deletions.

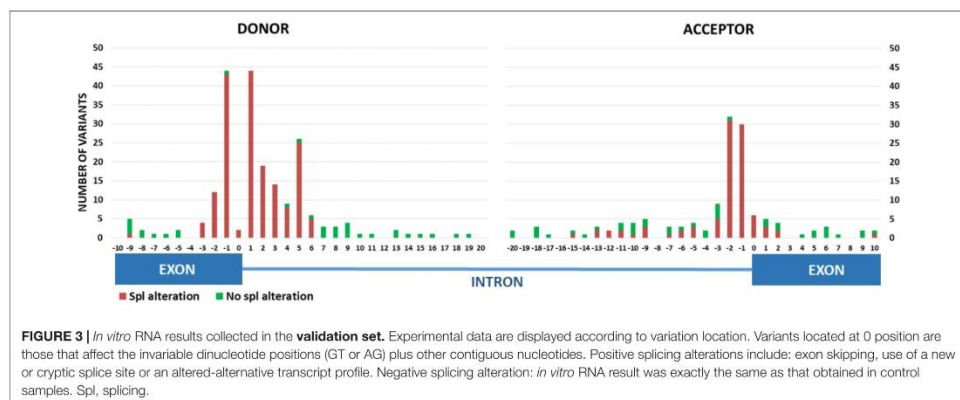
Overall, the *in silico* predictions in the validation dataset were more accurate for variants with effects on donor splice sites than acceptor sites (**Table 3** and **Figure 4**). These findings were in agreement with those results obtained with the discovery set (**Table 2**).

The data analysis indicated that for 5’ donor sites the best combinations, with 98.57% accuracy, 99.44% of sensitivity and 96.88% of NPV, are HSF+SSF-like or HSF+SSF-like+MES (**Table 3**) with very slight differences in performance, between the estimations of splicing effects for all variants (including variants placed at invariable dinucleotides) and for the group of variants located outside the two invariable nucleotides. For acceptor sites, the sequential combination of MES and SSF-like (Houdayer et al., 2012) and SSF-like stand-alone reached a performance with the same score of accuracy, 92.65%, but SSF-like showed a highest NPV (**Table 3**). Unlike the donor site, the accuracy of these predictors decreased (to 85.29%) when the variants analyzed did not include those at the two nucleotide invariables (AG) of the 3’ acceptor splice site (**Table 3**). For predictions of variants outside these dinucleotides, the rate of false negatives showed by SSF-like is slightly lower than those rates of MES and SSF-like sequential combination (25% versus 28.57%, respectively, **Table 3**).

**DISCUSSION**

The use of massive parallel sequencing in clinical diagnostics is leading to a significant increase in data and the detection of a high number of variants of uncertain significance (VUS) with potential effect on splicing which need interpretation. Therefore, prediction of the effect of DNA sequence variations on splicing using *in silico* tools has become a common approach. Several studies have been published on the performance and reliability of *in silico* predictions of the splicing impact of variants (Jian et al., 2014b). **Table 1** details the results obtained in these studies and shows that the recommendations provided about the most appropriate to be used are not concordant. However, the studies that give clear recommendations, always include one of the HSF, SSF, or MES programs, alternatively.

We have evaluated the reliability of *in silico* splicing effect predictions of six programs (MES, HSF, SSF-like, SPANR, NNSplice, and dbcsSNV) comparing their scores with splicing



*in vitro* analysis outcomes of variants identified in hereditary cancer related genes. We elaborated the study in two stages, discovery and validation, to identify the best predictors or the best combination for their application in routine clinical testing, taking into account the percentages reached for sensitivity, specificity, accuracy and NPV as well as the score of Mathews Coefficient Correlation (MCC).

In the discovery stage, significant performance differences were appreciated among individual tools (Table 2). For global, as well as for 5', and 3' splice sites, low accuracies of SPANR and NNSplice contrasted with the high performance achieved by SSF, MES, and HSF, while dbscSNV demonstrated an intermediate accuracy.

At the second stage of our study, we validated the combinations of HSF with SSF-like or HSF+SSF-like+MES as the highest performance for splicing aberrations at donor sites, and SSF-like stand-alone at acceptor sites (Table 3). All these results are in agreement with the trend observed in the previous published results, where HSF or SSF or MES outperformed other methods (Table 1). Of note, besides high accuracy and sensitivity, these validated tools, combined or as stand-alone, also had high NPV. This is relevant in a clinical setting, since it allows to separate the variants with an extremely low or non-existent probability of being abnormally spliceogenic from those variants in which *in vitro* RNA studies are of interest, with the consequent saving of resources in the laboratory.

All of the three predictors are available through Alamut Visual 2.10 (Interactive Biosoftware, Rouen), allowing a high throughput analysis, which is essential in a massive parallel sequencing annotation pipeline. Yet, in the newest version of Alamut Visual (2.11) the HSF predictor is not included in its splicing module, it is freely available at Human Splice Finder website<sup>3</sup> or through VarAFT software<sup>4</sup>, which allows the annotation of a large batch of variants. MES program is also freely

accessible via web<sup>5,6</sup>, although caution should be taken when obtaining predictions via Alamut or via web, since differences have been reported (Tang et al., 2016). SSF-like tool is currently only accessible through Alamut, yet it has been recently published a free program named Splicing Prediction in Consensus Elements (SPiCE<sup>7</sup>) that combines predictions from SSF-like and MES (Leman et al., 2018). On the other hand, SPANR and dbscSNV are free and could be easily implemented in a pipeline (Xiong et al., 2015; Liu et al., 2017), but these tools are not able to interpret splicing alterations caused by insertion or deletions (6.36% of validation set variants), which represents a limitation for their use compared to the other tools.

Non-canonical GC-AG and AT-AC sequences at the splice site invariant positions occur in 0.56 and 0.09% of the splice site pairs, respectively (Abramowicz and Gos, 2018). In the list of the genes that we analyzed, only six splice sites vary from the canonical splice site GT-AG: *ATM* exon 50 donor site (GC), *BRCA2* exon 17 donor site (GC), *MUTYH* exon 14 donor site (GC), *PALB2* exon 12 donor site (GC), *STK11* exon 2 donor site (AT) and exon 3 acceptor site (AC). In our validation dataset, we only had variants at atypical *BRCA2* exon 17 donor site (GC), and among the studied tools, only SSF-like and SPANR were able to identify these atypical splicing sites and made a prediction for variants located nearby. As the performance of SSF-like is better than SPANR, we suggest the use of SSF-like to analyze these non-canonical splicing sites.

The tools analyzed in this article have only been interrogated to predict alteration at donor and acceptor splice sites. However, alterations in RNA may be produced by variant effects on other factors in *cis* (branch points, polypyrimidine tract, intronic and exonic splicing silencers and enhancers) or create new splice sites or activate cryptic ones. At the stage of validation, the rate of false negative predictions is significantly higher for acceptor sites

<sup>3</sup><http://www.umd.be/HSF3/>

<sup>4</sup><https://varaft.eu/>

<sup>5</sup>[http://genes.mit.edu/burgelab/maxent/Xmaxentscan\\_scoreseq.html](http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html)

<sup>6</sup>[http://genes.mit.edu/burgelab/maxent/Xmaxentscan\\_scoreseq\\_acc.html](http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html)

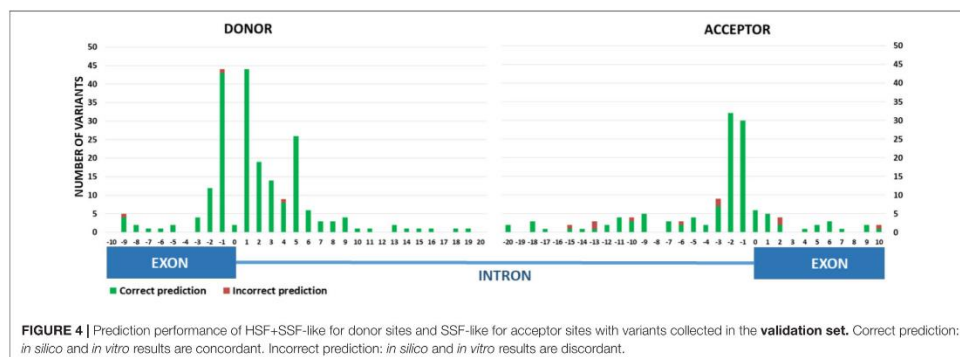
<sup>7</sup><https://sourceforge.net/projects/spicev2-1/>



TABLE 3 | Performance with the validation dataset of the best *in silico* tools previously selected from the results at discovery stage.

Donor	Sensitivity	Specificity	Accuracy	MCC	Positive predictive Value	Negative Predictive Value	False Negative Rate	False Positive Rate	False Discovery Rate	False Omission Rate
<b>HSF</b>										
All variants	96.045	90.909	95.238	0.831	98.266	81.081	3.955	9.091	1.734	18.919
Without invariable dinucleotides	94.643	90.909	93.793	0.830	97.248	83.333	5.357	9.091	2.752	16.667
<b>HSF+SSF-like</b>										
All variants	<b>99.435</b>	<b>93.939</b>	<b>98.571</b>	<b>0.946</b>	<b>98.876</b>	<b>96.875</b>	<b>0.565</b>	<b>6.061</b>	<b>1.124</b>	<b>3.125</b>
Without invariable dinucleotides	<b>99.107</b>	<b>93.939</b>	<b>97.931</b>	<b>0.941</b>	<b>98.230</b>	<b>96.875</b>	<b>0.893</b>	<b>6.061</b>	<b>1.770</b>	<b>3.125</b>
<b>HSF+SSF-like+MES</b>										
All variants	<b>99.435</b>	<b>93.939</b>	<b>98.571</b>	<b>0.946</b>	<b>98.876</b>	<b>96.875</b>	<b>0.565</b>	<b>6.061</b>	<b>1.124</b>	<b>3.125</b>
Without invariable dinucleotides	<b>99.107</b>	<b>93.939</b>	<b>97.931</b>	<b>0.941</b>	<b>98.230</b>	<b>96.875</b>	<b>0.893</b>	<b>6.061</b>	<b>1.770</b>	<b>3.125</b>
<b>Acceptor</b>										
<b>MES and SSF-like sequential</b>										
All variants	91.579	<b>95.122</b>	<b>92.647</b>	<b>0.837</b>	<b>97.753</b>	82.979	8.421	<b>4.878</b>	<b>2.247</b>	17.021
Without invariable dinucleotides	71.429	<b>95.000</b>	<b>85.294</b>	<b>0.699</b>	<b>90.909</b>	82.609	28.571	<b>5.000</b>	<b>9.091</b>	17.391
<b>SSF-like</b>										
All variants	<b>92.632</b>	92.683	<b>92.647</b>	0.832	96.703	<b>84.444</b>	<b>7.368</b>	7.317	3.297	<b>15.556</b>
Without invariable dinucleotides	<b>75.000</b>	92.500	<b>85.294</b>	0.695	87.500	<b>84.091</b>	<b>25.000</b>	7.500	12.500	<b>15.909</b>

The best performance scores are highlighted in bold. The atypical BRCA2 exon 17 native donor site (GC) was not estimated by HSF nor MES, and we have considered it as a failed prediction of the two tools for variants affecting this exon regardless of the *in vitro* splicing effect of the variant. False Discovery Rate represents the rate of false positives of the total variants positively predicted and False Omission Rate represents the rate of false negatives of the total negative predicted variants.



than for donor sites (Table 3). This difference may be due to the greater complexity of the sequence adjacent to the 3', with the presence of the branch point and the polypyrimidine tract. Therefore, variants located in these two last elements could alter RNA and not be detected as changes in the scores of the splicing sites computed by the predictors. For example, the variant c.1066-6T>G at *ATM* (included in the validation set), which is not predicted correctly by MES and SSF-like sequential combination (Supplementary Table 7), alters the polypyrimidine tract causing an aberrant transcript (Dörk et al., 2001).

Likewise, the *BRCA2* exonic variant c.467A>G, located nine nucleotides upstream from the 5' donor site, causes the loss of these last nine nucleotides, while the HSF and SSF-like predicts that their scores for the native donor splice site of 88.9 and 84.5, respectively, are not changed by the variant, which it is misinterpreted as a false negative (Supplementary Table 6). Using some of the tools analyzed in our study to identify enhanced cryptic sites or creation of new splice sites, the variant is predicted to cause a new donor site at nine nucleotides from 5', in concordance with *in vitro* results: SSF-like indicates a new donor site with a score of 96.9 against 84.5 of the natural splice site, MES 11.1 vs. 9.5 and HSF 98.2 vs. 88.9.

Furthermore, variants located in the exonic regions collected in our study could affect enhancer elements (ESEs) leading to an exon skipping, but they would not be correctly predicted by the analyzed tools. Although variants with specific experimental evidence of suffering this type of alteration were not included in our study, most articles consulted do not explicitly describe or exhaustively exclude the effect of ESEs. As an example, the *BRCA1* c.557C>A altering splicing variant gathered at validation set is not predicted to affect native acceptor site by SSF-like, but specific tools to predict splicing defect caused by regulatory sequence disruption indicates an ESE disturbance: ESRseq score of  $-1.567$  (Ke et al., 2011) and HEXplorer  $\Delta HZ_{EI} = -30.24$  (Erkelenz et al., 2014).

Computational tools or programs able to perform predictions on the disruption of all *cis* DNA elements would cover the whole landscape of aberrant RNA splicing yielded by spliceogenic VUS. Theoretically, SPANR is able to detect exon skipping caused by all

of the elements above mentioned, although our study indicated that this program has a low performance for at least to predict correctly alterations of donor and acceptor sites (Table 2). The HSF predictor accessed via its website<sup>8</sup>, also predicts the impact of genetic variations on branch point elements and has been improved for the identification of natural non-canonical splice sites (Oetting et al., 2018). The breast cancer genes PRIORS probabilities program<sup>9</sup>, gives MES estimations of disruption of natural splice sites and also computes the creation of new donor and acceptor splice sites using NNSplice, yet only for *BRCA1* and *BRCA2* genes (Vallée et al., 2016). However, the accuracy and performance of SPANR, HSF, and PRIORS predictions of variants placed in elements other than natural splice sites has not yet been evaluated.

To our knowledge, our study is the only that evaluates the accuracy of different tools separately for donor and acceptor sites, resulting in different recommendations for each one with high performance (Table 1).

One limitation of our study is the use of splicing *in silico* tools through a non-free commercial program, Alamut Visual 2.10, with the uncertainty of whether the predictions obtained through Visual Alamut are the same as those estimated directly by the tools in their respective free access websites. We have confirmed that HSF via web (see footnote 8; data not shown) and MES via SPICE (see footnote 7; Supplementary Table 8), at least for native splice sites, provide the same estimations than those provided by Alamut Visual 2.10. However, SSF-like predictions obtained through Alamut Visual 2.10 slightly differ from the predictions ascertained through SPICE (Supplementary Table 8). Therefore and considering our findings, we recommend as a free pipeline to use HSF accessed via web and MES via SPICE for donor and acceptor site predictions, respectively.

Another limitation is the higher number of variants causing splicing defects compared to the number of variants causing no

<sup>8</sup><http://www.umd.be/HSF3/>

<sup>9</sup><http://priors.hci.utah.edu/PRIORS/index.php>

splicing alteration in our validation dataset. This bias is due to a tendency to report only variants that cause splicing defects. Some studies, in order to avoid this bias, have included common single nucleotide polymorphisms (SNPs) from control dataset, assuming that they do not cause alterations (Table 1). Likewise, reports of RNA *in vitro* effects of variants in the two invariable dinucleotides GT-AG are overrepresented, while those located further from splice junctions are less frequently analyzed.

## CONCLUSION

In conclusion, to perform *in silico* analysis of VUS potentially affecting natural splice sites in hereditary cancer genes, we recommend the use of the HSF+SSF-like combination (with  $\Delta$ -2% and  $\Delta$ -5% as thresholds, respectively) for donor sites and SSF-like ( $\Delta$ -5%) stand-alone for acceptor sites. These tools have shown in the validation stage a high sensitivity and especially a high NPV. Although the *in vitro* study of RNA remains the gold standard to evaluate the process of splicing, and it is not recommended to use these predictions as the sole source of evidence to make clinical assertions (Richards et al., 2015), our results indicate that these combined tools can be used to filter out VUS with a very low probability of altering splicing without losing true spliceogenic variants that will need deeper experimental validation. Complementing the analysis using specific predictors to identify variants that could affect elements other than splice sites (such as branch points or ESEs), may be useful for the screening of the whole RNA defect landscape. Lastly, it is worth stating that (i) the aim of this work was not to classify variants but to provide an *in silico* algorithm with the highest performance to predict an altered *in vitro* splicing regardless of whether the variants are benign or pathogenic; and (ii) the detection of splicing defect does not automatically denote the pathogenicity of the variant for which a comprehensive qualitative and quantitative RNA analysis is warranted as highlighted in ENIGMA<sup>10</sup> or ACGM guidelines (Richards et al., 2015) for variant classification.

## AUTHOR CONTRIBUTIONS

AM-F, LD-L, SG-E, and OD: conception or design of the work. AM-F, LD-L, GM, SB, IL-P, MM, MS, RB, AB, EC, AL-F, NS, and

<sup>10</sup> <https://enigmaconsortium.org/>

## REFERENCES

- Abramowicz, A., and Gos, M. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.* 59, 253–268. doi: 10.1007/s13353-018-0444-7
- Cartegni, L., Chew, S. L., and Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.* 3, 285–298. doi: 10.1038/nrg775
- Colombo, M., Blok, M. J., Whiley, P., Santamariña, M., Gutiérrez-Enriquez, S., Romero, A., et al. (2014). Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the

MP: acquisition of data for the work. AM-F, AV, CL, MP, GC, MdH, JB, SG-E, and OD: data analysis and interpretation. AM-F, SG-E, and OD: drafting the work. All authors: critical revision of the article and final approval of the version to be published.

## FUNDING

This work was supported by Spanish Instituto de Salud Carlos III (ISCIII) funding, an initiative of the Spanish Ministry of Economy and Innovation partially supported by European Regional Development FEDER Funds: PI15/00355 (to OD), PI16/01218 (to SG-E), PI15/00059 (to MdH), PI16/00563 (to CL), SAF2015-68016-R (to GC and MP), CIBERONC (to GC), INT15/00070, INT16/00154, and INT17/00133 (to AV). The Catalan Institute of Oncology (ICO) work was supported by the Government of Catalonia [Pla estratègic de recerca i innovació en salut (PERIS), 2017SGR1282 and 2017SGR496]; and the Scientific Foundation Asociación Española Contra el Cáncer. ICO thanks CERCA Program/Generalitat de Catalunya for institutional support. This work was partially funded by CIBERER (ER17P1AC7112/2017) and Xunta de Galicia (IN607B) funds given to AV. SG-E and SB were supported by the Miguel Servet Program (CP10/00617) and Asociación Española Contra el Cáncer (AECC) contract, respectively. RB was supported by European Union's Horizon 2020 research and innovation program under grant agreement N° 634935.

## ACKNOWLEDGMENTS

We thank Xavier de la Cruz for helpful discussions and Leo Judkins for English language-editing. We acknowledge the Cellex Foundation for providing research facilities and equipment. We also thank the participating patients and families and all the members of the Units of Genetic Counselling and Genetic Diagnostic the Hereditary Cancer Program of the Catalan Institute of Oncology (ICO-IDIBELL).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2018.00366/full#supplementary-material>

ENIGMA consortium. *Hum. Mol. Genet.* 23, 3666–3680. doi: 10.1093/hmg/ddu075

- Colombo, M., de Vecchi, G., Caleca, L., Foglia, C., Ripamonti, C. B., Ficarazzi, F., et al. (2013). Comparative *in vitro* and *in silico* analyses of variants in splicing regions of BRCA1 and BRCA2 genes and characterization of novel pathogenic mutations. *PLoS One* 8:e57173. doi: 10.1371/journal.pone.0057173
- Desmet, F. O., Hamroun, D., Collod-Bérout, G., Claustres, M., and Bérout, C. (2010). "Bioinformatics identification of splice site signals and prediction of mutation effects," in *Research Advances in Nucleic Acids Research*, ed. R. M. Mohan (Kerala: Global Research Network), 1–16.



- Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Bérout, G., Claustres, M., and Bérout, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 37, 1–14. doi: 10.1093/nar/gkp215
- Dörk, T., Bendix, R., Bremer, M., Rades, D., Klöpffer, K., Bremer, M., et al. (2001). Spectrum of ATM gene mutations in a hospital-based series of unselected breast cancer patients. *Cancer Res.* 61, 7608–7615.
- Eccles, E. B., Mitchell, G., Monteiro, A. N. A., Schmutzler, R., Couch, F. J., Spurdle, A. B., et al. (2015). BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance. *Ann. Oncol.* 26, 2057–2065. doi: 10.1093/annonc/mdv278
- Erkelenz, S., Theiss, S., Otte, M., Widera, M., Peter, J. O., and Schaal, H. (2014). Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Res.* 42, 10681–10697. doi: 10.1093/nar/gk/ u736
- Holla, Ö.L., Nakken, S., Matningsdal, M., Ranheim, T., Berge, K. E., Defesche, J. C., et al. (2009). Effects of intronic mutations in the LDLR gene on pre-mRNA splicing: comparison of wet-lab and bioinformatics analyses. *Mol. Genet. Metab.* 96, 245–252. doi: 10.1016/j.ymgme.2008.12.014
- Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., et al. (2012). Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Hum. Mutat.* 33, 1228–1238. doi: 10.1002/humu. 22101
- Houdayer, C., Dehainault, C., Mattler, C., Michaux, D., Caux-Moncoutier, V., Pagès-Berhouet, S., et al. (2008). Evaluation of in silico splice tools for decision-making in molecular diagnosis. *Hum. Mutat.* 29, 975–982. doi: 10.1002/humu. 20765
- Jian, X., Boerwinkle, E., and Liu, X. (2014a). In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* 42, 13534–13544. doi: 10.1093/nar/gku1206
- Jian, X., Boerwinkle, E., and Liu, X. (2014b). In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet. Med.* 16, 497–503. doi: 10.1038/gim.2013.176
- Ke, S., Shang, S., Kalachikov, S. M., Morozova, I., Yu, L., Russo, J. J., et al. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* 21, 1360–1374. doi: 10.1101/gr.119628.110
- Leman, R., Gaildrat, P., Gac, G. L., Ka, C., Fichou, Y., Audrezet, M., et al. (2018). Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico / in vitro studies: an international collaborative effort. *Nucleic Acids Res.* doi: 10.1093/nar/gky372 [Epub ahead of print].
- Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2017). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human non-synonymous and splice site SNVs. *Hum. Mutat.* 37, 235–241. doi: 10.1002/humu.2 2932
- Moreno, L., Linossi, C., Esteban, I., Gadea, N., Carrasco, E., Bonache, S., et al. (2016). Germline BRCA testing is moving from cancer risk assessment to a predictive biomarker for targeting cancer therapeutics. *Clin. Transl. Oncol.* 18, 981–987. doi: 10.1007/s12094-015-1470-0
- Oetting, W. S., Bérout, C., Brenner, S. E., Greenblatt, M. S., Karchin, R., and Mooney, S. D. (2018). Methods and tools for assessing the impact of genetic variations. *Hum. Mutat.* 39, 454–458. doi: 10.1002/humu.23393
- Reese, M. G., Eckman, F. H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in gene. *J. Comput. Biol.* 4, 311–323. doi: 10.1089/cmb.1997.4.311
- Rhine, C. L., Cygan, K. J., Soemedi, R., Maguire, S., Murray, M. F., Monaghan, S. F., et al. (2018). Hereditary cancer genes are highly susceptible to splicing mutations. *PLoS Genet.* 14:e1007231. doi: 10.1371/journal.pgen.1007231
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., and Gastier-Foster, J. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the association for molecular pathology. *Genet. Med.* 17, 405–424. doi: 10.1038/gim.2015.30
- Shapiro, M. B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* 15, 7155–7174. doi: 10.1093/nar/15.17.7155
- Soukariéh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frébourg, T., et al. (2016). Exonic splicing mutations are more prevalent than currently estimated and can be predicted by using in silico tools. *PLoS Genet.* 12:e1005756. doi: 10.1371/journal.pgen.1005756
- Tang, R., Prosser, D. O., and Love, D. R. (2016). Evaluation of bioinformatic programmes for the analysis of variants within splice site consensus regions. *Adv. Bioinformatics* 2016:5614058. doi: 10.1155/2016/5614058
- Théry, J. C., Krieger, S., Gaildrat, P., Révillon, F., Buisine, M. P., Killian, A., et al. (2011). Contribution of bioinformatics predictions and functional splicing assays to the interpretation of unclassified variants of the BRCA genes. *Eur. J. Hum. Genet.* 19, 1052–1058. doi: 10.1038/ejhg.2011.100
- Vallée, M. P., Di Sera, T. L., Nix, D. A., Paquette, A. M., Parsons, M. T., Bell, R., et al. (2016). Adding in silico assessment of potential splice aberration to the integrated evaluation of BRCA gene unclassified variants. *Hum. Mutat.* 37, 627–639. doi: 10.1002/humu.22973
- Vreeswijk, M. P., Kraan, J. N., Van Der Klift, H. M., Vink, G. R., Cornelisse, C. J., Wijnen, J. T., et al. (2009). Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs. *Hum. Mutat.* 30, 107–114. doi: 10.1002/humu.20811
- Whiley, P. J., de la Hoya, M., Thomassen, M., Becker, A., Brandao, R., Pedersen, I. S., et al. (2014). Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. *Clin. Chem.* 60, 341–352. doi: 10.1001/jamasurg.2014.1086
- Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347:1254806. doi: 10.1126/science.1254806
- Yeo, G., and Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* 11, 377–394. doi: 10.1089/1066527041410418

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Moles-Fernández, Duran-Lozano, Montalban, Bonache, López-Perolico, Menéndez, Santamariña, Behar, Blanco, Carrasco, López-Fernández, Sijepanovic, Balmaña, Capellá, Pineda, Vega, Lázaro, de la Hoya, Diez and Gutiérrez-Enriquez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



## Article 2

### Role of Splicing Regulatory Elements and *In Silico* Tools Usage in the Identification of Deep Intronic Splicing Variants in Hereditary Breast/Ovarian Cancer Genes.

Moles-Fernández A, Domènech-Vivó J, Tenés A, Balmaña J, Diez O, Gutiérrez-Enríquez S.

**Cancers (Basel)**. 2021 Jul 3;13(13):3341.

doi: 10.3390/cancers13133341.

Article

# Role of Splicing Regulatory Elements and In Silico Tools Usage in the Identification of Deep Intronic Splicing Variants in Hereditary Breast/Ovarian Cancer Genes

Alejandro Moles-Fernández <sup>1,†</sup>, Joanna Domènech-Vivó <sup>1,†</sup> , Anna Tenés <sup>2</sup>, Judith Balmaña <sup>1,3</sup>, Orland Diez <sup>1,2,\*</sup> and Sara Gutiérrez-Enriquez <sup>1,\*</sup> 

<sup>1</sup> Hereditary Cancer Genetics Group, Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron Barcelona Hospital Campus, 08035 Barcelona, Spain; amoles@vhio.net (A.M.-F.); jdomenech@vhio.net (J.D.-V.); jbalmana@vhio.net (J.B.)

<sup>2</sup> Area of Clinical and Molecular Genetics, Vall d'Hebron Hospital Universitari, Vall d'Hebron Barcelona Hospital Campus, 08035 Barcelona, Spain; atenes@vhebron.net

<sup>3</sup> Medical Oncology Department, Vall d'Hebron Hospital Universitari, Vall d'Hebron Barcelona Hospital Campus, 08035 Barcelona, Spain

\* Correspondence: odiez@vhio.net (O.D.); sgutierrez@vhio.net (S.G.-E.)

† Co-first authors.

‡ Co-last authors.



**Citation:** Moles-Fernández, A.; Domènech-Vivó, J.; Tenés, A.; Balmaña, J.; Diez, O.; Gutiérrez-Enriquez, S. Role of Splicing Regulatory Elements and In Silico Tools Usage in the Identification of Deep Intronic Splicing Variants in Hereditary Breast/Ovarian Cancer Genes. *Cancers* **2021**, *13*, 3341. <https://doi.org/10.3390/cancers13133341>

Academic Editor: David Wong

Received: 25 May 2021

Accepted: 29 June 2021

Published: 3 July 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Simple Summary:** There is a significant percentage of hereditary breast and ovarian cancer (HBOC) cases that remain undiagnosed, because no pathogenic variant is detected through massively parallel sequencing of coding exons and exon-intron boundaries of high-moderate susceptibility risk genes. Deep intronic regions may contain variants affecting RNA splicing, leading ultimately to disease, and hence they may explain several cases where the genetic cause of HBOC is unknown. This study aims to characterize intronic regions to identify a landscape of “exonizable” zones and test the efficiency of two in silico tools to detect deep intronic variants affecting the mRNA splicing process.

**Abstract:** The contribution of deep intronic splice-altering variants to hereditary breast and ovarian cancer (HBOC) is unknown. Current computational in silico tools to predict spliceogenic variants leading to pseudoexons have limited efficiency. We assessed the performance of the SpliceAI tool combined with ESRseq scores to identify spliceogenic deep intronic variants by affecting cryptic sites or splicing regulatory elements (SREs) using literature and experimental datasets. Our results with 233 published deep intronic variants showed that SpliceAI, with a 0.05 threshold, predicts spliceogenic deep intronic variants affecting cryptic splice sites, but is less effective in detecting those affecting SREs. Next, we characterized the SRE profiles using ESRseq, showing that pseudoexons are significantly enriched in SRE-enhancers compared to adjacent intronic regions. Although the combination of SpliceAI with ESRseq scores (considering  $\Delta$ ESRseq and SRE landscape) showed higher sensitivity, the global performance did not improve because of the higher number of false positives. The combination of both tools was tested in a tumor RNA dataset with 207 intronic variants disrupting splicing, showing a sensitivity of 86%. Following the pipeline, five spliceogenic deep intronic variants were experimentally identified from 33 variants in HBOC genes. Overall, our results provide a framework to detect deep intronic variants disrupting splicing.

**Keywords:** spliceogenic deep intronic variants; pseudoexons; cryptic splice sites; splicing regulatory elements; hereditary breast ovarian cancer; in silico prediction tools

## 1. Introduction

Pathogenic variants in the tumor suppressor genes *BRCA1* and *BRCA2* (*BRCA1/2*) and other genes, mainly involved in DNA repair, have been linked to high or moderate risks of developing hereditary breast and ovarian cancer (HBOC) [1,2]. The identification of

pathogenic variants in these genes offers patients and families precise clinical management based on personalized prevention and therapeutic strategies [3]. However, there is still a significant fraction of cases for which the genetic analysis does not identify causative variants underlying their predisposition to breast and/or ovarian cancer [4–6].

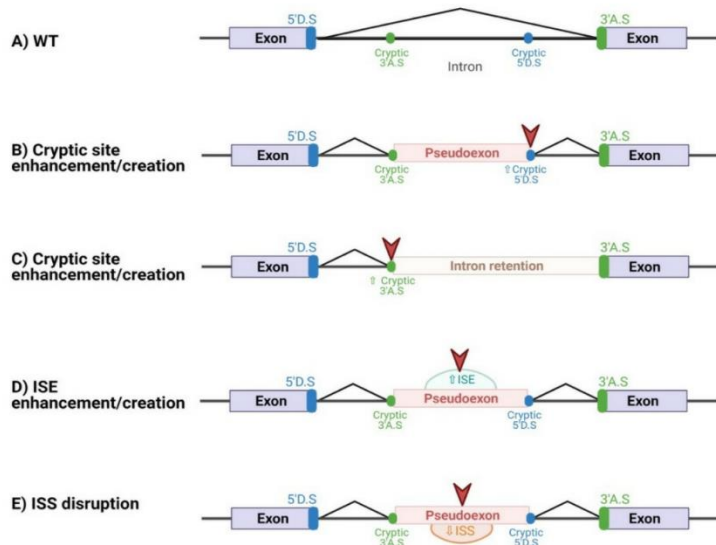
Currently, the detection of pathogenic variants is addressed mainly by massively parallel sequencing of high-moderate penetrance gene panels. An important number of identified deleterious variants affect pre-mRNA splicing and interestingly, hereditary cancer genes (including some HBOC and Lynch syndrome genes) are enriched for this type of variants [7]. The spliceogenic variants may occur in both introns and exons and disrupt consensus “cis” sequences such as canonical splice site nucleotides, branch point, polypyrimidine tract motifs, and Splicing Regulatory Elements (SREs). SREs are sequences that act as splicing intronic/exonic enhancers (ISE/ESE) or silencers (ISS/ESS), binding (SR)-rich proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs), respectively [8,9]. Exon definition is the initial step in pre-mRNA splicing, and it has been suggested that accurate splice site recognition resides in a SRE balance, i.e., exons enriched with Exonic Splicing Enhancers (ESEs) and introns with Intronic Splicing Silencers (ISSs) [10–12].

In contrast to several studies showing the spliceogenic effect of exonic variants, there is a lack of information about the frequency of deleterious variants occurring in deep intronic regions ( $\pm 20$  bp from canonical splicing sites) since conventional genetic diagnosis is usually restricted to coding exons and flanking intronic regions. However, nucleotide changes in these regions could generate aberrant transcripts by introducing intronic sequences in mature mRNA [13–15]. In fact, pathogenic deep intronic variants have been described in more than 75 disease-associated genes including monogenic disorders such as hereditary cancer syndromes [16]. Deep intronic variants can alter splicing by two different mechanisms [17]: the creation/enhancement of cryptic splice sites and the alteration of an intronic SRE by the disruption of an ISS or the creation/strengthening of an ISE (Figure 1). A few examples of diseases driven by the inclusion of a pseudoexon due to these phenomena are HBOC caused by the c.4185 + 4105C > T variant in *BRCA1*, the first reported deep intronic variant in this gene that activates a pre-existing cryptic donor site [18], and the Ataxia-telangiectasia disease due to the c.2839-581\_2839-578del variant in the *ATM* gene, which creates an ISE [19]. These examples highlight the relevance of screening deep intronic regions in HBOC patients to identify germline pathogenic variants leading to an aberrant RNA processing.

Given that experimental testing of all possible spliceogenic detected variants is currently not feasible in a clinical setting, multiple computational prediction tools have been developed to infer their effect and hence to prioritize variants to be experimentally evaluated [20]. Moreover, computational predictions of splicing variants are part of the supporting evidence included in the variant interpretation guidelines of the American College of Medical Genetics and Genomics (ACMG) [21]. Although different *in silico* tools have been published that accurately identify splicing exonic variants affecting canonical splice sites (MES, SSF, HSF) [21–24] or altering SREs (ESRseq, HZ<sub>EL</sub>, HAL) [11,25–27], there has been limited success in identifying deep intronic variants [18,27,28].

Deep learning tools such as SpliceAI, could outperform classical prediction approaches [29], but little is known about its performance for identifying deep intronic variants affecting splicing either by creating *de novo*/enhancing splice sites or SREs [27,28]. Our work assesses the performance of the SpliceAI tool combined with ESRseq scores to identify spliceogenic deep intronic variants using literature and experimental datasets.





**Figure 1.** Splicing effects caused by deep intronic variants. (A) Normal splicing using natural splicing sites. (B) Deep intronic variant creating/enhancing a cryptic splice site, resulting in the inclusion of a pseudoexon by using a complementary cryptic site. (C) Intronic retention caused by a deep intronic variant that creates/enhances a cryptic site, which is used instead of the canonical splice site. (D) Deep intronic variant creating/enhancing an ISE, resulting in the inclusion of a cryptic exon using two cryptic splice sites. (E) Deep intronic variant disrupting an ISS, resulting in the inclusion of a cryptic exon using two cryptic splice sites.

Moreover, the arrangement of *cis* splicing elements in deep intronic regions, especially those corresponding to regulatory elements, would configure sequences with structures similar to those of canonical exons [30,31]. This landscape of regulatory zones would favor the generation of pseudoexons if a new deep intronic variant helps to define these structures. Thus, in this study we also compare the SRE presence between canonical exons of HBOC genes and published pseudoexons in order to characterize an “exon-like landscape” across introns which would help to identify potentially “exonizable” intronic regions.

To our knowledge, we provide the first *in silico* framework to prioritize deep intronic variants for their experimental RNA analysis, taking into account the landscape of splicing elements and highlighting the importance of SRE balance in the inclusion of sequences in mature RNA.

## 2. Materials and Methods

### 2.1. Datasets

Four datasets were used in this study: (i) 233 blood-detected germline deep intronic variants (located >20 nt from known exon-intron boundaries) collected from literature using the keywords “pseudoexon” and “deep intronic” (Table S1). Their splicing effect had been assessed experimentally by gel electrophoresis or Sanger sequencing using blood, lymphoblastoid cell lines, midgenes, or minigenes. We defined as deep intronic variants those that were more than 20 nt away from the nearest exon-intron junction because these nucleotides are outside of the known splice site consensus sequences [8,9]; (ii) 1161 exonic variants compiled by Tubeuf et al. [27] to compare the accuracy of SREs-dedicated algorithms for predicting splicing alteration by affecting SREs (Table S2A); (iii) an additional supportive dataset of 207 somatic deep intronic splice-altering variants detected

and characterized in tumor by both whole-genome sequencing and RNA-sequencing, retrieved from Jung et al. [32] (Table S3); and (iv) an experimental dataset from patients of Hospital Universitari Vall d'Hebron, comprised of 33 unique intronic germline variants and selected according to their blood RNA availability (Table S4). The splicing impact of all 33 variants was assessed in silico with SpliceAI and ESRseq and experimentally characterized using whole blood RNA.

Additionally, we retrieved the sequences of all exons and adjacent 100 intronic nucleotides of HBOC and Lynch genes (*BRCA1*, *BRCA2*, *PALB2*, *BRIP1*, *RAD51C*, *RAD51D*, *PTEN*, *TP53*, *CDH1*, *CHEK2*, *BARD1*, *STK11*, *MSH2*, *MSH6* and *MLH1*) according to their NCBI reference transcripts (Table S5), to compare the splicing regulatory balance between canonical exons and pseudoexons from the literature dataset. Taking into account that most exons contain less than 300 nt, and that the mean length is 147 nt [33], exceptionally long exons (e.g., exons 10 of *BRCA1* and 10 and 11 of *BRCA2*) were not included.

## 2.2. In Silico Variant Annotation and Analysis for All Datasets

SpliceAI (v1.3; <https://github.com/Illumina/SpliceAI> accessed on 1 January 2021) was run to obtain the delta ( $\Delta$ ) score of a variant (DS), defined as the maximum of DS for acceptor gain (DS\_AG) and DS for donor gain (DS\_DG) for deep intronic variants and the maximum of DS for acceptor loss (DS\_AL) and DS for donor loss (DS\_DL) for exonic variants [34]. The DS value ranges from 0 to 1 and can be interpreted as the probability of a variant being splice-altering. We considered the SpliceAI estimations of a gained/lost splice site in the 4999 nucleotides located on each side of the variant.

For the prediction of variant-induced SREs alterations, the  $\Delta$ ESRseq value was calculated according to Ke., et al. [11] as the difference between the ESRseq scores of a variant sequence of 11 nucleotides (5 nucleotides at each side of the variant) and wild type sequence scores (ESRseq VAR- ESRseq WT).

To get the SRE distribution along any genomic region, we focused exclusively on the ESRseq scores, which were also calculated according to Ke., et al., obtaining individual nucleotide scores of each one of the positions of a sequence [11] (Table S6). The sum of the scores for each nucleotide in a given sequence was defined as "area". To account for the differences in size between exons, the area was divided by the number of nucleotides of each exon obtaining a normalized SRE area value (Normalized SRE area =  $\sum$  ESRseq scores of all nt of region of interest/size of region of interest). This value was used to compare the SREs of all constitutive exonic sequences from HBOC and Lynch genes, with all pseudoexons collected in the literature dataset.

Alamut Visual software v.2.10 (Interactive Biosoftware) was used for annotation of variants included in the patients' experimental validation dataset, providing data of allele frequencies in general population from the Genome Aggregation Database (gnomAD 2.1) and variant classification reported in ClinVar database (<https://www.ncbi.nlm.nih.gov/clinvar/> accessed on 1 March 2021), considering the number of variant submitters and reviews by expert panel.

## 2.3. Statistical Analysis

Performance values of sensitivity, specificity, overall accuracy, Positive and Negative Predictive Values, False and Positive Discovery Rates, and Matthews correlation coefficient (MCC) were calculated with different in silico tool thresholds individually and in sequential combinations. The statistical measures used for evaluation of the performance are depicted in Table S7.

The SpliceAI (4999 bp) optimized threshold was calculated based on the highest MCC using the literature dataset. In the case of ESRseq, the cut-off optimization was estimated by maximizing the sum of specificity and sensitivity since in the optimization based on MCC, the sensitivity/specificity was too unbalanced.

Analysis of variance (ANOVA) and T-test were used to compare the means of SRE scores between canonical exons, adjacent intronic sequences and pseudoexons with their

adjacent sequences. *T*-test was used to compare the absolute SRE differences between the group of variants causing pseudoexons and variants without any effect. All tests were performed using Graphpad Prism 6.

#### 2.4. Experimental RNA Analysis in Patient's Data Set

The patient dataset included unrelated cases from HBOC families ascertained through the Familiar Cancer Unit of Hospital Universitari Vall d'Hebron, HUVH (Barcelona, Spain). A total of 33 unique intronic germline variants were selected based on blood RNA availability in *ATM*, *BARD1*, *BRCA2*, *FAM175A*, *MLH1*, *MSH2*, *MUTYH*, *NF1*, *PTEN*, *RAD51C*, *RBBP8*, and *TP53* (Table S4). All variants were identified in DNA by massively parallel sequencing using Illumina technology with a diagnostic routine panel of coding exons and exon-intron boundaries or by a research panel specifically designed to sequence whole intronic regions and confirmed by Sanger sequencing [4,5]. Healthy individuals without familial cancer history were included as negative controls.

##### 2.4.1. Reverse Transcription-PCR (RT-PCR) and Sanger Sequencing

Total RNA from variant carriers and controls was isolated from 10 mL of peripheral blood using Trizol reagent (Invitrogen, Waltham MS, USA) following the manufacturer's protocol. RNA was cleaned-up using RNeasy Mini Kit (QIAGEN, Hilden, Germany) with an additional step of DNase digestion using RNase-Free DNase Set (QIAGEN) or Ambion™ DNase I RNase-free (ThermoFisher, Waltham, MS, USA) in samples with a low RNA concentration. A total of 100 ng of RNA were retrotranscribed to yield cDNA using PrimeScript RT reagent kit (Takara Bio, Shiga, Japan), combining random and oligo-dT primers. PCR primers were designed to amplify a whole exon upstream and downstream from the intron containing the variant of interest. PCR assays were performed in 25 µL reaction volume containing 50 ng of cDNA as template, using BioTaq DNA Polymerase (Meridian Bioscience, Cincinnati OH, USA). Samples were denatured at 95 °C for 10 min, followed by 35 cycles consisting of 95 °C for 30 s, 56–62 °C for 30 s, and 72 °C for 1–7 min; and a final extension step at 72 °C for 7 min. All primers used in this study and amplification conditions are detailed in Table S8.

Capillary electrophoresis using the 4200 TapeStation device (Agilent, Santa Clara CA, USA) with High Sensitivity D1000 ScreenTape reagents (Agilent) was used to assess the quality of PCR products. These products were enzymatically cleaned using ExoSAP-IT™ PCR Product Cleanup (Affimetrix, ThermoFisher Scientific, Waltham, MS, USA) and bidirectionally sequenced using BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Waltham, MS, USA). Sequencing products were run in an ABI3130xl Genetic Analyzer (Applied Biosystems) and were analyzed using Sequencing Analysis v6.0 software (Applied Biosystems). The reference transcripts based on GRCh37 (hg19) genome and listed in the Table S8 were used for sequence alignment and transcript annotation.

##### 2.4.2. Qualitative Analysis by Capillary Electrophoresis of Fluorescent Amplicons

RT-PCRs using primers labelled with 6-Carboxyfluorescein (6-FAM) at the 5' end were performed in triplicate (see labelled primers in Table S8). These fluorescent products were assessed by high-resolution capillary electrophoresis to detect and annotate all amplified transcripts. Specifically, 0.5 µL of the PCR products were run in an ABI3130xl Genetic Analyzer instrument (Applied Biosystems) for fragment analysis. GeneScan 500 and 1000 ROX (Applied Biosystems) was used as internal size-standard. Electrophoresis conditions were the same for all samples: 60 °C, 12 s injection at 1.2 KV and 2000 s run at 12 KV. Data visualization and peak size-calling was performed using GeneMapper software v5.0 (Applied Biosystems).

The maximum fragment size that could be detected was 946 bp, using the internal size-standard 1000 ROX.



### 2.5. Editorial Policies and Ethical Considerations

This study was approved by the Clinical Research Ethics Committee (CEIC) of Hospital Universitari Vall d'Hebron, Barcelona, Spain. All individuals received genetic counseling and signed written informed consent for HBOC panel genetic testing and research studies.

## 3. Results

### 3.1. SpliceAI Optimally Predicts Deep Intronic Splice-Altering Variants but with Less Sensitivity Those Affecting Splicing by Altering Regulatory Elements

To establish the performance of SpliceAI in predicting deep intronic pseudoexon-generating variants, we interrogated a set of variants collected from the literature, after searching for variants located beyond 20 nucleotides from exon-intron boundaries and for which RNA data was available (Table S1). This collection contains 233 deep intronic variants from 80 HBOC and other rare mendelian disease genes, including 133 variants that promote the creation of pseudoexons or intron retention events and 100 variants that do not alter splicing. Once the delta ( $\Delta$ ) score for each variant was obtained by running SpliceAI (v1.3), it was compared with the experimental results in RNA to estimate sensitivity, specificity, and accuracy. Then, we estimated the threshold at which the performance of the tool was more optimal, obtaining at  $\Delta$  score of 0.05 the highest MCC of 0.86 (Table 1 and Figure S1).

**Table 1.** Performance of SpliceAI for literature database with 233 deep intronic variants using the optimized threshold of 0.05.

Dataset	Splicing Altering Variants	No Splicing Altering Variants	Sensitivity	Specificity	Accuracy	MCC
All variants	133	100	93.99	92.00	93.13	0.86
Cryptic splice	117	100	95.73	92.00	94.01	0.88
SREs altering	16	100	81.25	92.00	90.52	0.66

Since SpliceAI was specifically developed to detect altering splicing variants by activating or creating cryptic splice sites [34], we reassessed its predictions separately considering two groups of intronic-splice altering variants according to the *cis* element affected. We obtained better results with the group of cryptic splice variants than with the SRE disruptive variants (0.88 vs. 0.66 of MCC, respectively; Table 1 and Figure S1).

The lower sensitivity showed by SpliceAI in predicting the impact of only 16 deep intronic variants on SREs prompted us to assess its performance with a large previous published dataset, composed of 360 exonic variants that affect splicing by altering SREs and 801 exonic variants without effect on splicing (Table S2A) [27]. The performance with our pre-established 0.05 cut-off was 69.16% sensitivity, 84.27% specificity and 0.53 MCC, while with the cut-off of 0.06 showing the highest MCC the prediction improves, reaching 0.548 MCC (Table S2B and Figure S2).

To supplement the performance of SpliceAI in identifying deep intronic variants disrupting SREs, we consecutively added the ESRseq evaluation. ESRseq is a computational algorithm specifically developed to predict SREs disruption that showed the best performance in predicting both variant-induced exon skipping and exon inclusion in a recent benchmarking study [27]. To do this, the  $\Delta$ ESRseq values (differences between wild type (WT) ESRseq score and variant ESRseq score) were calculated as described in Ke et al. [11] for intronic variants negatively predicted by SpliceAI ( $<0.05$ ). The threshold optimization by maximizing the sum of specificity and sensitivity, indicated that variants with a score change equal or higher than 0.63 were predicted to promote pseudoexon inclusion by altering SREs with higher sensitivity and specificity (Table S9). The performance of the sequential pipeline applying to those variants with a SpliceAI  $\Delta$  score of  $<0.05$  and the  $\Delta$ ESRseq threshold of  $\geq 0.63$ , showed higher sensitivity (96.24%) than those obtained by SpliceAI alone, but lower specificity (69%) and MCC (0.69) (Table 2). We next sought to

know if the optimized SpliceAI cut-off of 0.05 with the whole set of variants, regardless of the *cis*-element affected, would be higher if they were evaluated using ESRseq primarily, i.e., SpliceAI would exclusively compute for variants affecting cryptic sites. However, this analysis of assessing the 233 intronic variants firstly with ESRseq and then with SpliceAI, led to the same previous optimized SpliceAI threshold of 0.05.

**Table 2.** Prediction performances of SpliceAI alone and sequentially combined with ESRseq tool with 233 intronic variants from the literature database (133 altering and 100 non altering splicing). Abs. Dif.: absolute difference.

Pipeline	Sensitivity	Specificity	Accuracy	MCC
(1) $\Delta$ SpliceAI $\geq$ 0.05	93.98	92.00	93.13	0.86
(2) $\Delta$ SpliceAI $\geq$ 0.05 + SpliceAI $<$ 0.05 and $\Delta$ ESRseq $\geq$ 0.63	96.24	69.00	84.55	0.69
(3) $\Delta$ SpliceAI $\geq$ 0.05 + SpliceAI $<$ 0.05 and $\Delta$ ESRseq $\geq$ 0.63 and Abs. Dif. 0.51	95.49	86.00	91.42	0.83

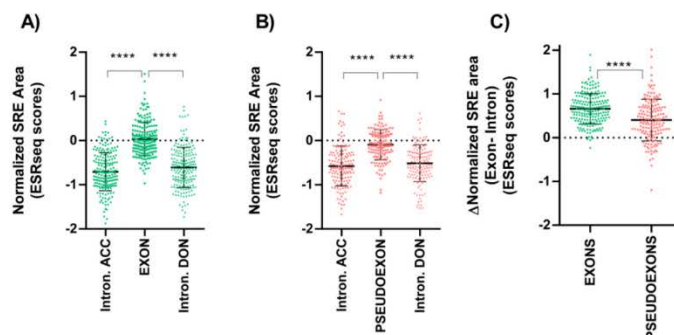
### 3.2. Splicing Regulatory Elements Balance Is Similar between Pseudoexons and Canonical Exons

To further investigate the role of splicing regulatory elements (SREs) in RNA included sequences, we compared the SREs landscape between constitutive exons and pseudoexons.

First, we extracted, from canonical transcripts of HBOC and Lynch genes (Table S5), the respective constitutional exon sequences and the 100 adjacent intron nucleotides for each gene. Next, an ESRseq value was assigned to each nucleotide according to Ke et al. [11], thus obtaining a map of the distribution of regulatory elements along the different exons. In addition, we calculated the sum of the total values (area), and the normalized SRE area score (sum of total scores/number of region length nucleotides), of each exonic and adjacent intronic region for each gene. In Table S6, we show an example of the obtained data for *BARD1* gene; data for all other genes is available upon request. Comparing the values obtained from canonical exons and adjacent upstream and downstream introns, significant differences were observed. Exons were enriched in positive values, corresponding to exonic splicing enhancers (ESE), while in intronic regions predominated the negative values, indicating an abundance of intronic splicing silencers (ISS) (Figure 2A).

We then determined the presence and proportion of SREs in the pseudoexons in our literature dataset. The sequences of the pseudoexonized regions or intron retentions and the adjacent up and downstream 100 nucleotides were analyzed, following the cDNA position indicated in the corresponding publication (Table S1). These sequences were mapped with ESRseq scores obtained for each nucleotide, thus obtaining the area and the normalized SRE area values. Next, we compared the ESRseq values obtained between the region included in the mRNA as a pseudoexon and in upstream and downstream intronic regions. As with the canonical exons, significant differences were also observed. The regions included in the mRNA as a pseudoexon due to the variant presented a higher percentage of positive values and for those that remained as introns, a greater proportion of negative values (Figure 2B; data in Table S1).

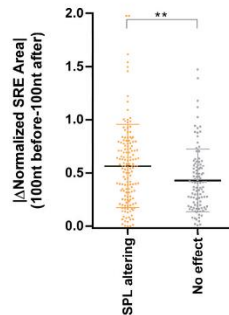
However, these differences were more pronounced between canonical exons and their surrounding intronic regions (Figure 2C), since the presence of enhancer or silencer regulatory elements is more abundant in exonic and intronic canonical regions respectively (Figure S3).



**Figure 2.** Definition of SREs abundance in different genomic regions, using the normalized SRE area calculated with ESRseq scores. (A) Normalized SRE area using ESRseq scores of exonic and adjacent 100 intronic nucleotides located upstream and downstream of canonical exons of HBOC and Lynch genes. Significant differences were identified between exons and intronic regions (Pair-wise significance levels calculated by Tukey test, \*\*\*\*  $p$ -value < 0.0001). (B) Normalized SRE Area using ESRseq scores of pseudoexons and adjacent 100 intronic nucleotides located upstream and downstream of pseudoexons listed in the literature dataset. Significant differences were identified between pseudoexons and intronic regions (Pair-wise significance levels calculated by Tukey test, \*\*\*\*  $p$ -value < 0.0001). (C) Comparison of the exon-intron difference of normalized SRE areas between canonical exons and pseudoexons. First, the mean of normalized SRE area of adjacent donor and acceptor site intronic regions was calculated. Then, this mean was subtracted from the exon and pseudoexon normalized SRE area value. The difference between exonic and intronic regions in canonical exons was higher than in the case of pseudoexons, suggesting that they are more defined by a SRE balance. ( $t$ -test, \*\*\*\*  $p$ -value < 0.0001). Mean  $\pm$  standard deviation is represented in each graph. Intron. ACC: intronic sequence adjacent to acceptor site; Intron. DON: intronic sequence adjacent to donor site.

We also estimated for each variant of the literature dataset, regardless of its splicing effect, the ESRseq values for each position of the 100 intronic bases before and after the variant (without including the hexamers affected by the variant), and then the absolute difference of normalized SRE area between these two regions was calculated (Table S1). This value was used to compare the two groups of deep intronic variants: 133 spliceogenic vs. 100 no effect variants. Our results demonstrated that variants with no effect presented a very low difference between adjacent regions, compared to variants with a spliceogenic effect (Figure 3). This is consistent with spliceogenic intronic variants being in regions with a significant difference in the balance of SRE between exons and introns (in line with the pseudoexon landscape showed in Figure 2B), while variants that do not cause alteration are in intronic regions with no SREs that could make them susceptible to be pseudoexonized. Overall, these results suggest that: (i) intronic regions with a similar SRE balance to that of exons are more susceptible to be included in mature RNA; (ii) a SRE balance is relevant in the RNA misplacing caused by deep intronic variants and; (iii) bearing in mind this balance will facilitate the in silico identification of intronic variants leading to pseudoexon inclusions.





**Figure 3.** Comparison of the absolute difference of normalized SRE area from 100 nucleotides before and after each deep intronic variant compiled from the literature. Absolute values of normalized SRE area difference from 100 nucleotides upstream and downstream of each variant were used to compare those spliceogenic with those without any effect. Splicing variants (SPL altering) showed higher differences between previous and posterior sequences (*t*-test, \*\* *p*-value  $\leq 0.01$ ). Mean  $\pm$  standard deviation is represented.

### 3.3. Inclusion of SRE Landscape in the In Silico Detection of Deep Intronic Splice-Altering Variants

Although SRE in silico tools (such as ESRseq) can detect variant-induced SRE alterations, they are not able to identify whether the SRE landscape where the variant is located presents similarities to those of an exon, and ignore the relevance of the SRE balance to the pipeline of SpliceAI ( $\Delta$  score cut-off of  $</\geq 0.05$ ) and ESRseq ( $\Delta$  score cut-off of  $+0.63$ ), the estimation of the absolute difference between 100 nucleotides up and downstream for each variant (Table S1). With the absolute difference threshold of 0.51 we obtained a MCC of 0.83, with 95.49% sensitivity, 86% specificity, and 91.42% accuracy (Table 2 and Table S10). Although this pipeline would theoretically detect variants that alter SREs, the sensitivity of 95.49% only slightly improves that observed using SpliceAI alone (93.98%) (Table 2).

The performance of the last pipeline (SpliceAI  $\Delta$  score 0.05  $\rightarrow$   $\Delta$ ESRseq 0.63  $\rightarrow$  difference in absolute values 0.51), was tested with a set of 207 splicing-disrupting deep intronic variants identified in tumors by RNAseq [32] with 86% sensitivity which again slightly improves that observed using SpliceAI alone (85.5%) (Table S3).

### 3.4. Experimental Analysis of Hereditary Cancer Gene Variants

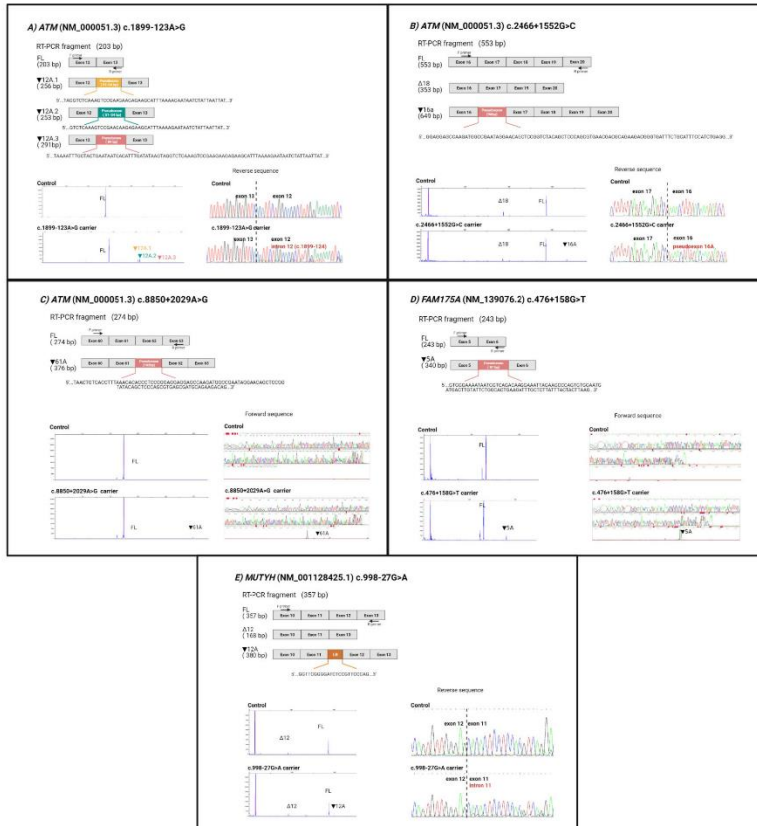
Thirty-three unique variants were experimentally assessed. Thirteen variants passed the 0.05 SpliceAI threshold and six presented  $\Delta$ ESRseq equal or greater than 0.63 and an absolute difference value greater than 0.51 (Table S4). The remaining 14 variants were not predicted as spliceogenic (Table S4). We characterized the variant effect by RT-PCR assays comparing their splicing profiles (by high-resolution electrophoresis) with those in the healthy controls, and posterior Sanger sequencing. This analysis detected the inclusion of intronic regions in mature mRNA in only 5 of the 13 variants prioritized by SpliceAI alone (Table 3 and Figure 4): *ATM* variants c.1899-123A  $>$  G, c.2466 + 1552G  $>$  C, c.8850 + 2029A  $>$  G, *FAM175A* variant c.476 + 158G  $>$  T and *MUTYH* variant c.998-27G  $>$  A. The six variants with a  $\Delta$ ESRseq equal or greater than 0.63 and absolute difference prediction in favor of altering a SRE region did not show an aberrant splicing. All 14 variants with a negative splicing alteration prediction presented a normal splicing pattern. Table 4 shows that sensitivity and specificity using SpliceAI alone for all variants with the cut-off of  $\geq 0.05$  is higher than the pipeline of applying  $\Delta$ ESRseq equal or greater than 0.63, and an absolute difference value greater than 0.51 for those variants with  $\Delta$  SpliceAI scores of  $<0.05$ .

Table 3. Deep intronic variants with a spliceogenic effect detected in hereditary cancer genes.

Gene *	cNomenclature **	Intron	SpliceAI † (Position of Predicted Splice site)	ΔESReeq ‡	ABS dif.	Splicing Outcome §	Population Variant Frequencies (GnomAD)	ClinVar Review Status ¶
ATM	c.1899-123A > G	12	AG 0.15 (-51 bp) and AG 0.74 (-90 bp)/DG 0.71 (-1 bp)	0.633	1.217	(r.1899_1900hns1899-174_1899-124), ▼12A.1 ▼12A.2 (r.1899_1900hns1899-177_1899-124) ▼12A.3	0.000032	NR
			AG 0.93 (3 bp)/DG 0.69 (97 bp)	-0.144	1.816	(r.1899_1900hns1899-213_1899-124) Pseudoxon ▼16A	NR	Likely pathogenic
			AG 0.22(1 bp)/DG 0.16 (102 bp)	-0.187	0.765	(r.2466_2467hns2466 + 1555_2466 + 1650) Pseudoxon ▼61A	NR	NR
FAM175A	c.476+158G > T	5	AG 0.17 (2 bp)/DG 0.22 (-94 bp)	0.559	0.591	(r.8850_8851hns8850 + 2030_8850 + 2131) Pseudoxon ▼5A (r.476_477hns476 + 156_476 + 252)	0.000446	NR
MUTYH	c.998-27G > A	11	AG 0.41 (-4 bp)/DG 0.05 (-215 bp)	0.344	1.000	Intron retention ▼12A (c.997_998hns998-23 + 998)	0.001192	Likely benign    Likely benign    Likely

\* Reference sequences for annotating variants: NM\_000651.3 for ATM; NM\_139076.2 for FAM175A; NM\_001128425.1 for MUTYH. \*\* HGVS nomenclature guidelines were used for variant annotation (<http://varnomen.hgvs.org/>, accessed on 1 September 2020). † SpliceAI predictor: AG = Acceptor Gain/DG = Donor Gain. The AScore ranges from 0 to 1 and can be interpreted as the probability that the variant affects splicing at any position around it. For each variant, SpliceAI evaluates a nucleotide window (±/- 4000 in this case) to see how the variant affects the probabilities of different positions in the pre-mRNA being splice acceptors or donors. The values in brackets represent positions with the biggest probability of being used as splicing sites within the window. Negative values are upstream (5') of the variant, and positive are downstream (3') of the variant. In bold are the scores of >0.05. ‡ ΔESReeq score was calculated as described by Ke, et al., and in bold are indicated values >0.63. Absolute difference (ABS dif.): absolute value of the difference between the normalized SRE area of 100 bp before and after a variant. In bold are indicated values >0.51. § RNA splicing profiles were compared between carriers and controls by capillary electrophoresis and Sanger sequencing. ¶ Variant classification reported in ClinVar database (<https://www.ncbi.nlm.nih.gov/clinvar/>, accessed on 1 March 2021) according to number of variant submitters. NR: not reported.





**Figure 4.** Spliceogenic variants characterization in patients’ RNA. For each variant, there is an RT-PCR assay graphical representation, the results of capillary electrophoresis of 6-FAM labelled amplicons and Sanger sequencing to confirm the expression of additional transcripts. (A) The *ATM*.c.1899-123A > G variant activates a cryptic donor site, which is used to yield three different pseudoexons: ▼12A.1, ▼12A.2, and ▼12A.3, each generated as result of the usage of different cryptic acceptor sites (c.1899-174, c.1899-177, and c.1899-213) and the cryptic donor site created by the variant. The ▼12A.1 and ▼12A.2 transcripts were equally expressed, and their abundance was greater than the ▼12A.3. (B) The *ATM* c.2466 + 1552G > C variant generates the ▼16A additional transcript. This pseudoexon comprises nucleotides from the acceptor site created by the variant and the cryptic donor at c.2466 + 1650. (C) The *ATM* c.8850 + 2029A > G variant presents an additional transcript (▼61A), from the cryptic acceptor site created by the variant to the cryptic donor at c.8850 + 2131. It was not possible to clearly read the sequence of the aberrant transcript because of its low expression levels, but in the Sanger sequence, we could detect the additional transcript with the insertion since it is marked by the FAM signal at the end of the fragment. (D) The *FAM175A* c.476 + 156G > T variant leads to the inclusion of a pseudoexon (▼5A), which results in the usage of the cryptic acceptor site activated by the variant and the cryptic donor at c.476 + 252. Its abundance was very low, but the transcript with the insertion was also detected in the Sanger sequence since it is marked by the FAM signal at the end of the fragment. (E) The *MUTYH* c.998-27G > A variant creates/enhances a cryptic acceptor site which is used instead of the natural acceptor site of exon 12, generating an intronic retention (▼12A transcript).

**Table 4.** Performance pipeline (experimental variants dataset) with 33 variants: 5 splicing altering and 28 non-altering.

Pipeline	Sensitivity	Specificity	Accuracy	MCC
Splice $\geq$ 0.05	100	71.42	75.75	0.62
Splice $\geq$ 0.05; ESRseq $\geq$ 0.63; Abs. Dif. 0.51	100	50.00	57.57	0.43

#### 4. Discussion

The contribution of deep intronic variants to HBOC disease is not well known due to their location in poorly screened regions, but their potential effect on transcript splicing including intron sequences in mature RNA may be clinically significant [10,26]. For this reason, the identification and subsequent RNA characterization of this type of variants should be considered when conventional genetic analysis focused on coding regions and exon/intron boundaries does not lead to the identification of pathogenic variants [35].

However, the identification of deep intronic variants is challenging due to the lack of specific in silico pipelines [28]. Recently, some published studies using SpliceAI, a deep learning-derived algorithm [34], suggest its utility to identify with high efficiency intronic and exonic variants creating or enhancing cryptic splice sites and leading to splicing alterations [36–39]. Nevertheless, to our knowledge, only small datasets of deep intronic variants have been used to test the performance of the SpliceAI tool for identifying this type of variants [27,28,34,37].

Our work provides a large dataset of deep intronic variants that are clinically relevant, as they were tested in a clinical setting using blood, mini, or midgene assays and is well balanced with 133 altering and 100 non-altering splicing events. Hence, this data can be used as a positive control training set for further improvements of computational prediction tools. With this data, we confirmed that SpliceAI with a threshold of  $\geq 0.05$  has an optimal predictive value in the identification of spliceogenic deep intronic variants, obtaining a MCC of 0.86. Interestingly, Riepe et al. [37] with an optimized SpliceAI cut-off score of 0.18, also showed a high performance of 0.84 MCC for predicting 81 deep intronic variants in the *ABCA4* gene, that are also included in our literature dataset. The authors further demonstrated that SpliceAI was the best tool for these 81 deep intronic variants compared with other deep-learning based algorithms [37]. In this line, Jaganathan et al. [34] in their SpliceAI development article, demonstrated that by applying SpliceAI with a cut-off of  $\geq 0.5$  to GTEx RNAseq data, it achieved a sensitivity of 71% when the variants were near exons (82 variants, overlapping exons or  $\leq 50$  nt from exon-intron boundaries), but dropped to 41% when the variants were in deep intronic regions (37 variants,  $>50$  nt from exons). In sum, our study together with the two last works mentioned above, support that a low SpliceAI threshold is needed to especially detect deep intronic splice-altering variants. In our experimental dataset with clinical variants, using our optimized  $\geq 0.05$  threshold, SpliceAI attained a performance of 0.62 MCC, predicting all five experimentally confirmed spliceogenic variants, but with a low specificity of 71.42% due to a high number of false-positives (Table 4).

Besides creating or enhancing cryptic splice sites, the intronic variants can lead to the inclusion of pseudoexons by creating or disrupting intronic SREs. Our evaluation of the prediction of splice alteration through SRE involvement of both deep intronic (Table 1 and Figure 1) and exonic (Table S2B and Figure S2) variants pointed out that for SpliceAI, it is more challenging to predict the impact of this type of variants. This is possibly due to the fact that the deep learning network approach used for SpliceAI development was not able to account for the SREs, denoting that the performance of SpliceAI can still be improved. To note, this is the first study testing the prediction capacity of a deep learning method such as SpliceAI of exonic variants disrupting SREs leading to exon skipping.

To supplement lower SpliceAI performance for detecting SREs altering variants, we added the ESRseq, which has a high capacity to recognize this type of variants [27], obtaining an increase of sensitivity but a lower specificity (Table 2). We reasoned that this

limitation was due to the fact that ESRseq evaluates on a hexamer local level, without accounting for a SREs landscape that defines a region to be included as a pseudoexon. This prompted us to characterize the landscape of SREs in pseudoexons using SRE scores, showing that the relation of the SRE landscape between the pseudoexon and flanking introns is similar to that of canonical exons, but less defined (Figure 2C and Figure S3). In contrast, the ESRseq developers in Ke et al. [11] reported that the pseudoexons did not present a different balance of SRE concerning the adjacent intronic regions. This discrepancy could be due to the fact that the pseudoexons analyzed in the above-mentioned work were theoretically defined, without an experimental RNA evaluation, as the intronic sequences had lengths between 50 and 250 nt and consensus values based on the Shapiro-Senapathy algorithm, of  $\geq 75$  for 3' splice sites and  $\geq 78$  for 5' splice sites, and were located beyond 100 bp from the exons [11]. Instead, we collected 133 pseudoexons from literature, experimentally confirmed using patient, mini, or midgene-derived RNA. Notably, similar findings to our results using approaches other than ESRseq tools have been recently reported, in 42 pseudoexons experimentally validated in the *DMD* gene, showing a smaller density of exonic splicing enhancers (ESEs) together with a higher density of exonic splicing silencers (ESSs) compared to canonical exons, which suggested that the pseudoexons presented a weaker exon profile in terms of SREs [30]. Interestingly, these differences have also been observed between alternative and canonical exons. The effect of variants altering the balance of SREs appears to be greater in alternative exons, which have fewer redundant enhancer elements, compared to constitutive ones [40]. Therefore, we suggest that deep intronic variants that strengthen an enhancer or even decrease a silencer will have a greater chance of being spliceogenic provided they are located in intron regions with an exon-like SRE landscape, similar to what happens in alternative exons.

Given the role of an exon-like SRE landscape in the inclusion of pseudoexons, we combined the  $\Delta$ SpliceAI 0.05 and 0.63  $\Delta$ ESRseq optimized thresholds, and the absolute difference of SRE balance between the regions before and after a variant. This last value, with the variants compiled from literature, indicates clear differences between splicing altering variants and those without any effect (Figure 3) and helps to identify spliceogenic variants, with high sensitivity and specificity (Table 2). The combination was assayed in a tumor RNA-seq dataset (Table S3) [32] obtaining a sensitivity of 86%. Moreover, applying the same in silico combination to a set of deep intronic variants from a cohort of HBOC patients allowed us to identify five spliceogenic variants that were predicted by SpliceAI to activate a cryptic splice site nearby, while assaying eight false positives (Table S4).

The low accuracy of the in-silico strategy combining  $\Delta$ SpliceAI,  $\Delta$ ESRseq, and SRE landscape using ESRseq scores can be explained by two reasons. First, our literature dataset only contains 16 splice altering variants by affecting a regulatory element. Second, the splicing aberrations of the literature and experimental datasets were assayed using RT-PCR, which has an intrinsic bias towards smaller amplicons. This is especially relevant in the case of pseudoexons as they generate larger fragments than normal transcripts and are also less expressed if they cause a premature termination codon and are targeted by non-sense mediated decay. Massively long-read RNA sequencing, such as those using Oxford Nanopore Technologies, could address this limitation by allowing simultaneous detection and quantification of all RNA transcripts avoiding the PCR amplification step [41].

Additionally, it is worth stating that the main purpose of this work was not the clinical classification of the variants assessed but rather to investigate how to improve the in silico identification of deep intronic splicing variants. Qualitative analysis can only detect aberrant transcripts, but additional quantification of the functional transcripts, co-segregation data or other functional assays are needed to classify the variants that induce pseudoexons [42,43].

Overall, the results obtained with the three sets of deep intronic variants (literature, tumor, and experimental) demonstrated that SpliceAI alone is able to identify variants causing pseudoexons and that the addition of ESRseq increases the number of false positives. Moreover, our use of ESRseq values to map the SRE balance in canonical exons



and pseudoexons differentiating exonic landscapes from intronic ones suggest that this approach might be systematically used to identify exon-like landscapes in introns of HBOC and Lynch genes, thus helping to interpret whether an intronic variant makes a region much more exonizable.

## 5. Conclusions

We have provided evidence that SpliceAI, a deep learning-based *in silico* tool, can predict splicing altering deep intronic variants with high-performance. However, its accuracy is limited with variants affecting SREs, either with intronic variants introducing pseudoexons or exonic variants inducing exon skipping. The addition of ESRseq, a specific bioinformatic tool to detect SRE disruption/enhancement, did not increase the accuracy of the deep intronic splicing-altering variants prediction. However, our findings show that pseudoexons have a “SRE landscape” similar to that of exons. This indicates that intronic regions with a high potential to be included as pseudoexons can be systematically identified throughout the HBOC genes, facilitating the *in silico* detection of spliceogenic deep intronic variants.

**Supplementary Materials:** The following are available online at <https://www.mdpi.com/article/10.3390/cancers13133341/s1>, Figure S1:  $\Delta$ SpliceAI scores of spliceogenic and not spliceogenic variants collected in the literature database, taking predictions of cryptic sites in a window of 50 bp or 4999 bp, Figure S2: Comparison of  $\Delta$ SpliceAI scores between spliceogenic and not spliceogenic exonic variants collected from Tubeuf et al. [27], Figure S3: Comparison of SREs abundance in different genomic regions, using the Normalized SRE Area calculated by ESRseq scores, Table S1: Deep intronic variants collected from literature (literature dataset), Table S2A: Exonic variants altering splicing by affecting SREs and causing exon skipping and exonic non-splice altering variants obtained from Tubeuf et al. [27] Hum Mut., and annotated using SpliceAI *in silico* tool with two windows, 4999 and 50 nt before and after the variant, Table S2B: Optimized threshold of SpliceAI (4999 nt window) scores (MAX acceptor loss-AL donor loss-DL) for SREs disrupting variants identification (variants database obtained from Tubeuf et al., Hum Mut [27]), Table S3: Tumor RNA-seq data (Jung et al. [32]) analyzed with SpliceAI v1.3 (4999 window), Table S4: Deep intronic experimentally assessed variants, Table S5: List of HBOC risk and Lynch genes and their respective NCBI reference transcripts used to annotate with ESRseq scores canonical exons and adjacent intronic nucleotides, Table S6: ESRseq scores from all BARD1 exons and adjacent intronic regions, obtained according Ke et al. [11]; data for the other genes available upon request, Table S7: Formulas for performance evaluation of *in silico* tools used stand-alone or in sequential combination, Table S8: Primers and PCR conditions for characterizing the splicing effect of deep intronic variants with RNA extracted from patient’s blood, Table S9: ESRseq threshold optimization (sensitivity + specificity), Table S10: Pipeline performance values.

**Author Contributions:** Conceptualization, A.M.-F., J.D.-V., O.D. and S.G.-E.; methodology, A.M.-F., J.D.-V. and A.T.; software, A.M.-F. and J.D.-V.; validation, A.M.-F., J.D.-V. and S.G.-E.; formal analysis, A.M.-F. and J.D.-V.; investigation, A.M.-F., J.D.-V. and A.T.; resources, A.T., J.B. and O.D.; data curation, A.M.-F. and J.D.-V.; writing—original draft preparation, A.M.-F. and J.D.-V.; writing—review and editing, A.M.-F., J.D.-V., J.B., O.D. and S.G.-E.; visualization, A.M.-F. and J.D.-V.; supervision, O.D. and S.G.-E.; project administration, S.G.-E.; funding acquisition, S.G.-E. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Spanish Instituto de Salud Carlos III (ISCIII) funding an initiative of the Spanish Ministry of Economy and Innovation, partially supported by European Regional Development FEDER Funds, grant numbers PI16/01218 and PI19/01303. AM-F contract is supported by the award ERAPERMED2019-215 granted by AECC FC and by ISCIII through AES 2019, both within the ERAPERMed framework”. J.D.-V. contract is supported by the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia and the European Social Fund.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Clinical Research Ethics Committee (CEIC) from Hospital Universitari Vall d’Hebron, Barcelona, Spain (project n° PR(AG)415/2016 with date of approval of 11 July 2017).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** All data analyzed in this study are included in this article and its supplementary information files.

**Acknowledgments:** We acknowledge the Cellex Foundation for providing research facilities and thank CERCA Programme/Generalitat de Catalunya for institutional support. We also thank the help and technical support of Sara Hermosa-García and Cristina Zamarreño-Pastor.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Dorling, L.; Carvalho, S.; Allen, J.; González-Neira, A.; Luccarini, C.; Wahlström, C.; Pooley, K.A.; Parsons, M.T.; Fortuno, C.; Wang, Q.; et al. Breast Cancer Risk Genes—Association Analysis in More than 113,000 Women. *N. Engl. J. Med.* **2021**, *384*, 428–439. [\[CrossRef\]](#)
2. Hu, C.; Hart, S.N.; Gnanaolivu, R.; Huang, H.; Lee, K.Y.; Na, J.; Gao, C.; Lilyquist, J.; Yadav, S.; Boddicker, N.J.; et al. A Population-Based Study of Genes Previously Implicated in Breast Cancer. *N. Engl. J. Med.* **2021**, *384*, 440–451. [\[CrossRef\]](#)
3. Hasson, S.P.; Menes, T.; Sonnenblick, A. Comparison of patient susceptibility genes across breast cancer: Implications for prognosis and therapeutic outcomes. *Pharmacogenomics Personal. Med.* **2020**, *13*, 227–238. [\[CrossRef\]](#)
4. Bonache, S.; Esteban, I.; Moles-Fernández, A.; Tenés, A.; Duran-Lozano, L.; Montalban, G.; Bach, V.; Carrasco, E.; Gadea, N.; López-Fernández, A.; et al. Multigene panel testing beyond BRCA1/2 in breast/ovarian cancer Spanish families and clinical actionability of findings. *J. Cancer Res. Clin. Oncol.* **2018**, *144*, 2495–2513. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Feliubadaló, L.; López-Fernández, A.; Pineda, M.; Díez, O.; del Valle, J.; Gutiérrez-Enriquez, S.; Teulé, A.; González, S.; Stjepanovic, N.; Salinas, M.; et al. Opportunistic testing of BRCA1, BRCA2 and mismatch repair genes improves the yield of phenotype driven hereditary cancer gene panels. *Int. J. Cancer* **2019**, *145*, 2682–2691. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Couch, F.J.; Nathanson, K.L.; Offit, K. Two decades after BRCA: Setting paradigms in personalized cancer care and prevention. *Science* **2014**, *343*, 1466–1470. [\[CrossRef\]](#)
7. Rhine, C.L.; Cygan, K.J.; Soemedi, R.; Maguire, S.; Murray, M.F.; Monaghan, S.F.; Fairbrother, W.G. Hereditary cancer genes are highly susceptible to splicing mutations. *PLoS Genet.* **2018**, *14*, 1–18. [\[CrossRef\]](#)
8. Anna, A.; Monika, G. Splicing mutations in human genetic disorders: Examples, detection, and confirmation. *J. Appl. Genet.* **2018**, *59*, 253–268. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Cartegni, L.; Chew, S.L.; Krainer, A.R. Listening To Silence and Understanding Nonsense: Exonic Mutations That Affect Splicing. *Nat. Rev. Genet.* **2002**, *3*, 285–298. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Zhang, X.H.-F.; Chasin, L.A. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.* **2004**, *18*, 1241–1250. [\[CrossRef\]](#)
11. Ke, S.; Shang, S.; Kalachikov, S.M.; Morozova, I.; Yu, L.; Russo, J.J.; Ju, J.; Chasin, L.A. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* **2011**, *21*, 1360–1374. [\[CrossRef\]](#)
12. Baralle, D.; Baralle, M. Splicing in action: Assessing disease causing sequence changes. *J. Med. Genet.* **2005**, *42*, 737–748. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Dhir, A.; Buratti, E. Alternative splicing: Role of pseudoexons in human disease and potential therapeutic strategies: Minireview. *FEBS J.* **2010**, *277*, 841–855. [\[CrossRef\]](#)
14. Sironi, M.; Menozzi, G.; Riva, L.; Cagliani, R.; Comi, G.P.; Bresolin, N.; Giorda, R.; Pozzoli, U. Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Res.* **2004**, *32*, 1783–1791. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Pozzoli, U.; Sironi, M. Silencers regulate both constitutive and alternative splicing events in mammals. *Cell. Mol. Life Sci.* **2005**, *62*, 1579–1604. [\[CrossRef\]](#)
16. Vaz-Drágo, R.; Custódio, N.; Carmo-Fonseca, M. Deep intronic mutations and human disease. *Hum. Genet.* **2017**, *136*, 1093–1111. [\[CrossRef\]](#)
17. Romano, M.; Buratti, E.; Baralle, D. Role of Pseudoexons and Pseudointrons in Human Cancer. *Int. J. Cell Biol.* **2013**, *2013*, 810572. [\[CrossRef\]](#) [\[PubMed\]](#)

18. Montalban, G.; Bonache, S.; Moles-fernández, A.; Gisbert-beamud, A.; Tenés, A.; Bach, V.; Carrasco, E.; López-fernández, A.; Stjepanovic, N.; Balmaña, J.; et al. Screening of BRCA1/2 deep intronic regions by targeted gene sequencing identifies the first germline BRCA1 variant causing pseudoexon activation in a patient with breast/ovarian cancer. *J. Med. Genet.* **2018**, *56*, 63–74. [[CrossRef](#)] [[PubMed](#)]
19. Pagani, F.; Buratti, E.; Stuani, C.; Bendix, R.; Dörk, T.; Baralle, F.E. A new type of mutation causes a splicing defect in ATM. *Nat. Genet.* **2002**, *30*, 426–429. [[CrossRef](#)]
20. Özkan, S.; Padilla, N.; Moles-Fernández, A.; Diez, O.; Gutiérrez-Enríquez, S.; de la Cruz, X. The computational approach to variant interpretation: Principles, results, and applicability. In *Clinical DNA Variant Interpretation: Theory and Practice*; Lázaro, C., Lerner-Ellis, J., Spurdle, A., Eds.; Academic Press: Cambridge, MA, USA, 2021; Volume 1, ISBN 9780128205198.
21. Desmet, F.O.; Hamroun, D.; Lalande, M.; Collod-Bèroud, G.; Claustres, M.; Bèroud, C. Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **2009**, *37*, 1–14. [[CrossRef](#)]
22. Shapiro, M.B.; Senapathy, P. RNA splice junctions of different classes of eukaryotes: Sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* **1987**, *15*, 7155–7174. [[CrossRef](#)]
23. Yeo, G.; Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **2004**, *11*, 377–394. [[CrossRef](#)]
24. Moles-Fernández, A.; Duran-Lozano, L.; Montalban, G.; Bonache, S.; López-Perolio, I.; Menéndez, M.; Santamariña, M.; Behar, R.; Blanco, A.; Carrasco, E.; et al. Computational tools for splicing defect prediction in breast/ovarian cancer genes: How efficient are they at predicting RNA alterations? *Front. Genet.* **2018**, *9*, 366. [[CrossRef](#)] [[PubMed](#)]
25. Rosenberg, A.B.; Patwardhan, R.P.; Shendure, J.; Seelig, G. Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell* **2015**, *163*, 698–711. [[CrossRef](#)]
26. Erkelenz, S.; Theiss, S.; Otte, M.; Widera, M.; Peter, J.O.; Schaal, H. Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Res.* **2014**, *42*, 10681–10697. [[CrossRef](#)] [[PubMed](#)]
27. Tubeuf, H.; Charbonnier, C.; Soukariéh, O.; Blavier, A.; Lefebvre, A.; Dauchel, H.; Frebourg, T.; Gaildrat, P.; Martins, A. Large-scale comparative evaluation of user-friendly tools for predicting variant-induced alterations of splicing regulatory elements. *Hum. Mutat.* **2020**, *41*, 1811–1829. [[CrossRef](#)]
28. Canson, D.; Glubb, D.; Spurdle, A.B. Variant effect on splicing regulatory elements, branchpoint usage, and pseudoexonization: Strategies to enhance bioinformatic prediction using hereditary cancer genes as exemplars. *Hum. Mutat.* **2020**, *41*, 1705–1721. [[CrossRef](#)]
29. Rowlands, C.F.; Baralle, D.; Ellingford, J.M. Machine Learning Approaches for the Prioritization of Genomic Variants Impacting Pre-mRNA Splicing. *Cells* **2019**, *8*, 1513. [[CrossRef](#)] [[PubMed](#)]
30. Xie, Z.; Tang, L.; Xie, Z.; Sun, C.; Shuai, H.; Zhou, C.; Liu, Y.; Yu, M.; Zheng, Y.; Meng, L.; et al. Splicing characteristics of dystrophin pseudoexons and identification of a novel pathogenic intronic variant in the DMD gene. *Genes* **2020**, *11*, 1180. [[CrossRef](#)] [[PubMed](#)]
31. Dhir, A.; Buratti, E.; Van Santen, M.A.; Lührmann, R.; Baralle, F.E. The intronic splicing code: Multiple factors involved in ATM pseudoexon definition. *EMBO J.* **2010**, *29*, 749–760. [[CrossRef](#)] [[PubMed](#)]
32. Jung, H.; Lee, K.S.; Choi, J.K. Comprehensive characterisation of intronic mis-splicing mutations in human cancers. *Oncogene* **2021**, *40*, 1347–1361. [[CrossRef](#)] [[PubMed](#)]
33. Bolisetty, M.T.; Beemon, K.L. Splicing of internal large exons is defined by novel cis-acting sequence elements. *Nucleic Acids Res.* **2012**, *40*, 9244–9254. [[CrossRef](#)]
34. Jaganathan, K.; Panagiotopoulou, S.K.; McRae, J.F.; Darbandi, S.F.; Knowles, D.; Li, Y.L.; Kosmicki, J.A.; Arbelaez, J.; Cui, W.; Schwartz, G.B.; et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **2019**, *176*, 535–548.e24. [[CrossRef](#)]
35. Sakaguchi, N.; Suyama, M. In silico identification of pseudo-exon activation events in personal genome and transcriptome data. *RNA Biol.* **2021**, *18*, 382–390. [[CrossRef](#)] [[PubMed](#)]
36. Qian, X.; Wang, J.; Wang, M.; Igelman, A.D.; Jones, K.D.; Li, Y.; Wang, K.; Goetz, K.E.; Birch, D.G.; Yang, P.; et al. Identification of Deep-Intronic Splice Mutations in a Large Cohort of Patients With Inherited Retinal Diseases. *Front. Genet.* **2021**, *12*, 647400. [[CrossRef](#)]
37. Riepe, T.V.; Khan, M.; Roosing, S.; Cremers, F.P.M.; Hoen, P.A.C. Benchmarking deep learning splice prediction tools using functional splice assays. *Hum. Mutat.* **2021**. [[CrossRef](#)]
38. Wai, H.A.; Lord, J.; Lyon, M.; Gunning, A.; Kelly, H.; Cibin, P.; Seaby, E.G.; Spiers-Fitzgerald, K.; Lye, J.; Ellard, S.; et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet. Med.* **2020**, *22*, 1005–1014. [[CrossRef](#)]
39. Rentzsch, P.; Schubach, M.; Shendure, J.; Kircher, M. CADD-Splice—Improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **2021**, *13*, 1–12. [[CrossRef](#)]
40. Baeza-Centurion, P.; Miñana, B.; Valcárcel, J.; Lehner, B. Mutations primarily alter the inclusion of alternatively spliced exons. *eLife* **2020**, *9*, 1–74. [[CrossRef](#)]
41. de Jong, L.C.; Cree, S.; Lattimore, V.; Wiggins, G.A.R.; Spurdle, A.B.; Miller, A.; Kennedy, M.A.; Walker, L.C. Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. *Breast Cancer Res.* **2017**, *19*, 1–9. [[CrossRef](#)] [[PubMed](#)]



42. Anczuków, O.; Buisson, M.; Léoñe, M.; Coutanson, C.; Lasset, C.; Calender, A.; Sinilnikova, O.M.; Mazoyer, S. BRCA2 deep intronic mutation causing activation of a cryptic exon: Opening toward a new preventive therapeutic strategy. *Clin. Cancer Res.* **2012**, *18*, 4903–4909. [[CrossRef](#)]
43. Montalban, G.; Bonache, S.; Moles-Fernández, A.; Gadea, N.; Tenés, A.; Torres-Esquius, S.; Carrasco, E.; Balmaña, J.; Diez, O.; Gutiérrez-Enriquez, S. Incorporation of semi-quantitative analysis of splicing alterations for the clinical interpretation of variants in BRCA1 and BRCA2 genes. *Hum. Mutat.* **2019**, *40*, 2296–2317. [[CrossRef](#)]

## Article 3

### Collaborative Effort to Define Classification Criteria for ATM Variants in Hereditary Cancer Patients

Feliubadaló L, Moles-Fernández A, Santamariña-Pena M, Sánchez AT, López-Novo A, Porras LM, Blanco A, Capellá G, de la Hoya M, Molina IJ, Osorio A, Pineda M, Rueda D, de la Cruz X, Díez O, Ruiz-Ponte C, Gutiérrez-Enríquez S, Vega A, Lázaro C. A

**Clinical Chemistry.** 2021 Mar 1;67(3):518-533.

doi: 10.1093/clinchem/hvaa250.



## A Collaborative Effort to Define Classification Criteria for *ATM* Variants in Hereditary Cancer Patients

Lidia Feliubadaló,<sup>a,b,c,\*</sup> Alejandro Moles-Fernández,<sup>d,t</sup> Marta Santamariña-Pena,<sup>e,f,g,t</sup> Alysso T. Sánchez,<sup>a,b,t</sup> Anael López-Novo,<sup>e,f</sup> Luz-Marina Porras,<sup>h</sup> Ana Blanco,<sup>e,f,g</sup> Gabriel Capellá,<sup>a,b,c</sup> Miguel de la Hoya,<sup>c,i</sup> Ignacio J. Molina,<sup>j</sup> Ana Osorio,<sup>g,k</sup> Marta Pineda,<sup>a,b,c</sup> Daniel Rueda,<sup>l</sup> Xavier de la Cruz,<sup>h,m</sup> Orland Diez,<sup>d,n</sup> Clara Ruiz-Ponte,<sup>e,f,g</sup> Sara Gutiérrez-Enriquez,<sup>d</sup> Ana Vega,<sup>e,f,g</sup> and Conxi Lázaro<sup>a,b,c,\*</sup>

**BACKGROUND:** Gene panel testing by massive parallel sequencing has increased the diagnostic yield but also the number of variants of uncertain significance. Clinical interpretation of genomic data requires expertise for each gene and disease. Heterozygous *ATM* pathogenic variants increase the risk of cancer, particularly breast cancer. For this reason, *ATM* is included in most hereditary cancer panels. It is a large gene, showing a high number of variants, most of them of uncertain significance. Hence, we initiated a collaborative effort to improve and standardize variant classification for the *ATM* gene.

**METHODS:** Six independent laboratories collected information from 766 *ATM* variant carriers harboring 283 different variants. Data were submitted in a consensus template form, variant nomenclature and clinical information were curated, and monthly team conferences were established to review and adapt American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) criteria to *ATM*, which were used to classify 50 representative variants.

**RESULTS:** Amid 283 different variants, 99 appeared more than once, 35 had differences in classification among laboratories. Refinement of ACMG/AMP criteria to *ATM* involved specification for twenty-one criteria and adjustment of strength for fourteen others. Afterwards, 50 variants carried by 254 index cases were

classified with the established framework resulting in a consensus classification for all of them and a reduction in the number of variants of uncertain significance from 58% to 42%.

**CONCLUSIONS:** Our results highlight the relevance of data sharing and data curation by multidisciplinary experts to achieve improved variant classification that will eventually improve clinical management.

### Introduction

Genetic diagnosis for hereditary cancers (HC) has changed over the past decade thanks to the introduction of massive parallel sequencing (MPS) technologies which allow the screening of multiple genes outright. MPS diagnostic panels increase sensitivity but also the number of variants of uncertain clinical significance (VUS) identified; application of MPS panels poses a significant challenge in the clinical management of patients and evidences the need for standardization in variant classification. The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have provided a general framework for classification of genetic variants (1). However, these universal guidelines need to be tuned according to the disease and the specific gene by a consensus of experts. Currently ACMG/AMP guidelines

<sup>a</sup>Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL, Hospitalet de Llobregat, Barcelona, Spain; <sup>b</sup>Oncobell Program, IDIBELL, Hospitalet de Llobregat, Barcelona, Spain; <sup>c</sup>Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain; <sup>d</sup>Hereditary Cancer Genetics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain; <sup>e</sup>Fundación Pública Galega Medicina Xenómica (PPGMX), SERGAS, Santiago de Compostela, Spain; <sup>f</sup>Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Santiago de Compostela, Spain; <sup>g</sup>Centro de Investigación en Red de Enfermedades Raras (CIBERER), Madrid, Spain; <sup>h</sup>Research Unit in Clinical and Translational Bioinformatics, Vall d'Hebron Institute of Research (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain; <sup>i</sup>Molecular Oncology Laboratory, Hospital Clínico San Carlos, IdISSC (Instituto de Investigación Sanitaria del Hospital Clínico San Carlos), Madrid, Spain; <sup>j</sup>Institute of Biopathology and Regenerative Medicine, Center for Biomedical Research, Health Sciences Technology Park, University of Granada, Granada, Spain; <sup>k</sup>Human Genetics Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain; <sup>l</sup>Hereditary Cancer Laboratory, Doce de Octubre University

Hospital, i+12 Research Institute, Madrid, Spain; <sup>m</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain; <sup>n</sup>Clinical and Molecular Genetics Area, University Hospital Vall d'Hebron, Barcelona, Spain.

\* Address correspondence to: L.F. at Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL, Hospitalet de Llobregat, Av. Gran Via 199-203, 08908 Barcelona, Spain. Fax +34-932607466; e-mail lfeliubadaló@iconcologia.net. C.L. at Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL, Hospitalet de Llobregat, Av. Gran Via 199-203, 08908 Barcelona, Spain. E-mail clazaró@iconcologia.net.

<sup>†</sup>These authors contributed equally to this work and should be considered second co-authors

<sup>‡</sup>These authors should be considered corresponding co-authors

Received June 17, 2020; accepted September 29, 2020.

DOI: 10.1093/clinchem/hvaa250

### Classification of *ATM* Variants

have been adapted for some hereditary cancer genes such as *PTEN* (2), *CDH1* (3), and *TP53* (4).

Most of the currently used HC panels include the *ATM* gene, mainly because heterozygous *ATM* mutations increase the risk of cancer, particularly breast cancer (BC) (5), and have also been associated with colorectal, prostate, and pancreatic cancer predisposition (6–8). A moderate breast cancer risk of about 2.4-fold was estimated from breast cancer families with *ATM* pathogenic variants (9). In this sense, *ATM* loss-of-function variants confer an increase in breast cancer risk 10 times greater than that of missense variants (10); however, the p.(Val2424Gly) missense variant seems to confer a higher risk, comparable to that of *BRCA2* variants (11). For other cancers, an overall risk of 2.23 (94% CI 1.26–4.28) has been suggested to increase to 4.94 (95% CI = 1.90 to 12.9) in carriers under 50 years of age (8).

*ATM* is also responsible for the autosomal recessive genetic disorder ataxia telangiectasia (AT) (MIM# 208900) (12). AT is a pleiotropic neurodegenerative disease whose symptoms include malignancy and genome instability, often accompanied by immunodeficiencies, premature aging, insulin resistance, and infertility (13, 14). Most AT patients bear compound heterozygous pathogenic variants from over 800 currently registered in the Human Genome Mutation Database (15). A recent study in Spanish AT patients identified disease-causing mutations in 96% of the alleles studied, frameshift being the most common type of variant (16). The *ATM* protein is a member of the phosphatidylinositol-3' kinase-related protein kinase (PIKK) family, which phosphorylates hundreds of targets containing Ser/Thr-Gln motifs, and plays critical roles in double-strand break (DSB) DNA repair and cell cycle (14). DNA breaks recruit inactive *ATM* dimers through the Mre11-Rad50-NBS1 (MRN) sensor complex, which allows *ATM* dissociation into Ser1981-autophosphorylated active monomers, able to act upon a number of direct substrates such as TP53 or indirect such as histone H2AX (14). These events are key indicators of *ATM* functional activity, and coordinated activity of phosphorylated downstream targets determines whether the genomic instability resulting from DNA damage can be prevented (17).

With the aim of improving and standardizing variant classification for HC genes in Spain, 6 independent molecular laboratories using MPS panels agreed to create a common variant database. *ATM* was chosen for the pilot study because it is a large gene included in the majority of HC panels and shows a remarkable number of VUS (18). After adapting ACMG/AMP classification guidelines to *ATM*, 50 variants were designated for classification with the established consensus.

### Materials and Methods

A detailed description of the methodology used can be found in the Supplemental Patients and Methods. Briefly, a multidisciplinary group was built with complementary expertise. Most of the members are molecular geneticists with experience in hereditary cancer and RNA splicing. Some are members of gene-specific international endeavors such as Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA, <https://enigmaconsortium.org/>), International Society for Gastrointestinal Hereditary Tumours (InSiGHT, <https://www.insight-group.org/>), and Clinical Genome Resource (ClinGen, <https://clinicalgenome.org/>). In addition, the team had a Spanish expert in AT and *ATM* functional assays and 2 experts in computational biology and bioinformatics. Patients included in this study were seen in the different genetic counseling units of each reference laboratory. All patients had a clinical suspicion of HC and were tested by gene panel sequencing. All variants detected in the *ATM* coding sequence and 20-bp surrounding regions with minor allele frequency lower than 1% were collected in the Spanish Hereditary Cancer Variant Database (DB hereinafter) created for this purpose.

Cut-offs for allele frequency calculations, as well as the selection of different splicing and protein prediction assessment tools and the adjustment of the corresponding threshold values, are described in the Supplemental Methods. This section also details the process of functional study type selection and the strategy for variant classification of 50 pilot variants from our DB.

### Results

#### *ATM* VARIANT DATABASE

In total, we collected information from 769 individuals carrying 283 different *ATM* variants; 104 index cases carried more than one *ATM* variant. Hereditary breast and/or ovarian cancer was the most common clinical indication in the whole cohort (67%) (Supplemental Fig. 1), being women 85% of individuals (Supplemental Table 1). Ninety-nine of the 283 different variants collected appeared in more than one family; 78 were found in more than one laboratory, and 20 appeared in 10 or more families (Supplemental Fig. 1). The 5-tier pathogenicity classification given by each laboratory was recorded, and 35 of the 78 variants detected by more than one laboratory had discordant classifications (45%). Thirty of these discordances were due to the variant being classified as VUS vs. likely benign (LB); the remaining 5 discordances were as follows: 3 from likely pathogenic (LP) vs. pathogenic (P), 1 from VUS vs. benign (B), and 1 from VUS vs. LP.



#### ATM-SPECIFIC REFINEMENT OF ACMG-AMP CLASSIFICATION GUIDELINES

Based on previous studies of other HC genes (2, 3), we decided to adapt the widely used ACMG/AMP classification guidelines (1) to the characteristics of the *ATM* gene and its associated phenotypes. From the 28 criteria listed in the guidelines, several have been modified, restricted, rejected as non-applicable or expanded to diverse strengths. The resulting criteria proposed are detailed in Table 1. Criteria where modifications were based on *ATM*-specific data and unpublished modifications are justified in the following sections. Some criteria are not applied because they overlap with others (PP4), ClinGen itself has discarded them (PP5 and BP6) (22) or are not applicable to *ATM* (PP2, BP1, BP3 BP5, see Table 1 footnotes). Combination rules are kept from Richards et al. (1).

#### POPULATION EVIDENCE

Since allelic frequencies in general populations are powerful tools for identifying common benign variants, we used the statistical framework defined by Whiffin et al. (23) to calculate the maximum credible population allele frequency (MCPAF) for *ATM* pathogenic variants with AT data. We obtained a cut-off allele frequency in the general population of 0.005 for BA1 and 0.0005 for BS1. We translated the threshold to population datasets as the lower boundary of their 99% confidence interval and propose to use any of the nonfounder GnomAD v2.1.1 non-cancer populations (24). Due to the low penetrance of *ATM* pathogenic variants for breast cancer, we cannot apply BS2 to healthy heterozygous variant carriers. BS2 is met if we find one homozygous carrier without AT affection. BS2\_Supporting will be applied to 2 homozygous observations with no clinical data provided. As the main manifestations of AT are neurologic, we propose to use the GnomAD v2.1.1 non-neuro dataset.

#### PREDICTIVE EVIDENCE

Regarding splicing alterations, our performance assessment of *in silico* predictors supports the election of the predictor SPiCE (25) for variants affecting the canonical donor splice site, applying PP3 when they exceed the threshold of 0.240 (100% sensitivity), and BP4 when they are below it (with a sensitivity of 89.9% to identify variants not affecting splicing). For variants affecting the canonical acceptor splice site, PP3 is assigned when exceeding the threshold of 0.789 (sensitivity 87.6%) and BP4 when they are under 0.282 (with a sensitivity of 86.3% to identify variants not affecting splicing), no evidence is considered for acceptor variants with scores between 0.282 and 0.789. No called variants account for 6.2% of splicing altering and 10.8% of splicing neutral variants in our dataset (Supplemental Fig. 2). For

activation or creation of splicing sites, we used a combination of predictors such as SpliceSiteFinder-like, MaxEntScan and GeneSplicer, as detailed in Table 1.

In relation to protein predictors for missense variants, we performed a comparative analysis of different tools for the two *ATM* halves (see Supplemental Methods and Supplemental Table 2). Our results sustain the use of the following combinations of two predictors: REVEL plus VEST4 for the N-terminal half (residues 1–1959) and REVEL plus PROVEAN for the C-terminal half (residues 1960–3056). We proposed the same procedure for both halves: PP3 or BP4 is awarded when the 2 predictors assigned to the protein half agree on a damaging effect or an absence of effect, respectively; otherwise, the contribution of *in silico* evidence is not considered.

#### FUNCTIONAL EVIDENCE

Spliceogenic variants are usually confirmed by the study of the RNA of carriers or by mini-gene assays. Splicing analysis in RNA from a carrier, if well designed and performed quantitatively with the appropriate controls, can demonstrate that a variant produces only aberrant transcripts with premature termination codons undergoing nonsense-mediated decay (NMD). We consider that such cases deserve to be very strong pathogenic evidence, PS3\_VeryStrong, analogous to the strength bestowed in ClinGen's PVS1 decision tree (19). We propose a gradual decrease of PS3 strength when the damaging effect is less certain or less severe (Tables 1 and 2).

Protein function assays are quite specific to the gene and associated conditions. AT-patient cells show hypersensitivity to ionizing radiation and other DSB-DNA-inducing agents manifesting as absence of ATM serine 1981 phosphorylation (26), decrease in cell survival, an increased rate of chromosomal aberrations and defects in cell cycle checkpoints (14, 27, 28). Null variants that result in the absence or loss of ATM expression or prevent the Ser1981-mediated activation of ATM, reduce the phosphorylation of numerous substrates and increase the sensitivity to DNA damaging agents have been associated with classical AT phenotypes. On the other hand, missense and splicing variants allowing some ATM expression, thus presenting residual kinase activity and/or intermediate sensitivity to agents that damage DNA, have been associated with AT patients with milder or atypical phenotypes (16, 29). Consequently, we consider that these 3 functional assays are useful for investigating the pathogenicity of *ATM* variants for the 2 phenotypes (Fig. 1, A). We propose to confer different strengths to PS3 depending on how many of the 3 assays are found to be altered, and the extent of the alteration. Thereby, PS3 will be met when the 3 assays are completely altered,

## Classification of ATM Variants

Table 1. ATM variant classification proposed criteria.		
PATHOGENIC CRITERIA		
Criteria	Criteria description	Specification
<b>STAND-ALONE CRITERION</b>		
PVS1_StandAlone	For a full gene deletion, a pathogenic classification is warranted (in the absence of conflicting data). <sup>a</sup>	None
<b>VERY-STRONG CRITERIA</b>		
PVS1	-Null variant (nonsense, frameshift, canonical $\pm 1$ or 2 splice sites, single or multi-exon deletion or tandem duplication) predicted to undergo NMD. <sup>a</sup> OR -Variants disrupting the initiation codon. <sup>b</sup>	None
PS2_VeryStrong or PM6_VeryStrong	AT patients with de novo score $\geq 4.0$ as per ClinGen SVI Recommendation for de novo Criteria (PS2 & PM6) - Version 1.0. <sup>c</sup>	Strength
PS3_VeryStrong	Splicing analysis in RNA from a carrier quantitatively proves that the variant produces a splicing alteration predicted to undergo NMD, and the variant allele does not produce any full-length transcript. See text and Table 2 for details.	Strength
PS4_VeryStrong	Sixteen AT families. <sup>d</sup> It can only be applied to AT families and NOT in: breast cancer families, breast cancer case-control studies, variants that meet BA1 or BS1, nor together with PM3 at any strength.	Strength
PM3_VeryStrong	AT probands with <i>in trans</i> score $\geq 4.0$ as per ClinGen SVI Recommendation for <i>in trans</i> Criterion (PM3) - Version 1.0. <sup>e</sup> It cannot be applied to variants that meet BA1 or BS1, nor together with PS4 at any strength.	Strength
<b>STRONG CRITERIA</b>		
PS1	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change (none of the variants affect splicing according to predictors).	None
PS2 or PM6_Strong	AT patients with a de novo score 2.0-3.75 as per ClinGen SVI Recommendation for de novo Criteria (PS2 & PM6) - Version 1.0. <sup>c</sup>	Strength
PS3	-SPlicing analysis in carrier RNA quantitatively proves that: · the variant alters splicing resulting in a deletion or insertion NOT predicted to undergo NMD but to alter/truncate a region critical to protein function or remove >10% of protein, and the variant allele does not produce any full-length transcript. See text and Table 2 for details. OR - The three following PROTEIN studies performed in AT patients or transfected cells show a strong alteration: · levels of ATM protein (or ATM phosphorylated in Ser1981) · levels of phosphorylation of two ATM substrates · sensitivity to DNA damaging agents. See text and Fig. 1 for details.	ATM-specific
PS4	Four to 15 AT probands. <sup>d</sup> It can only be applied to AT families and NOT in: breast cancer families, breast cancer case-control studies, variants that meet BA1 or BS1, nor together with PM3 at any strength.	Strength
PVS1_Strong	-Nonsense, frameshift, canonical $\pm 1$ or 2 splice sites, single or multi-exon deletion NOT predicted to undergo NMD but to alter/truncate a region critical to protein function or remove >10% of protein. <sup>a</sup> -Also single or multi-exon duplication presumed in tandem with prediction of NMD. <sup>a</sup>	None
PM3_Strong	AT probands with <i>in trans</i> score 2.0-3.75 as per ClinGen SVI Recommendation for <i>in trans</i> Criterion (PM3) - Version 1.0. <sup>e</sup> It cannot be applied to variants that meet BA1 or BS1, nor together with PS4 at any strength.	Strength
<i>Continued</i>		

Downloaded from https://academic.oup.com/clinchem/article/67/3/518/6024902 by guest on 13 July 2021

**Table 1. (continued)**

PATHOGENIC CRITERIA		
Criteria	Criteria description	Specification
PP1_Strong	Co-segregation with AT in multiple affected family members, with $\geq 7$ meioses observed across at least two families. <sup>d,f</sup>	Strength
MODERATE CRITERIA		
PM1	Variant affecting the mutational hotspot codon p. R3008 (NP_000042.3; see Results section) or the autophosphorylation codon p. S1981. See text for reasoning.	ATM-specific
PM2	Absent, or present at $< 0.00001$ (0.001%) allele frequency in gnomAD or another large sequenced population. If multiple alleles are present within any subpopulation, allele frequency in that subpopulation must be $< 0.00002$ (0.002%). <sup>1</sup>	ATM-specific
PM3	AT probands with <i>in trans</i> score 1.0–1.75 as per ClinGen SVI Recommendation for <i>in trans</i> Criterion (PM3) - Version 1.0. <sup>g</sup> It cannot be applied to variants that meet BA1 or BS1, nor together with PS4 at any strength.	Strength
PM4	Protein length changes as a result of in-frame deletions/insertions impacting at least one residue in a critical functional region (see PM1)	ATM-specific
PM5	Missense change at an amino acid residue where a different missense change determined to be pathogenic or likely pathogenic has been seen before. In addition, variant being interrogated must have a BLOSUM62 score equal to or less than the known variant. <sup>1,g</sup>	Restrictive
PM6 or PS2_Moderate	AT patients with de novo score 1.0–1.75 as per ClinGen SVI Recommendation for de novo Criteria (PS2 & PM6) - Version 1.0. <sup>c</sup>	Strength
PVS1_Moderate	Nonsense, frameshift, canonical $\pm 1$ or 2 splice sites, single or multi-exon deletion NEITHER predicted to result in NMD NOR to alter/truncate a region critical to protein function, removing $< 10\%$ of protein. <sup>g</sup>	None
PS3_Moderate	<p>- SPLICING analysis:</p> <ul style="list-style-type: none"> <li>· in patient RNA quantitatively proves that the variant alters splicing resulting in a deletion or insertion NOT predicted to result in NMD but to remove <math>&lt; 10\%</math> of protein, and the variant allele does not produce any full-length transcript; OR</li> <li>· in patient RNA quantitatively proves that the variant produces 90%–99% of altered transcript predicted to undergo NMD; OR</li> <li>· with a mini-gene quantitatively proves that the variant alters splicing resulting in NMD, and the variant allele does not produce any full-length transcript; OR</li> <li>· in patient RNA with NMD inhibition, semi-quantitatively shows with similar band intensity that the variant alters splicing resulting in NMD, without evidence that the variant allele produces any full-length transcript. See text and Table 2 for details.</li> </ul> <p>OR</p> <ul style="list-style-type: none"> <li>· Two of the following PROTEIN studies in AT patients or transfected cells show a strong alteration and the other one shows an intermediate alteration or has not been performed:                             <ul style="list-style-type: none"> <li>· levels of ATM protein (or ATM phosphorylated in Ser1981)</li> <li>· levels of phosphorylation of two ATM substrates</li> <li>· sensitivity to DNA damaging agents. See text and Fig. 1 for details.</li> </ul> </li> </ul>	Strength; ATM-specific
PS4_Moderate	Two to three AT probands. <sup>d</sup> It can only be applied to AT families and NOT in: breast cancer families, breast cancer case-control studies, variants that meet BA1 or BS1, nor together with PM3 at any strength.	Strength

*Continued*

Downloaded from https://academic.oup.com/clinchem/article/67/3/518/6024902 by guest on 13 July 2021



## Classification of ATM Variants

Table 1. (continued)		
PATHOGENIC CRITERIA		
Criteria	Criteria description	Specification
PP1_Moderate	Co-segregation with AT in multiple affected family members, with 5-6 meioses observed. <sup>d,f</sup>	Strength
SUPPORTING CRITERIA		
PP1	Co-segregation with AT in multiple affected family members, with 3-4 meioses observed. <sup>d,f</sup>	ATM-specific
PP2	Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease.	<b>N/A</b> <sup>h</sup>
PP3	-Probability of splicing alteration of the closest natural site predicted with Splice 2.1 is $\geq 0.240$ for donor sites or $\geq 0.789$ for acceptor sites, OR a splicing site is created/activated according to at least 2 splicing predictors of the set SpliceSiteFinderlike-MaxEntScan-NNSplice, with a score higher than the score of the natural site in the mutated allele. <sup>i</sup> OR -Only for missense variants, when the above splicing predictors indicate no impact, but protein predictors do. For variants affecting codons 1-1959, PP3 is met when VEST4 and REVEL predict damaging effects (scores $>0.5$ ). For variants affecting codons 1960-3056, PP3 is met when PROVEAN (score $<-2.5$ ) and REVEL (score $>0.5$ ) predict damaging effects. <sup>j</sup>	ATM-specific
PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic etiology.	<b>N/A</b> (use PS4 instead) <sup>j</sup>
PP5	Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation.	<b>N/A</b> <sup>k</sup>
PS1_Supporting	Different variant at same nucleotide position as a pathogenic SPLICING variant, where <i>in silico</i> models predict impact equal to or greater than the known pathogenic variant.	ATM-specific
PS2_Supporting or PM6_Supporting	AT patients with <i>de novo</i> score 0.5-0.75 as per ClinGen SVI Recommendation for <i>de novo</i> Criteria (PS2 & PM6) - Version 1.0. <sup>c</sup>	ATM-specific
PS3_Supporting	- SPLICING analysis: - with NMD inhibition in carrier RNA shows by visual inspection that the altered and wild-type electrophoretic bands have similar intensity, and the altered transcript is predicted to undergo NMD; OR - found in peer-reviewed article(s), without gel shown or quantitation mentioned, where authors declare that the variant produces a splicing alteration predicted to undergo NMD. See text and Table 2 for details. OR - One of the following PROTEIN studies in AT patients or transfected cells shows a strong alteration and the other two show an intermediate alteration or have not been performed: - levels of ATM protein (or ATM phosphorylated in Ser1981) - levels of phosphorylation of two ATM substrates - sensitivity to DNA damaging agents. See text and Fig. 1 for details.	Strength; ATM-specific
PS4_Supporting	One AT proband. <sup>d</sup> It can only be applied to AT families and NOT in: breast cancer families, breast cancer case-control studies, variants that meet BA1 or BS1, nor together with PM3 at any strength.	ATM-specific
PM1_supporting	Missense or small in-frame deletion or insertion located in the kinase (residues 2712-2962) or FATC (residues 3024-3056) functional domains (NP_000042.3; see results section).	ATM-specific
PM3_Supporting	AT probands with <i>in trans</i> score 0.5-0.75 as per ClinGen SVI Recommendation for <i>in trans</i> Criterion (PM3) - Version 1.0. <sup>a</sup> It cannot be applied to variants that meet BA1 or BS1, nor together with PS4 at any strength.	None

Continued

Downloaded from https://academic.oup.com/clinchem/article/67/3/518/6024902 by guest on 13 July 2021

Table 1. (continued)		
<b>PATHOGENIC CRITERIA</b>		
<b>Criteria</b>	<b>Criteria description</b>	<b>Specification</b>
<b>BENIGN CRITERIA</b>		
<b>Criteria</b>	<b>Criteria description</b>	<b>Specification</b>
<b>STAND-ALONE CRITERION</b>		
BA1	99% confidence interval of the variant allele frequency in any of the NFE, AFR, LAT, EAS, SAS GnomAD v2.1 (non-cancer) populations is > 0.5%. <sup>l</sup>	ATM-specific
<b>STRONG CRITERIA</b>		
BS1	99% confidence interval of the variant allele frequency in any of the NFE, AFR, LAT, EAS, SAS GnomAD v2.1 (non-cancer) populations is > 0.05%. <sup>l</sup>	ATM-specific
BS2	Observed in the homozygous state in a healthy or AT-unaffected individual. One observation if homozygous status confirmed; two if not confirmed.  Note that if BS1 is applied, BS2 must be downgraded to BS2_Supporting. <sup>l</sup>	ATM-specific
BS3	- SPLICING analysis in carrier RNA demonstrate (by Sanger sequencing or a quantitative technique) biallelic expression of the full-length transcript by an exonic SNV. See text and Table 2 for details.  OR -In a variant not predicted or proven to alter RNA splicing, the three following PROTEIN studies in AT patients or transfected cells show results similar to a wild-type control: · levels of ATM protein phosphorylated in Ser1981 · levels of phosphorylation of 2ATM substrates · sensitivity to DNA damaging agents. See text and Fig. 1 for details.	ATM-specific
BS4	Lack of segregation in affected members of 2 or more AT families. <sup>f</sup>	None
<b>SUPPORTING CRITERIA</b>		
BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease.	<b>N/A</b> <sup>h</sup>
BP2	Co-occurrence in trans of the variant with a pathogenic or likely pathogenic ATM variant in well phenotyped AT-unaffected individual from internal cohort or the literature.	ATM-specific
BP3	In-frame deletions/insertions in a repetitive region without a known function	<b>N/A</b> <sup>m</sup>
BP4	-For synonymous and intronic variants, probability of splicing alteration of the closest natural site predicted with SPiCE 2.1 is < 0.240 for donor sites or < 0.282 for acceptor sites, AND no splicing site is created/activated according to at least 2 splicing predictors of the set SpliceSiteFinderlike-MaxEntScan-NNSplice (if a site is recognized, the score is lower than the score of the natural site in the variant allele). <sup>l</sup>  -For coding non-synonymous variants, NEITHER splicing predictors as above NOR protein predictors predict any impact. The latter is established for variants affecting codons 1–1959 when both VEST4 and REVEL (scores <0.5) predict NO alteration, and for variants affecting codons 1960–3056 when both PROVEAN (score >-2.5) and REVEL (score <0.5) predict NO alteration. <sup>l</sup>	ATM-specific
BP5	Variant found in a case with an alternate molecular basis for disease.	<b>N/A</b> <sup>n</sup>
BP6	Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation.	<b>N/A</b> <sup>k</sup>
BP7	Synonymous variant where nucleotide is not highly conserved (100 vertebrates basewise conservation PhyloP score < 6.66, available at the UCSC Browser). <sup>o</sup>  This evidence can be used with BP4, as appropriate, to classify variants meeting both criteria as likely benign.	ATM-specific
<i>Continued</i>		

Downloaded from https://academic.oup.com/clinchem/article/67/3/518/6024902 by guest on 13 July 2021

Classification of *ATM* Variants

Table 1. (continued)		
PATHOGENIC CRITERIA		
Criteria	Criteria description	Specification
BS2_Supporting	Two homozygous observations with no clinical data provided, or meets criteria for BS2 but BS1 is also applied. <sup>f</sup> Observations without clinical data provided can be retrieved from the GnomAD non-neuro dataset.	ATM-specific
BS3_Supporting	- SPLICING analysis in carrier RNA with NMD inhibition and proper controls, shows only the wild-type transcript although do not demonstrate biallelic expression by an exonic SNV. See text and Table 2 for details. OR - In a variant not predicted or proven to alter RNA splicing, two of the following PROTEIN studies in AT patients or transfected cells show results similar to a wild-type control and the other one shows an intermediate alteration or has not been performed: · levels of ATM protein phosphorylated in Ser1981 · levels of phosphorylation of two ATM substrates · sensitivity to DNA damaging agents. See text and Fig. 1 for details.	ATM-specific
BS4_Supporting	Lack of segregation in affected members of one AT family. <sup>f</sup>	ATM-specific

NMD, nonsense-mediated decay; AT, ataxia-telangiectasia; **N/A, Not applicable to ATM.**

<sup>f</sup>Following Tayoun et al., decision tree (19, 20).

<sup>g</sup>Initiation codon variants have been shown to cause (classic or atypical) AT and absence of ATM kinase (21). Expression studies performed in these patients show a shorter underexpressed protein probably starting at the next in-frame methionine at codon 94 (21).

<sup>h</sup>Point-based system to determine the strength of de novo evidence based upon confirmed versus assumed status, phenotypic consistency and number of de novo observations, available at <https://clinicalgenome.org/working-groups/sequence-variant-interpretation/> (20).

<sup>i</sup>Applied analogously to *CDH1* ClinGen Specifications (3, 20).

<sup>j</sup>Point-based system to determine the strength of homozygous and *in trans* observations based upon variant phasing and classification of the variant occurring on the other allele, available at <https://clinicalgenome.org/working-groups/sequence-variant-interpretation/>.

<sup>k</sup>Applied analogously to *PTEN* ClinGen Specifications (2, 20).

<sup>l</sup>If the other missense change is determined to be likely pathogenic, the variant being classified should not reach pathogenic classification.

<sup>m</sup>Both missense and frameshift variants contribute with comparable frequency to ATM-related diseases.

<sup>n</sup>Splicing predictor assessment is detailed in the text. SPICE 2.1 predictions can be found at <https://sourceforge.net/projects/spicev2-1/#>.

<sup>o</sup>Protein predictor assessment is detailed in the text. VEST4 predictions can be found at <http://cravat.us/CRAVAIT/>, REVEL predictions at <https://sites.google.com/site/jpopgen/dbNSFP> and PROVEAN predictions at [http://provean.jcvi.org/genome\\_submit\\_2.php?species=human](http://provean.jcvi.org/genome_submit_2.php?species=human).

<sup>p</sup>Following Biesecker et al., recommendations (20, 22).

<sup>q</sup>The 99% confidence intervals can be calculated in the INVERSE AF tab of the website <http://cardiodb.org/allelefrequencyapp/>; see Materials and methods and Results sections for details on cut-offs.

<sup>r</sup>A repetitive region without a known function has not been found in ATM.

<sup>s</sup>The frequency of pathogenic variants in ATM and other breast cancer predisposing genes is high enough to allow such combinations and a lethal or strikingly stronger phenotype is not anticipated.

<sup>t</sup>The 100 vertebrates base-wise conservation PhyloP score can be seen as a graphic track at the UCSC Genome Browser (<https://genome.ucsc.edu/>), and the scores can be downloaded for each position.

PS3\_Moderate when 2 are altered and 1 has not been performed or gives intermediate results, and PS3\_Supporting when 1 is altered and the other 2 have not been performed or give intermediate results, as depicted in Fig. 1, B.

In cases where experimental data from RNA and protein support the same damaging effect, the evidence of higher strength will be used. When RNA data do not support an effect in splicing, the protein data prevail to reflect other defects in protein function.

Benign criterion BS3 has a similar approach (Tables 1 and 2). If assay(s) in carrier RNA demonstrate biallelic expression of the variant or an exonic single

nucleotide variation (SNV), quantitatively or with similar peak height by Sanger sequencing, BS3 will be met. BS3\_Supporting is achieved when no additional band to wild-type is detected in electrophoresis of carrier RNA, although biallelic expression cannot be demonstrated by an exonic variant. BS3 can be achieved by protein assays when the 3 assays yield the same results as the wild-type control (Fig. 1, B). BS3\_Supporting is met when 2 assays give the same results as wild-type and the other one gives intermediate results or has not been performed.

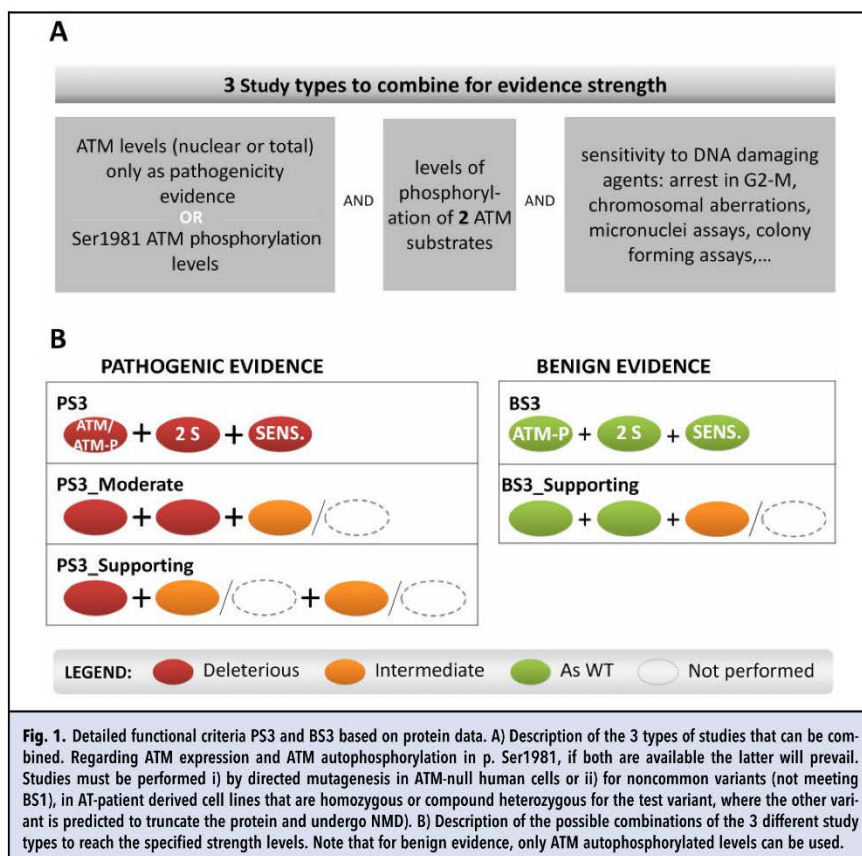
We have found germline deleterious missense variants in AT patients, located throughout the ATM



**Table 2. Detailed criteria for functional evidences PS3 and BS3 based on RNA splicing studies.**

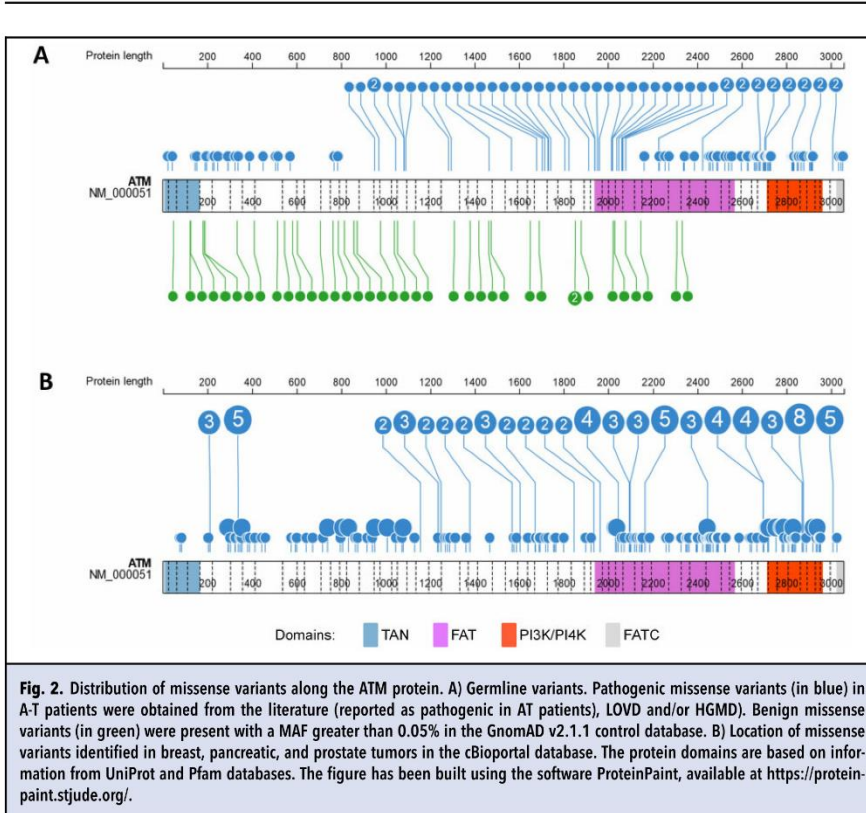
criteria	Assay	NMD	Protein effect / experimental evidence	Transcript effect	Qualitative / quantitative method
<b>PS3_Very Strong</b>	carrier RNA	NMD predicted		variant allele does not produce full-length transcript	quantitative allele specific expression
<b>PS3</b>	carrier RNA	frame-shift or in frame; NMD not predicted	truncates OR alters critical region OR removes >10% protein	variant allele does not produce full-length transcript	quantitative allele specific expression
<b>PS3_Moderate<sup>a</sup></b>	carrier RNA	frame-shift or in frame; NMD not predicted	removes <10% protein (no critical regions)	variant allele does not produce full-length transcript	quantitative allele specific expression
	mingene	NMD predicted		variant allele does not produce full-length transcript	quantitative expression
	carrier RNA	NMD predicted	variant band with similar intensity as WT	NO evidence that the variant allele produces or not full-length transcript	semiquantitative transcript specific expression
	carrier RNA	NMD predicted		variant allele produces 1%-10% full-length transcript	quantitative allele specific expression
<b>PS3_Supporting<sup>a</sup></b>	carrier RNA	NMD predicted	variant band with similar intensity as WT	NO evidence that the variant allele produces or not full-length transcript	visual inspection (gel + Sanger seq)
	carrier RNA	NMD predicted	no electrophoresis result or quantitation shown in peer-reviewed article		
	mingene	NMD predicted		variant allele does not produce full-length transcript	visual inspection (gel + Sanger seq)
<b>BS3</b>	carrier RNA		no additional band is seen / similar peaks by Sanger seq	biallelic expression is demonstrated by an exonic SNV	quantitative / visual inspection
<b>BS3_Supporting</b>	carrier RNA		no additional band is seen	biallelic expression is NOT demonstrated by an exonic SNV	visual inspection

<sup>a</sup>This criterion is met when any of the options represented by the corresponding rows are true.

Classification of *ATM* Variants

protein and a similar distribution was observed in breast, pancreatic, and prostate tumors (Fig. 2, A and B). In contrast, missense variants with a MAF  $\geq 0.05\%$  (gnomAD 2.1 controls) are distributed throughout the ATM regions except for phosphoinositide 3-kinase (PI3K) and FRAP, ATM, TRRAP C-terminal (FATC) domains (Fig. 2, A). The absence of germline frequent variants in PI3K and FATC C-terminal domains suggests their critical role. For this, we propose applying PM1\_supporting to variants located in these specific domains. We have also found some candidate codons for a PM1 hotspot, according to the ClinGen Germline/Somatic Variant Curation Subcommittee (30). Two *ATM* codons accumulate >10 somatic

missense occurrences in cancerhotspots.org (v2) (Supplemental Fig. 3). Codon 337 has 31 observations of p.(Arg337His) and 9 of p.(Arg337Cys); codon 3008 has 15 observations, distributed between p.(Arg3008Cys), p.(Arg3008His), and p.(Arg3008Leu). In GnomAD v2.1.1 (noncancer) variants c.1009C>T p.(Arg337Cys) and c.1010G>A p.(Arg337His) have 26 and 20 counts, respectively, whereas only variants c.9022C>T p.(Arg3008Cys) and c.9023G>A p.(Arg3008His) have been detected, with 3 and 2 counts, respectively. For this reason, we only consider codon 3008 as a hotspot. In addition, we propose applying PM1 to codon p. Ser1981, since autophosphorylation of this residue has been found to be



required for sustained retention of ATM at DSBs. Furthermore, its directed mutagenesis affects the ability of ATM to phosphorylate its downstream targets after DNA damage and correct the radiosensitivity of an AT cell line (31).

**DE NOVO, ALLELIC AND SEGREGATION EVIDENCE**

De novo criteria (PS2, PM6) and allelic evidence PM3 are applied following the ClinGen SVI recommendations (32, 33); segregation criteria (PP1, BS4) are formulated as in published guidelines (2) only for AT families, whereas the benign allelic evidence BP2 has been simplified (Table 1).

**PILOT CLASSIFICATION OF 50 ATM VARIANTS**

We performed a pilot classification of 50 ATM variants from our database which were selected to represent the variant type proportions of the whole set. The evidence

assigned to each variant, the data and publications on which they are based and the resulting pathogenicity classes are displayed in Table 3 and Supplemental Table 3. All this information together with the clinical information (Supplemental Table 4) will be submitted to ClinVar database (34) to be made publicly available to the whole community.

The pilot reclassification of 50 variants with the adapted criteria allowed us to reassign 18 cases from VUS to a more clinically meaningful class; of the remaining cases, 4 were moved to class 3 and 28 were left unchanged (Table 3, Supplemental Fig. 1). Of note, establishing ATM-adjusted cut-offs for BA1 and BS1 favored the classification of several recurrent variants as class 1 or 2. The BS2\_supporting criterion, applied to variants with at least 2 appearances in the GnomAD non-neuro dataset in the homozygous state, supported by the high penetrance and young age of onset observed

## Classification of ATM Variants

Table 3. Result of the 50-variant pilot classification.					
cDNA name	Protein name	Nr carriers	Initial Submitted Classification	Consensus classification	Evidence combination <sup>a</sup>
c.61A>G	p.(Thr21Ala)	1	3	3	PM2 + BP4
c.162T>C	p.(Tyr54=)	9	2, 3	2	BS1 + BP4 + BP7
c.496 + 4T>C	p.?	3	3	2	BP4 + BS3_P
c.609C>T	p.(Asp203=)	14	2, 3	2	BS1 + BP4 + BP7
c.826A>G	p.(Lys276Glu)	1	3	3	PM2
c.998C>T	p.(Ser333Phe)	60	2, 3	2	BS1 + BS2_P
c.1380G>C	p.(Thr460=)	2	2, 3	2	BS1 + BP4 + BP7
c.1463G>A	p.(Trp488*)	1	5	5	PVS1 + PM2 + PM3 + PS3_P
c.1564_1565del	p.(Glu522Ilefs*43)	2	5	5	PVS1 + PM3_VS
c.1810C>T	p.(Pro604Ser)	48	2, 3	1	BA1 (+ BS1 + BS2_P)
c.1899T>G	p.(Cys633Trp)	1	3	3	PM2
c.2012T>A	p.(Ile671Lys)	1	3	3	BP4
c.2250G>A	p.(Lys750=)	1	4	5	PP3 + PS3_M + PM3_VS + BP7
c.2362A>C	p.(Ser788Arg)	1	2	1	BA1 (+ BS1 + BS2_P)
c.2386A>C	p.(Asn796His)	1	3	3	PM2 + BP4
c.2839-2A>G	p.?	1	4	4	PVS1 + PM2
c.2921 + 1G>A	p.?	2	4, 5	5	PVS1 + PM3_VS + PS3_P
c.2921 + 1G>T	p.?	1	5	5	PVS1 + PM2 + PS1_P
c.3747-1G>C	p.?	2	4, 5	5	PVS1 + PS4_P + PM2 + PS3_M
c.3802del	p.(Val1268*)	3	4, 5	5	PVS1 + PM3_VS + PS3
c.4060C>A	p.(Pro1354Thr)	2	3	3	BP4
c.4110-9C>G	p.?	1	3	3	PS3_P + PM3_P + PP3 + PM2
c.4396C>G	p.(Arg1466Gly)	4	3	3	PM2 + PP3
c.4802G>A	p.(Ser1601Asn)	2	2, 3	3	BP4
c.4852C>T	p.(Arg1618*)	1	4	5	PVS1 + PS4_M + PM2
c.5071A>C	p.(Ser1691Arg)	9	2, 3	1	BS1 + BS2_P + BS3
c.5373T>C	p.(Asp1791=)	1	2	3	PM2 + BP4 + BP7
c.5558A>T	p.(Asp1853Val)	32	2, 3	1	BA1 (+ BS1 + BS2_P + PP3)
c.5623C>T	p.(Arg1875*)	2	5	5	PVS1 + PM3_S + PS3_M
c.6067G>A	p.(Gly2023Arg)	19	2, 3	3	BS1 + PP3
c.6115G>A	p.(Glu2039Lys)	1	3	3	PS4_P + PM2 + PP3
c.6203T>C	p.(Leu2068Ser)	1	3	4	PM2 + PS4_M + PS3_M + PP3
c.6315G>C	p.(Arg2105Ser)	1	3	3	PP3
c.6679C>T	p.(Arg2227Cys)	1	4	5	PM2 + PS4 + PP3 + PS3_M + PP1
c.6848C>T	p.(Ser2283Leu)	2	3	3	PM2 + BP4
c.6860G>C	p.(Gly2287Ala)	1	3	3	BP4
c.7135C>G	p.(Leu2379Val)	1	4	3	PS3_M + PP3
c.7191A>G	p.(Gln2397=)	2	3	2	BP4 + BP7
c.7375C>G	p.(Arg2459Gly)	10	3	3	PP3
c.7381C>T	p.(Arg2461Cys)	1	3	3	PP3

Continued

Downloaded from https://academic.oup.com/clinchem/article/67/3/518/6024902 by guest on 13 July 2021



cDNA name	Protein name	Nr carriers	Initial Submitted Classification	Consensus classification	Evidence combination <sup>a</sup>
c.7390T>C	p.(Cys2464Arg)	1	3	3	BS1 + PP3
c.7788 + 3A>G	p.?	1	4	4	PM2 + PM3 + PP3 + PS3_P
c.8122G>A	p.(Asp2708Asn)	1	4	4	PM2 + PP3 + PS3_M + PM3_S
c.8269-5T>G	p.?	1	3	3	PM2
c.8734A>G	p.(Arg2912Gly)	4	3	3	PP3 + PM1_P
c.8786 + 1G>T	p.?	1	5	5	PVS1 + PS3_M + PM2
c.8876_8879del	p.(Asp2959Glyfs*3)	2	5	5	PVS1 + PS3_P + PM3
c.9007_9034del	p.(Asn3003Aspfs*6)	2	4	5	PVS1_S + PS3 + PM2 + PM3
c.9023G>A	p.(Arg3008His)	1	4	4	PM1 + PM2 + PS4_M
c.9079dup	p.(Ser3027Lysfs*36)	1	4	4	PVS1_S + PM2 + PS4_M

See evidence details in Supplemental Table 3.  
<sup>a</sup>code for evidence strength modifications: VS, Very Strong; S, Strong; M, Moderate; P, Supporting.

in AT patients, allowed the classification of 5 variants as likely benign by its combination with BS1, without any other evidence needed. Similarly, our DB recorded the appearance of 3 variants in homozygosis in well-phenotyped individuals not presenting AT. This information allowed us to classify variant c.998C>T as likely benign; this variant was present in 60 patients in our DB but did not reach BS2\_P requirements with GnomAD data.

Nineteen out of 50 variants of the pilot study were classified as class 4 or 5 being present in 27 patients from our Spanish cohort. Most of these patients had breast/ovarian cancer although there were cases of other tumors (Supplemental Table 5). Cosegregation data was available in a few of these families and, as expected for a moderate penetrance cancer risk gene, was not very informative (Supplemental Table 5).

## Discussion

Variant classification is one of the main clinical challenges in the MPS era, being an enormous bottle neck in most genetic testing laboratories. In this article, we present the seed for a Spanish database of hereditary cancer variants, beginning with 6 laboratories and 1 gene, *ATM*. We identified *ATM* as a good candidate since it is one of the genes with more identified VUS (18) and it has been associated with different cancer syndromes (6–8, 10), making it well worth the joint effort to refine variant classification. Since there were no specific criteria for *ATM* variant classification, we also made an effort to adapt the ACMG/AMP guidelines (1) to *ATM*.

In our pilot classification study, the use of *ATM*-specific guidelines and data sharing amongst experts and

clinical laboratories led to a decrease in VUS from 58% to 42%, with the identification of 27 carriers of *ATM* (likely) pathogenic variants. Because pathogenic *ATM* variants predispose to potentially lethal cancers for which there are clinical management recommendations (35), these findings are clearly clinically actionable for carrier individuals and their relatives.

Since the ACMG/AMP classification guidelines were proposed for high-penetrance genes in classical Mendelian disorders (1), their adaptation to moderate/low-penetrance genes, such as *ATM*, is challenging and requires collaborative efforts. In this respect, we analyzed every ACMG/AMP classification criterion in the context of reported knowledge about the *ATM* gene, ATM protein function and *ATM* related phenotypes, with the aim of better adjusting each criterion and eventually facilitating variant classification in routine clinical laboratories. In this process, we took advantage of the fact that biallelic *ATM* variant carriers present the highly penetrant AT disease, allowing the use of criteria for recessive phenotypes. Our adjusted cut-offs for population variant frequency enabled the classification of a large quantity of recurrent variants as (likely) benign that would have been classified as VUS with the original general ACMG/AMP thresholds (1). In this sense, although our DB only contains variants below the common population frequency cut-off of 1%, 39 out of 283 unique variants meet the adjusted BS1 and 12 of these also meet the adjusted BA1. The 39 BS1 variants account for 51% of the individual entries in (448 out of 882, data not shown).

We found it especially challenging to establish functional evidence for or against pathogenicity. At the protein level, the selected assays were based on relevant functional characteristics of ATM that are altered in AT patients and are involved in oncogenic mechanisms, such as double-

### Classification of *ATM* Variants

strand break signaling (presence and activation of *ATM*, phosphorylation of its substrates), and mitosis checkpoints and chromosomal stability. We set the splicing-related criteria at the RNA level on the basis of previous ENIGMA and InSiGHT recommendations and the authors' own experience. Additional considerations will be required if *ATM* naturally occurring in-frame transcripts are described that would rescue the variant allele effect. Functional evidence has helped us to classify 18 out of 50 variants. Luckily, splicing data from 5 of the 8 variants came from our own laboratories. Unfortunately, protein functional studies are not available in our teams. In this sense, the development of calibrated high-throughput *ATM* functional assays, similar to the saturation genome editing study published for *BRCA1* (36) will provide more power to PS3 and BS3 criteria.

The presence of a rare variant in AT families or *in trans* with a (likely) pathogenic variant in an AT patient has allowed us to classify 12 of 50 variants, while cosegregation AT data has turned out to be scarce in the literature. Conversely, *ATM* variant classification is most frequently requested for breast cancer risk assessment, but the great heterogeneity and numerous phenocopies of breast cancer impairs its use in cosegregation or family counting evidence. Large case-control studies by international hereditary cancer consortia like BRIDGES (37) in Europe and CARRIERS (38) in US will hopefully help to classify some of these variants.

An underlying assumption of this and other studies in the field is that the very same spectrum of *ATM* variants causing autosomal recessive AT disease when present in both alleles, cause increased BC risk when present in one allele. Overall, the assumption is probably true, and as far as we know, it holds true for premature termination codon variants expected to cause NMD. Nevertheless, some evidence suggests that subtle differences might also exist. For instance, variant p.(Val2424Gly) is associated with a 6-fold increased BC risk, much higher than average truncating variants. Conversely, the same variant does not cause classical AT, but an attenuated form (11, 39). Another study suggested that the risk of malignancies is higher in individuals with mild *ATM* missense variants producing proteins with residual kinase activity (40).

In summary, by pooling variant information currently stored in individual clinical laboratories, we have developed a general framework for homogeneous and clinically useful variant interpretation in our country. It will also serve for the identification of Spanish founder/recurrent variants and analysis of their associated cancer risk. Moreover, it will facilitate sharing of curated data to international databases. In recent years, similar initiatives focused on the generation of clinical-grade genetic variant databases have been conducted in other countries (41–44). In our case, we have started by adjusting

general ACMG/AMP guidelines to a single gene, *ATM*, with the aim of using them within the framework of molecular diagnostics for HC. In our joint effort we performed a pilot study and classified 50 *ATM* variants carried by 257 index cases. Our results highlight the relevance of data sharing and data curation by multidisciplinary experts to achieve improved variant classification that will eventually improve clinical management.

### Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

**Nonstandard Abbreviations:** HC, hereditary cancer; MPS, massive parallel sequencing; VUS, variants of uncertain clinical significance; ACMG, American College of Medical Genetics and Genomics; AMP, Association for Molecular Pathology; BC, breast cancer; AT, ataxia telangiectasia; PI3K, phosphatidylinositol-3' kinase-related protein kinase; DSB, double-strand break; MRN, Mre11-Rad50-NBS1; DB, Spanish Hereditary Cancer Variant Database; LB, likely benign; LP, likely pathogenic; P, pathogenic; B, benign; NMD, nonsense mediated decay; SNV, single nucleotide variation; PI3K, phosphoinositide 3-kinase; FATC, FRAP, *ATM*, TRRAP C-terminal (domain)

**Human Genes:** *ATM*, *ATM* serine/threonine kinase; *BRCA1*, *BRCA1* DNA repair associated; *BRCA2*, *BRCA2* DNA repair associated; *PTEN*, phosphatase and tensin homolog; *CDH1*, cadherin 1; *TP53*, tumor protein p53

**Author Contributions:** All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.

T. Sánchez, administrative support; G. Capellá, financial support, provision of study material or patients; M. de la Hoya, administrative support; A. Osorio, provision of study material or patients; M. Pineda, provision of study material or patients; A. Vega, provision of study material or patients; C. Lázaro, financial support, provision of study material or patients.

**Authors' Disclosures or Potential Conflicts of Interest:** Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

**Employment or Leadership:** None declared.

**Consultant or Advisory Role:** None declared.

**Stock Ownership:** None declared.

**Honoraria:** None declared.

**Research Funding:** L. Feliubadaló, A-T. Sánchez, G. Capellá, M. Pineda and C. Lázaro, Carlos III National Institute of Health (Spain) funded by 19 FEDER funds—a way to build Europe—(PI19/00553; PI16/00563; PI16/01898; SAF2015-68016-R and CIBERONC), Government of Catalonia (Pla estratègic de recerca i innovació en salut (PERIS\_MedPerCan and URDCat projects: 2017SGR1282, 2017SGR496), CERCA Program: Government of Catalonia to institution. A. López-Novo, Fellowship GAIN, Xunta de Galicia; D. Rueda, Instituto de Salud Carlos III. AES 2019 (PI19/00340); L-M Porras and

X. de la Cruz, Ministerio de Economía y Competitividad (Grant number: SAF2016-80255-R), European Regional Development Fund, Grant/Award Numbers: Interreg program POCTEFA (Grant number: Pirepred, EFA086/15); C. Ruiz-Ponte, the Instituto de Salud Carlos III and co-funded by the European Regional Development Fund (ERDF) (PI17/00509); A. Moles-Fernández, O. Díez and S. Gutiérrez-Enríquez, the Spanish Instituto de Salud Carlos III (ISCIII) funding, an initiative of the Spanish Ministry of Economy and Innovation, partially supported by the European Regional Development FEDER Funds: PI 15/00355, PI16/01218 and PI19/01303 grant numbers, the ISCIII Miguel Servet Program: CP16/00034 contract number; M. Santamarina-Pena, A. Blanco and A. Vega, Spanish Health Research Foundation, Instituto de Salud Carlos III (ISCIII), partially supported by FEDER funds through Research Activity Intensification Program (contract grant numbers: INT15/00070, INT16/00154, INT17/00133), Centro de Investigación Biomédica en Red de Enfermedades Raras CIBERER (ACCI 2016; ER17PIAC7112/2018), Autonomous Government of Galicia (Consolidation and structuring program: IN607B), Fundación Mutua Madrileña (call 2018), L-M Porras, Scholarship from the Departamento

de Santander Colciencias, Colombia. M. de la Hoya, Spanish Instituto de Salud Carlos III (ISCIII) funding (grant PI15/00059), an initiative of the Spanish Ministry of Economy and Innovation partially supported by European Regional Development FEDER Funds. I. J. Molina, Grant AAT- 8GRA02 from Action for A-T, United Kingdom. A. Osorio, Instituto de Salud Carlos III, cofunded by European Regional Development Fund (ERDF), and partially supported by project PI19/00640.

**Expert Testimony:** None declared.

**Patents:** None declared.

**Role of Sponsor:** The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

**Acknowledgments:** We thank the participating patients, and all the staff from our Genetic Diagnostics and Genetic Counseling Units. The VHIO authors acknowledge the Cellex Foundation for providing research facilities.

## References

- Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405-24.
- Mester JL, Ghosh R, Pesaran T, Huether R, Karam R, Hruska KS, et al. Gene-specific criteria for PTEN variant curation: recommendations from the ClinGen PTEN Expert Panel. *Hum Mutat* 2018;39:1581-92.
- Lee K, Krempley K, Roberts ME, Anderson MJ, Carneiro F, Chao E, et al. Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants. *Hum Mutat* 2018;39:1553-68.
- ClinGen TP53 Expert Panel Specifications to the ACMG/AMP Variant Interpretation Guidelines Version 1. TP53 Variant Curation Expert Panel. [https://clinicalgenome.org/site/assets/files/3876/dingen\\_tp53\\_acmg\\_specifications\\_v1.pdf](https://clinicalgenome.org/site/assets/files/3876/dingen_tp53_acmg_specifications_v1.pdf) (Accessed June 2020).
- Swift M, Morrell D, Cromartie E, Chamberlin AR, Skolnick MH, Bishop DT. The incidence and gene frequency of ataxia-telangiectasia in the United States. *Am J Hum Genet* 1986;39:5:573-83.
- Pritchard CC, Mateo J, Walsh MF, De Sarkar N, Abida W, Beltran H, et al. DNA-repair gene mutations in metastatic prostate cancer. *N Engl J Med* 2016;375:443-5.
- Roberts NJ, Jiao Y, Yu J, Kopelovich L, Petersen GM, Bondy ML, et al. ATM mutations in patients with hereditary pancreatic cancer. *Cancer Discov* 2012;2: 41-6.
- Thompson D, Duedal S, Kirner J, McGuffog L, Last J, Reiman A, et al. Cancer risks and mortality in heterozygous ATM mutation carriers. *J Natl Cancer Inst* 2005;97: 813-22.
- Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* 2006;38:873-5.
- Girard E, Eon-Marchais S, Olaso R, Renault AL, Damiola F, Dondon MG, et al. Familial breast cancer and DNA repair genes: Insights into known and novel susceptibility genes from the GENESIS study, and implications for multigene panel testing. *Int J Cancer* 2019;144:1962-74.
- Goldgar DE, Healey S, Dowty JG, Da Silva L, Chen X, Spurdle AB, et al. Rare variants in the ATM gene and risk of breast cancer. *Breast Cancer Res* 2011;13:R73.
- Savitsky K, Bar-Shira A, Gilad S, Rotman G, Ziv Y, Vanagaite L, et al. A single ataxia telangiectasia gene with a product similar to PI-3 kinase. *Science* 1995;268: 1749-53.
- Ambrose M, Gatti RA. Pathogenesis of ataxia-telangiectasia: the next generation of ATM functions. *Blood* 2013;121: 4036-45.
- Lavin MF. Ataxia-telangiectasia: from a rare disorder to a paradigm for cell signalling and cancer. *Nat Rev Mol Cell Biol* 2008;9: 759-69.
- Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 2003;21:577-81.
- Carranza D, Vega AK, Torres-Rusillo S, Montero E, Martínez LJ, Santamaria M, et al. Molecular and functional characterization of a cohort of Spanish patients with ataxia-telangiectasia. *Neuromol Med* 2017;19: 161-74.
- Lee JH, Paull TT. Activation and regulation of ATM kinase activity in response to DNA double-strand breaks. *Oncogene* 2007;26:7741-8.
- Yurgelun MB, Allen B, Kaldate RR, Bowles KR, Judkins T, Kaushik P, et al. Identification of a variety of mutations in cancer predisposition genes in patients with suspected lynch syndrome. *Gastroenterology* 2015;149: 604-13.e20.
- Abou Tayoun AN, Pesaran T, DiSiefano MT, Oza A, Rehm HL, Biesecker LG, Harrison SM. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat* 2018;39:1517-24.
- SVI General Recommendations for Using ACMG/AMP Criteria. ClinGen Sequence Variant Interpretation Working Group. <https://clinicalgenome.org/working-groups/sequence-variant-interpretation/> (Accessed September 2020).
- Byrd PJ, Srinivasan V, Last JJ, Smith A, Biggs P, Carney EF, et al. Severe reaction to radiotherapy for breast cancer as the presenting feature of ataxia telangiectasia. *Br J Cancer* 2012;106:262-8.
- Biesecker LG, Harrison SM. The ACMG/AMP reputable source criteria for the interpretation of sequence variants. *Genet Med* 2018;20:1687-8.
- Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med* 2017;19:1151-8.
- Karczewski KJ, Franciolli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434-43.
- Leman R, Gaildrat P, Le Gac G, Ka C, Fichou Y, Audrezet MP, et al. Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort. *Nucleic Acids Res* 2018;46:7913-23.
- Bakkenist CJ, Kastan MB. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature* 2003;421:499-506.
- Khanna KK, Keating KE, Kozlov S, Scott S, Gatei M, Hobson K, Taya Y, et al. ATM associates with and phosphorylates p53: mapping the region of interaction. *Nat Genet* 1998;20:398-400.
- Scott SP, Bendix R, Chen P, Clark R, Dork T, Lavin MF. Missense mutations but not allelic variants alter the function of ATM by dominant interference in patients with breast cancer. *Proc Natl Acad Sci USA* 2002;99: 925-30.
- Fievet A, Bellanger D, Rieunier G, Dubois d'Enghien C, Sophie J, Calvas P, et al. Functional classification of ATM variants in ataxia-telangiectasia patients. *Hum Mutat* 2019;40:1713-30.
- Walsh MF, Ritter DJ, Kesserwan C, Sonkin D, Chakravarty D, Chao E, et al. Integrating somatic variant data and biomarkers for germline variant classification in cancer predisposition genes. *Hum Mutat* 2018;39:1542-52.
- So S, Davis AJ, Chen DJ. Autophosphorylation at serine 1981 stabilizes ATM at DNA damage sites. *J Cell Biol* 2009;187:977-90.
- SVI Recommendation for in Trans Criterion (PM3)-Version 1.0. ClinGen Sequence Variant Interpretation Working Group. 2019. [https://clinicalgenome.org/site/assets/files/37117/svi\\_proposal\\_for\\_pm3\\_criterion\\_-\\_version\\_1.pdf](https://clinicalgenome.org/site/assets/files/37117/svi_proposal_for_pm3_criterion_-_version_1.pdf) (Accessed September 2020).
- SVI Recommendation for De Novo Criteria (PS2 & PM6)-Version 1.0. ClinGen Sequence Variant Interpretation Working Group. 2018. [https://clinicalgenome.org/site/assets/files/3461/svi\\_proposal\\_for\\_de\\_novo\\_criteria\\_v1\\_0.pdf](https://clinicalgenome.org/site/assets/files/3461/svi_proposal_for_de_novo_criteria_v1_0.pdf) (Accessed September 2020).

## Classification of ATM Variants

34. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;46:D1062-D7.
35. National Comprehensive Cancer Network. Genetic/Familial High-Risk Assessment: Breast, Ovarian and Pancreatic (Version 1.2021). [https://www.nccn.org/professionals/physician\\_gls/pdf/genetics\\_bop.pdf](https://www.nccn.org/professionals/physician_gls/pdf/genetics_bop.pdf) (Accessed September 2020).
36. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, Janizek JD, et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 2018;562:217-22.
37. Breast Cancer Risk after Diagnostic Gene Sequencing (BRIDGES project). <https://bridges-research.eu/> (Accessed September 2020).
38. Liljquist J, Kraft P, Hart SM, Hallberg EJ, Hu C, Moore R, et al. 2016. The CARRIERS consortium: Establishing refined breast cancer risk estimates in known predisposition genes. 2016 AACR 107th Annual Meeting 2016. New Orleans, LA.
39. Decker B, Allen J, Luccarini C, Pooley KA, Shah M, Bolla MK, et al. Rare, protein-truncating variants in ATM, CHEK2 and PALB2, but not XRCC2, are associated with increased breast cancer risks. *J Med Genet* 2017;54:732-41.
40. Schon K, Os N, Oscrift N, Baxendale H, Coffings D, Ray J, et al. Genotype, extrapyramidal features, and severity of variant ataxia-telangiectasia. *Ann Neurol* 2018;85:170-80.
41. Garrett A, Callaway A, Durkie M, Cubuk C, Alikian M, Burghel GJ, et al. Cancer Variant Interpretation Group UK (CanVIG-UK): an exemplar national subspecialty multidisciplinary network. [Epub ahead of print] *J Med Genet* March 13, 2020 as doi:10.1136/jmedgenet-2019-106759.
42. Kamada M, Nakatsui M, Kojima R, Nohara S, Uchino E, Tanishima S, et al. MGenD: an integrated database for Japanese clinical and genomic information. *Hum Genome Var* 2019;6:53.
43. Lerner-Ellis J, Wang M, White S, Lebo MS; Canadian Open Genetics Repository Group. Canadian Open Genetics Repository (COGR): a unified clinical genomics database as a community resource for standardising and sharing genetic interpretations. *J Med Genet* 2015;52:438-45.
44. Pan M, Cong P, Wang Y, Lin C, Yuan Y, Dong J, et al. Novel LOVD databases for hereditary breast cancer and colorectal cancer genes in the Chinese population. *Hum Mutat* 2011;32:1335-40.





## Article 4

Unravelling genetic predisposition to familial breast and ovarian cancer: identification of new susceptibility genes by case-control study

A. Moles-Fernández, E. Aguado-Flor, C. Zamarreño-Pastor, Tu Nguyen-Dumont, Melissa C Southey, M. Antolín, S. Bonache, A. López-Fernández, L. Feliubadaló, J. Fernández-Navarro, C. Lázaro, J. Balmaña, O. Díez, S. Gutiérrez-Enríquez

*Article in preparation*

**“Unravelling genetic predisposition to familial breast and ovarian cancer: identification of new susceptibility genes by case-control study”**

A. Moles-Fernández<sup>1</sup>, E. Aguado-Flor<sup>1</sup>, C. Zamarreño-Pastor<sup>1</sup>, Tu Nguyen-Dumont<sup>2</sup>, Melissa C Southey<sup>2</sup>, M. Antolín<sup>3</sup>, S. Bonache<sup>1</sup>, A. López-Fernández<sup>1</sup>, L. Feliubadaló<sup>4,5</sup>, J. Fernández-Navarro<sup>6</sup>, C. Lázaro<sup>4,5</sup>, J. Balmaña<sup>1</sup>, O. Díez<sup>1,3</sup>, S. Gutiérrez-Enríquez<sup>1</sup>

<sup>1</sup>Hereditary Cancer Genetics, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain,

<sup>2</sup>Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Australia

<sup>3</sup>Vall d'Hebron University Hospital, Barcelona, Spain,

<sup>4</sup>Catalan Institute of Oncology (ICO), Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Hospitalet de Llobregat, Barcelona, Spain,

<sup>5</sup>Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain.

<sup>6</sup>Bioinformatics Unit, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain.

*Manuscript in preparation*

## Introduction

The identification of new susceptibility-related genes to hereditary breast and ovarian cancer (HBOC) could explain the missing heritability in this disease. The advent of massively parallel sequencing has led to testing of multiple genes using panels, whole-exome sequencing (WES) or whole-genome sequencing (WGS) with the objective of uncovering the genetic landscape underlying Mendelian diseases as well as complex traits. WES has become a common approach to identify rare deleterious variants by performing a staged study starting with the sequencing of small cohorts of cases with strong familial aggregation of cancer, highlighting potential candidate genes (Rotunno et al., 2020). Functional analyses or mutational tumor signatures indicating the relevance of candidate genes in developing the disease are also valuable approaches to identify potential risk genes (Polak et al., 2017; Hernández et al., 2018). Following these approaches, a number of genes have been identified as potential candidates in colorectal cancer and polyposis (Te Paske et al., 2020) and hereditary breast and ovarian cancer over the last decade (Rotunno et al., 2020; Subramanian et al., 2020). However, these genes have to be validated in large case-control studies to verify if they are associated with the disease.

Published studies reporting candidate genes comprehend both the identification and the validation using a different cohort of patients and sequencing healthy controls or using public controls databases. For example, germline *RBBP8* variants have recently been associated with early-onset breast cancer by sequencing and functional approaches, firstly identifying potentially deleterious variants in a small cohort of patients, and secondly validating their association by sequencing a large cohort of patients and by functional assays (Zarrizi et al., 2020). In addition, *RECQL5* was highlighted as a related gene by identifying a deleterious variant by WES in an HBOC family and observing an enrichment of deleterious variants in affected patients after comparing with healthy controls (Tavera-Tapia et al., 2019). Other genes such as *NTHL1* or *EDC4*, have also been also associated with a low to moderate risk to develop breast cancer (Hernández et al., 2018; Li et al., 2021b). These are examples of how using massively parallel approaches in patients enables the identification of candidate genes. Our goal is to identify candidate genes in Spanish HBOC families testing negative for *BRCA1* and *BRCA2* genes and their posterior validation by the analysis of cases and controls.

### Patients, materials and methods

#### Selection of patients

- i) Candidate gene identification set: 25 HBOC female patients from 13 Spanish families without pathogenic variants detected in *BRCA1* and *BRCA2* recruited from 2013 to 2015 were selected to be analyzed using exome sequencing. All cases were assessed by High Risk and Cancer Prevention Unit at Vall d'Hebron University Hospital to be eligible for clinical genetic testing for HBOC (Supplementary table 1).
- ii) Candidate gene validation dataset 1: was composed by 1,012 HBOC female index patients from Vall d'Hebron University Hospital (without pathogenic variants identified in *BRCA1* and *BRCA2*) and 488 healthy female control samples. All cases were assessed by High Risk and Cancer Prevention Unit at Vall d'Hebron University Hospital to be eligible for clinical genetic testing for HBOC (Supplementary table 1).

Control samples were gathered from the Spanish National DNA Bank (Salamanca, Spain) and selected randomly from a population of healthy women with an average age of 58.16 (range 47 - 93) years of age with no personal or family history of cancer.

- iii) Candidate gene validation dataset 2: at the moment of the thesis submission is composed of a different group of 638 HBOC patients from Vall d'Hebron University Hospital and Catalan Institute of Oncology (with no identified pathogenic variants in HBOC genes, such as *BRCA1*, *BRCA2*, *PALB2*, *ATM*, *CHEK2*, *TP53*, *RAD51C*, *RAD51D*, *PTEN*, *CDH1*, *BRIP1* and *STK11*), and 202 healthy control samples from the Spanish National DNA Bank (Salamanca, Spain). At the end of the project, it is planned to include a total of 1,100 patient samples and 500 healthy controls. The criteria for cases and controls selection were the same for than candidate gene validation dataset 1.

#### Massive parallel sequencing

The patient's genomic DNA was extracted from whole blood using Puregene Genome DNA purification Kit (Gentra System, Minneapolis, MN, USA), according to the manufacture's standard protocol. DNA concentration was assessed with the Qubit dsDNA BR Assay kit (ThermoFisher Scientific, Florida, USA).

**i) Library preparation for the identification of candidate genes**

Genomic library was prepared from 1 µg of DNA. Exome enrichment was performed using the SureSelect XT HumanAllExon 50Mb (Agilent). Sequencing was performed on a HiSeq2000 instrument (Illumina) with paired-end (2x100) reads, with a coverage of 72% 10X (Fig 1).

**ii) Library preparation for candidate gene validation dataset 1**

All exons and ten bp into each exon-intron boundary of *ALKBH1*, *ALKBH3*, *BLM*, *CAMKK1*, *FANCD2*, *FANCF*, *FANCM*, *NEIL3*, *NTHL1*, *PER1*, *RASSF7*, *RBL1*, *RECQL*, *RECQL4*, *RINT1*, *RUVBL1*, *SALL2*, *SLX4*, *STRADA*, *WRN*, *XRCC4*, and *ZNHIT1* genes, selected by COMPLEXO consortium (Southey et al., 2013), were sequenced using a customized, targeted HaloPlex HS Targeted Enrichment Assay panel (Agilent Technologies, Santa Clara, CA) (Hammet et al., 2019). Library pools were sequenced, 2x100 bp paired-end reads, by the Australian Genome Research Facility (North Melbourne, VIC, Australia) on an Illumina HiSeq2500 sequencer (Illumina, San Diego, CA) with a minimum read depth target of 10x coverage of 94.44% in cases (1,186 mean reads) and 92.86% in controls (1,097 mean reads).

**iii) Library preparation for candidate gene validation dataset 2**

All exon and 50 bp into each exon-intron boundary of *DMC1*, *EDC4*, *MACROD1*, *RALGDS*, *RBBP8*, *RECQL5*, *TDP2*, and *TPMT* selected candidate genes enrichment were sequenced using PARAGON CleanPlex Custom Amplicon Sequencing Targeted Panel. Sequencing, paired-end 2x250, was performed on a MiSeq Genome Analyzer (Illumina, San Diego, CA) with a resulting coverage of 30x in the 98% of targeted sequences.

**Bioinformatic analysis**

**i) Candidate genes identification.** FASTQs were first checked for quality using the tool fastQC and then paired-end aligned to the human genome (hg19) using Burrows-Wheeler Alignment (BWA v. 0.6.2) (Li and Durbin, 2009) with default settings. GATK and VarScan2 (v.2.3.7) (Koboldt et al., 2009) were used to generate variant calls. Finally, ANNOVAR (Wang et al., 2010) was used to annotate single nucleotide variants (SNVs) and insertions/deletions, report functional importance scores, and identify variants reported in the 1000 Genomes Project and dbSNP (Fig 1).

- ii) **Candidate genes validation dataset 1.** Sequencing data were processed, aligned, and analyzed through a pipeline constructed using Seqliner v0.1a (<http://bioinformatics.petermac.org/seqliner>) by Bioinformatic Core Facility of Peter MacCallum Cancer Centre. GATK Unified Genotyper v2.4 (Broad Institute, Cambridge, MA), HaplotypeCaller, and PLATYPUS were used for variant calling.
- iii) **Candidate genes validation dataset 2.** FASTQs were first checked for quality using the tool fastQC and then paired-end aligned to the human genome (hg38) using Burrows-Wheeler Alignment (BWA v. 0.6.2) (Li and Durbin, 2009) with default settings. Strelka2 and HaplotypeCaller were used to perform the variant calling.

#### Variant filtering and annotation

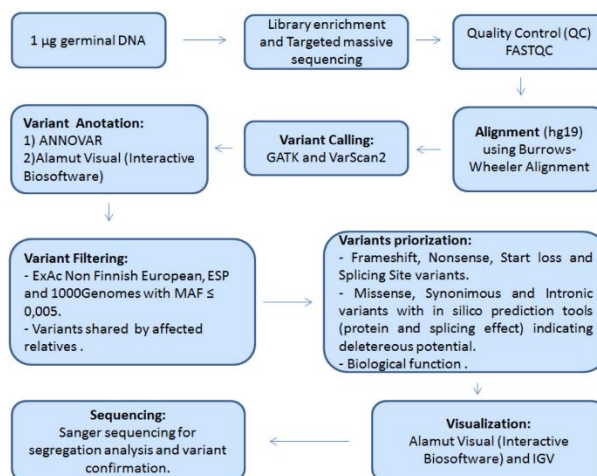
##### i) Exome candidate identification

Variants with >10 reads and alternative allele read frequency higher than 8% were considered. Then, variants were annotated using Alamut Batch software (Interactive Biosoftware). For the *in silico* protein effect, we used AGVD ( $\geq C45$ ), SIFT, and MutationTaster through Alamut Visual. In addition, missense variants were manually annotated with PolyPhen2 (<http://genetics.bwh.harvard.edu/pph2/>). The missense variants with at least three deleterious predictions out of the four *in silico* were prioritized as VUS potentially leading to a functional effect. For variants located in the surrounding exon-intron junction, the *in silico* splicing tools Splice Site Finder (SSF) and Human Splice Finder (HSF) were used as described in Moles-Fernández et al., 2018. (Fig 1).

##### ii) Candidate genes validation 1 and 2

Variants with at least 20 reads and alternative allele read frequency >15% were annotated with Alamut Visual 2.10. For removing sequencing errors and common variants in the data sets, we excluded all variants detected in  $\geq 1\%$  in the cases or control sequenced individuals, a frequency that can be regarded as common and therefore not likely to be related to HBOC. *In silico* splicing tools Splice Site Finder (SSF) and Human Splice Finder (HSF) were used as described in Moles-Fernández et al., 2018 for testing disruption of natural splice sites by variants outside intronic di-nucleotides of donor and acceptor sites. In addition, SpliceAI (threshold 0.2) tool was used to identify exonic and intronic variants creating new splicing sites outside  $\pm 1$  and 2 intronic splicing site positions.

Carrier frequencies in the HBOC cases were compared for each candidate gene in the two sequencing panels with those from GnomAD v2.1.1 healthy non-cancer controls non-Finnish European (NFE) ethnic background (Karczewski et al., 2020). Hence, the variants occurring in each candidate gene were downloaded from GnomAD, and were annotated following the same criteria as those applied for the candidate genes validation sets.



**Figure 1.** Candidate gene identification pipeline through exome sequencing of 25 individuals from 13 families negative for BRCA1 and BRCA2 genes.

### Statistical analysis

Odds ratios (ORs),  $p$  and confidence intervals for every gene for the case-control comparisons were calculated using a two-tailed Fisher's exact test in GraphPad Prism v6. All graphs were plotted using GraphPad Prism v6. Given that in GnomAD 2.1 control database the number of called high quality genotypes, in which the alternate allele count is annotated, is usually not the same among variants, we sought to account for this variability by using the mean of the total number of alleles listed for that gene, instead of using the maximum number of alleles from the population. Then, the sum of these normalised averaged alleles counts in each gene from GnomAD database (for LoF, potentially spliceogenic or missense) was compared with the incidence of variants of interest in cases. To note, that the number of subject carriers was considered to compute ORs.

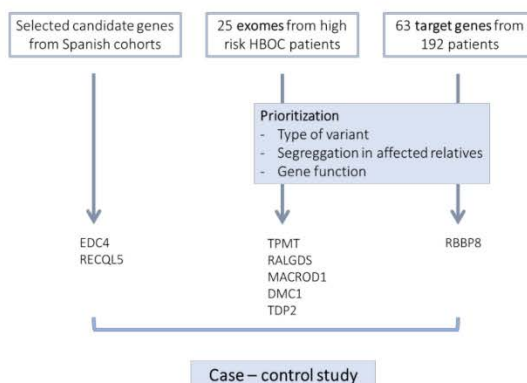


## Results

### Candidate genes identification

Whole exome sequencing was performed in 25 members of 13 multiple-case families affected with early-onset BC without identified *BRCA1* or *BRCA2* pathogenic variant. Rare variants (0.5% minor allele frequency in the non-Finnish European population according to ExAc) shared between the affected sequenced relatives were selected. Then, truncating (except these located in last exons of the genes), potentially spliceogenic or missense variants highlighted as deleterious by *in silico* tools and located in functional protein domains were retained. Then, we selected variants in genes with gene ontology functions related to DNA repair, cellular maintenance, apoptosis or DNA binding domains obtaining a set of 35 potentially deleterious variants (Supplementary Table 2). To prioritize the initial identified variants, we performed a co-segregation analysis (when it was feasible) in affected relatives. The variant c.238G>C in *TPMT* gene showed a co-segregation pattern in two families and the c.82C>G in *TDP2* gene in one family. *DMC1*, *MACROD1* and *RALGDS* genes were further selected because were the top deleterious candidate genes in their respective families. Taking together the segregation evidence, type of variant, and gene function, we selected a list of genes that could be linked with the disease: *DMC1*, *MACROD1*, *RALGDS*, *TDP2*, and *TPMT* (Figure 2; Supplementary Table 2).

Moreover, in previous research performed in our group, an extended panel of 34 known high/moderate-risk cancer genes and 63 "promising candidate" genes related to HBOC was sequenced in 192 *BRCA1* and *BRCA2* negative patients, identifying two *RBBP8* truncating variants in two early onset BC patients without family history of BC (Bonache et al., 2018). The protein coded by this gene is associated with *BRCA1* and with a role in homologous recombination repair and transcriptional regulation of cell cycle checkpoint control (Mozaffari et al., 2021). Thus, we selected *RBBP8* as a candidate to perform a validation analysis. In addition, *RECQL5* and *EDC4* were also selected as candidate genes due to promising results obtained in other Spanish HBOC cohorts (Hernández et al., 2018; Tavera-Tapia et al., 2019). Taking all this evidence together, we decided to explore the association of these eight genes, *DMC1*, *MACROD1*, *RALGDS*, *TDP2*, *TPMT*, *RBBP8*, *RECQL5* and *EDC4*, with HBOC disease in a cohort of cases and healthy controls (Figure 2).



**Figure 2.** Summary of the process followed to select the candidate genes dataset 2 for case-control validation.

Separately, in collaboration with the COMPLEXO consortium (Southey et al., 2013), a set of 22 candidate genes was selected due to previous results suggestive of their association with HBOC or with cellular functions similar to those of HBOC risk genes: *ALKBH1*, *ALKBH3*, *BLM*, *CAMKK1*, *FANCD2*, *FANCF*, *FANCM*, *NEIL3*, *NTHL1*, *PER1*, *RASSF7*, *RBL1*, *RECQL*, *RECQL4*, *RINT1*, *RUVBL1*, *SALL2*, *SLX4*, *STRADA*, *WRN*, *XRCC4*, and *ZNHIT1*.

#### Candidate genes dataset 1 validation

Coding and adjacent intronic regions of candidate genes *ALKBH1*, *ALKBH3*, *BLM*, *CAMKK1*, *FANCD2*, *FANCF*, *FANCM*, *NEIL3*, *NTHL1*, *PER1*, *RASSF7*, *RBL1*, *RECQL*, *RECQL4*, *RINT1*, *RUVBL1*, *SALL2*, *SLX4*, *STRADA*, *WRN*, *XRCC4*, *ZNHIT1* were sequenced in 1,012 HBOC patients and 488 healthy controls. Stop gain, frameshift and +1,2 intronic splicing donor and acceptor variants, considered as protein-coding loss-of-function (LoF), were identified in all genes. After variant annotation, the incidence of this type of variants in both groups was compared.

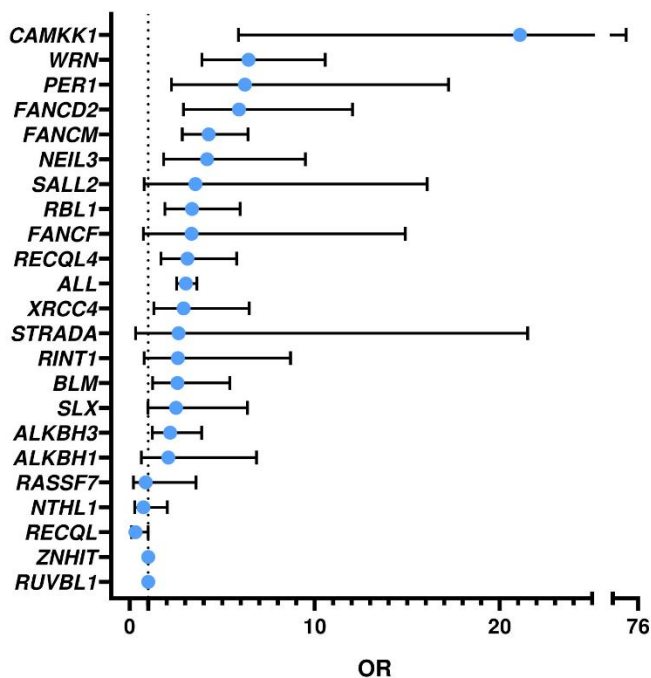
Due to the small number of identified LoF variants in the healthy Spanish controls, variant frequencies were compared between patients and GnomAD v2.1.1 controls. Protein truncating variants in *CAMKK1*, *WRN*, *PER1*, *FANCD2*, *FANCM*, *NEIL3*, *RBL1*, *XRCC4*, *BLM*, and *ALKBH3* were associated with a significant ( $P < 0.05$ ) risk of breast cancer, reaching an OR above 2 (Table 1 comparative OR and Fig 3).

## RESULTS

**Table 1.** Overall hereditary breast ovarian cancer risk associated with LoF Variants in candidate gene dataset 1. No LoF variants were detected in *RUVBL1* and *ZNHIT* genes. The genes are listed in decreasing order of estimated odds ratio for breast cancer.

Genes	n° of carriers of LoF variants in cases (%) (n=1,012)	n° of carriers of LoF variants in GnomAD controls* (%)	OR (95%CI) cases and GnomAD controls	p	n° of carriers of LoF variants in controls (%) (n=488)
<i>CAMKK1</i>	5 (0.494)	4.493 (0.023)	21.09 (5.890 - 75.53)	< 0,0001	0 (0)
<i>WRN</i>	21 (2.075)	64.71 (0.328)	6.432 (3.915 - 10.56)	< 0,0001	7 (1.434)
<i>PER1</i>	5 (0.494)	14.71 (0.079)	6.235 (2.256 - 17.23)	0.0005	0 (0)
<i>FANCD2</i>	10 (0.988)	32.70 (0.168)	5.915 (2.905 - 12.04)	< 0.0001	0 (0)
<i>FANCM</i>	29 (2.865)	135.0 (0.686)	4.265 (2.841 - 6.402)	< 0.0001	1 (0.205)
<i>NEIL3</i>	7 (0.691)	32.37 (0.166)	4.184 (1.844 - 9.496)	0.0007	0 (0)
<i>SALL2</i>	2 (0.197)	10.92 (0.055)	3.556 (0.786 - 16.07)	0.2675	0 (0)
<i>RBL1</i>	14 (1.383)	76.50 (0.414)	3.368 (1.898 - 5.976)	< 0.0001	0 (0)
<i>FANCF</i>	2 (0.197)	12.37 (0.059)	3.340 (0.749 - 14.89)	0.2774	0 (0)
<i>RECQL4</i>	12 (1.185)	69.79 (0.382)	3.128 (1.689 - 5.790)	< 0.0001	1 (0.205)
<i>XRCC4</i>	7 (0.691)	46.39 (0.238)	2.915 (1.313 - 6.470)	0.0138	0 (0)
<i>STRADA</i>	1 (0.098)	6.902 (0.037)	2.640 (0.323 - 21.52)	0.8941	0 (0)
<i>RINT1</i>	3 (0.296)	22.21 (0.114)	2.601 (0.777 - 8.700)	0.2433	0 (0)
<i>BLM</i>	8 (0.790)	60.32 (0.307)	2.582 (1.231 - 5.413)	0.0192	0 (0)
<i>SLX</i>	5 (0.494)	39.89 (0.197)	2.505 (0.986 - 6.363)	0.1	0 (0)
<i>ALKBH3</i>	13 (1.284)	120.7 (0.592)	2.183 (1.227 - 3.882)	0.0119	0 (0)
<i>ALKBH1</i>	3 (0.296)	28.46 (0.142)	2.082 (0.632 - 6.855)	0.3994	0 (0)
<i>RASSF7</i>	2 (0.197)	36.87 (0.228)	0.863 (0.207 - 3.586)	0.8946	1 (0.205)
<i>NTHL1</i>	4 (0.395)	98.56 (0.527)	0.748 (0.274 - 2.038)	0.7235	3 (0.615)
<i>RECQL</i>	3 (0.296)	171.5 (0.926)	0.318 (0.101 - 0.997)	0.0566	0 (0)
<b>ALL</b>	<b>158 (15.61)</b>	<b>1106.45 (5.748)</b>	<b>3.033 (2.533 - 3.632)</b>	<b>&lt; 0.0001</b>	<b>14 (2.869)</b>

\* GnomAD LoF carriers are normalised values (described in methods section).



**Figure 1.** Risk of hereditary breast ovarian cancer associated with LoF variants in candidate dataset 1. Odds ratios and 95% confidence intervals (CIs) for breast cancer associated with LoF variants in 22 genes and its combination (ALL). The genes are listed in decreasing order of estimated odds ratios for breast cancer. ZNHIT and RUVBL1 is indicated as OR = 1, due to no LoF variant was identified in cases.

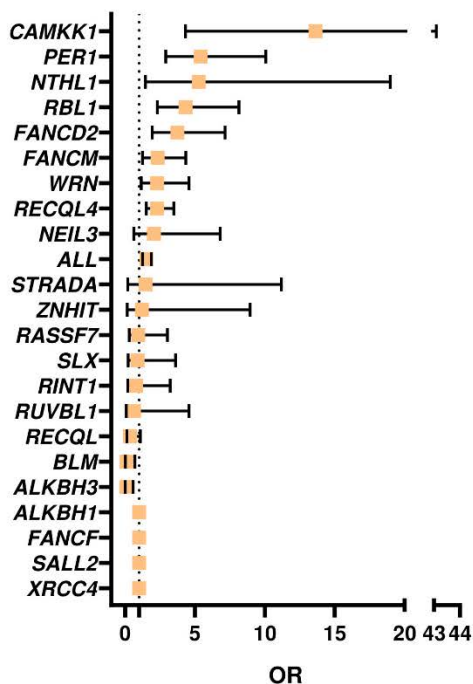
In addition, after the identification of variants potentially altering splicing sites located outside  $\pm 1,2$  intronic splicing donor and acceptor positions or creating new splicing sites (using *in silico* tools and thresholds described in methods), the incidence of predicted spliceogenic variants were compared between cases and GnomAD controls (Table 2, Fig 4). CAMKK1, PER1, NTHL1, RBL1, FANCD2, FANCM, WRN, and RECQL4, showed significant ORs higher than 2.

## RESULTS

**Table 2.** Overall risk of hereditary breast ovarian cancer associated with predicted spliceogenic variants, located outside +1,2 intronic splicing donor and acceptor positions in candidate dataset 1. No predicted spliceogenic variants were identified in *ALKBH1*, *FANCF*, *SALL2*, and *ZNHIT1*. Genes are ranked according their estimated ORs.

Genes	n° of carriers of spliceogenic variants in cases (%) (n=1,012)	n° of carriers of spliceogenic variants in GnomAD controls* (%)	OR (95%CI)	p	n° of carriers of spliceogenic variants in controls (%) (n=488)
<i>CAMKK1</i>	5 (0.494)	6.952 (0.036)	13.63 (4.311 - 43.09)	0.0002	1 (0.205)
<i>PER1</i>	13 (1.284)	44.34 (0.239)	5.412 (2.907 - 10.07)	<0.0001	3 (0.615)
<i>NTHL1</i>	3 (0.296)	10.56 (0.056)	5.256 (1.456 - 18.97)	0.0322	0 (0)
<i>RBL1</i>	12 (1.185)	50.98 (0.276)	4.329 (2.301 - 8.146)	<0.0001	1 (0.205)
<i>FANCD2</i>	11 (1.086)	56.95 (0.293)	3.735 (1.953 - 7.146)	0.0005	4 (0.820)
<i>FANCM</i>	11 (1.086)	92.83 (0.472)	2.316 (1.235 - 4.341)	0.0176	6 (1.230)
<i>WRN</i>	9 (0.889)	77.02 (0.390)	2.287 (1.143 - 4.574)	0.0377	0 (0)
<i>RECQL4</i>	25 (2.470)	200.1 (1.096)	2.285 (1.500 - 3.480)	0.0004	5 (1.025)
<i>NEIL3</i>	3 (0.296)	28.05 (0.144)	2.061 (0.625 - 6.792)	0.1957	1 (0.205)
<i>STRADA</i>	1 (0.098)	12.50 (0.067)	1.456 (0.189 - 11.17)	0.5270	1 (0.205)
<i>ZNHIT</i>	1 (0.098)	16.77 (0.083)	1.188 (0.157 - 8.944)	>0.9999	0 (0)
<i>RASSF7</i>	3 (0.296)	50.80 (0.315)	0.939 (0.292 - 3.016)	>0.9999	0 (0)
<i>SLX4</i>	2 (0.197)	45.47 (0.225)	0.876 (0.212 - 3.616)	>0.9999	3 (0.615)
<i>RINT1</i>	2 (0.197)	49.27 (0.253)	0.779 (0.189 - 3.211)	>0.9999	0 (0)
<i>RUVBL1</i>	1 (0.098)	30.67 (0.158)	0.623 (0.084 - 4.571)	>0.9999	1 (0.205)
<i>RECQL</i>	3 (0.296)	158.1 (0.853)	0.345 (0.109 - 1.083)	0.0702	0 (0)
<i>BLM</i>	1 (0.098)	193.7 (0.988)	0.099 (0.013 - 0.707)	0.0012	0 (0)
<i>ALKBH3</i>	1 (0.098)	245.5 (1.204)	0.081 (0.011 - 0.578)	0.0002	1 (0.205)
<b>ALL</b>	<b>107 (10.57)</b>	<b>1386.59 (7.203)</b>	<b>1.522 (1.237 - 1.874)</b>	<b>0.0002</b>	<b>27 (0.055)</b>

\* GnomAD individuals with spliceogenic variants are normalised values (described in methods section)



**Figure 2.** Risk of hereditary breast ovarian cancer associated with predicted spliceogenic variants in candidate dataset 1. Odds ratios and 95% confidence intervals (CIs) for breast cancer associated with potentially spliceogenic variants in 22 genes and its combination (ALL). The genes are listed decreasing order of estimated odds ratios for breast cancer. ALKBH1, FANCF, SALL2 and ZNHIT are indicated as OR = 1, since no predicted spliceogenic variant was identified in these genes.

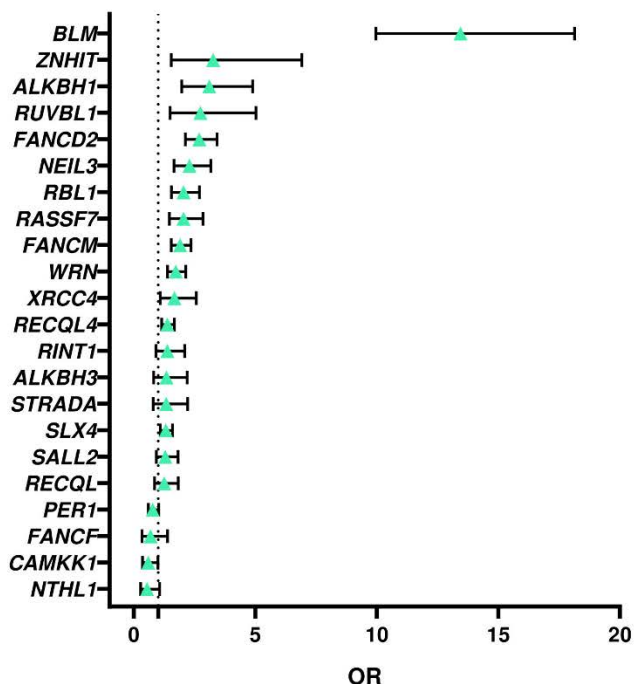
Furthermore, incidence of rare missense (GnomAD controls NFEF  $\leq 0.5\%$ ) were compared between 1,012 cases and 488 healthy Spanish controls, and GnomAD controls 2.1 (Table 3 and Figure 5).

## RESULTS

**Table 3.** Overall risk of hereditary breast ovarian cancer associated with missense variants in candidate dataset 1. Genes are ranked according their estimated ORs.

Genes	n° of carriers of rare missense variants in cases (%) (n=1,012)	N° of carriers of rare missense variants in Spanish controls (%)	OR (95% CI)	p	n° of carriers of rare missense variants in controls (%) (n=488)
<i>BLM</i>	74 (7.312)	114.436 (0.583)	13.44 (9.961 - 18.13)	<0.0001	17 (3.483)
<i>ZNHIT</i>	8 (0.790)	49.1049 (0.243)	3.265 (1.542 - 6.912)	0.0054	0 (0)
<i>ALKBH1</i>	22 (2.173)	141.801 (0.710)	3.106 (1.973 - 4.890)	<0.0001	7 (1.434)
<i>RUVBL1</i>	12 (1.185)	84.4582 (0.436)	2.738 (1.491 - 5.030)	0.0027	2 (0.409)
<i>FANCD2</i>	82 (8.102)	614.908 (3.166)	2.696 (2.121 - 3.426)	<0.0001	17 (3.483)
<i>NEIL3</i>	42 (4.150)	361.457 (1.855)	2.290 (1.653 - 3.173)	<0.0001	8 (1.639)
<i>RBL1</i>	59 (5.830)	541.719 (2.936)	2.046 (1.552 - 2.698)	<0.0001	11 (2.254)
<i>FANCM</i>	41 (4.051)	326.093 (2.024)	2.043 (1.467 - 2.846)	<0.0001	10 (2.049)
<i>RASSF7</i>	101 (9.980)	1081.99 (5.503)	1.903 (1.536 - 2.359)	<0.0001	16 (3.278)
<i>WRN</i>	98 (9.683)	1154.66 (5.858)	1.722 (1.387 - 2.139)	<0.0001	29 (5.942)
<i>XRCC4</i>	23 (2.272)	266.872 (1.370)	1.673 (1.087 - 2.573)	0.0191	5 (1.024)
<i>RECQL4</i>	135 (13.33)	1829.03 (10.01)	1.383 (1.146 - 1.668)	<0.0001	43 (8.811)
<i>SLX4</i>	24 (2.371)	336.872 (1.731)	1.378 (0.906 - 2.096)	0.1083	6 (1.229)
<i>RINT1</i>	17 (1.679)	256.941 (1.260)	1.338 (0.816 - 2.195)	0.2451	4 (0.819)
<i>STRADA</i>	16 (1.581)	220.207 (1.194)	1.328 (0.796 - 2.215)	0.2359	1 (0.204)
<i>ALKBH3</i>	123 (12.15)	1919.92 (9.516)	1.315 (1.083 - 1.597)	0.0003	39 (7.991)
<i>RECQL</i>	37 (3.656)	557.969 (2.841)	1.297 (0.924 - 1.821)	0.0964	15 (3.073)
<i>SALL2</i>	29 (2.865)	426.938 (2.304)	1.250 (0.853 - 1.831)	0.1943	5 (1.024)
<i>PER1</i>	59 (5.830)	1360.21 (7.356)	0.779 (0.595 - 1.019)	0.224	18 (3.688)
<i>FANCF</i>	8 (0.790)	240.111 (1.149)	0.685 (0.337 - 1.389)	0.361	7 (1.434)
<i>CAMKK1</i>	15 (1.482)	472.691 (2.475)	0.592 (0.353 - 0.994)	0.0562	5 (1.024)
<i>NTHL1</i>	9 (0.889)	303.074 (1.621)	0.544 (0.279 - 1.059)	0.0889	2 (0.409)





**Figure 5.** Risk of hereditary breast ovarian cancer associated with missense variants in candidate genes dataset 1. Odds ratios and 95% confidence intervals (CIs) for breast cancer associated with predicted deleterious missense variants in 22 genes and its combination (ALL). The genes are listed in decreasing order of their estimated odds ratios for breast cancer. Only showed ORs of genes with identified variants in the sequenced cohort.

**Candidate genes dataset 2**

Coding and adjacent intronic regions of selected candidate genes *DMC1*, *EDC4*, *MACROD1*, *RALGDS*, *RBBP8*, *RECQL5*, *TDP2*, and *TPMT* were sequenced in 638 HBOC patients and 206 healthy controls. Stop gain, frameshift, and +1,2 intronic splicing donor and acceptor variants considered as protein-coding loss-of-function (LoF), were identified in *DMC1*, *MACROD1*, *RALGDS*, and *RECQL5* (Table 4). No LoF variant was identified in the 206 healthy controls.



**Table 4.** LoF variants identified in cases in candidate genes dataset 2 validation.

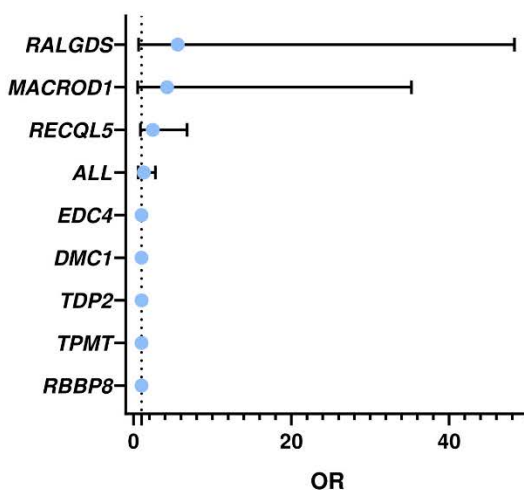
Gene	Coding Effect	cNomenclature	pNomenclature	Nº patients
<b>DMC1</b>	frameshift	c.804_807del	p.(Asn268Lysfs*2)	1
<b>MACROD1</b>	frameshift	c.866_867del	p.(Glu289Valfs*58)	1
<b>RALGDS</b>	frameshift	c.2183_2184del	p.(Pro728Argfs*4)	1
<b>RECQL5</b>	stop gain	c.2308C>T	p.(Arg770*)	2
	splicing	c.1812+2T>C		1
	splicing	c.1948-1G>A		1

Since no LoF variants were detected in controls due to the small number of sequenced DNAs, the number of individuals with LoF variants were compared between patients and GnomAD v2.1.1 controls (Table 5, Fig 6). LoF variants in *RALGDS*, *MACROD1* and *RECQL5* showed an OR value above 2 (Table 5 and Figure 6). However, none of the genes reached a significant ( $p < 0.05$ ) association. A higher number of patients will be analyzed to complete this validation dataset study.

**Table 5.** Risk of hereditary breast ovarian cancer associated with protein-coding LoF variants in candidate dataset 2. No LoF variants were detected in *EDC4*, *TDP2*, *TPMT* and *RBBP8*. The genes are ranked according to their OR estimated value. ALL includes *EDC4*, *TDP2*, *TPMT* and *RBBP8* genes.

Genes	nº of carriers of LoF variants in cases (%) (n=638)	nº of carriers of LoF variants in GnomAD controls* (%)	OR (95%CI)	p	nº of carriers of LoF variants in controls (%) (n=206)
<b>RALGDS</b>	1 (0.156)	4.798 (0.028)	5.593 (0.647 - 48.30)	0.197	0 (0)
<b>MACROD1</b>	1 (0.156)	6.231 (0.036)	4.258 (0.514 - 35.22)	0.228	0 (0)
<b>RECQL5</b>	4 (0.626)	47.202 (0.258)	2.435 (0.874 - 6.778)	0.092	0 (0)
<b>DMC1</b>	1 (0.156)	33.712 (0.168)	0.929 (0.127 - 6.806)	>0.999	0 (0)
<b>ALL</b>	7 (1.097)	158.299 (0.849)	1.293 (0.604 - 2.767)	0.5064	0 (0)

\* GnomAD LoF variant carriers are normalised values (described in methods section).



**Figure 6.** Risk of hereditary breast ovarian cancer associated with LoF variants in candidate dataset 2. Odds ratios and 95% confidence intervals (CIs) for breast cancer associated LoF variants in 8 genes and its combination (ALL). The genes are listed in decreasing order of estimated odds ratio for breast cancer. EDC4, DMC1, TDP2, TPMT and RBBP8 are indicated as OR = 1, due to no LoF were identified in cases.

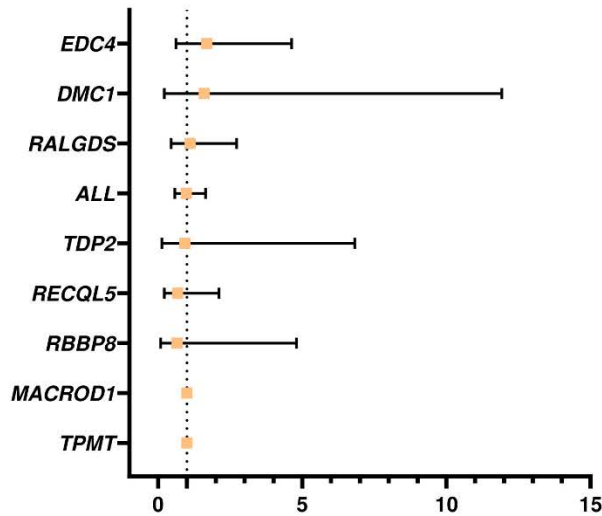
In addition, after the identification of variants potentially altering splicing sites or creating new splicing site (in silico tools used and threshold described in methods), the incidence of spliceogenic variants were compared between cases and GnomAD controls. The results show no statistical ORs for the analysed genes (Table 6, Fig 7).

RESULTS

**Table 6.** Table Risk of hereditary breast ovarian cancer associated with variants predicted as spliceogenic in candidate dataset 2 (truncating variants excluded). No spliceogenic variants were predicted in TPMT and MACROD1. The genes are ranked according their OR values.

Genes	n° of carriers of spliceogenic * variants in cases (%) (n=638)	n° of carriers of spliceogenic variants in GnomAD controls (%)	OR (95%CI)	p	n° of carriers of spliceogenic variants in controls (%) (n=206)
<b>EDC4</b>	4 (0.626)	74.08 (0.372)	1.686 (0.614 - 4.626)	0.3092	0 (0)
<b>DMC1</b>	1 (0.156)	19.65 (0.098)	1.596 (0.213 - 11.92)	0.4665	0 (0)
<b>RALGDS</b>	5 (0.783)	120.86 (0.706)	1.109 (0.452 - 2.724)	0.8066	0 (0)
<b>TDP2</b>	1 (0.156)	30.31 (0.168)	0.929 (0.126 - 6.825)	>0.9999	2 (0.971)
<b>RECQL5</b>	3 (0.470)	127.54 (0.698)	0.671 (0.213 - 2.116)	0.8046	1 (0.485)
<b>RBBP8</b>	1 (0.156)	46.1 (0.237)	0.660 (0.090 - 4.798)	>0.9999	0 (0)
<b>ALL</b>	15 (2.351)	448.10 (2.404)	0.977 (0.580 - 1.645)	>0.9999	4 (1.941)

\*Spliceogenic refer to those variants located outside intonic -1 -2 / +1 +2 constitutive positions of the acceptor and donor splice sites, respectively.



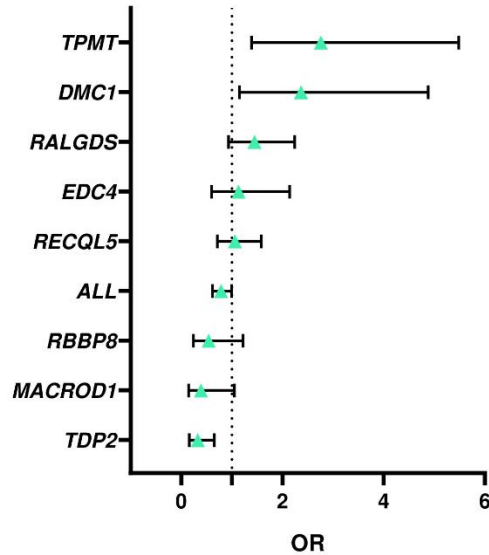
**Figure 7.** Risk of hereditary breast ovarian cancer associated with variants *in silico* predicted as spliceogenic (outside +1,2 intronic splicing donor and acceptor positions) in candidate dataset 2. Odds ratios and 95% confidence intervals (CIs) for breast cancer associated spliceogenic variants in eight genes and its combination (ALL). The genes are listed in order of decreasing estimated odds ratios for breast cancer. *TPMT* and *MACROD1* are indicated as OR = 1, due to no spliceogenic variant was identified in cases.

Regarding missense variants, rare variants in *TPMT* and *DMC1* were associated with a risk of cancer (Table 7 and Figure 8).

**Table 7.** Overall risk of hereditary breast ovarian cancer associated with missense variants in candidate dataset 2. Genes are ranked according their estimated ORs.

Genes	n° of carriers of missense variants		OR (95%CI)	p	n° of carriers of missense variants in controls (%) (n=206)
	in cases (n=638)	variants in GnomAD controls* (%)			
<i>TPMT</i>	9 (1.410)	100.712 (0.515)	2.759 (1.388 - 5.482)	0.0078	0 (0)
<i>DMC1</i>	8 (1.253)	106.705 (0.533)	2.367 (1.149 - 4.878)	0.0252	0 (0)
<i>RALGDS</i>	22 (3.448)	411.292 (2.405)	1.449 (0.936 - 2.242)	0.1144	7 (3.398)
<i>EDC4</i>	10 (1.567)	275.000 (1.383)	1.135 (0.600 - 2.143)	0.6074	7 (3.398)
<i>RECQL5</i>	26 (4.075)	702.360 (3.845)	1.062 (0.712 - 1.584)	0.7531	9 (4.368)
<i>RBBP8</i>	6 (0.940)	334.191 (1.718)	0.542 (0.241 - 1.221)	0.1592	3 (1.456)
<i>MACROD1</i>	4 (0.626)	268.299 (1.586)	0.391 (0.145 - 1.053)	0.05	2 (0.970)
<i>TDP2</i>	8 (1.253)	674.265 (3.750)	0.325 (0.161 - 0.657)	0.0003	5 (2.427)

## RESULTS



**Figure 8.** Risk of hereditary breast ovarian Cancer Associated with missense variants in candidate dataset 2. Odds ratios and 95% confidence intervals (CIs) for breast cancer associated with predicted deleterious missense variants in *n* genes and its combination (ALL). The genes are listed in decreasing order of their estimated odds ratios for breast cancer. Only showed ORs of genes with identified variants in the sequenced cohort.

### Discussion

Since up to 50% of HBOC patients remain without a genetic diagnosis, identifying new susceptibility-related genes to HBOC could explain the missing heritability in this disease. For this reason, the aim of this work was to identify candidate genes related to the HBOC and validate them using a case-control approach. For this purpose, two different sets of candidate genes were separately analyzed. To maximize the identification of deleterious variants, *in silico* splicing tools were used and their incidence was compared between cases and controls. To our knowledge, this is the first case-control study that uses *in silico* splicing tools to enrich the genetic landscape with variants potentially affecting splicing outside the intronic +1,2 of donor and acceptor sites.

In candidate genes dataset 1, twenty-two genes selected in collaboration with COMPLEXO consortium (Southey et al., 2013) were sequenced in 1,012 HBOC Spanish patients without

identified *BRCA1* or *BRCA2* pathogenic variants and 488 Spanish healthy women. The ORs calculation for classic LoF variants indicates that *ALKBH3*, *BLM*, *CAMKK1*, *FANCD2*, *FANCM*, *NEIL3*, *PER1*, *RBL1*, *RECQL4*, *WRN* and *XRCC4* were associated with a significant ( $p < 0.05$ ) risk of breast cancer overall, reaching an OR above 2. The identification of an enrichment in cases for deleterious variants in these genes is consistent with the possibility that an impaired function of these genes may predispose to breast cancer.

Of note, *ALKBH3*, *BLM*, *CAMKK1*, *FANCD2*, *FANCM*, *NEIL3*, *PER1*, *RBL1*, *RECQL4*, *WRN* and *XRCC4* genes showed significant ORs above 2 ( $p < 0.05$ ), in at least two of the three assessed groups of variants with potential deleterious effect: LoF, potentially spliceogenic or missense variants. Of the genes mentioned above, *BLM*, *FANCD2*, *FANCM*, *NEIL3*, *RECQL4* and *WRN* are involved in DNA repair, and some publications have suggested association with cancer risk. *BLM* has been associated with prostate cancer and mesothelioma risk (Ledet et al., 2019; Bononi et al., 2020). *FANCD2* gene has been linked to hereditary breast cancer susceptibility (Mantere et al., 2017). Interestingly, *FANCM* showed some evidence of an association as a risk factor for BC (Peterlongo et al., 2015), and with ER-negative breast cancer in Dorling et al., 2021. In addition, *WRN* helicase gene has been associated with recessive Werner Syndrome, a genetic instability/cancer predisposition disorder (Lebel and Monnat, 2018). On the other hand, limited results have been published related to *NEIL3* and cancer risk (Rolseth et al., 2017; Li et al., 2020). The inactivation of *RBL1*, a tumor suppressor gene involved in cell cycle regulation, has been associated with retinoblastoma susceptibility (Di Fiore et al., 2013). Furthermore, some evidence suggested a relationship between *RINT1* gene (interactor of *RAD50* gene) and breast cancer and Lynch syndrome risk (Park et al., 2014). *PER1* (involved in circadian rhythm maintenance) and *CAMKK1* (calcium/calmodulin-dependent kinase cascade) are genes not related to DNA repair and have not been previously associated to breast cancer risk.

Recently, Li et al 2021 in a large multicenter study with Australian familial breast cancer analyzed 26 of the 30 genes included in this study (Suppl table 3). They sequenced candidate genes in 3,892 BC cases and controls, and then validated 145 shortlisted genes in 7,619 subjects (cases and controls), identifying an overall excess of LoF and rare missense variants in cases. However, candidate genes with LoF variants with ORs of 2-4 did not account for even a 1% of cases in Li et al., study. This result suggests that much of the remaining genetic causes of high-risk BC families are due to genes in which pathogenic variants are both infrequent and convey only low to moderate risk. Larger cases and controls must be sequenced to determine the relevance and association between these genes and the disease. In addition, there are discrepancies between

## RESULTS

our work and results observed in Li et al., 2021, being *BLM* and *WRN* the only ones of 26 genes analyzed in both studies that reached significant ORs above 2 (Suppl. table 3).

In candidate gene dataset 2, *DMC1*, *MACROD1*, *RALGDS*, *TPMT*, *TDP2*, *RBBP8*, *EDC4* and *RECQL5* genes have been sequenced so far in DNA samples of 638 HBOC patients without pathogenic variants identified in known genes and 206 healthy women. ORs were calculated using LoF and potentially splicing variants in both groups and GnomAD controls database v2.1.1 (Karczewski et al., 2020). The final objective of the study is to analyze the candidate genes in up to 1,100 cases and 500 control samples. Our first results to date indicate that *TPMT* and *TDP2* genes are either not related to HBOC genetic susceptibility or they might confer a low risk that cannot be confirmed by the size of the study. In a recently published article (Li et al., 2021), the *DMC1*, *RBBP8*, *RECQL5* and *TDP2* genes were sequenced in at least 1,027 patients and 943 controls, and the obtained results did not supported a role of these genes in the susceptibility to familial breast cancer.

In this article, we have identified candidate genes and suggested their association by a case-control approach. In the last decade, more than 200 candidate genes in HBOC and other hereditary cancer types have been evaluated (Rotunno et al., 2020). However, only a few genes are recognized as HBOC susceptibility cancer by the consensus of the scientific community (Dorling et al., 2021; Hu et al., 2021).

Various limitations have already been noted in previously published candidate genes identification and validation articles (Rotunno et al., 2020). Most of the studies focused their efforts on sequencing only cases and use the GnomAD database to compare the incidence of variants and calculate their ORs. Usually, the highest number of sequenced individuals in GnomAD for a specific gene is used to obtain an OR estimation. However, not all alleles for each variant has been sequenced in all subjects (i.e. due to sequencing failure or differences between dataset that make up GnomAD). Consequently, assuming the higher value of sequenced patients will lead to underestimating the allele frequency in controls, leading to an overestimation of ORs. To address this bias, we calculated a normalised variant allele count which was obtained by adjusting its GnomAD reported allele frequency to the average of the total list of number of alleles (termed in GnomAD as number of called high quality genotypes) specified for that gene of interest, and then the sum of these normalised variant alleles and the allele mean number were taken to compare cases with controls and estimate the respective ORs.

Another common limitation is the size of studied cohorts. For example, Rotunno et al., 2020 indicate that 53% of studies selected candidate genes after analyzing ten or fewer families

(thirteen in this article), and 43% of publications did not perform an independent validation analysis. Finally, another limitation that must be noted is the bias on the candidate gene selection, where genes in specific pathways related to DNA repair or replication pathways are selected.

In summary, this study provides evidences of potentially association to HBOC susceptibility of *ALKBH3*, *BLM*, *CAMKK1*, *FANCD2*, *FANCM*, *NEIL3*, *PER1*, *RBL1*, *RECQL4*, *RINT1*, *WRN* and *XRCC4* genes. However, sequencing in more cases and controls will provide further data for or against risk association. Our results suggest that deleterious variants identified in various genes related but not limited to DNA repair could explain a small percentage of the genetic landscape of affected HBOC patients. Furthermore, they support the hypothesis that the genetic susceptibility of a significant fraction of families with breast/ovarian cancer is heterogeneous with multiple *loci* of rare frequency and penetrance.

#### References

Bonache S, Esteban I, Moles-Fernández A, Tenés A, Duran-Lozano L, Montalban G, Bach V, Carrasco E, Gadea N, López-Fernández A, Torres-Esquius S, Mancuso F, et al. 2018. Multigene panel testing beyond BRCA1/2 in breast/ovarian cancer Spanish families and clinical actionability of findings. *J Cancer Res Clin Oncol* 144:2495–2513.

Bononi A, Goto K, Ak G, Yoshikawa Y, Emi M, Pastorino S, Carparelli L, Ferro A, Nasu M, Kim J-H, Suarez JS, Xu R, et al. 2020. Heterozygous germline BLM mutations increase susceptibility to asbestos and mesothelioma. *Proc Natl Acad Sci* 117:33466 LP – 33473.

Dorling L, Carvalho S, Allen J, González-Neira A, Luccarini C, Wahlström C, Pooley KA, Parsons MT, Fortuno C, Wang Q, Bolla MK, Dennis J, et al. 2021. Breast Cancer Risk Genes — Association Analysis in More than 113,000 Women. *N Engl J Med* 384:428–439.

Fiore R Di, D’Anneo A, Tesoriere G, Vento R. 2013. RB1 in cancer: different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis. *J Cell Physiol* 228:1676–1687.

Hammet F, Mahmood K, Green TR, Nguyen-Dumont T, Southey MC, Buchanan DD, Lonie A, Nathanson KL, Couch FJ, Pope BJ, Park DJ. 2019. Hi-Plex2: a simple and robust approach to targeted sequencing-based genetic screening. *Biotechniques* 67:118–122.

Hernández G, Ramírez MJ, Minguillón J, Quiles P, Ruiz de Garibay G, Aza-Carmona M, Bogliolo M, Pujol R, Prados-Carvajal R, Fernández J, García N, López A, et al. 2018. Decapping protein EDC4 regulates DNA repair and phenocopies BRCA1. *Nat Commun* 9:967.

Hu C, Hart SN, Gnanaolivu R, Huang H, Lee KY, Na J, Gao C, Lilyquist J, Yadav S, Boddicker



NJ, Samara R, Klebba J, et al. 2021. A Population-Based Study of Genes Previously Implicated in Breast Cancer. *N Engl J Med* 384:440–451.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581:434–443.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25:2283–2285.

Lebel M, Monnat RJJ. 2018. Werner syndrome (WRN) gene variants and their association with altered function and age-associated diseases. *Ageing Res Rev* 41:82–97.

Ledet E, Antonarakis ES, Pritchard C, Isaacs WB, Sartor AO. 2019. Germline heterozygous BLM mutations and prostate cancer risk. *J Clin Oncol* 37:321.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Li N, Lim BWX, Thompson ER, McInerney S, Zethoven M, Cheasley D, Rowley SM, Wong-Brown MW, Devereux L, Goringe KL, Sloan EK, Trainer A, et al. 2021a. Investigation of monogenic causes of familial breast cancer: data from the BEACCON case-control study. *NPJ breast cancer* 7:76.

Li N, Wang J, Wallace SS, Chen J, Zhou J, D’Andrea AD. 2020. Cooperation of the NEIL3 and Fanconi anemia/BRCA pathways in interstrand crosslink repair. *Nucleic Acids Res* 48:3014–3028.

Li N, Zethoven M, McInerney S, Devereux L, Huang Y-K, Thio N, Cheasley D, Gutiérrez-Enríquez S, Moles-Fernández A, Diez O. 2021b. Evaluation of the association of heterozygous germline variants in NTHL1 with breast cancer predisposition: an international multi-center study of 47,180 subjects. *NPJ Breast Cancer* 7:1–12.

Mantere T, Tervasmäki A, Nurmi A, Rapakko K, Kauppila S, Tang J, Schleutker J, Kallioniemi A, Hartikainen JM, Mannermaa A, Nieminen P, Hanhisalo R, et al. 2017. Case-control analysis of truncating mutations in DNA damage response genes connects TEX15 and FANCD2 with hereditary breast cancer susceptibility. *Sci Rep* 7:681.

Moles-Fernández A, Duran-Lozano L, Montalban G, Bonache S, López-Perolio I, Menéndez M, Santamariña M, Behar R, Blanco A, Carrasco E, López-Fernández A, Stjepanovic N, et al. 2018.

Mozaffari NL, Pagliarulo F, Sartori AA. 2021. Human CtIP: A “double agent” in DNA repair and tumorigenesis. *Semin Cell Dev Biol* 113:47–56.

Park DJ, Tao K, Calvez-Kelm F Le, Nguyen-Dumont T, Robinot N, Hammet F, Odefrey F,

Tsimiklis H, Teo ZL, Thingholm LB, Young EL, Voegelé C, et al. 2014. Rare mutations in RINT1 predispose carriers to breast and Lynch syndrome-spectrum cancers. *Cancer Discov* 4:804–815.

Paske IBAW Te, Ligtenberg MJL, Hoogerbrugge N, Voer RM de. 2020. Candidate Gene Discovery in Hereditary Colorectal Cancer and Polyposis Syndromes-Considerations for Future Studies. *Int J Mol Sci* 21:.

Peterlongo P, Catucci I, Colombo M, Caleca L, Mucaki E, Bogliolo M, Marin M, Damiola F, Bernard L, Pensotti V, Volorio S, Dall'Olio V, et al. 2015. FANCM c.5791C>T nonsense mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor. *Hum Mol Genet* 24:5345–5355.

Polak P, Kim J, Braunstein LZ, Karlic R, Haradhavala NJ, Tiao G, Rosebrock D, Livitz D, Kübler K, Mouw KW, Kamburov A, Maruvka YE, et al. 2017. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat Genet* 49:1476–1486.

Rolseth V, Luna L, Olsen AK, Suganthan R, Scheffler K, Neurauter CG, Esbensen Y, Kuśnierczyk A, Hildrestrand GA, Graupner A, Andersen JM, Slupphaug G, et al. 2017. No cancer predisposition or increased spontaneous mutation frequencies in NEIL DNA glycosylases-deficient mice. *Sci Rep* 7:4384.

Rotunno M, Barajas R, Clyne M, Hoover E, Simonds NI, Lam TK, Mechanic LE, Goldstein AM, Gillanders EM. 2020. A systematic literature review of whole exome and genome sequencing population studies of genetic susceptibility to cancer. *Cancer Epidemiol Biomarkers Prev* 29:1519–1534.

Southey MC, Park DJ, Nguyen-Dumont T, Campbell I, Thompson E, Trainer AH, Chenevix-Trench G, Simard J, Dumont M, Soucy P, Thomassen M, Jønson L, et al. 2013.

Subramanian DN, Zethoven M, McInerney S, Morgan JA, Rowley SM, Lee JEA, Li N, Goringe KL, James PA, Campbell IG. 2020. Exome sequencing of familial high-grade serous ovarian carcinoma reveals heterogeneity for rare candidate susceptibility genes. *Nat Commun* 11:1640.

Tavera-Tapia A, la Hoya M de, Calvete O, Martín-Gimeno P, Fernández V, Macías JA, Alonso B, Pombo L, Diego C de, Alonso R, Pita G, Barroso A, et al. 2019. RECQL5: Another DNA helicase potentially involved in hereditary breast cancer susceptibility. *Hum Mutat* 40:566–577.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164.

Zarrizi R, Higgs MR, Voßgröne K, Rossing M, Bertelsen B, Bose M, Kousholt AN, Rösner H, Network TC, Ejlersen B, Stewart GS, Nielsen FC, et al. 2020. Germline RBBP8 variants associated with early-onset breast cancer compromise replication fork stability. *J Clin Invest* 130:4069–4080.

**Supplementary material**

**Supplementary table 1.** SEOM clinical criteria for germline genetic analysis in hereditary breast and ovarian cancer patients. Adapted from Gonzalez-Santiago et al., 2020.

Selection criteria for germline testing
<b>Regardless of family history:</b>
Women with synchronous or metachronous breast and ovarian cancer
Breast cancer ≤ 40 years
Bilateral breast cancer (the first diagnosed ≤ 50 years)
Triple-negative breast cancer ≤ 60 years
High-grade epithelial non-mucinous ovarian cancer (or fallopian tube or primary peritoneal cancer)
Ancestry with founder mutations
BRCA somatic mutation detected in any tumor type with a tumor allele frequency > 30% (if it is known)
Metastatic HER2-negative breast cancer patients eligible to consider PARP inhibitor therapy
<b>2 or more first degree relatives with any combination of the following high-risk features:</b>
Bilateral breast cancer + another breast cancer < 60 years
Breast cancer < 50 years and prostate or pancreatic cancer < 60 years
Male breast cancer
Breast and ovarian cancer
Two cases of breast cancer diagnosed before age 50 years
<b>3 or more direct relatives with breast cancer (at least one premenopausal) and/or ovarian cancer and/or, pancreatic cancer or high Gleason (≥ 7) prostate cancer</b>

**Supplementary table 2.** Genes with prioritized variants in candidate genes identification dataset.

Gene	Var type	Coding effect	Biological process	Segregation study
<i>AHCTF1</i>	S	M	Cytokinesis, Mitotic Cycle, mRNA transport, etc.	NA
<i>ALKBH3</i>	D	F	DNA alkylation damage repair, Cell proliferation, etc.	No
<i>APOBEC1</i>	S	SL	DNA demethylation, RNA processing, Cell proliferation regulation, etc	No
<i>ATM</i>	S	M	DNA damage check point, DNA damage response, Cell Cycle control, etc.	No
<i>CEP164</i>	S	M	G2 DNA damage control, Involved in ATM/ATR pathway on DNA damage response.	NA
<i>DHTKD1</i>	S	M	Hematopoietic progenitor cell differentiation factor, Glycolitic process, etc.	NA
<i>DMC1</i>	S	M	DNA Meiotic Recombinase	NA
<i>DOT1L</i>	S	M	Chromatine organization and silencing, Cell Cycle regulation, Hystone metilation, etc.	NA
<i>DPPA4</i>	D	F	Mesenchymal development, Transcription regulation, etc.	No
<i>DTHD1</i>	S	SE	Proteosome regulation.	NA
<i>ELP5</i>	S	SE	Chromatine organization, Positive regulation of cellular migration, etc.	No
<i>ERCC6</i>	S	M	DNA repair, Apoptotic pathway involved in DNA damage response, etc.	No
<i>FCGR2A</i>	S	SE	Fc-gamma pathway receptor.	NA
<i>INSRR</i>	S	M	Cytoskeletal actine organization, Alkalyne pH cellular response, etc.	NA
<i>LARP1B</i>	D	F	Mitophagia activation as response to mitochondrial depolarization, etc.	NA
<i>MACROD1</i>	S	N	DNA damage response, purine metabolism, etc.	NA
<i>MED23</i>	S	SL	Protein ubiquitination, Expression regulation, etc.	NA

## RESULTS

<i>MZF1</i>	Dup	F	Negative regulation of transcription of RNA pol II promotor.	NA
<i>NOB1</i>	S	M	rRNA Processing, Visual perception, etc.	NA
<i>NOTCH3</i>	S	M	Cell differentiation, Notch pathway component, etc.	NA
<i>PFKM</i>	S	SE	Glucose metabolism, Regulation of insuline secretion, etc.	NA
<i>POLR2A</i>	S	M	RNA polymerization DNA-dependent.	NA
<i>POLR3B</i>	S	M	RNA Pol III promotor transcription.	NA
<i>PSMD3</i>	S	M	DNA damage response, G1 DNA damage control, etc.	NA
<i>RALGDS</i>	S	N	RAS metabolic pathway signal transduction.	NA
<i>RNF216</i>	S	N	Apoptotic process, etc.	NA
<i>SLCO1B3</i>	S	M	Transport of organic anions.	NA
<i>TDP2</i>	S	M	Double strand DNA repair, Nucleic acid Phosphodiester bonds hydrolisis, etc.	Yes
<i>TPMT</i>	S	M	Methilation, Xenobiotic metabolism, etc	Yes
<i>TSHZ3</i>	S	M	Transcription negative regulation, CASP4 Inhibition, etc.	NA
<i>ZFHX3</i>	S	M	Cell Cycle regulation, Positive regulation of myoblast differentiation, etc.	NA
<i>ZNF268</i>	D	F	Cell differentiation, Negative regulation of apoptosis, Negative response to necrosis tumoral factor, etc.	No
<i>ZNF439</i>	S	N	Transcription regulation.	NA
<i>ZNF491</i>	S	N	Transcription regulation.	NA
<i>ZNF626</i>	S	N	Transcription regulation.	NA

Highlighted in red genes selected for validation; D, deletion; S, single nucleotide variation; Dup, duplication, F, frameshift; N, nonsense; M, missense. NA, not available.

**Supplementary table 3.** LoF incidence in cases and controls, and OR values for shared genes between Li et al., 2021 and this study. *EDC4, TPMT, MACROD1 and RALGDS* were not included. **Bold:** genes with LoF variants which reached significant OR. **Bold and underlined:** genes with LoF variants which reached significant OR in our study and Li et al., 2021.

Genes	n° of LoF alleles in cases (%)	n° of LoF alleles in GnomAD controls* (%)	OR (95%CI)	Li et al., 2021 n° of LoF alleles in cases (%)	Li et al., 2021 n° of LoF alleles in GnomAD controls (%)	Li et al., 2021 LoF_OR (95%CI)	Cases sample size	Controls sample size
<i>ALKBH1</i>	3 (0.296)	28.46 (0.142)	2.082 (0.632 - 6.855)	14 (0.250)	12 (0.250)	1.16 (0.5-2.75)	4807	4782
<b><i>ALKBH3</i></b>	13 (1.284)	120.7 (0.592)	2.183 (1.227 - 3.882)	20 (0.278)	16 (0.278)	1.24 (0.61-2.57)	5770	5741
<b><u><i>BLM</i></u></b>	8 (0.790)	60.32 (0.307)	2.582 (1.231 - 5.413)	20 (0.139)	8 (0.139)	2.49 (1.05-6.55)	5770	5741
<b><i>CAMKK1</i></b>	5 (0.494)	4.493 (0.023)	21.09 (5.890 - 75.53)	0 (0.026)	1 (0.026)	0.34 (0-39.61)	3780	3839
<i>DMC1</i>	1 (0.156)	33.712 (0.168)	0.929 (0.127 - 6.806)	0 (0)	0 (0)	0.92 (0-Inf)	1027	943
<b><i>FANCD2</i></b>	10 (0.988)	32.70 (0.168)	5.915 (2.905 - 12.04)	7 (0.083)	4 (0.083)	1.74 (0.44-8.12)	4807	4782
<b><i>FANCF</i></b>	2 (0.197)	12.37 (0.059)	3.340 (0.749 - 14.89)	8 (0.209)	10 (0.209)	0.8 (0.27-2.24)	4807	4782
<b><i>FANCM</i></b>	30 (2.964)	135.0 (0.686)	4.416 (2.958 - 6.594)	27 (0.439)	21 (0.439)	1.28 (0.7-2.39)	4807	4782
<b><i>NEIL3</i></b>	7 (0.691)	32.37 (0.166)	4.184 (1.844 - 9.496)	5 (0)	0 (0)	10.95 (0.91-Inf)	4807	4782
<b><i>NTHL1</i></b>	4 (0.395)	98.56 (0.527)	0.748 (0.274 - 2.038)	38 (0.313)	15 (0.313)	2.53 (1.36-4.96)	4807	4782
<b><i>PER1</i></b>	5 (0.494)	14.71 (0.079)	6.235 (2.256 - 17.23)	1 (0.026)	1 (0.026)	1.02 (0.01-79.7)	3780	3839
<b><i>RASSF7</i></b>	2 (0.197)	36.87 (0.228)	0.863 (0.207 - 3.586)	4 (0.156)	6 (0.156)	0.68 (0.14-2.86)	3780	3839
<b><i>RBBP8</i></b>	NA	NA	NA	2 (0.105)	2 (0.105)	0.96 (0.07-13.19)	1990	1902
<b><i>RBL1</i></b>	14 (1.383)	76.50 (0.414)	3.368 (1.898 - 5.976)	13 (0.260)	10 (0.260)	1.32 (0.53-3.37)	3780	3839
<b><i>RECQL</i></b>	3 (0.296)	171.5 (0.926)	0.318 (0.101 - 0.997)	12 (0.480)	23 (0.480)	0.52 (0.23-1.09)	4807	4782

RESULTS

<b>RECQL4</b>	12 (1.185)	69.79 (0.382)	3.128 (1.689 - 5.790)	16 (0.355)	17 (0.355)	0.94 (0.44- 1.97)	4807	4782
<b>RECQL5</b>	4 (0.626)	47.202 (0.258)	2.435 (0.874 - 6.778)	1 (0.212)	2 (0.212)	0.46 (0.01- 8.83)	1027	943
<b>RINT1</b>	3 (0.296)	22.21 (0.114)	2.601 (0.777 - 8.700)	2 (0.062)	3 (0.062)	0.66 (0.06- 5.79)	4807	4782
<b>RUVBL1</b>	1 (0.098)	0.5 (0.002)	38.29 (1.284 - 1142.)	0 (0)	0 (0)	0.92 (0-Inf)	1027	943
<b>SALL2</b>	2 (0.197)	10.92 (0.055)	3.556 (0.786 - 16.07)	1 (0.041)	2 (0.041)	0.5 (0.01- 9.56)	4807	4782
<b>SLX</b>	5 (0.494)	39.89 (0.197)	2.505 (0.986 - 6.363)	9 (0.188)	9 (0.188)	0.99 (0.35- 2.83)	4807	4782
<b>STRADA</b>	1 (0.098)	6.902 (0.037)	2.640 (0.323 - 21.52)	1 (0.078)	3 (0.078)	0.34 (0.01- 4.22)	3780	3839
<b>TDP2</b>	NA	NA	NA	0 (0)	0 (0)	0.92 (0-Inf)	1027	943
<b>WRN</b>	21 (2.075)	64.71 (0.328)	6.432 (3.915 - 10.56)	34 (0.355)	17 (0.355)	2 (1.08- 3.82)	4807	4782
<b>XRCC4</b>	7 (0.691)	46.39 (0.238)	2.915 (1.313 - 6.470)	1 (0.106)	1 (0.106)	0.92 (0.01- 72.11)	1027	943
<b>ZNHIT1</b>	NA	NA	NA	1 (0.062)	3 (0.062)	0.33 (0.01- 4.13)	4807	4782







**DISCUSSION**



Hereditary breast and ovarian cancer (HBOC) patients without pathogenic variants identified in susceptibility genes represent a mis-opportunity to benefit from a personalized and precise medical management. The lack of a definitive genetic diagnosis after a germline molecular test in patients with a suspected HBOC disorder is driven mainly by an inconclusive result with the detection of variants of unknown significance (VUS) or a negative result with no detected pathogenic variant in known risk genes. This thesis proposes that an improved genetic diagnostic will be achieved through the optimization of the use of computational *in silico* algorithms for interpreting VUS spliceogenic effects as well as of the variant classification process (**Articles 1, 2 and 3**) and the identification of new HBOC susceptibility genes (**Article 4 in preparation**).

## **1. *In silico* tools for spliceogenic variants identification in HBOC genes**

The use of massive parallel sequencing in clinical diagnostics is leading to a significant increase in genomic data and the detection of a high number of variants of uncertain significance (VUS) with potential effects on splicing that need interpretation. DNA variant disrupting any of the *cis*-acting core or regulating elements may lead to incorrect splicing, generating aberrant transcripts and hence non-functional proteins. Therefore, prediction of the effect of DNA sequence variations on splicing using *in silico* tools has become a common approach. In ACMG guidelines, the likely consequences predicted by *in silico* tools are essential for application of the supporting evidence PP3 (multiple lines of computational evidence support a deleterious effect) and BP4 (multiple lines of computational evidence suggest no impact on gene or gene product). However, there is no consensus or unified way about which tool has to be used and how to identify the effect caused by variants disrupting the different *cis*-elements.

### 1.1 Identification of variants altering donor and acceptor splice sites in HBOC using *in silico* tools

Splicing acceptor and donor sites are critical elements for the correct exon inclusion in RNA, delimiting exons and introns. These regions are recognized by DNA binding proteins of the spliceosome complex, and variants located in these highly conserved sequences could impede this recognition leading to a splicing alteration. Due to the importance of these elements, several *in silico* tools for their analysis have been developed and a few studies have been published on their reliability in predicting the impact of variants on splicing sites. Their results show that the recommendations provided on the most appropriate use are not concordant (Jian et al., 2014a and Table 1 in **Article 1**).

The main objective in **Article 1** was to provide a framework to detect exonic and intronic spliceogenic variants affecting acceptor and donor splicing sites (-10 to +20 and -20 to +10 in donor and acceptor, respectively) using *in silico* tools. We collected variants identified in hereditary cancer-related genes and compared the *in silico* predicted effect of six programs (MES, HSF, SSF-like, SPANR, NNSplice, and dbSNV), with splicing *in vitro* outcomes, thus evaluating the reliability of the predictions. We elaborated the study in two stages, discovery and validation, to identify the best predictors or the best combination for their application in routine clinical testing, taking into account sensitivity, specificity, accuracy, and negative predictive value as well as the score of Matthews Coefficient Correlation (MCC). In the discovery stage, significant performance differences were appreciated among individual tools. Globally, as well as for 5', and 3' splice sites, low accuracies of SPANR and NNSplice contrasted with the high performance achieved by SSF, MES, and HSF.

At the second stage of the study, we validated the combinations of HSF with SSF-like or HSF+SSF-like+MES (at least one of them indicating alteration) as the highest performances for splicing aberrations at donor sites, and SSF-like stand-alone at acceptor sites.

These results provided recommendations for identifying splicing site altering variants using *in silico* tools with a high level of confidence, based on MCC, sensitivity and negative predictive values. This framework is relevant in a clinical setting since it allows to separate the variants with an extremely low or non-existent probability of being abnormally spliceogenic from those variants in which *in vitro* RNA studies are of interest. For example, the use of **Article 1** *in silico* splicing recommendations helped us (Duran-Lozano et al., 2019; Montalban et al., 2019) (Articles 5 and 6 in the Appendix section) and other groups (Sanoguera-Miralles et al., 2020) to classify pathogenic variants altering splicing in HBOC related genes.

The recommendations provided in our study are partially in concordance with previously published papers, most of them indicating SSF-like, MES, or HSF as high-performance tools. In Houdayer et al., 2012, using a dataset of HBOC spliceogenic variants affecting splicing sites and not spliceogenic variants, the authors recommended a sequential approach for both acceptor and donor sites, using MES first and SSF-like second. However, this combination reached a lower performance with variants located in the donor site in the discovery step of our **Article 1**. On the other hand, SPiCE tool (Leman et al., 2018) showed high sensitivity and specificity in a dataset of 395 HBOC variants. This user-friendly and freely available tool combines SSF (modified from original Position Weigh Matrix published by Shapiro and Senapathy, 1987) and MES, providing a unique probability score, allowing a high throughput variant analysis. Interestingly, SPiCE showed a similar performance than our recommended tools (inferred from Supplementary table 8 of **Article 1**). Finally, according to previously published studies, NNS and GS showed low performances (Houdayer

## DISCUSSION

et al., 2012; Tang et al., 2016), and their use should be avoided due to the high rate of false positives and negatives.

To note, the rate of false-negative predictions in our study was significantly higher for acceptor sites than donor sites. This difference may be due to the greater complexity of the sequence adjacent to the 3', with the presence of the branch point and the polypyrimidine tract. Moreover, conserved splicing site sequences are different between acceptor and donor sites, i.e., 11 bases for the 5' splice site (from the three last exonic to the eight first intronic bases) and 14 bases for the 3' splice site (from the 12 last intronic to the first two exonic bases) (Burge et al., 1999). To our knowledge, our study is one of the few that evaluates the accuracy of different tools separately for donor and acceptor sites, resulting in different recommendations for each one with high performance. Interestingly, Danis et al. (2021) reported the development of a new machine learning method for predicting splicing alterations of non-canonical variants located outside AG/GT intronic dinucleotides, considering and training two site-specific models to differentiate splice variants from neutral variants, one for the donor variants and the other for the acceptor variants.

The tools analysed in **Article 1** have only been interrogated to predict alterations at donor and acceptor splice sites. However, alterations in RNA may be produced by variants that affect other factors in *cis* (branch point, polypyrimidine tract, intronic and exonic splicing silencers and enhancers), create new splice sites or activate cryptic ones. *In silico* tools able to analyse different *cis*-splicing sequences could increase the spliceogenic variants detection. For example, SPANR, included in our performance analysis, is the first approach of a machine learning splicing tool integrating different conserved elements. Moreover, after **Article 1** was published, new *in silico* tools addressing this point have been developed, by combining in separated modules different tools in a "meta-predictor" approach, such as SPiP

(<https://github.com/raphaelleman/SPiP>) or machine learning tools integrating the identification of potentially spliceogenic variants related to different *cis*-elements (SpliceAI or MMSplice). SPiP is a freely available and user-friendly tool that provides a probability score obtained from the separate and independent analysis of each splicing element by its dedicated tool. For example, SPiP includes SPiCE for splicing sites and new or cryptic sites creation, ESRseq for exonic splicing silencers and enhancers, and BPP *in silico* tool for the identification of branch point regions. In other words, it is a predictor that analyses different elements without integrating its balance or interdependence between regions.

One of the most noteworthy aspects of the new generation tools is the use of machine learning approaches, which enables the consideration of the fact that the functionality of a splicing element depends on its interactions with the other *cis*-elements. Hence, these predictors take into account large sequences to assess the effect of a variant to accurately predict which splicing elements are altered. SpliceAI (Jaganathan et al., 2019), for example, is able to analyse the effect of a variant taking into account the surrounding 10,000 nucleotides to the variant. In addition, MMSplice (Cheng et al., 2019b) considers competitive interactions among close splicing sites together with changes in splicing efficiency. This tool ranked first at the Critical Assessment of Genome Interpretation 5 (CAGI5) exon skipping prediction challenge (Cheng et al., 2019a), in which two splicing prediction challenges were proposed based on two experimental perturbation minigene high-throughput assays: Vex-seq, assessing exon skipping, and MaPSy, assessing splicing efficiency. Using these pre-established datasets, the performance of *in silico* tools to correctly discriminate altering and not altering variants were compared. These challenges are an unbiased option to compare the utility of different tools with an independent set of variants and an opportunity to train and improve different predictors.



## DISCUSSION

The CAGI 5 also included a challenge for predicting which of the about 400 *BRCA1* or *BRCA2* variants was associated with increased risk for breast cancer (CAGI 5 ENIGMA challenge) (Cline et al., 2019) Article 7, Appendix). The **Article 1** recommendations were included in the "BRCA1- and BRCA2-specific *in silico* tools for variant interpretation in the CAGI 5 ENIGMA challenge" (Padilla et al., 2019), which ranked second being the only participant that included splicing interpretation, that supports the importance of splicing analysis in identifying pathogenic variants (Article 8, Appendix).

In a clinical setting, the selection of splicing algorithms should be based on the reliability of their predictions of the variant functional impact, facility of their implementation and output interpretation (i.e., what sequence features lead to the prediction score that reflects the probability that a given variant is spliceogenic) (Lord and Baralle, 2021). However, to date there is no defined process of how to establish the precise degree of confidence that *in silico* predictions must have for their clinical application. For example, the selection of a score as cut-off to distinguish a variant as splice disrupter is usually arbitrary or estimated from the evaluation of relatively small number of variants with known splicing effect (Lord and Baralle, 2021). Since an evidence-based, unified splicing *in silico* approach is still needed in the clinical setting, further independent studies with a high number of variants are required, comparing the performance of an increasing list of tools, to establish which ones have to be used (Lord and Baralle, 2021). Highlighting the still need of benchmarking splicing *in silico* predictions with experimental data to better handle their reliability, is the recent launch (on June 2021) of the CAGI 6 challenge for predicting splicing disruption from variants of unknown significance. In this project, participants are asked to provide a prediction "score" in the range of 0 to 1 to distinguish splicing altering variants among a set of variants of unknown significance clinically ascertained and experimentally assessed by the organizers (<https://genomeinterpretation.org/cagi6-splicing-vus.html>).

## 1.2 Deep intronic variant identification using *in silico* tools

Despite the advances in sequencing technologies, there is still an important fraction of HBOC cases without a genetic diagnosis. A percentage of this fraction may include variants in non-coding deep intronic regions. The contribution of deep intronic variants to HBOC disease is not well known due to their location in poorly screened regions, but their potential effect on transcript splicing, including intron sequences in mature RNA, may be clinically significant since several spliceogenic variants have been detected (Vaz-Drago et al., 2017; Montalban et al., 2018a). However, identifying these variants is challenging due to the lack of specific *in silico* pipelines (Canson et al., 2020). For this reason, the work developed in this thesis aimed to provide a framework to identify spliceogenic variants in regions historically under-analysed (Toland et al., 2018).

Jaganathan et al., 2019, in their SpliceAI development article, demonstrated that applying SpliceAI with a cut-off of  $\geq 0.5$  (Jaganathan et al., 2019), achieved a sensitivity of 71% when the analysed variants were near to exons, but fell to 41% when the variants were in deep intronic regions (37 variants, >50 nt from exons). In **Article 2**, using a large dataset of deep intronic variants clinically relevant, we confirmed that SpliceAI *in silico* tool with a threshold of  $\geq 0.05$  reaches an optimal predictive value in identifying spliceogenic deep intronic variants.

Moreover, recently published papers such as Riepe et al., 2021, with an optimized SpliceAI cut-off score of 0.18, also showed a high performance, with a 0.84 MCC for predicting 81 deep intronic variants in the *ABCA4* gene; these variants had also been included in our dataset. In addition, Riepe et al.

## DISCUSSION

demonstrated that SpliceAI was the best tool to identify spliceogenic deep intronic variants compared with other deep-learning based algorithms such as SPANR (Xiong et al., 2015) and the "classical" tools SSF-like or MES, based on Position Weight Matrix or Maximum Entropy SSF-like, respectively.

Most of the spliceogenic deep intronic variants reported create sequences similar to splicing sites or activate cryptic splicing sites (Vaz-Drago et al., 2017). However, variants disrupting or creating intronic SREs can also lead to the inclusion of pseudoexons in RNA. In Table 1 and Supplementary Figure 1 of the **Article 2**, we can observe a lower performance of SpliceAI detecting SREs-related variants (MCC=0.66) than intronic variants that create or activate cryptic splicing sites (MCC=0.88). Also, we show that SpliceAI performance in predicting the impact of SREs by exonic variants is limited (MCC=0.53). This low performance is possibly due to the fact that the deep learning network approach used for SpliceAI is not able to account for the SREs because of their limited presence in the tool training dataset (Jaganathan et al., 2019). This indicates that the performance of SpliceAI to identify SREs-related variants can still be improved, or other *in silico* tools could be used for this purpose. Particularly, ESRseq, HAL, and HEXPLORER tools have been developed to identify variants affecting splicing regulatory elements, but most of the performance studies have been done using exonic variants (Canson et al., 2020; Tubeuf et al., 2020). In Tubeuf et al., 2020, the authors used these tools with a large set of exonic variants that only affected SREs, altering or not exonic splicing. They showed that ESRseq achieved the highest performance to detect exon skipping after optimizing the threshold to -0.50 score. Moreover, focusing on identifying variants that increase exon inclusion by creating or enhancing ESEs using a dedicated dataset, they adapted the ESRseq threshold, optimizing it at +0.36. The authors selected this threshold to use ESRseq to detect deep intronic variants creating pseudoexons (n = 13), due to the impossibility of HAL and SPANR to analyse in deep intronic regions (Tubeuf et al., 2020). Ten variants

were correctly identified, and the authors suggested that this tool may be useful for predicting the creation of variant-induced pseudoexons. This result is slightly different from that obtained in the **Article 2** of the present thesis, in which the addition of ESRseq to SpliceAI analysis improves the sensitivity values but did not show improvement in the MCC. This discordance could be explained by the absence of no-spliceogenic variants in Tubeuf et al., 2020 dataset, and also because the authors aimed to detect only SREs altering intronic variants, in contrast to **Article 2**. In addition, we also reasoned that the limitation of improvement in the identification of SRE-altering variants using ESRseq, was due to the fact that this tool evaluates at a local hexamer level, without accounting for the SREs landscape that defines a region to be included as a pseudoexon.

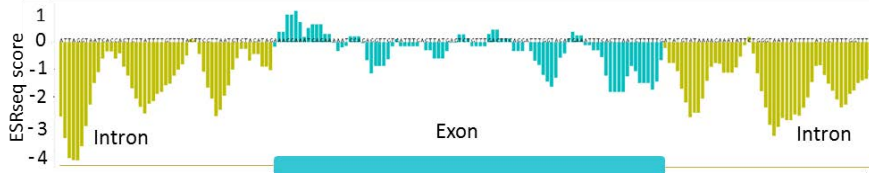
In summary, SpliceAI alone is able to identify variants causing pseudoexons with a high performance, and the addition of ESRseq has limited success in the identification of SREs-altering deep intronic variants.

### **1.3 Importance of SREs balance in the pseudoexon inclusion caused by deep intronic variants**

The lack of improvement in the detection of spliceogenic deep intronic variants following the addition of ESRseq tool scores to SpliceAI led us to further investigate the characteristics and relevance of the SREs. Using ESRseq scores inferred from experimental results in Ke et al., 2011, we characterized the landscape of SREs along the exonic and adjacent intronic regions of HBOC genes (**Article 2**, Figure 2A) by scoring and mapping each nucleotide and its hexamer surrounding region. As expected, an exonic enhancers (ESEs) enrichment was observed in exons compared to intronic regions, and conversely a lower exonic

## DISCUSSION

splicing silencers (ESSs) density in the exonic regions than in introns. These results are in agreement with other articles (Wang et al., 2005; Cáceres and Hurst, 2013; Erkelenz et al., 2014), indicating that ESRseq mapping is an interesting option to identify SREs along DNA sequences (Fig 7).



*Figure 7. SRE mapping of exonic and intronic regions. Representative image showing ESRseq scores used to map SREs sequences along exon and adjacent intronic regions. Nucleotides with negative values mean that they may act as silencer elements. Nucleotides with positive values may act as enhancer elements.*

Then, we characterized the pseudoexonized regions included in RNA caused by deep intronic spliceogenic variants collected in **Article 2** and their surrounding intronic sequences. We observed that the relation of the splicing regulatory elements (SREs) landscape between the pseudoexon and flanking introns is similar to that of canonical exons. In contrast, the ESRseq developers reported that the pseudoexons did not present a different balance of SRE than the adjacent intronic regions (Ke et al., 2011). This discrepancy could be because the pseudoexons analysed in the above-mentioned work were theoretically defined, without an experimental RNA evaluation, instead of using experimentally confirmed variants, as collected in **Article 2**. In addition, surrounding sequences to non spliceogenic deep intronic variants showed poor differences in the SREs balance compared with spliceogenic ones (**Article 2**, Figure 3). To our knowledge, up to date, this is the most extensive characterization of SREs in pseudoexonized regions caused by deep intronic variants experimentally assessed.

The analysis of **Article 2** shows that the balance of SREs between exons and introns was less defined in pseudoexons than in HBOC canonical exons. Similar

findings to our results using approaches other than ESRseq tools have been recently reported, showing that in pseudoexons there is a smaller density of ESEs together with a higher density of ESSs compared to canonical exons, that is, the pseudoexons presented a weaker exon profile in terms of SREs (Xie et al., 2020). Collectively, these results suggest that SREs balance is critical for the exon inclusion or the recognition of an intronic region to be included into the RNA together with other splicing elements. It is worth to note that the presence in deep intronic regions of cryptic branch points or cryptic polypyrimidine tracts might also have a role in the inclusion of pseudoexons and their consideration in the *in silico* prediction of deep intronic splice altering variants has not been yet addressed.

The density of ESEs in exons and ISSs in adjacent introns can be variable (Figure 2A in **Article 2**). This observation agrees with Tubeuf et al., 2020 analysis, in which the *in silico* tools performances was separately analysed in specific exons using a pre-established threshold. Their results showed differences in MCC between groups, and they decided to optimize the threshold to each specific exon, improving the tool performance. This specific exon effect could be driven by differences in density, balance, or strength SREs landscape, and indicates the need of considering these features for a correct detection of variants disrupting all the regulatory *cis*-splicing elements.

In this sense and in favour of the SRE density importance, Baeza-Centurion et al. suggested that the effect of variants altering the balance of SREs appears to be greater in alternative exons, which have fewer redundant enhancer elements, compared to constitutive ones (Baeza-Centurion et al., 2020). Therefore, we suggest that deep intronic variants that strengthen an enhancer or even decrease a silencer will have a greater chance of being spliceogenic provided that they are located in intron regions with an exon-like SRE landscape, similar to what happens in alternative exons. Moreover, in an exonic context, Baeza-Centurion et al., 2020 and Tubeuf et al., 2020 results suggest

## DISCUSSION

that there exist a potentially sequence redundancy of ESEs in some exons (high ESEs density), which results in a "tolerance" to SREs disruption avoiding exon skipping. On the contrary, constitutive exons with a lower ESEs density would be more susceptible to suffer splicing alteration, due to the essentialness of these elements.

Regarding the importance of an interplay between SRE and other *cis*-elements, Tubeuf et al., 2020 showed (in supplementary figure 12 of their manuscript), that exon-skipping SREs altering variants were more frequently found in exons with weaker 3'ss, suggesting a relevant interdependence between splicing conserved elements, and that this effect can play a role in the *in silico* tools usage.

Taken together our results generated in **Article 2** and those mentioned above, we hypothesize that mapping and scoring SREs using ESRseq in exons and introns, could find out regions susceptible to be altered by SREs-related variants (for example exons with low presence of ESEs), and sequences with "tolerance" to be splicing-altered (exons with ESE "redundancy"). This information will facilitate the detection of splice altering variants using "classical" *in silico* approaches or even be used as a variable to be included in machine learning algorithms for improving its performance. To address this point, at the moment of the thesis presentation, a Python-based *script* is being developed by the Hereditary Cancer Genetics Group at VHIO, that will allow to i) map and score each nucleotide of a sequence located within an initial and final genomic position, and ii) calculate  $\Delta$ ESRseq scores caused by a variant in comparison with the wild type sequence in a high-throughput way with genomic coordinates as input. This initial resource, together with future improvements will be published to be used by the scientific community. One example of its utility could be to map and target exonic regions of interest, identifying enhancers and silencers, thereby, assisting to the experimental identification of

SREs sequences in exons using minigenes (such as Sanoguera-Miralles et al., 2020 and Bueno-Martínez et al., 2021).

#### **1.4 Future of splicing variants identification: through a unified *in silico* pipeline and *in vitro* RNA sequencing**

The development of integrative tools including different splicing *cis*-elements, considering their interdependences, would improve the current identification of spliceogenic variants by *in silico* tools. It seems that this complexity might be tackled using the most advanced machine learning techniques as it has been recently demonstrated with the development of tools like SpliceAI or SQUIRLS algorithms (Jaganathan et al., 2019; Danis et al., 2021) or instead using meta-predictor *in silico* tools like SPIP (<https://github.com/raphaelleman/SPIP>). Thus, it could be included in the ACMG guidelines as computational splicing evidence (PP3 and BP4), which use are still very limited and without established consensus recommendations.

*In silico* tools capable of analysing stand-alone elements or more recently multiple elements, have been published. However, there is currently no single *in silico* tool with a verified performance of detection of splicing disrupting variants due to alteration or creation of any *cis* DNA elements. Thanks to the experience and knowledge gained along this thesis, we proposed a pipeline of *in silico* splicing analysis (Table 6) that covers the detection of spliceogenic variants located in all *cis*-splicing elements. This pipeline needs to be refined and validated with large variant datasets from HBOC patients before applying it in a real clinical setting.



DISCUSSION

Table 6.: *In silico* splicing analysis proposed pipeline. Different tools with its respective thresholds recommended to analyse potentially spliceogenic variants.

Cis-element/ type of alteration	Covered nucleotides	Tools	Threshold	URL link / reference
<b>Splicing Site</b>	Donor site: from -3 exonic+8 intronic	Donor S: SSF-like or HSF / Acceptor S: SSF- like (Alamut Visual)	-5% (SSF) or - 2% (HSF) / -5% (SSF)	(Shapiro and Senapathy, 1987; Desmet et al., 2009)
	Acceptor site: - 14 intronic to +2 exonic	Donor S: SPICE / Acceptor S: SPiCE	0.24 / 0.282 or 0.789	(Leman et al., 2018)
<b>New sites/Cryptic activation</b>	All the exonic and intronic region	SpliceAI	≥0.05	(Jaganathan et al., 2019)
<b>Polypyrimidin e tract</b>	From -18 to -12 intronic nucleotides adjacent to the Acceptor site	MaxEntScan (via SPIP)	-15% (MES)	(Yeo and Burge, 2004) / ( <a href="https://github.com/raphaellemann/SPIP">https://github.c om/raphaellem ann/SPIP</a> )
<b>Branch Point</b>	From -44 to -18 intronic nucleotides adjacent to the Acceptor site	BPP	Indicated by SPiP as located in the branch point (motif: TRAY)	(Zhang et al., 2017) / ( <a href="https://github.com/raphaellemann/SPIP">https://github.c om/raphaellem ann/SPIP</a> )
<b>Splicing Regulatory Element</b>	All the exonic and intronic region	ESRseq	-0.5	(Ke et al., 2011)
<b>Pseudoexons by deep intronic variants</b>	All the intronic region	SpliceAI	≥0.05	(Jaganathan et al., 2019)

The development of an integrative *in silico* tool pipeline, considering splicing element interdependences, will enhance the identification and analysis of potentially altering RNA variants in the massively parallel sequencing approach, particularly in a clinical context, where variants outside di-nucleotide acceptor and donor sites are not usually explored, and will result in a significant improvement in diagnosis in HBOC (Wai et al., 2020), but also in rare diseases (Lord and Baralle, 2021). In fact, the spliceogenic potential of synonymous and

intronic variants outside of dinucleotide consensus splice sites is frequently overlooked in bioinformatic pipelines, both in somatic and germline settings.

Predictions made by *in silico* tools assist to the identification of variants in DNA (as evidenced **Articles 1 and 2**), but the potential effect in the RNA of the patient has to be verified by *in vitro* approaches. The recommendations for mRNA analysis best practice in clinical testing published by ENIGMA (Whiley et al., 2014) include a qualitative RNA analysis in order to find aberrant splicing profiles. In addition, the use of minigenes to functionally test variants (in absence of patient samples) is widely extended (Gaildrat et al., 2010; Sanoguera-Miralles et al., 2020; Bueno-Martínez et al., 2021). ENIGMA recommends the use of RT-PCR and digital or capillary electrophoresis to detect transcripts with abnormal length, followed by cloning and Sanger sequencing to characterize their sequence (Whiley et al., 2014). However, it is essential to know the level at which these transcripts are expressed in order to determine their functional significance. Thus, a combination of qualitative and quantitative analysis is needed to provide proper characterization of spliceogenic variants (Montalban et al., 2019).

However, this approach includes several time-consuming assays that diminishes their feasibility in a clinical setting. A promising alternative is the long-read RNA massively sequencing approach which allows the parallel evaluation of alterations in RNA splicing, RNA expression and changes in the RNA sequence in one assay (Sedlazeck et al., 2018; Sakamoto et al., 2020). This approach would simplify and accelerate the RNA analysis, coupled to the potentially spliceogenic variant identification by *in silico* tools.

## 2. Adaptation of ACMG guidelines to the *ATM* gene

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have provided a general framework for the classification of genetic variants (Richards et al., 2015). However, to improve classification and reduce the number of variants of unknown significance (VUS), these universal guidelines need to be tuned according to the disease and the specific gene. With this aim, ClinGene Variant Curation Expert Panels ([https://clinicalgenome.org/affiliation/vcep/#ep\\_table\\_heading](https://clinicalgenome.org/affiliation/vcep/#ep_table_heading)) are focused on developing adapted guidelines for specific genes or diseases. Given that the ACMG/AMP classification guidelines were proposed for high-penetrance genes in classical Mendelian disorders (Richards et al., 2015), their adaptation to moderate/low-penetrance genes, such as *ATM*, is challenging and requires multidisciplinary collaborative efforts. *ATM* gene is included in hereditary cancer panels, mainly because heterozygous *ATM* deleterious mutations increase the risk of cancer, particularly breast cancer (BC), and have also been associated with colorectal, prostate, and pancreatic cancer predisposition (Roberts et al., 2012; Na et al., 2017; Jerzak et al., 2018). Moreover, biallelic *ATM* loss-of-function variant carriers present the highly severe Ataxia Telangiectasia disease.

**Article 3** tackled the necessity of adapting the ACMG variant interpretation guideline to the *ATM* gene to ultimately reduce the VUS rate. The author of this thesis besides to be involved in the whole process of rules adaptation, focused also on the specific adjustment of the splicing predictive evidence: multiple lines of computational data support a deleterious effect (pathogenic supporting, PP3) or suggest no impact (benign supporting, BP4). Thus, considering that public access to the SSF-like and HSF tools is limited, since they

are included in Alamut Visual (Interactive Biosoftware), the SPiCE *in silico* tool (Leman et al., 2018) was selected after a performance assessment in **Article 3**, due to its free availability, user-friendly interface and high performance. For variants affecting the canonical donor splice site, it is proposed applying PP3 when the SPiCE score exceed the threshold of 0.240 (100% sensitivity), and BP4 when they are below it (89.9% sensitivity). For variants affecting the canonical acceptor splice site, PP3 is assigned when exceeding the threshold of 0.789 (87.6% sensitivity) and BP4 when they are under 0.282 (86.3% sensitivity). No evidence is considered for acceptor variants with scores between 0.282 and 0.789 (**Article 3**, Supplementary Fig 2).

The general ACMG variant interpretation guideline (Richards et al., 2015), suggested the use of some tools for computational splicing evidence (such as MES, Gene Splicer, NNSplicer, HSF, NetGene2, or FSPLICE) and recommended to use them in combination (assuming potential pathogenicity if all the tools identify the variant as deleterious). However, it did not specify which tools or what cut-off had to be used, nor the sensitivity or specificity that should be reached if *in silico* splicing tool is applied in this evidence module.

There are already some HBOC genes adapted guidelines with detailed mention for computational splicing evidence: i) Mester et al. (2018) recommended for *PTEN* gene the same combination of tools indicated by Richards et al., 2015, based on a small dataset of 23 variants. Moreover, it specified the *in silico* splicing tools use to characteristics of *PTEN* gene (such as a non-canonical donor splice site in exon 1); ii) For *CDH1* gene Lee et al. (Lee et al., 2018) indicated that at least three *in silico* splicing predictors (such as HSF, MES, Berkeley Drosophila Genome Project (BDGP), and ESEfinder), must be in agreement to apply the supporting rule for variants likely to impact splicing. Interestingly, the authors proposed to analyse coding and non-coding variants predicted to either have an impact on the native site, or result in activation/creation of cryptic/novel splice sites. The use of these tools is supported after positive

## DISCUSSION

correlation between prediction and experimental validation with only three *CDH1* variants; iii) To identify spliceogenic variants in *TP53* Fortunato et al. (2021), suggested the use of SpliceAI or VarSEAK (<https://varseak.bio/>) tools.

As can be observed, there are different recommendations of use, and there is no clear indication of their performance. Interestingly, following the acquired knowledge in **Article 1**, a separate analysis of donor and acceptor splicing sites was recommended for *ATM* *in silico* analysis (**Article 3**). These indications together with other adapted evidence, led to a VUS reduction from 58% to 42% in the pilot set of classified *ATM* variants. The use of SPICE in *ATM* gene allows an accurate splicing variant identification, although limited to splicing sites, still lacking a proven recommendation for the remaining *cis*-elements. Therefore, there is still room for improvement in *ATM* adapted *in silico* evidence. However, taken together, our *ATM* proposal and *PTEN*, *CHD1* and *TP53* examples, highlight the importance of developing specific and adapted ACMG guidelines.

Finally, despite of the improvement of the variant classification system for HBOC genes achieved over last years, there is still a main challenge that requires innovative solutions: a validated automatization of variant classification including adapted gene evidence. In this sense, an attractive approach is the proposal recently published named RENOVO (Favalli et al., 2021), a machine learning-based tool, that classifies variants as pathogenic or benign on the basis of publicly available information and provides a pathogenicity likelihood score (PLS). This tool was trained with ClinVar pathogenic and benign variants, and validated by the authors with *BRCA1*, *BRCA2* and *SCN5A* variants. The authors proposed a reclassification for 67% with >90% estimated precision after analyse all ClinVar VUS with RENOVO (Favalli et al., 2021). However, independent studies must validate its performance.

### 3. Candidate genes identification and validation in a case-control analysis

Since up to a 50% of HBOC patients remains without genetic diagnosis, the identification of new susceptibility genes could explain the missing heritability in this disease, giving the opportunity to develop preventive and therapeutic strategies for the benefit of patients.

For this reason, the aim of **Article 4** was to identify candidate genes related to the HBOC and validate them using a case-control approach. Two different sets of candidate genes were separately analysed.

One of the sets was sequenced in collaboration with COMPLEXO consortium (Southey et al., 2013), evaluating twenty-two genes, previously selected as candidate by its members, in 1,012 HBOC patients without pathogenic variants identified and 488 healthy women. After truncating and potentially spliceogenic variants detection, OR was calculated comparing cases with gnomAD database controls. Our results show that *CAMKK1*, *WRN*, *PER1*, *FANCD2*, *FANCM*, *NEIL3*, *RBL1*, *XRCC4*, *BLM*, and *ALKBH3* were associated with a significant ( $p < 0.05$ ) risk of breast cancer, reaching an OR above 2.

The genes included in the second set were selected from three different sources:

- i) Whole exome sequencing analysis of thirteen HBOC genetically undiagnosed families. *DMC1*, *MACROD1*, *RALGDS*, *TPMT* and *TDP2* were selected after the identification of deleterious or potentially deleterious variants shared in affected members of the families.

## DISCUSSION

- ii) Extended massively parallel sequencing targeted panel (including known susceptibility and “promising candidate” genes) performed in 192 HBOC patients. *RBBP8* was selected due the identification of truncating variants in two unrelated patients (Article 9, Appendix) (Bonache et al., 2018).
- iii) Candidate genes *EDC4* and *RECQL5* were selected due to the promising results showed in Spanish cohorts (Hernández et al., 2018; Tavera-Tapia et al., 2019).

These eight genes were sequenced in DNA samples of 638 HBOC patients without pathogenic variants identified and 206 healthy women. After bioinformatic analysis and variant annotation, ORs were calculated considering truncating and potentially splicing variants in both groups and GnomAD controls database v2.1.1 (Karczewski et al., 2020). Loss-of-function variants in *RALGDS*, *MACROD1* and *RECQL5* showed an OR value above 2. However, none of the genes reached a significant ( $p < 0.05$ ) association. More patients and controls are being sequenced to validate the association of the candidate genes.

Interestingly, Subramanian et al., 2020, after analysing WES in more than 500 high risk ovarian cancer patients, showed an enrichment of LoF variants in forty-three genes compared to gnomAD. The genes act in diverse functional pathways and relatively few were involved in DNA repair, suggesting that much of the remaining heritability is explained by previously underexplored genes and pathways.

In addition, Li et al., 2021, in a well-conducted study, sequenced candidate genes in 3,892 BC cases and controls, and then validated 145 shortlisted genes in 7,619 subjects. Their results identified an overall excess of LoF and missense variants in cases. However, candidate genes with LoF variants with ORs of 2-4 did not account even 1% of cases. This suggests that much of the remaining

genetic causes of high-risk BC families are due to genes in which pathogenic variants are both very rare and convey only low to moderate risk.

Considering the results of Li et al., 2021, Subramanian et al., 2020, articles reviewed in Rotunno et al., 2020, and **Article 4**, we expected that newly associated genes will be part of a diversity of pathways, not only DNA repair. Moreover, we hypothesize that the landscape of gene susceptibility will be completed by various genes explaining a reduced percentage of patients, with a variable penetrance between moderate and low. These genes will have to be validated in large cases and controls studies similar to Dorling et al., 2021 and Hu et al., 2021, where thousands of unselected and familial breast cancer cases were sequenced together with thousands of non-cancer controls and using multiethnicity cases and control populations.

In addition, most of the candidate genes or validation studies are focused on LoF and missense variants. Considering the expertise acquired in **Articles 1** and **2**, we decided to analyse *in silico* potentially splicing variants in **Article 4** (outside +-1 and 2 intronic positions) and perform a separate ORs analysis. Up to our knowledge, this is the first study that considers this kind of variants, that may outperform the identification of potentially deleterious variants.

During the last years, more than 200 candidate genes in HBOC and other hereditary cancer types have been highlighted (Rotunno et al., 2020). However, only a few are recognized as susceptibility cancer genes by the scientific community (Dorling et al., 2021; Hu et al., 2021). In **Article 4**, we have identified genes and suggested risk association by a case control approach, being the size the main limitation of this study, in particular that of the control group.

There exist various limitations in the investigations to identify and validate candidate genes (Rotunno et al., 2020). Most of the studies focus their efforts



## DISCUSSION

on sequencing cases rather than controls and use gnomAD controls database to compare the incidence of variants and calculate ORs. Authors use for comparisons the highest number of sequenced patients for the set of genes. However, not all alleles have been annotated in all patients of gnomAD (due to sequencing failure or differences between dataset that make up gnomAD). Consequently, assuming the highest value of alleles included in gnomAD leads to an underestimation of the allele frequency in controls, and therefore to an overestimation of ORs. To address this point, in **Article 4**, the ORs have been calculated by estimating the number of each allele according to the mean of all annotated alleles for that gene in gnomAD database.

Other weakness is the size of analysed cohorts. For example, Rotunno et al., 2020 indicates that the 53% of studies selected candidate genes after analysing ten or less families, and a 43% of articles did not perform an independent validation analysis, although some recent articles, such as Li et al., 2021 analysed a total of 11,511 samples for a selected number of genes, avoiding this limitation. Also, to note the existence of bias on the candidate gene selection in many studies (including **Article 4**), in favour of genes of specific pathways related with DNA repair or replication pathways.

Finally, recent studies indicate that non-conventional strategies by analysing the genomic tumoral landscape of HBOC patients could be an effective way to identify new related genes. Two recent examples combine breast and ovarian cancer GWAS datasets with transcriptome imputation from normal and tumour breast and ovarian tissues (Kar et al., 2021), or somatic whole-exome sequencing was performed to identify candidate genes in serrated polyposis syndrome (Soares de Lima et al., 2021)

Collectively, the findings from this thesis on the performance of *in silico* splicing tools, the optimization of variant classification guidelines, and the indication of

new susceptibility genes will contribute to the more precise diagnosis of familial breast and ovarian cancer, ensuring that more patients and their families can benefit of preventive measures to reduce the risk of developing cancer, as well as of personalised anti-cancer therapies.



**CONCLUSIONS**



- Analysing the Donor and Acceptor splice sites separately improve the identification of spliceogenic variants.
- The use of *in silico* tools SSF-like, and SSF-like and/or HSF in Acceptor and Donor sites respectively, allows to discriminate spliceogenic variants with a high performance.
- SpliceAI is an efficient *in silico* tool to identify deep intronic variants that create pseudoexonizations.
- The balance of splicing regulatory elements is essential for the pseudoexon formation.
- Mapping splicing regulatory elements is a promising way to identify regions susceptible to be pseudoexonized.
- The adaptation of the variant classification guidelines in the *ATM* gene, together with a validated *in silico* analysis of potential alterations in splicing, reduces the number of variants of uncertain significance.
- The significant identification of loss-of-function variants in *ALKBH3*, *BLM*, *CAMKK1*, *FANCD2*, *FANCM*, *NEIL3*, *PER1*, *RBL1*, *RECQL4*, *WRN* and *XRCC4* genes in patients with HBOC suggests that they may be susceptibility genes.
- Caution should be exercised when comparing allele frequencies of patient cohorts with those of the gnomAD control population.



## **BIBLIOGRAPHY**





## References

- Alenezi WM, Fierheller CT, Recio N, Tonin PN. 2020. Literature review of BARD1 as a cancer predisposing gene with a focus on breast and ovarian cancers. *Genes (Basel)* 11:1–24.
- Anna A, Monika G. 2018. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet* 59:253–268.
- Baeza-Centurion P, Miñana B, Valcárcel J, Lehner B. 2020. Mutations primarily alter the inclusion of alternatively spliced exons. *Elife* 9:1–74.
- Baralle D, Buratti E. 2017. RNA splicing in human disease and in the clinic. *Clin Sci* 131:355–368.
- Barnes DR, Rookus MA, McGuffog L, Leslie G, Mooij TM, Dennis J, Mavaddat N, Adlard J, Ahmed M, Aittomäki K, Andrieu N, Andrulis IL, et al. 2020. Polygenic risk scores and breast and epithelial ovarian cancer risks for carriers of BRCA1 and BRCA2 pathogenic variants. *Genet Med* 22:1653–1666.
- Beitsch PD, Whitworth PW, Hughes K, Patel R, Rosen B, Compagnoni G, Baron P, Simmons R, Smith LA, Grady I, Kinney M, Coomer C, et al. 2019. Underdiagnosis of Hereditary Breast Cancer: Are Genetic Testing Guidelines a Tool or an Obstacle? *J Clin Oncol Off J Am Soc Clin Oncol* 37:453–460.
- Berneburg M, Lehmann AR. 2001. Xeroderma pigmentosum and related disorders: defects in DNA repair and transcription. *Adv Genet* 43:71–102.
- Bonache S, Esteban I, Moles-Fernández A, Tenés A, Duran-Lozano L, Montalban G, Bach V, Carrasco E, Gadea N, López-Fernández A, Torres-Esquius S, Mancuso F, et al. 2018. Multigene panel testing beyond BRCA1/2 in breast/ovarian cancer Spanish families and clinical actionability of findings. *J Cancer Res Clin Oncol* 144:2495–2513.
- Bonadona V, Bonaïti B, Olschwang S, Grandjouan S, Huiart L, Longy M, Guimbaud R, Buecher B, Bignon Y-J, Caron O, Colas C, Noguès C, et al. 2011. Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome. *JAMA* 305:2304–2310.
- Bueno-Martínez E, Sanoguera-Miralles L, Valenzuela-Palomo A, Lorca V, Gómez-Sanz A, Carvalho S, Allen J, Infante M, Pérez-Segura P, Lázaro C, Easton DF, Devilee P, et al. 2021. RAD51D Aberrant Splicing in Breast Cancer: Identification of Splicing Regulatory Elements and Minigene-Based Evaluation of 53 DNA Variants. *Cancers (Basel)* 13:.

## BIBLIOGRAPHY

Burge CB, Tuschl T, Sharp PA. 1999. Splicing of Precursors to MRNAs by the Spliceosomes. *The RNA world*. Cold Spring Harbor Laboratory Press, Plainview, NY. 525–560 p.

Cáceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* 14:R143.

Canson D, Glubb D, Spurdle AB. 2020. Variant effect on splicing regulatory elements, branchpoint usage, and pseudoexonization: Strategies to enhance bioinformatic prediction using hereditary cancer genes as exemplars. *Hum Mutat* 41:1705–1721.

Cartegni L, Chew SL, Krainer AR. 2002. Listening To Silence and Understanding Nonsense: Exonic Mutations That Affect Splicing. *Nat Rev Genet* 3:285–298.

Chasin LA. 2007. Searching for splicing motifs. *Adv Exp Med Biol* 623:85–106.

Cheng J, Çelik MH, Nguyen TYD, Avsec Ž, Gagneur J. 2019a. CAGI 5 splicing challenge: Improved exon skipping and intron retention predictions with MMSplice. *Hum Mutat* 40:1243–1251.

Cheng J, Nguyen TYD, Cygan KJ, Çelik MH, Fairbrother WG, Avsec Ž, Gagneur J. 2019b. MMSplice: Modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol* 20:48.

Cline MS, Babbi G, Bonache S, Cao Y, Casadio R, Cruz X, Díez O, Gutiérrez-Enríquez S, Katsonis P, Lai C, Lichtarge O, Martelli PL, et al. 2019. Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants. *Hum Mutat* 40:1546–1556.

Colombo M, Radice P, la Hoya M de. 2021. Chapter 7 - Functional evidence (I) transcripts and RNA-splicing outline. In: Lázaro C, Lerner-Ellis J, Spurdle A, editors. *Clinical DNA Variant Interpretation*, Academic Press, p 121–144.

Cook SA, Tinker A V. 2019. PARP Inhibitors and the Evolving Landscape of Ovarian Cancer Management: A Review. *BioDrugs* 33:255–273.

Corvelo A, Hallegger M, Smith CWJ, Eyras E. 2010. Genome-Wide Association between Branch Point Properties and Alternative Splicing. *PLOS Comput Biol* 6:e1001016.

Cotton RGH, Scriver CR. 1998. Proof of “disease causing” mutation. *Hum Mutat* 12:1–3.

Couch FJ, Nathanson KL, Offit K. 2014. Two decades after BRCA: setting

paradigms in personalized cancer care and prevention. *Science* (80- ) 343:1466–70.

Danis D, Jacobsen JOB, Carmody LC, Gargano MA, McMurry JA, Hegde A, Haendel MA, Valentini G, Smedley D, Robinson PN. 2021. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *Am J Hum Genet*.

Desmet F-O, Hamroun D, Lalande M, Collod-Bérout G, Claustres M, Bérout C. 2009. Human Splicing Finder : an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 37:1–14.

Díez O, Osorio A, Durán M, Martínez-Ferrandis JI, la Hoya M de, Salazar R, Vega A, Campos B, Rodríguez-López R, Velasco E, Chaves J, Díaz-Rubio E, et al. 2003. Analysis of BRCA1 and BRCA2 genes in Spanish breast/ovarian cancer patients: a high proportion of mutations unique to Spain and evidence of founder effects. *Hum Mutat* 22:301–312.

Domchek SM, Robson ME. 2019. Update on Genetic Testing in Gynecologic Cancer. *J Clin Oncol Off J Am Soc Clin Oncol* 37:2501–2509.

Dorling L, Carvalho S, Allen J, González-Neira A, Luccarini C, Wahlström C, Pooley KA, Parsons MT, Fortunato C, Wang Q, Bolla MK, Dennis J, et al. 2021. Breast Cancer Risk Genes — Association Analysis in More than 113,000 Women. *N Engl J Med* 384:428–439.

Dufner-Almeida LG, Carmo RT do, Masotti C, Haddad LA. 2019. Understanding human DNA variants affecting pre-mRNA splicing in the NGS era. *Adv Genet* 103:39–90.

Duran-Lozano L, Montalban G, Bonache S, Moles-Fernández A, Tenés A, Castroviejo-Bermejo M, Carrasco E, López-Fernández A, Torres-Esquius S, Gadea N, Stjepanovic N, Balmaña J, et al. 2019. Alternative transcript imbalance underlying breast cancer susceptibility in a family carrying PALB2 c.3201+5G>T. *Breast Cancer Res Treat* 174:543–550.

Erkelenz S, Theiss S, Otte M, Widera M, Peter JO, Schaal H. 2014. Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Res* 42:10681–10697.

Fachal L, Blanco A, Santamariña M, Carracedo A, Vega A. 2014. Large genomic rearrangements of BRCA1 and BRCA2 among patients referred for genetic analysis in Galicia (NW Spain): delimitation and mechanism of three novel BRCA1 rearrangements. *PLoS One* 9:e93306.

Favalli V, Tini G, Bonetti E, Voza G, Guida A, Gandini S, Pelicci PG, Mazzarella

## BIBLIOGRAPHY

L. 2021. Machine learning-based reclassification of germline variants of unknown significance: The RENOVO algorithm. *Am J Hum Genet* 108:682–695.

Federici G, Soddu S. 2020. Variants of uncertain significance in the era of high-throughput genome sequencing: A lesson from breast and ovary cancers. *J Exp Clin Cancer Res* 39:1–12.

Feliubadaló L, López-Fernández A, Pineda M, Díez O, Valle J del, Gutiérrez-Enríquez S, Teulé A, González S, Stjepanovic N, Salinas M, Capellá G, Brunet J, et al. 2019. Opportunistic testing of BRCA1, BRCA2 and mismatch repair genes improves the yield of phenotype driven hereditary cancer gene panels. *Int J Cancer* 145:2682–2691.

Ferla R, Calò V, Cascio S, Rinaldi G, Badalamenti G, Carreca I, Surmacz E, Colucci G, Bazan V, Russo A. 2007. Founder mutations in BRCA1 and BRCA2 genes. *Ann Oncol* 18:vi93–vi98.

Fitzgerald RC, Hardwick R, Huntsman D, Carneiro F, Guilford P, Blair V, Chung DC, Norton J, Ragnath K, Krieken JH Van, Dwerryhouse S, Caldas C. 2010. Hereditary diffuse gastric cancer: updated consensus guidelines for clinical management and directions for future research. *J Med Genet* 47:436–444.

Fortuno C, Lee K, Olivier M, Pesaran T, Mai PL, Andrade KC de, Attardi LD, Crowley S, Evans DG, Feng B-J, Foreman AKM, Frone MN, et al. 2021. Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. *Hum Mutat* 42:223–236.

Gaildrat P, Killian A, Martins A, Tournier I, Frébourg T, Tosi M. 2010. Use of Splicing Reporter Minigene Assay to Evaluate the Effect on Splicing of Unclassified Genetic Variants. In: Webb M, editor. *Cancer Susceptibility: Methods and Protocols*, Totowa, NJ: Humana Press, p 249–257.

Gallagher S, Hughes E, Wagner S, Tshiaba P, Rosenthal E, Roa BB, Kurian AW, Domchek SM, Garber J, Lancaster J, Weitzel JN, Gutin A, et al. 2020. Association of a Polygenic Risk Score With Breast Cancer Among Women Carriers of High- and Moderate-Risk Breast Cancer Genes. *JAMA Netw open* 3:e208501.

Gao K, Masuda A, Matsuura T, Ohno K. 2008. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res* 36:2257–2267.

González-Santiago S, Ramón y Cajal T, Aguirre E, Alés-Martínez JE, Andrés R, Balmaña J, Graña B, Herrero A, Llorca G, González-del-Alba A, the SEOM Hereditary Cancer Working Group. 2020. SEOM clinical guidelines in hereditary breast and ovarian

cancer (2019). *Clin Transl Oncol* 22:193–200.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC. 1990. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* (80-) 250:1684–1689.

Harrison SM, Pesaran TF, Mester JL. 2021. Chapter 3 - International consensus guidelines for constitutional sequence variant interpretation. In: Lázaro C, Lerner-Ellis J, Spurdle A, editors. *Clinical DNA Variant Interpretation*, Academic Press, p 29–40.

Hauke J, Horvath J, Groß E, Gehrig A, Honisch E, Hackmann K, Schmidt G, Arnold N, Faust U, Sutter C, Hentschel J, Wang-Gohrke S, et al. 2018. Gene panel testing of 5589 BRCA1/2-negative index patients with breast cancer in a routine diagnostic setting: results of the German Consortium for Hereditary Breast and Ovarian Cancer. *Cancer Med* 7:1349–1358.

Hernández G, Ramírez MJ, Minguillón J, Quiles P, Ruiz de Garibay G, Aza-Carmona M, Bogliolo M, Pujol R, Prados-Carvajal R, Fernández J, García N, López A, et al. 2018. Decapping protein EDC4 regulates DNA repair and phenocopies BRCA1. *Nat Commun* 9:967.

Hoang LN, Gilks BC. 2018. Hereditary Breast and Ovarian Cancer Syndrome: Moving Beyond BRCA1 and BRCA2. *Adv Anat Pathol* 25:85–95.

Houdayer C, Caux-Moncoutier V, Krieger S, Barrois M, Bonnet F, Bourdon V, Bronner M, Buisson M, Coulet F, Gaildrat P, Lefol C, Léone M, et al. 2012. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Hum Mutat* 33:1228–1238.

Hu C, Hart SN, Gnanaolivu R, Huang H, Lee KY, Na J, Gao C, Lilyquist J, Yadav S, Boddicker NJ, Samara R, Klebba J, et al. 2021. A Population-Based Study of Genes Previously Implicated in Breast Cancer. *N Engl J Med* 384:440–451.

Huber D, Seitz S, Kast K, Emons G, Ortmann O. 2020. Use of oral contraceptives in BRCA mutation carriers and risk for ovarian and breast cancer: a systematic review. *Arch Gynecol Obstet* 301:875–884.

Jagadeesh KA, Paggi JM, Ye JS, Stenson PD, Cooper DN, Bernstein JA, Bejerano G. 2019. S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat Genet* 51:755–763.

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, et

## BIBLIOGRAPHY

al. 2019. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 176:535–548.e24.

Jerzak KJ, Mancuso T, Eisen A. 2018. Ataxia-telangiectasia gene (ATM) mutation heterozygosity in breast cancer: a narrative review. *Curr Oncol* 25:e176–e180.

Jian X, Boerwinkle E, Liu X. 2014. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res* 42:13534–13544.

Johansen Øystein, Ryen T, Eftesøl T, Kjosmoen T, Ruoff P. 2009. Splice Site Prediction Using Artificial Neural Networks.

Jr Kazazian CD, Boehm WKS. 2000. ACMG recommendations for standards for interpretation of sequence variations. *Genet Med* 2:302–303.

Kar SP, Considine DPC, Tyrer JP, Plummer JT, Chen S, Dezem FS, Barbeira AN, Rajagopal PS, Rosenow WT, Moreno F, Bodelon C, Chang-Claude J, et al. 2021. Pleiotropy-guided transcriptome imputation from normal and tumor tissues identifies candidate susceptibility genes for breast and ovarian cancer. *Hum Genet Genomics Adv* 2:100042.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581:434–443.

Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* 21:1360–1374.

Kerkhof J, Schenkel LC, Reilly J, McRobbie S, Aref-Eshghi E, Stuart A, Rupar CA, Adams P, Hegele RA, Lin H, Rodenhiser D, Knoll J, et al. 2017. Clinical Validation of Copy Number Variant Detection from Targeted Next-Generation Sequencing Panels. *J Mol Diagnostics* 19:905–920.

Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips K-A, Mooij TM, Roos-Blom M-J, Jervis S, Leeuwen FE van, Milne RL, Andrieu N, Goldgar DE, Terry MB, et al. 2017. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* 317:2402–2416.

Ledermann JA. 2016. PARP inhibitors in ovarian cancer. *Ann Oncol* 27:i40–i44.  
Lee K, Krempely K, Roberts ME, Anderson MJ, Carneiro F, Chao E, Dixon K,

Figueiredo J, Ghosh R, Huntsman D, Kaurah P, Kesserwan C, et al. 2018. Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants. *Hum Mutat* 39:1553–1568.

Leman R, Gaildrat P, Gac GL, Ka C, Fichou Y, Audrezet M, Caux-Moncoutier V, Caputo SM, Boutry-Kryza N, Léone M, Mazoyer S, Bonnet-Dorion F, et al. 2018. Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico / in vitro studies : an international collaborative effort. *Nucleic Acids Res* 1–11.

Leman R, Tubeuf H, Raad S, Tournier I, Derambure C, Lanos R, Gaildrat P, Castelain G, Hauchard J, Killian A, Baert-Desurmont S, Legros A, et al. 2020. Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants. *BMC Genomics* 21:1–12.

Li X, You R, Wang X, Liu C, Xu Z, Zhou J, Yu B, Xu T, Cai H, Zou Q. 2016. Effectiveness of Prophylactic Surgeries in BRCA1 or BRCA2 Mutation Carriers: A Meta-analysis and Systematic Review. *Clin cancer Res an Off J Am Assoc Cancer Res* 22:3971–3981.

Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, Zheng Z, Bose S, Call KM, Tsou HC, Peacocke M, Eng C, Parsons R. 1997. Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat Genet* 16:64–67.

Lier MGF van, Westerman AM, Wagner A, Looman CWN, Wilson JHP, Rooij FWM de, Lemmens VEPP, Kuipers EJ, Mathus-Vliegen EMH, Leerdam ME van. 2011. High cancer risk and increased mortality in patients with Peutz-Jeghers syndrome. *Gut* 60:141–147.

Litton JK, Rugo HS, Ettl J, Hurvitz SA, Gonçalves A, Lee K-H, Fehrenbacher L, Yerushalmi R, Mina LA, Martin M, Roché H, Im Y-H, et al. 2018. Talazoparib in Patients with Advanced Breast Cancer and a Germline BRCA Mutation. *N Engl J Med* 379:753–763.

Lord J, Baralle D. 2021. Splicing in the Diagnosis of Rare Disease: Advances and Challenges. *Front Genet* 12:1146.

Loveday C, Turnbull C, Ramsay E, Hughes D, Ruark E, Frankum JR, Bowden G, Kalmyrzaev B, Warren-Perry M, Snape K, Adlard JW, Barwell J, et al. 2011. Germline mutations in RAD51D confer susceptibility to ovarian cancer. *Nat Genet* 43:879–882.

Lumish HS, Steinfeld H, Koval C, Russo D, Levinson E, Wynn J, Duong J, Chung



## BIBLIOGRAPHY

WK. 2017. Impact of Panel Gene Testing for Hereditary Breast and Ovarian Cancer on Patients. *J Genet Couns* 26:1116–1129.

Malkin D. 1993. p53 and the Li-Fraumeni syndrome. *Cancer Genet Cytogenet* 66:83–92.

Malkin D. 2011. Li-fraumeni syndrome. *Genes Cancer* 2:475–484.

Meijers-Heijboer H, Ouweland A van den, Klijn J, Wasielewski M, Snoo A de, Oldenburg R, Hollestelle A, Houben M, Crepin E, Veghel-Plandsoen M van, Elstrodt F, Duijn C van, et al. 2002. Low-penetrance susceptibility to breast cancer due to CHEK2(\*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet* 31:55–59.

Meindl A, Hellebrand H, Wiek C, Erven V, Wappenschmidt B, Niederacher D, Freund M, Lichtner P, Hartmann L, Schaal H, Ramser J, Honisch E, et al. 2010. Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. *Nat Genet* 42:410–414.

Mersch J, Jackson MA, Park M, Nebgen D, Peterson SK, Singletary C, Arun BK, Litton JK. 2015. Cancers associated with BRCA1 and BRCA2 mutations other than breast and ovarian. *Cancer* 121:269–275.

Mester J, Pesaran T. 2019. The Evolution of Constitutional Sequence Variant Interpretation. *Adv Mol Pathol* 2:1–11.

Mester JL, Ghosh R, Pesaran T, Huether R, Karam R, Hruska KS, Costa HA, Lachlan K, Ngeow J, Barnholtz-Sloan J, Sesock K, Hernandez F, et al. 2018. Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel. *Hum Mutat* 39:1581–1592.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 266:66–71.

Mirza MR, Coleman RL, González-Martín A, Moore KN, Colombo N, Ray-Coquard I, Pignata S. 2020. The forefront of ovarian cancer therapy: update on PARP inhibitors. *Ann Oncol* 31:1148–1159.

Moghadasi S, Eccles DM, Devilee P, Vreeswijk MPG, Asperen CJ van. 2016. Classification and Clinical Management of Variants of Uncertain Significance in High Penetrance Cancer Predisposition Genes. *Hum Mutat* 37:331–336.

Montalban G, Bonache S, Moles-Fernández A, Gadea N, Tenés A, Torres-Esquius S, Carrasco E, Balmaña J, Diez O, Gutiérrez-Enríquez S. 2019. Incorporation of

semi-quantitative analysis of splicing alterations for the clinical interpretation of variants in BRCA1 and BRCA2 genes. *Hum Mutat* 40:2296–2317.

Montalban G, Bonache S, Moles-fernández A, Gisbert-beamud A, Tenés A, Bach V, Carrasco E, López-fernández A, Stjepanovic N, Balmaña J, Díez O, Gutiérrez-enríquez S. 2018a. Screening of BRCA1 / 2 deep intronic regions by targeted gene sequencing identifies the first germline BRCA1 variant causing pseudoexon activation in a patient with breast / ovarian cancer. *J Med Genet* 56:63–74.

Montalban G, Fraile-Bethencourt E, Lopez-Perolio I, Perez-Segura P, Infante M, Duran M, Alonso-Cerezo MC, Lopez-Fernandez A, Díez O, la Hoya M de, Velasco EA, Gutierrez-Enriquez S. 2018b. Characterization of spliceogenic variants located in regions linked to high levels of alternative splicing: BRCA2 c.7976+5G > T as a case study. *Hum Mutat* 39:1155–1160.

Na R, Zheng SL, Han M, Yu H, Jiang D, Shah S, Ewing CM, Zhang L, Novakovic K, Petkewicz J, Gulukota K, Helseth DLJ, et al. 2017. Germline Mutations in ATM and BRCA1/2 Distinguish Risk for Lethal and Indolent Prostate Cancer and are Associated with Early Age at Death. *Eur Urol* 71:740–747.

Ngeow J, Eng C. 2016. Precision medicine in heritable cancer: When somatic tumour testing and germline mutations meet. *npj Genomic Med* 1:2015–2017.

Nyberg T, Frost D, Barrowdale D, Evans DG, Bancroft E, Adlard J, Ahmed M, Barwell J, Brady AF, Brewer C, Cook J, Davidson R, et al. 2020. Prostate Cancer Risks for Male BRCA1 and BRCA2 Mutation Carriers: A Prospective Cohort Study. *Eur Urol* 77:24–35.

Ohno K, Takeda J, Masuda A. 2018. Rules and tools to predict the splicing effects of exonic and intronic mutations. *WIREs RNA* 9:e1451.

Padilla N, Moles-Fernández A, Riera C, Montalban G, Özkan S, Ootes L, Bonache S, Díez O, Gutiérrez-Enríquez S, la Cruz X de. 2019. BRCA1- and BRCA2-specific in silico tools for variant interpretation in the CAGI 5 ENIGMA challenge. *Hum Mutat* 40:1593–1611.

Parsons MT, Tadini E, Li H, Hahnen E, Wappenschmidt B, Feliubadaló L, Aalfs CM, Agata S, Aittomäki K, Alducci E, Alonso-Cerezo MC, Arnold N, et al. 2019. Large scale multifactorial likelihood quantitative analysis of BRCA1 and BRCA2 variants: An ENIGMA resource to support clinical variant classification. *Hum Mutat* 40:.

Pashayan N, Antoniou AC, Ivanus U, Esserman LJ, Easton DF, French D, Sroczynski G,

## BIBLIOGRAPHY

Hall P, Cuzick J, Evans DG, Simard J, Garcia-Closas M, et al. 2020. Personalized early detection and prevention of breast cancer: ENVISION consensus statement. *Nat Rev Clin Oncol* 17:687–705.

Paske IBAW Te, Ligtenberg MJL, Hoogerbrugge N, Voer RM de. 2020. Candidate Gene Discovery in Hereditary Colorectal Cancer and Polyposis Syndromes- Considerations for Future Studies. *Int J Mol Sci* 21:.

Petrucelli N, Daly MB, Pal T. 2016. BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, Amemiya A, editors. *Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, Amemiya A, Editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993–2021., Seattle (WA):.*

Pilarski R, Burt R, Kohlman W, Pho L, Shannon KM, Swisher E. 2013. Cowden Syndrome and the PTEN Hamartoma Tumor Syndrome: Systematic Review and Revised Diagnostic Criteria . *JNCI J Natl Cancer Inst* 105:1607–1616.

Plazzer JP, Sijmons RH, Woods MO, Peltomäki P, Thompson B, Dunnen JT Den, Macrae F. 2013. The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. *Fam Cancer* 12:175—180.

Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FBL, Hoogerbrugge N, Spurdle AB, Tavtigian S V. 2008. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat* 29:1282–1291.

Polak P, Kim J, Braunstein LZ, Karlic R, Haradhavala NJ, Tiao G, Rosebrock D, Livitz D, Kübler K, Mouw KW, Kamburov A, Maruvka YE, et al. 2017. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat Genet* 49:1476–1486.

Pujol P, Barberis M, Beer P, Friedman E, Piulats JM, Capoluongo ED, Garcia Foncillas J, Ray-Coquard I, Penault-Llorca F, Foulkes WD, Turnbull C, Hanson H, et al. 2021. Clinical practice guidelines for BRCA1 and BRCA2 genetic testing. *Eur J Cancer* 146:30–47.

Qian Y, Mancini-DiNardo D, Judkins T, Cox HC, Brown K, Elias M, Singh N, Daniels C, Holladay J, Coffee B, Bowles KR, Roa BB. 2017. Identification of pathogenic retrotransposon insertions in cancer predisposition genes. *Cancer Genet* 216–217:159–169.

Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, Jayatilake H, McGuffog L, et al. 2007. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* 39:165–167.

Reese MG, Eeckman FH, Kulp D, Haussler D. 1997. Improved Splice Site Detection in Genie MARTIN. *J Comput Biol* 4:311–323.

Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, et al. 2015. ClinGen — The Clinical Genome Resource. *N Engl J Med* 372:2235–2242.

Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, McGuffog L, Evans DG, et al. 2006. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* 38:873–875.

Rhine CL, Cygan KJ, Soemedi R, Maguire S, Murray MF, Monaghan SF, Fairbrother WG. 2018. Hereditary cancer genes are highly susceptible to splicing mutations. *PLoS Genet* 14:1–18.

Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, Ward BE. 2008. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet Med* 10:294–300.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J. 2015. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17:405–424.

Riepe T V, Khan M, Roosing S, Cremers FPM, 't Hoen PAC. 2021. Benchmarking deep learning splice prediction tools using functional splice assays. *Hum Mutat*.

Rivera-Muñoz EA, Milko L V, Harrison SM, Azzariti DR, Kurtz CL, Lee K, Mester JL, Weaver MA, Currey E, Craigen W, Eng C, Funke B, et al. 2018. ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Hum Mutat* 39:1614–1622.

Roberts NJ, Jiao Y, Yu J, Kopelovich L, Petersen GM, Bondy ML, Gallinger S, Schwartz AG, Syngal S, Cote ML, Axilbund J, Schulick R, et al. 2012. ATM mutations in patients with hereditary pancreatic cancer. *Cancer Discov* 2:41–46.

Robson M, Im S-A, Senkus E, Xu B, Domchek SM, Masuda N, Delaloge S, Li W,

## BIBLIOGRAPHY

Tung N, Armstrong A, Wu W, Goessl C, et al. 2017. Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. *N Engl J Med* 377:523–533.

Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. 2015. Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell* 163:698–711.

Rotunno M, Barajas R, Clyne M, Hoover E, Simonds NI, Lam TK, Mechanic LE, Goldstein AM, Gillanders EM. 2020. A systematic literature review of whole exome and genome sequencing population studies of genetic susceptibility to cancer. *Cancer Epidemiol Biomarkers Prev* 29:1519–1534.

Rowlands CF, Baralle D, Ellingford JM. 2019. Machine Learning Approaches for the Prioritization of Genomic Variants Impacting Pre-mRNA Splicing. *Cells* 8:.

Sakamoto Y, Sereewattanawoot S, Suzuki A. 2020. A new era of long-read sequencing for cancer genomics. *J Hum Genet* 65:3–10.

Samadder NJ, Giridhar K V., Baffy N, Riegert-Johnson D, Couch FJ. 2019. Hereditary Cancer Syndromes—A Primer on Diagnosis and Management: Part 1: Breast-Ovarian Cancer Syndromes. *Mayo Clin Proc* 94:1084–1098.

Sanoguera-Miralles L, Valenzuela-Palomo A, Bueno-Martínez E, Llovet P, Díez-Gómez B, Caloca MJ, Pérez-Segura P, Fraile-Bethencourt E, Colmena M, Carvalho S, Allen J, Easton DF, et al. 2020. Comprehensive Functional Characterization and Clinical Interpretation of 20 Splice-Site Variants of the RAD51C Gene. *Cancers (Basel)* 12:.

Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, North B, McGuffog L, et al. 2006. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet* 38:1239–1241.

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet* 19:329–346.

Shapiro MB, Senapathy P. 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res* 15:7155–7174.

Signal B, Gloss BS, Dinger ME, Mercer TR. 2018. Machine learning annotation of human branchpoints. *Bioinformatics* 34:920–927.

Soares de Lima Y, Arnau-Collell C, Díaz-Gay M, Bonjoch L, Franch-Expósito S, Muñoz J, Moreira L, Ocaña T, Cuatrecasas M, Herrera-Pariente C, Carballal S, Moreno L,

et al. 2021. Germline and Somatic Whole-Exome Sequencing Identifies New Candidate Genes Involved in Familial Predisposition to Serrated Polyposis Syndrome. *Cancers (Basel)* 13:.

Sociedad Española de Oncología Médica. 2019. *Cáncer Hereditario (3ª Edición)*. 470 p.

Southey MC, Park DJ, Nguyen-Dumont T, Campbell I, Thompson E, Trainer AH, Chenevix-Trench G, Simard J, Dumont M, Soucy P, Thomassen M, Jønson L, et al. 2013.

Subramanian DN, Zethoven M, McInerney S, Morgan JA, Rowley SM, Lee JEA, Li N, Goringe KL, James PA, Campbell IG. 2020. Exome sequencing of familial high-grade serous ovarian carcinoma reveals heterogeneity for rare candidate susceptibility genes. *Nat Commun* 11:1640.

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 0:1–41.

Tang R, Prosser DO, Love DR. 2016. Evaluation of Bioinformatic Programmes for the Analysis of Variants within Splice Site Consensus Regions. *Adv Bioinformatics*.

Tavera-Tapia A, la Hoya M de, Calvete O, Martín-Gimeno P, Fernández V, Macías JA, Alonso B, Pombo L, Diego C de, Alonso R, Pita G, Barroso A, et al. 2019. RECQL5: Another DNA helicase potentially involved in hereditary breast cancer susceptibility. *Hum Mutat* 40:566–577.

Thompson BA, Spurdle AB, Plazzer J-P, Greenblatt MS, Akagi K, Al-Mulla F, Bapat B, Bernstein I, Capellá G, Dunnen JT den, Sart D du, Fabre A, et al. 2014. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat Genet* 46:107–115.

Tischkowitz MD, Hodgson S V. 2003. Fanconi anaemia. *J Med Genet* 40:1–10. Toland AE, Forman A, Couch FJ, Culver JO, Eccles DM, Foulkes WD, Hogervorst FBL, Houdayer C, Levy-Lahad E, Monteiro AN, Neuhausen SL, Plon SE, et al. 2018. Clinical testing of BRCA1 and BRCA2: a worldwide snapshot of technological practices. *npj Genomic Med* 3:7.

Tsaousis GN, Papadopoulou E, Apeessos A, Agiannitopoulos K, Pepe G, Kampouri S, Diamantopoulos N, Floros T, Iosifidou R, Katopodi O, Koumariou A, Markopoulos C, et al. 2019. Analysis of hereditary cancer syndromes by using a panel of genes: novel

## BIBLIOGRAPHY

and multiple pathogenic mutations. *BMC Cancer* 19:535.

Tubeuf H, Charbonnier C, Soukarieh O, Blavier A, Lefebvre A, Dauchel H, Frebourg T, Gaildrat P, Martins A. 2020. Large-scale comparative evaluation of user-friendly tools for predicting variant-induced alterations of splicing regulatory elements. *Hum Mutat* 1–19.

Tung N, Lin NU, Kidd J, Allen BA, Singh N, Wenstrup RJ, Hartman A-R, Winer EP, Garber JE. 2016. Frequency of Germline Mutations in 25 Cancer Susceptibility Genes in a Sequential Series of Patients With Breast Cancer. *J Clin Oncol Off J Am Soc Clin Oncol* 34:1460–1468.

Ule J, Blencowe BJ. 2019. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Mol Cell* 76:329–345.

Vaz-Drago R, Custódio N, Carmo-Fonseca M. 2017. Deep intronic mutations and human disease. *Hum Genet* 136:1093–1111.

Wai HA, Lord J, Lyon M, Gunning A, Kelly H, Cibir P, Seaby EG, Spiers-Fitzgerald K, Lye J, Ellard S, Thomas NS, Bunyan DJ, et al. 2020. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med* 22:1005–1014.

Wang E, Aifantis I. 2020. RNA Splicing and Cancer. *Trends in cancer* 6:631–644.

Wang G-S, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 8:749–761.

Wang J, Smith PJ, Krainer AR, Zhang MQ. 2005. Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res* 33:5053–5062.

Wang Z, Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 14:802–813.

Warner E. 2018. Screening BRCA1 and BRCA2 Mutation Carriers for Breast Cancer. *Cancers (Basel)* 10:.

Whiley PJ, la Hoya M de, Thomassen M, Becker A, Brandao R, Pedersen IS, Montagna M, Menendez M, Quiles F, Gutierrez-Enriquez S, Leeneer K De, Tenes A, et al. 2014. Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. *Clin Chem* 60:341–352.

Wilkinson ME, Charenton C, Nagai K. 2020. RNA Splicing by the Spliceosome.

Annu Rev Biochem 89:359–388.

Woodward ER, Veen EM Van, Evans DG. 2021. From BRCA1 to Polygenic Risk Scores: Mutation-Associated Risks in Breast Cancer-Related Genes. *Breast Care*.

Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D. 1994. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* 265:2088–2090.

Xie Z, Tang L, Xie Z, Sun C, Shuai H, Zhou C, Liu Y, Yu M, Zheng Y, Meng L, Zhang W, Leal SM, et al. 2020. Splicing characteristics of dystrophin pseudoexons and identification of a novel pathogenic intronic variant in the DMD gene. *Genes (Basel)* 11:1–13.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, et al. 2015. The human splicing code reveals new insights into the genetic determinants of disease. *Science* (80-) 347:1254806–1.

Yanes T, Young M-A, Meiser B, James PA. 2020. Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field. *Breast Cancer Res* 22:21.

Yang X, Song H, Leslie G, Engel C, Hahnen E, Auber B, Horváth J, Kast K, Niederacher D, Turnbull C, Houlston R, Hanson H, et al. 2020. Ovarian and Breast Cancer Risks Associated With Pathogenic Variants in RAD51C and RAD51D. *JNCI J Natl Cancer Inst* 112:1242–1250.

Yeo G, Burge CB. 2004. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J Comput Biol* 11:377–394.

Zhang Q, Fan X, Wang Y, Sun M-A, Shao J, Guo D. 2017. BPP: a sequence-based algorithm for branch point prediction. *Bioinformatics* 33:3166–3172.



## BIBLIOGRAPHY

### **URL of online resources and tools**

ClinVar database: <https://www.ncbi.nlm.nih.gov/clinvar/>

BRCAExchange: <https://brcaexchange.org/>

SpliceAI: <https://github.com/Illumina/SpliceAI>

gnomAD database: <https://gnomAD.broadinstitute.org/>

ENIGMA: <https://enigmaconsortium.org/>

INSIGHT: <https://www.insight-group.org/variants/databases/>

**APPENDIX**



## Additional publications

This section includes additional works in which the doctoral student participated as co-author. These publications were not strictly related with the main topic of this thesis, but were relevant for its development

### Article 5

Alternative transcript imbalance underlying breast cancer susceptibility in a family carrying PALB2 c.3201+5G>T.

Duran-Lozano L, Montalban G, Bonache S, Moles-Fernández A, Tenés A, Castroviejo-Bermejo M, Carrasco E, López-Fernández A, Torres-Esquius S, Gadea N, Stjepanovic N, Balmaña J, Gutiérrez-Enríquez S, Diez O.

**Breast Cancer Res Treat.** 2019 Apr;174(2):543-550.

doi: 10.1007/s10549-018-05094-8.



## Alternative transcript imbalance underlying breast cancer susceptibility in a family carrying *PALB2* c.3201+5G>T

Laura Duran-Lozano<sup>1</sup> · Gemma Montalban<sup>1</sup> · Sandra Bonache<sup>1</sup> · Alejandro Moles-Fernández<sup>1</sup> · Anna Tenés<sup>2</sup> · Marta Castroviejo-Bermejo<sup>3</sup> · Estela Carrasco<sup>4</sup> · Adrià López-Fernández<sup>4</sup> · Sara Torres-Esquius<sup>4</sup> · Neus Gadea<sup>4,5</sup> · Neda Stjepanovic<sup>4,5</sup> · Judith Balmaña<sup>4,5</sup> · Sara Gutiérrez-Enríquez<sup>1</sup> · Orland Diez<sup>1,2</sup>

Received: 8 June 2018 / Accepted: 7 December 2018  
© Springer Science+Business Media, LLC, part of Springer Nature 2018

### Abstract

**Purpose** Disruption of splicing motifs by genetic variants can affect the correct generation of mature mRNA molecules leading to aberrant transcripts. In some cases, variants may alter the physiological transcription profile composed of several transcripts, and an accurate in vitro evaluation is crucial to establish their pathogenicity. In this study, we have characterized a novel *PALB2* variant c.3201+5G>T identified in a breast cancer family.

**Methods** Peripheral blood RNA was analyzed in two carriers and ten controls by RT-PCR and Sanger sequencing. The splicing profile was also characterized by semi-quantitative capillary electrophoresis and quantitative PCR. RAD51 foci formation and *PALB2* LOH status were evaluated in primary breast tumor samples from the carriers.

**Results** *PALB2* c.3201+5G>T disrupts intron 11 donor splice site and modifies the abundance of several alternative transcripts ( $\Delta 11$ ,  $\Delta 12$ , and  $\Delta 11,12$ ), also present in control samples. All transcripts are predicted to encode for non-functional proteins. Semi-quantitative and quantitative analysis of *PALB2* full-length transcript indicated haploinsufficiency in carriers. One tumor exhibited *PALB2* LOH and RAD51 assay indicated homologous recombination deficiency in both tumors.

**Conclusions** Our results support a pathogenic classification for *PALB2* c.3201+5G>T, highlighting the impact of variants causing an imbalanced expression of natural RNA isoforms in cancer susceptibility.

**Keywords** Hereditary breast cancer · *PALB2* · Alternative splicing · RNA isoforms

Laura Duran-Lozano and Gemma Montalban have contributed equally to this work.

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s10549-018-05094-8>) contains supplementary material, which is available to authorized users.

✉ Sara Gutiérrez-Enríquez  
sgutierrez@vhio.net

✉ Orland Diez  
odiez@vhio.net

<sup>1</sup> Oncogenetics Group, Vall d'Hebron Institute of Oncology, VHIO, 08035 Barcelona, Spain

<sup>2</sup> Area of Clinical and Molecular Genetics, University Hospital of Vall d'Hebron, 08035 Barcelona, Spain

<sup>3</sup> Experimental Therapeutics Group, Vall d'Hebron Institute of Oncology, VHIO, 08035 Barcelona, Spain

<sup>4</sup> High Risk and Cancer Prevention Group, VHIO, 08035 Barcelona, Spain

<sup>5</sup> Medical Oncology Department, University Hospital of Vall d'Hebron, 08035 Barcelona, Spain

### Introduction

Genetic variants that disrupt splicing motifs can lead to RNA mis-splicing, contributing to human hereditary diseases. The most common type of variants that alter splicing are located in highly conserved GT and AG dinucleotides (positions +1 and +2 of the 5' donor site, and positions -2 and -1 of the 3' acceptor site, respectively). Other exonic and intronic nucleotides surrounding these positions are also conserved and critical for a correct splice site selection [1, 2]. Yet, their potential effects on splicing are scarcely assayed in the clinical setting.

Routine splicing analysis in clinical diagnostics is usually performed by RT-PCR, agarose gel examination and Sanger sequencing. However, splicing profiles can be complex to interpret when several alternative transcripts are present, especially if these transcripts are in-frame events that might rescue gene functionality [3]. Moreover, the variant allele may still be able to produce full-length transcript [4], which

may affect the resulting cell phenotype. High-resolution electrophoresis and expression assays can provide a comprehensive qualitative and quantitative screening of the whole mRNA landscape [5].

*PALB2* (Partner and Localizer of *BRCA2*) encodes for a nuclear protein of 1186 amino acids, that interacts with *BRCA1*, *BRCA2*, and other DNA repair proteins like *RAD51* paralogs, to promote DNA double-strand break repair by homologous recombination (HR) [6–8]. *PALB2* is now unquestionably present in multi-gene panel testing for hereditary breast cancer (BC) individuals, since large studies of patients and controls found that *PALB2* pathogenic variants conferred high risk of developing BC [9, 10]. *PALB2* is also linked with male BC and pancreatic cancer [11, 12], and recent works also found an association with colorectal cancer [13], although larger epidemiological studies are still needed to definitely consider *PALB2* as a colorectal cancer susceptibility gene.

In this work, we have characterized at RNA level a novel *PALB2* c.3201+5G>T variant, located outside the canonical donor splice site from intron 11. Our study highlights the complexity to interpret the pathogenicity of variants causing an imbalanced expression of natural RNA isoforms.

## Materials and methods

The proband was diagnosed with a breast invasive ductal carcinoma (IDC) at age 45 and with a second primary lung cancer at age 62. Two first-degree relatives (brother and mother) were also affected with breast IDCs at the ages of 54 and 82, respectively. Hormonal status from the three IDCs was ER+, PR+, and HER2-. A paternal cousin was affected with BC at age 66, and two paternal uncles had stomach cancer (age unknown). The mother was also diagnosed with colorectal cancer at age 86, and the brother had colon polyposis at age 40 (Fig. 1a). Patients received genetic counseling and written informed consent was obtained for further genetic and research studies.

*PALB2* c.3201+5G>T variant was identified by multi-gene panel testing as described in a previous work from our laboratory [14], and confirmed by Sanger sequencing. Protocols for germline DNA, RNA, and tumor DNA extraction, as well as in silico splicing analysis, RT-PCR experiments, sequencing methodologies and immunofluorescence for *RAD51* foci detection, are extensively described in Supplementary Material. In brief, RT-PCR primers were designed to amplify the *PALB2* region comprised between exons 9 and 13 (Supplementary Table 1) in two carriers (proband and brother) and 10 healthy controls. RT-PCR products were qualitatively analyzed by capillary electrophoresis in a QIAxcel Advanced system (QIAGEN) and by Sanger

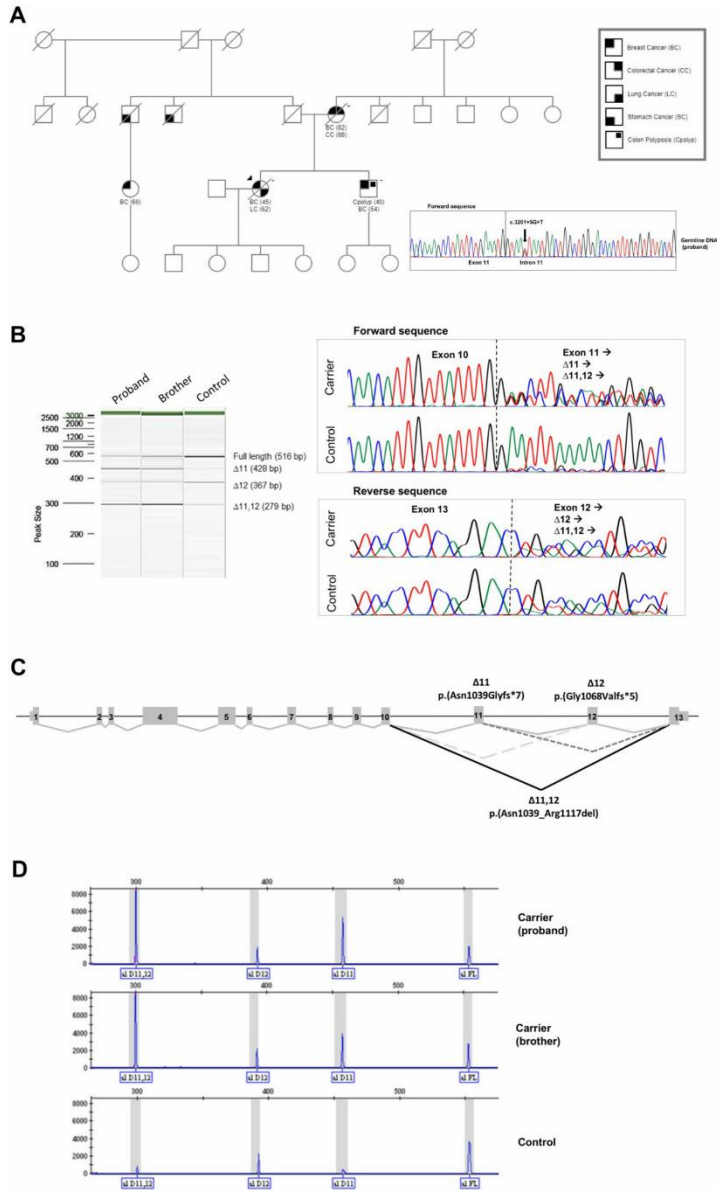
**Fig. 1** Family pedigree and qualitative analysis of the RNA effect caused by *PALB2* c.3201+5G>T variant. **a** The index case is indicated with an arrow head and ages of cancer onset are specified between brackets. Confirmed *PALB2* c.3201+5G>T carriers are marked with a + symbol. Sanger electropherogram confirming the presence of the variant in proband's germline DNA is shown. **b** Capillary electrophoresis in a QIAxcel instrument of RT-PCR products from two carriers (proband and brother) and one control showing different abundance of *PALB2* isoforms. Forward and reverse electropherograms show cDNA sequences of *PALB2* c.3201+5G>T carrier and a healthy control. Isoforms lacking exon 11, exon 12 and both 11 and 12 were confirmed. **c** Diagram showing the different *PALB2* isoforms detected: reference full-length isoform (solid gray lines),  $\Delta 11$  (discontinuous line),  $\Delta 12$  (spotted line) and  $\Delta 11,12$  (solid black line). **d** Fluorescence fragment profiles of the two carriers and one control showing the presence of  $\Delta 11$ ,  $\Delta 12$  and  $\Delta 11,12$  in all samples but at different expression levels

sequencing. Semi-quantitative analysis was performed by capillary electrophoresis of FAM-labeled amplicons in a Genetic Analyzer ABI3130xl (Applied Biosystems); and *PALB2* expression was measured by quantitative real-time PCR (qPCR) using predesigned human-specific primers and TaqMan probes. *PALB2* loss of heterozygosity (LOH) was determined in primary breast tumors from the proband's mother and brother by Sanger sequencing and targeted gene sequencing. Homologous recombination repair activity was also assessed in the tumor specimens by immunofluorescence detection of *RAD51* foci as previously described [15] (see Supplementary Material for further details).

## Results

*PALB2* variant c.3201+5G>T was identified in a breast cancer family by multi-gene panel testing. This variant was confirmed by Sanger sequencing in blood DNA from the proband and her brother, also affected with BC (Fig. 1a). To our knowledge, this variant is not present in ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>), LOVD (Leiden Open Variation Database, <http://www.lovd.nl/3.0/home>), gnomAD (Genome Aggregation Database, <http://gnomad.broadinstitute.org/>), FLOSSIES (<https://whi.color.com/>) and HGMD (The Human Gene Mutation Database, <http://www.hgmd.cf.ac.uk/ac/index.php>) databases, and it is not reported in the literature. In silico analysis showed a reduction of the natural donor splice site, indicating a potential splicing alteration (Supplementary Table 2).

Qualitative cDNA study of the two carriers and 10 controls by QIAxcel capillary electrophoresis revealed in all samples the presence of four transcripts corresponding to the reference full-length transcript (516 bp),  $\Delta 11$  (428 bp),  $\Delta 12$  (367 bp), and  $\Delta 11,12$  (279 bp) (Fig. 1b, c). All transcripts were confirmed by Sanger sequencing (Fig. 1b). Capillary electrophoresis of FAM-labeled amplicons revealed a variation in the proportion of the alternative transcripts





$\Delta 11$ ,  $\Delta 11,12$ , and  $\Delta 12$  in variant carriers compared to controls (Fig. 1d). Semi-quantitative data obtained from QIAxcel electrophoresis showed a  $>0.5$  reduction of full-length transcript levels in both carriers compared to controls, suggesting that the variant allele is not producing normal transcripts (Fig. 2a). Splicing fraction estimation obtained from capillary electrophoresis of FAM-labeled amplicons indicated that isoform  $\Delta 11,12$  substantially contributes to the total splicing fraction in carriers: this isoform was present at higher proportion in both carriers (48.50%) compared to controls (11.94%) (Fig. 2b), and it causes an in-frame deletion resulting in a PALB2 protein lacking 79 aa (p.Asn1039\_Arg1117del) that are part of the WD40 domain (Supplementary Fig. 1). Isoform  $\Delta 11$  was also increased in carriers (24.14%) compared to controls (5.27%), and it is predicted to introduce a premature stop codon (p.Asn1039Glyfs\*7). Finally, the proportion of isoform  $\Delta 12$ , also predicted to introduce a premature stop codon (p.Gly1068Valfs\*5), diminishes in carriers (11.88%) compared to controls (24.83%) (Fig. 2b). Quantitative measurement of *PALB2* global expression and *PALB2* full-length expression using Taqman assays targeting exon 5–6 and exon 11–12 junctions, respectively, showed a significant reduction in carriers compared to controls: carriers =  $0.67 \pm 0.18$  vs. controls =  $1.42 \pm 0.11$ ,  $p = 0.0175$ ; carriers =  $0.36 \pm 0.07$  vs. controls =  $1.90 \pm 0.19$ ,  $p = 0.0059$ , respectively (Mean  $\pm$  95% CI, *t-test* of unpaired samples) (Fig. 2c).

We also examined *PALB2* LOH status in breast tumor samples from the proband's mother and brother. Targeted gene sequencing and Sanger sequencing revealed the loss of the wild-type allele in the mother's tumor (80% cellularity; Variant Allele Frequency (VAF) = 81.82%), whereas the brother did not exhibit LOH (70% cellularity; VAF = 56.52%) (Fig. 3a). Immunofluorescence assay did not show RAD51 foci in breast tumor-FFPE from carriers (see Fig. 3b for mother's tumor; refer to [16] for brother's tumor), indicating homologous recombination deficiency in both carriers.

## Discussion

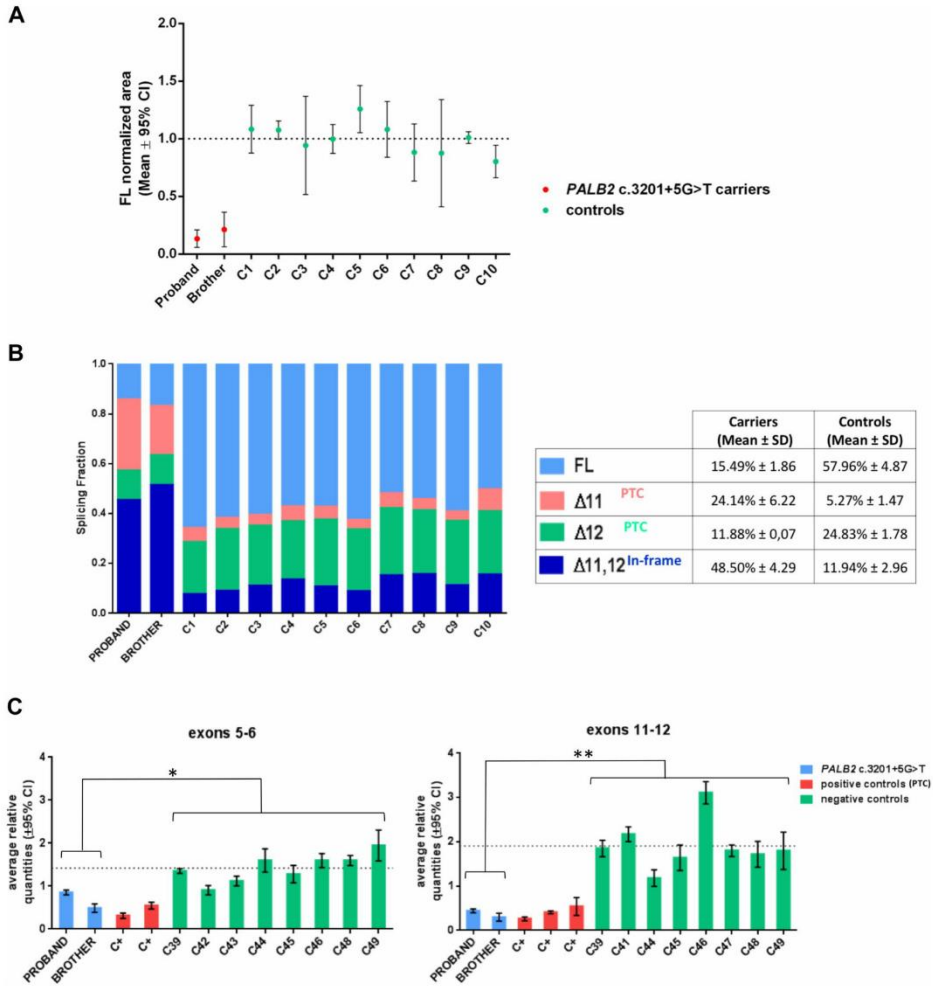
We have characterized a novel *PALB2* variant c.3201+5G>T identified in a family with breast cancer history. The variant is located at +5 position from intron 11 donor splice site (CAG/GTAAAGTAT) and in silico analysis predict the disruption of the splice site (Supplementary Table 2). Results obtained from RNA analysis confirmed a splicing alteration consisting of an imbalanced expression of several *PALB2* alternative RNA isoforms. The variant up-regulates isoforms  $\Delta 11,12$  (in-frame) and  $\Delta 11$  (frameshift), and down-regulates isoform  $\Delta 12$  (frameshift). The splicing profiles detected in

peripheral blood from control samples are consistent with RNAseq results obtained from lymphoblastoid cell lines [17], i.e.,  $\Delta 12$  isoform is predominant over  $\Delta 11$  and  $\Delta 11,12$  (Fig. 2b). Isoform  $\Delta 11,12$  presumably contributes to variant pathogenicity by encoding a PALB2 protein lacking 79aa of the WD40 domain. Protein alignment of the deleted region in 10 reference model species, shows high conservation of this region with  $\approx 50\%$  of the residues conserved across species (Supplementary Fig. 1). WD40 domain mediates direct interactions between PALB2 and key proteins involved in homologous recombination (HR), such as BRCA2 and RAD51 [18], RAD51C and XRCC3 [8], as well as the binding to POLH which mediates recombination-associated DNA synthesis [19]. Hence, the loss of amino acids 1039 to 1117 from WD40, comprising the PALB2 key residues (Leu1046, Lys1047, Leu1070, Pro1097, and Lys1098) that provide interaction with BRCA2 [18], is probably related to HR repair impairment. In fact, functional studies evaluating interactions between PALB2 WD40 domain and HR proteins found that breast cancer-associated missense variants (L939W, T1030I, L1143P) gave rise to unstable PALB2 proteins that altered the binding to BRCA2, RAD51C, and RAD51 [20]. In addition, WD40 domain contains a nuclear export sequence (NES) (amino acids 852–987) that would be exposed to an export protein if a premature stop codon is present after aa 987, leading to an unusual cytoplasmic localization and aberrant function of PALB2 [21]. Therefore, isoforms  $\Delta 11$  and  $\Delta 12$  would also contribute to variant pathogenicity by producing truncated proteins (p.Asn1039Glyfs\*7 and p.Gly1068Valfs\*5, respectively) exposing NES and in consequence inducing a cytoplasmic mislocalization of PALB2 along with its interacting proteins, preventing their access to sites of DNA damage [21].

Semi-quantitative analysis showed a drastic reduction of full-length transcript levels in carriers. Unfortunately, none of the carriers had informative heterozygous exonic variants to perform allele-specific assays and formally exclude the possibility that the variant allele produces a certain amount of full-length transcript. Alternatively, specific amplification and measurement of full-length transcript by qPCR (exons 11–12 probe), showed a significant reduction in carriers compared to controls, supporting that the variant allele is not transcribing full-length transcripts (Fig. 2c).

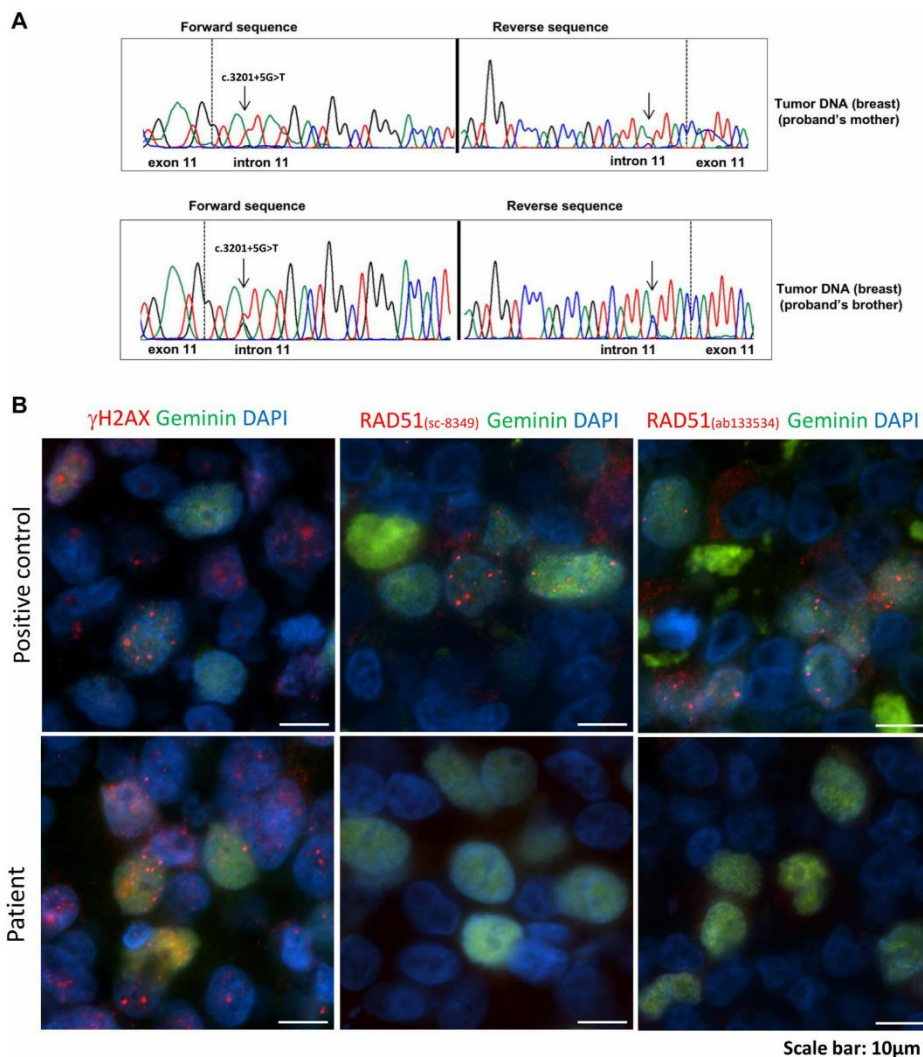
The loss of the wild-type allele in one tumor sample also supports a potential causality of the variant. However, the second tumor did not exhibit LOH and massive sequencing of *PALB2* whole coding region ruled out the presence of somatic deleterious mutations. In this regard, some *PALB2* heterozygous tumors with no LOH and high HR deficiency scores have been described, suggesting alternative mechanisms of PALB2 functional loss [22] or a dominant negative effect of PALB2 mutated proteins [23].





**Fig. 2** Semi-quantitative and quantitative evaluation of the effect on RNA yielded by *PALB2* c.3201+5G>T. **a** Full-length relative amounts measured in the two carriers (red dots) and 10 non-carrier controls (green dots). Normalized full-length mean obtained from controls is indicated in dotted line. Error bars indicate 95% confidence interval. **b** Splicing fraction measured by capillary electrophoresis analysis of FAM-labeled fragments. Bar graphs show the splicing fraction mean of each transcript, which is indicated by different

colors: full-length (FL) in light blue, Δ11 in orange and Δ11,12 in dark blue. **c** *PALB2* expression measured by qPCR in peripheral blood from *PALB2* c.3201+5G>T carriers (blue) and non-carrier controls (green). Carriers of *PALB2* frameshift variants were used as positive controls (in red). *B2M* and *GAPDH* were used as reference genes for data normalization. Mean ±95% CI of technical replicates is denoted. Unpaired *t*-test in carriers vs. controls (\* $p < 0.05$ ; \*\* $p < 0.01$ )



**Fig. 3** LOH analysis and RAD51 assay in tumor samples. **a** Sanger analysis in primary breast tumor samples detected LOH in one tumor (mother). **b** Detection of  $\gamma$ H2AX and RAD51 foci (red) in S/G2-phase cells (geminin-positive; green) by immunofluorescence.

RAD51 foci were not detected in mother's tumor (patient). Nuclei were visualized with DAPI (blue). A *PALB2* wild-type patient-derived xenograft model of a breast tumor sample was used as positive control

Previous works have proposed RAD51 foci formation as predictive biomarker of response to PARP inhibitors (PARPi) in homologous recombination (HR) deficient

tumors [15, 24] and demonstrated that RAD51 foci detection is feasible in formalin-fixed paraffin-embedded breast cancer samples to accurately detect HR activity [15].

Interestingly, tumor samples from carriers did not display RAD51 foci, supporting that HR-deficiency is due to non-functional PALB2 proteins generated from *PALB2* c.3201+5G>T alleles.

*PALB2* has been defined as a crucial mediator of HR in human cells, and PALB2-deficient cells have been shown to be sensitive to PARPi [25]. In this regard, the promising clinical applicability of PARP inhibitors in HR-deficient tumors is likely to be feasible in the short term for carriers of germline *PALB2* pathogenic variants [26, 27].

According to ACMG (American College of Medical Genetics and Genomics) guidelines for variant interpretation [28], *PALB2* c.3201+5G>T variant should be classified as likely pathogenic: there are well-established in vitro functional studies supporting a damaging effect (PS3), the variant is present in affected individuals and absent in gnomAD controls (PM2), it cosegregates with disease in multiple-affected family members (PP1) and there is computational evidence supporting a deleterious effect (PP3).

Other *PALB2* variants affecting intron 11 donor site (intronic + 1,+2 positions, or beyond) have been described in breast cancer families and received different types of classification. Some variants were predicted to alter splicing, but to our knowledge, no experimental characterization has been performed (Supplementary Table 2).

The lack of studies evaluating the functionality of proteins generated from in-frame RNA isoforms, questions whether *PALB2* isoform  $\Delta 11,12$  could retain some functionality and modulate c.3201+5G>T cancer risk. However, our study supports a pathogenic role for the variant based on: (i) segregation in three relatives affected with BC (Fig. 1a); (ii) reduction of *PALB2* global expression in carriers (exons 5–6 probe; Fig. 2c) which would indirectly indicate less amount of PALB2 proteins; (iii) reduction of *PALB2* full-length transcript levels in carriers (exons 11–12 probe; Fig. 2c); (iv) detection of isoforms  $\Delta 11,12$ ,  $\Delta 11$  and  $\Delta 12$  that encode proteins lacking totally or partially an important functional domain (WD40); (v) absence of RAD51 foci in tumor samples from the variant carriers, indicating homologous recombination deficiency (Fig. 3b).

In all, our study shows how an imbalanced expression of natural occurring *PALB2* RNA isoforms can predispose to breast cancer disease, and highlights the use of accurate qualitative, quantitative and functional assays as a key procedure to correctly interpret genetic variants that generate complex splicing landscapes.

**Acknowledgements** The authors thank Cristina Cruz and Violeta Serra from the Experimental Therapeutics Group at VHIO for kindly providing immunofluorescence protocols and helpful discussions. The authors also acknowledge the Cellex Foundation for providing research facilities and equipment, and Leo Judkins for English language editing.

**Funding** This work was supported by Spanish Instituto de Salud Carlos III (ISCIII) funding, an initiative of the Spanish Ministry of Economy and Innovation partially supported by European Regional Development FEDER Funds: FIS P112/02585 and P115/00355 (to O Diez), P113/01711 and P116/01218 (to S. Gutiérrez-Enríquez). S. Gutiérrez-Enríquez is supported by a Miguel Servet contract (CP10/00617). M. Castroviejo-Bermejo is awarded with a Junta Provincial de Barcelona, Fundació Científica Asociación Española Contra el Cáncer (AECC) fellowship. S. Bonache is recipient of an Asociación Española Contra el Cáncer (AECC) contract.

### Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed consent** informed consent was obtained from all individual participants included in the study.

### References

1. Scotti MM, Swanson MS (2016) RNA mis-splicing in disease. *Nat Rev Genet* 17:19–32. <https://doi.org/10.1038/nrg.2015.3>
2. Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 3:285–298. <https://doi.org/10.1038/nrg775>
3. de la Hoya M, Soukariéh O, López-Perollo I et al (2016) Combined genetic and splicing analysis of BRCA1 c.[594-2A> C; 641A> G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. *Hum Mol Genet* 25:2256–2268. <https://doi.org/10.1093/hmg/ddw094>
4. Bonnet C, Krieger S, Vezain M et al (2008) Screening BRCA1 and BRCA2 unclassified variants for splicing mutations using reverse transcription PCR on patient RNA and an ex vivo assay based on a splicing reporter minigene. *J Med Genet* 45:438–446. <https://doi.org/10.1136/jmg.2007.056895>
5. Baralle D, Buratti E (2017) RNA splicing in human disease and in the clinic. *Clin Sci* 131:355–368. <https://doi.org/10.1042/CS20160211>
6. Zhang F, Ma J, Wu J et al (2009) PALB2 Links BRCA1 and BRCA2 in the DNA-Damage Response. *Curr Biol* 19:524–529. <https://doi.org/10.1016/j.cub.2009.02.018>
7. Xia B, Sheng Q, Nakanishi K et al (2006) Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. *Mol Cell* 22:719–729. <https://doi.org/10.1016/j.molcel.2006.05.022>
8. Park JY, Zhang F, Andreassen PR (2014) PALB2: The hub of a network of tumor suppressors involved in DNA damage responses. *Biochim Biophys Acta* 1846:263–275. <https://doi.org/10.1016/j.bbcan.2014.06.003>
9. Tung N, Domchek SM, Stadler Z et al (2016) Counselling framework for moderate-penetrance cancer-susceptibility mutations. *Nat Rev Clin Oncol* 13:581–588. <https://doi.org/10.1038/nrcli.nonc.2016.90>
10. Antoniou AC, Casadei S, Heikkinen T et al (2014) Breast-cancer risk in families with mutations in PALB2. *N Engl J Med* 371:497–506. <https://doi.org/10.1056/NEJMoa1400382>



11. Pritzlaff M, Summerour P, McFarland R et al (2017) Male breast cancer in a multi-gene panel testing cohort: insights and unexpected results. *Breast Cancer Res Treat* 161:575–586. <https://doi.org/10.1007/s10549-016-4085-4>
12. Zhen DB, Rabe KG, Gallinger S et al (2015) BRCA1, BRCA2, PALB2, and CDKN2A mutations in familial pancreatic cancer: A PACGENE study. *Genet Med* 17:569–577. <https://doi.org/10.1038/gim.2014.153>
13. AlDubayan SH, Giannakis M, Moore ND et al (2018) Inherited DNA-Repair Defects in Colorectal Cancer. *Am J Hum Genet* 102:401–414. <https://doi.org/10.1016/j.ajhg.2018.01.018>
14. Bonache S, Esteban I, Moles-Fernández A et al (2018) Multigene panel testing beyond BRCA1/2 in breast/ovarian cancer Spanish families and clinical actionability of findings. *J Cancer Res Clin Oncol* 144:2495–2513
15. Cruz C, Castroviejo-Bermejo M, Gutiérrez-Enríquez S et al (2018) RAD51 foci as a functional biomarker of homologous recombination repair and PARP inhibitor resistance in germline BRCA-mutated breast cancer. *Ann Oncol* 29:1203–1210. <https://doi.org/10.1093/annonc/mdy099>
16. Castroviejo-Bermejo M, Cruz C, Llop-Guevara A et al (2018) A RAD51 assay feasible in routine tumor samples calls PARP inhibitor response beyond BRCA mutation. *EMBO Mol Med* 10:e9172
17. Davy G, Rousselin A, Goardon N et al (2017) Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer. *Eur J Hum Genet* 25:1147–1154. <https://doi.org/10.1038/ejhg.2017.116>
18. Oliver AW, Swift S, Lord CJ et al (2009) Structural basis for recruitment of BRCA2 by PALB2. *EMBO Rep* 10:990–996. <https://doi.org/10.1038/embor.2009.126>
19. Buisson R, Niraj J, Pauty J et al (2014) Breast cancer proteins PALB2 and BRCA2 stimulate polymerase  $\eta$  in recombination-associated DNA Synthesis At Blocked Replication Forks. *Cell Rep* 6:553–564. <https://doi.org/10.1016/j.celrep.2014.01.009>
20. Park JY, Singh TR, Nassar N et al (2014) Breast cancer-associated missense mutants of the PALB2 WD40 domain, which directly binds RAD51C, RAD51 and BRCA2, disrupt DNA repair. *Oncogene* 33:4803–4812. <https://doi.org/10.1038/onc.2013.421>
21. Pauty J, Couturier AM, Rodrigue A et al (2017) Cancer-causing mutations in the tumor suppressor PALB2 reveal a novel cancer mechanism using a hidden nuclear export signal in the WD40 repeat motif. *Nucleic Acids Res* 45:2644–2657. <https://doi.org/10.1093/nar/gkx011>
22. Lee J, Li N, Rowley S et al (2018) Molecular analysis of PALB2 associated breast cancers. *J Pathol* 245:53–60. <https://doi.org/10.1111/peps.12055>
23. Buisson R, Masson J-Y (2012) PALB2 self-interaction controls homologous recombination. *Nucleic Acids Res* 40:10312–10323. <https://doi.org/10.1093/nar/gks807>
24. Naipal KAT, Verkaik NS, Ameziane N et al (2014) Functional ex vivo assay to select homologous recombination-deficient breast tumors for PARP inhibitor treatment. *Clin Cancer Res* 20:4816–4826. <https://doi.org/10.1158/1078-0432.CCR-14-0571>
25. Buisson R, Dion-Côté A-M, Coulombe Y et al (2010) Cooperation of breast cancer proteins PALB2 and piccolo BRCA2 in stimulating homologous recombination. *Nat Struct Mol Biol* 17:1247–1254. <https://doi.org/10.1038/nsmb.1915>
26. Southey MC, Winship I, Nguyen-Dumont T (2016) PALB2: research reaching to clinical outcomes for women with breast cancer. *Hered Cancer Clin Pract* 14:9. <https://doi.org/10.1186/s13053-016-0049-2>
27. Nepomuceno TC, De Gregoriis G, de Oliveira FMB et al (2017) The role of PALB2 in the DNA Damage response and cancer predisposition. *Int J Mol Sci*. <https://doi.org/10.3390/ijms18091886>
28. Richards S, Aziz N, Bale S et al (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17:405–423. <https://doi.org/10.1038/gim.2015.30>



## Article 6

Screening of *BRCA1/2* deep intronic regions by targeted gene sequencing identifies the first germline *BRCA1* variant causing pseudoexon activation in a patient with breast/ovarian cancer.

Montalban G, Bonache S, Moles-Fernández A, Gisbert-Beamud A, Tenés A, Bach V, Carrasco E, López-Fernández A, Stjepanovic N, Balmaña J, Diez O, Gutiérrez-Enríquez S.

**J Med Genet.** 2019 Feb;56(2):63-74.

doi: 10.1136/jmedgenet-2018-105606.

## ORIGINAL ARTICLE

# Screening of *BRCA1/2* deep intronic regions by targeted gene sequencing identifies the first germline *BRCA1* variant causing pseudoexon activation in a patient with breast/ovarian cancer

Gemma Montalban,<sup>1</sup> Sandra Bonache,<sup>1</sup> Alejandro Moles-Fernández,<sup>1</sup> Alexandra Gisbert-Beamud,<sup>1</sup> Anna Tenés,<sup>2</sup> Vanessa Bach,<sup>1</sup> Estela Carrasco,<sup>3</sup> Adrià López-Fernández,<sup>3</sup> Neda Stjepanovic,<sup>3,4</sup> Judith Balmaña,<sup>3,4</sup> Orland Diez,<sup>1,2</sup> Sara Gutiérrez-Enríquez<sup>1</sup>

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/jmedgenet-2018-105606>).

<sup>1</sup>Oncogenetics Group, Vall d'Hebron Institut d'Oncologia, Barcelona, Spain

<sup>2</sup>Area of Clinical and Molecular Genetics, University Hospital of Vall d'Hebron, Barcelona, Spain

<sup>3</sup>High Risk and Cancer Prevention Group, Vall d'Hebron Institut d'Oncologia, Barcelona, Spain

<sup>4</sup>Medical Oncology Department, University Hospital of Vall d'Hebron, Barcelona, Spain

## Correspondence to

Dr. Orland Diez and Dr. Sara Gutiérrez-Enríquez, Oncogenetics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, 08035, Spain; [odiez@vhio.net](mailto:odiez@vhio.net), [sgutierrez@vhio.net](mailto:sgutierrez@vhio.net)

Received 12 July 2018  
Revised 16 October 2018  
Accepted 28 October 2018



© Author(s) (or their employer(s)) 2018. No commercial re-use. See rights and permissions. Published by BMJ.

**To cite:** Montalban G, Bonache S, Moles-Fernández A, et al. *J Med Genet* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jmedgenet-2018-105606

## ABSTRACT

**Background** Genetic analysis of *BRCA1* and *BRCA2* for the diagnosis of hereditary breast and ovarian cancer (HBOC) is commonly restricted to coding regions and exon-intron boundaries. Although germline pathogenic variants in these regions explain about ~20% of HBOC cases, there is still an important fraction that remains undiagnosed. We have screened *BRCA1/2* deep intronic regions to identify potential spliceogenic variants that could explain part of the missing HBOC susceptibility.

**Methods** We analysed *BRCA1/2* deep intronic regions by targeted gene sequencing in 192 high-risk HBOC families testing negative for *BRCA1/2* during conventional analysis. Rare variants (MAF <0.005) predicted to create/activate splice sites were selected for further characterisation in patient RNA. The splicing outcome was analysed by RT-PCR and Sanger sequencing, and allelic imbalance was also determined when heterozygous exonic loci were present.

**Results** A novel transcript was detected in *BRCA1* c.4185+4105C>T variant carrier. This variant promotes the inclusion of a pseudoexon in mature mRNA, generating an aberrant transcript predicted to encode for a non-functional protein. Quantitative and allele-specific assays determined haploinsufficiency in the variant carrier, supporting a pathogenic effect for this variant. Genotyping of 1030 HBOC cases and 327 controls did not identify additional carriers in Spanish population.

**Conclusion** Screening of *BRCA1/2* intronic regions has identified the first *BRCA1* deep intronic variant associated with HBOC by pseudoexon activation. Although the frequency of deleterious variants in these regions appears to be low, our study highlights the importance of studying non-coding regions and performing comprehensive RNA assays to complement genetic diagnosis.

## INTRODUCTION

Pathogenic germline variants in the tumour suppressor genes *BRCA1* (MIM# 113705) and *BRCA2* (MIM# 600185) (*BRCA1/2*) predispose to breast and ovarian cancer (BC/OC). To date, more than ~3500 risk-associated variants in *BRCA1/2* have been reported in the ClinVar database (<https://www.ncbi.nlm.nih.gov/clinvar/>).

Carriers of pathogenic variants in *BRCA1* result in an increased cumulative risk of developing BC and OC that reaches 66% and 41% at age 70, respectively. Similarly, for *BRCA2*, the cumulative risks for BC and OC reach 61% and 15%, respectively.<sup>1</sup> The identification of pathogenic variants in *BRCA1/2* provides accurate clinical management of hereditary breast and ovarian cancer (HBOC) families based on personalised prevention and therapeutic strategies.<sup>2–4</sup> However, pathogenic mutations in these genes explain less than 20% of HBOC cases.

In recent years, causative genetic variants in other tumour suppressor genes involved in homologous recombination DNA repair have also been linked to moderate risks of developing BC/OC.<sup>5–8</sup> The application of massive sequencing technologies in the clinical setting allows the simultaneous screening of risk-associated HBOC genes, improving the effectiveness of identifying new families at risk.<sup>3,9,10</sup> However, there is still an important fraction of HBOC cases for which genetic analysis does not identify causative variants underlying the predisposition to BC/OC.<sup>2</sup>

Genetic testing commonly identifies variants that generate truncated proteins (nonsense, frame-shift, splicing variants) or alter functional domains (missense, in-frame variants). However, deep intronic regions may also contain nucleotide changes that could alter splicing by creating/activating splice sites or by disrupting cis-regulatory splicing elements (intronic enhancers/silencers). These variants have the potential to generate aberrant transcripts by introducing pseudoexons in the mature mRNA.<sup>11–13</sup> Such events have been reported in a variety of human genetic syndromes, including cancer,<sup>14,15</sup> but their association with HBOC remains largely unexplored. Given that conventional *BRCA1/2* analysis in the majority of clinical laboratories is restricted to coding regions and exon-intron boundaries, there is currently a lack of information about the frequency of deleterious spliceogenic variants occurring in deep intronic regions of these genes. Previous works based on RNA analysis of patients with HBOC with uninformative *BRCA1/2* results (i.e., no pathogenic variant identified) have



## Cancer genetics

also proposed the existence of spliceogenic variants in deep intronic regions that could explain part of the missing HBOC genetic susceptibility.<sup>16–18</sup> In the present study, we have analysed *BRCA1/2* introns by targeted gene sequencing in a group of Spanish patients with high-risk HBOC testing negative for *BRCA1/2*. Candidate spliceogenic variants have been selected using *in silico* approaches and their impact has been characterised in patient RNA. To our knowledge, this is the first study aiming to screen *BRCA1/2* deep intronic regions in clinical samples and determine the frequency of spliceogenic variants in these regions in the Spanish population.

## MATERIALS AND METHODS

## Patient and control samples acquisition

A total of 192 patients with hereditary BC/OC testing negative for *BRCA1/2* were selected according to the following inclusion criteria: personal history of BC before age 36 (n=77), BC with two or more first or second-degree relatives with BC/OC (n=60), personal history of OC before age 60 (n=38). Additionally, we included six patients diagnosed with two BC (bilateral or ipsilateral with or without BC family history); seven patients with BC diagnosed after age 36, with one male BC, OC or pancreatic cancer-affected relative and four patients with BC/OC history affected with colon, endometrium, sarcoma or stomach cancer. All patients were referred for genetic counselling at the High-risk and Prevention Cancer Unit from Vall d'Hebron Hospital, Barcelona and they provided written informed consent for *BRCA1/2* testing and research studies. Patients were screened for *BRCA1/2* point mutations and large genomic rearrangements by Sanger sequencing and Multiplex ligation-dependent probe amplification (MLPA) (MRC-Holland), respectively.

A total of 327 non-affected control samples were recruited from the Spanish National DNA Bank (Salamanca, Spain). Controls were selected randomly from a population of healthy women above 50 years of age with no personal or family history of cancer. Similarly, control RNAs were obtained from 20 healthy individuals without HBOC history and from four normal breast tissue samples supplied by Biochain (AMSBIO).

Patient DNA was obtained from 10 mL of peripheral blood and isolated using Gentra Puregene Blood Kit (QIAGEN), following manufacturer's protocol. DNA concentrations were determined using Qubit dsDNA BR Assay kit (ThermoFisher). RNA from variant carriers and controls (n=20) was isolated from 10 mL of peripheral blood samples using Trizol Reagent (ThermoFisher). RNA was cleaned up using RNeasy Minikit (QIAGEN) and treated with RNase-Free DNase Set (QIAGEN) to remove traces of genomic DNA. RNA integrity was determined in E-Gel Precast agarose gels (Invitrogen) and concentrations were measured using a NanoDrop Spectrophotometer (ThermoFisher).

Massively parallel sequencing of *BRCA1/2* intronic regions

Agilent SureDesign web-based tool (Agilent Technologies) was used to design a custom Agilent SureSelect bait library of probes targeting whole coding, non-coding and intronic sequences with additional flanking 10 kb genomic sequences of *BRCA1* and *BRCA2*. Captured genomic regions from *BRCA1* and *BRCA2* spanned chr17: 41,186,312–41,287,500 and chr13: 32,879,617–32,983,809, respectively (see online supplementary figure 1 for *BRCA1/2* genomic coverage). Deep sequencing was performed in a MiSeq Instrument (Illumina). DNA library preparation, sequencing protocols and bioinformatics pipeline for sequencing data alignment and variant calling have been extensively described in a previous work from our laboratory.<sup>19</sup>

*In silico* analysis and variant prioritisation

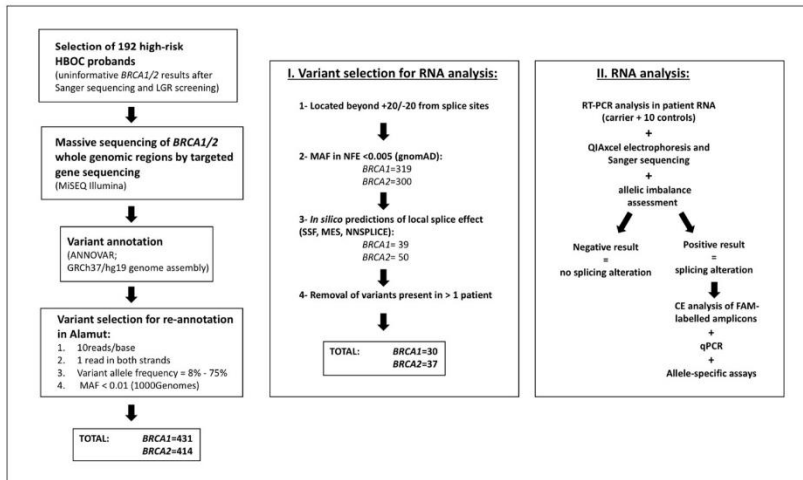
Variants were annotated with ANNOVAR tool using GRCh37/hg19 genome assembly. Variants with a minimum of 10 reads per base, with at least one read in sense (+) and antisense (-) strands, with a variant allele frequency between 75% and 8% and a Minor Allele Frequency (MAF) <0.01 at 1000 Genomes Project database, were included for reannotation using Alamut software v2.10 (Interactive Biosoftware). Reference transcripts NM\_007294.3 (*BRCA1*) and NM\_000059.3 (*BRCA2*) were used for variant reannotation. Population data from the Genome Aggregation Database (gnomAD) (<http://gnomad.broadinstitute.org/>) and splicing predictions from SpliceSiteFinder-like (SSF), MaxEntScan (MES), HumanSplicingFinder (HSF), GeneSplicer and Splice Site Prediction by Neural Network (NNSPLICE) were incorporated.

Variants located beyond +20/-20 positions from canonical donor (GT) and acceptor (AG) splice sites, respectively, were prioritised for RNA analysis when: (1) MAF in non-Finnish European population (NFE)<0.005; (2) a local splicing effect (i.e., creation of new splice sites or activation of existing cryptic sites) was predicted according to Alamut's interpretation algorithm, which uses splice site signal scores from MES, NNSPLICE and SSF tools (<https://www.interactive-biosoftware.com/alamut-visual/>). We applied a final filtering step that consisted of removing variants occurring in >1 patient. A diagram summarising the strategy followed for deep intronic regions screening and variant prioritisation is depicted in figure 1.

Characterisation of *BRCA1/2* variants in patient RNA

A total of 500 ng of RNA from carriers and controls was retrotranscribed using PrimeScript RT Reagent kit (Takara), combining oligo-dT and random primers. Long PCR fragments (1.5–6 kb) were obtained using Expand Long Range dNTPack (Roche), and short PCR fragments (up to 1.5 kb size) were obtained with EcoTaq (Ecogen). Amplified fragments covered the exons adjacent to intron containing the variant. Primer sequences and PCR conditions are described in online supplementary table 1A. We used 2–5 µL of cDNA to a final PCR reaction of 25–35 µL. Cycling conditions were performed according to manufacturer's instructions, with elongation times of 2 min for amplicons <1 kb and 7 min for amplicons >1 kb, to allow the amplification of potential long aberrant transcripts present in the samples. RT-PCR products were qualitatively assessed by capillary electrophoresis (CE) in a QIAxcel instrument, using QIAxcel DNA High-resolution kit (QIAGEN). Controls were run in parallel with patient samples and were used as reference to compare RNA patterns. RT-PCR products were purified using ExoSAP-IT PCR Product Cleanup Reagent (ThermoFisher) and sequenced using BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems). Sequencing products were run in a Genetic Analyzer ABI3130xl (Applied Biosystems) and Sanger electropherograms were visualised using SeqScape v2.6 and Sequencing Analysis v2.6 softwares (Applied Biosystems). Polymorphic exonic variants in *BRCA1* [c.4308T>C (rs1060915); c.4837A>G (rs1799966)] and *BRCA2* [c.-52A>G (rs206118); c.-26G>A (rs1799943); c.865A>C (rs766173); c.1114A>C (rs144848); c.7242A>G (rs1799955)] or other exonic variants present in samples were used to determine potential allelic imbalance derived from frameshift events degraded by the nonsense-mediated decay (NMD). PCR primers and internal sequencing primers used for allelic imbalance assessment are listed in online supplementary table 1B.





**Figure 1** Strategy followed for the study of *BRCA1/2* deep intronic regions in patients with HBOC. The diagram summarises the steps followed from patient selection to RNA analysis of the selected variants. The total number of unique *BRCA1/2* variants is denoted in each filtering step. CE, capillary electrophoresis; HBOC, hereditary breast and ovarian cancer; LGR, large genomic rearrangements; MAF, minor allele frequency; MES, MaxEntScan; NFE, non-Finnish European population; NNSPLICE, SpliceSitePrediction by Neural Network; SSF, SpliceSiteFinder-like.

**Qualitative analysis by capillary electrophoresis of fluorescent amplicons**

Analysis by CE of FAM-labelled amplicons was performed to characterise aberrant splicing patterns, providing higher sensitivity and resolution<sup>20,21</sup> to rule out the presence of aberrant products not detected with QIAxcel electrophoresis and Sanger sequencing. Amplicons were generated with primers labelled with a FAM molecule at the 5' end. RT-PCR products were diluted 1:30 and run into a Genetic Analyzer ABI3130xl (Applied Biosystems) under the following electrophoresis conditions: temperature 60°C, 12 s injection at 1.2 kV, 2000 s run at 12 kV. GeneScan ROX500 (Applied Biosystems) was used as internal size standard and peak electropherograms were visualised using GeneMapper Software 5 (Applied Biosystems).

**BRCA1 expression analysis by quantitative PCR**

Global *BRCA1* expression levels were measured in variant carrier and 10 controls in two independent quantitative PCR (qPCR) assays, using Taqman probes targeting exons 5–6 junction (Hs01556196\_m1) and exons 23–24 junction (Hs01556193\_m1), respectively. A known pathogenic *BRCA1* splicing mutation c.302-1G>A<sup>22</sup> was used as positive control. Taqman Universal Master Mix II (ThermoFisher) was used for qPCR reactions, and reference genes *GAPDH* (Hs02758991\_g1), *B2M* (Hs99999907\_m1), *ACTB* (Hs03023880\_g1) and *HPRT1* (Hs99999909\_m1) were used for data normalisation. Experiments were run in a 7900HT Fast Real-Time PCR System (Applied Biosystems) using default cycle conditions. *BRCA1* global expression levels were calculated using qBASE +software (Biogazelle), which applies the multiple-gene reference normalisation method.<sup>23</sup> All qPCR experiments were performed in triplicate.

**Allele-specific assessment of BRCA1 normal transcript**

The ability of the variant allele to generate normal transcripts was investigated with a specific RT-PCR assay. Full-length (FL) transcript specific amplification was performed using a forward primer complementary with the last 21 nucleotides of exon 12 and the first two nucleotides of exon 13 (5'-AGTGACATTTTAACCA CTCAGCA-3') and a reverse primer located in exon 18 (5'-TCCG TTCACACACAAACTCAG-3') (amplicon size=934 bp). PCR cycling conditions consisted in: denaturing step at 95°C for 5 min; 35 cycles of 95°C for 15 s, 56°C for 5 s and 72°C for 1 min 30 s and a final elongation step of 10 min at 72°C. Products were sequenced by Sanger and heterozygous exonic loci c.4308T>C (rs1060915; exon 13) and c.4837A>G (rs1799966; exon 16) were used to determine biallelic contribution to FL transcript expression.

Levels of *BRCA1* FL transcript were estimated using CE data obtained from QIAxcel instrument. Variant carrier and 10 controls were analysed in four RT-PCR experiments spanning exons 11–13 (see primers in online supplementary table 1) and peak areas corresponding to the FL transcript (275 bp) were used to estimate its relative abundance. FL levels in one carrier of the *BRCA2* c.6937+594T>G deep intronic variant<sup>16</sup> were also measured in three RT-PCR experiments, with a forward primer located in exon 12 (5'-AGGCTTCAAAAAGCACTCCA-3') and a reverse primer located in exon 14 (5'-TCATCAGAGCCATGTCCATC-3') (amplicon size=294 bp). FL data were normalised by dividing each FL peak area with the FL mean obtained from the control group. Data were analysed using GraphPad Prism software.

We also measured *BRCA1* FL transcripts by qPCR using a Taqman assay targeting exons 12–13 junction (Hs00173233\_m1), which is only present in the reference FL transcript but absent in the aberrant transcript. The splicing mutation c.302-1G>A and a large *BRCA1* deletion spanning exons 1–13 were used as positive controls. Cycling conditions and data analysis were performed as described in the previous section.

## Cancer genetics

**Variant genotyping in *BRCA1/2* negative families and controls**

We additionally genotyped *BRCA1* c.4185+4105C>T in 1030 HBOC Spanish families testing negative for *BRCA1/2* and 327 Spanish controls to determine the frequency of this variant in our population. Genotyping of 380 HBOC samples was performed by conventional PCR using intron 12 primers F-AAGCCCCCTGGAGTTGTCAA and R-TTGACAGAGTC-CCAAACCCA (amplicon size=184bp) and posterior Sanger sequencing. The remaining 650 samples and the controls were genotyped using a custom TaqMan SNP assay (ThermoFisher) containing unlabelled PCR primers (F-GTACCAGTATTCTCCACTTCTTCA, R-GCAAAGAGAGAAAAGGCCCTCTAAA), one VICdye-MGB-labelled probe to detect allele C and one FAMdye-MGB-labelled probe to detect allele T. Taqman Universal Master Mix II (ThermoFisher) was used and 5–20 ng DNA were loaded in each reaction. Allelic discrimination assays were run in a 7900HT Fast Real-Time PCR System (Applied Biosystems) under default cycling conditions. Allelic discrimination plots were obtained using SDS 2.4 software (Applied Biosystems). A positive control (variant carrier) was used in all assays.

**RESULTS****RNA analysis of *BRCA1/2* deep intronic variants with potential splicing effects**

The strategy followed for the study of *BRCA1/2* deep intronic regions in 192 high-risk HBOC families identified 30 *BRCA1* and 37 *BRCA2* candidate splicing variants in 53 patients (28%) (figure 1). Variants are listed in table 1 with detailed information from *in silico* predictions, gnomAD population frequencies, ClinVar review status and RNA results obtained in this study (splicing analysis + allelic imbalance assessment).

A total of 31 variants were not present in gnomAD, which includes whole-genome data from ≈15 500 unrelated individuals. According to Alamut's *in silico* predictions, 29 variants (43.3%) were predicted to create new splice sites (11 donor sites and 18 acceptor sites) and 38 variants (56.7%) were predicted to activate pre-existing splice sites (7 cryptic donor sites and 31 cryptic acceptor sites).

Patient RNA could be obtained to assess the effect of 27 *BRCA1* and 27 *BRCA2* variants. All variants were characterised by RT-PCR assays comparing splicing profiles with healthy controls and posterior Sanger sequencing. Visual inspection of RT-PCR products in QIAxcel and Sanger electropherograms did not detect splicing alterations in any variant carrier (see table 1 and online supplementary figures 2 and 3), with the exception of *BRCA1* c.4185+4105C>T carrier which generated an extra transcript absent in control samples (figure 2A). *BRCA1/2* allelic imbalance was ruled out in 28 patients (table 1) by inspection of Sanger electropherograms at heterozygous exonic loci (see online supplementary figure 4), but an allelic imbalance was detected in *BRCA1* c.4185+4105C>T carrier (figure 2B).

A total of 16 variants occurring in >1 patient (6 *BRCA1* and 10 *BRCA2*) were also identified but not prioritised for RNA analysis (online supplementary table 2). Among these, two individuals carried the *BRCA2* c.6937+594T>G variant previously detected in the French population and reported as the first *BRCA2* deep intronic variant generating an aberrant transcript by activation of a cryptic splice site.<sup>16</sup>

**Family origin and clinical features from *BRCA1* c.4185+4105C>T carrier**

The family is originally from Lleida (western Catalonia). The proband was diagnosed with a high-grade ovarian carcinoma

with papillary serous histology at age 58. The proband's father was diagnosed with prostate cancer at age 70, and a paternal female cousin was diagnosed with BC at age 40. After 2 years of follow-up, the patient was diagnosed with a grade 1 infiltrating ductal breast carcinoma, with positive hormonal receptors (ER+,PR+) and negative HER2 receptors (see family pedigree in online supplementary figure 5).

***In silico* splicing analysis of *BRCA1* c.4185+4105C>T**

*BRCA1* c.4185+4105C>T variant was detected in co-occurrence with *BRCA1* c.80+909T>C (MAF: ALL=0.11%, AFR=0.40%) (rs186169069). *In silico* analysis of *BRCA1* c.80+909T>C using Alamut visual v2.10 (including MES, NNSPLICE, GeneSplicer, Human Splicing Finder and SSF tools) predicted the activation of a pre-existing donor site. For *BRCA1* c.4185+4105C>T, only SSF-like predicted the activation of a pre-existing atypical GC donor site (wild-type=75.9 vs variant=78.5), and the remaining tools predicted the creation of a *de novo* donor site, probably due to their inability to detect GC sites (online supplementary figure 6). The Alamut's algorithm that uses SSF-like, NNSPLICE and MES to predict a local splice effect, defined this variant as creating a new donor splice site (table 1). To our knowledge, *BRCA1* c.4185+4105C>T variant is not present in genetic databases Leiden Open Variation Database (LOVD), Breast Cancer Information Core (BIC), BRCA Share, Human Gene Mutation Database (HGMD) and ClinVar, as of April 2018. Moreover, it is not present in gnomAD and it has not been reported before in the literature.

**Characterisation of *BRCA1* c.4185+4105C>T splicing effect in patient RNA**

RT-PCR experiments were performed covering exons 11–13 (275 bp) in variant carrier and 10 control samples. Experiments were performed in duplicate with mRNA from variant carrier drawn at two different time-points. Products visualised in QIAxcel instrument revealed an extra band in patient sample at ~400 bp (figure 2A). Sanger sequencing confirmed the insertion of a pseudoexon, consisting of 114 nucleotides from intron 12 (figure 2C). To determine whether this transcript could be a minor alternative *BRCA1* isoform, we performed high-sensitivity CE of fluorescent amplicons in variant carrier, 20 blood and 4 breast tissue samples from healthy controls. The novel transcript was only present in variant carrier and was detected in the two mRNA extractions (figure 3).

*In vitro* results were concordant with *in silico* predictions, indicating that *BRCA1* c.4185+4105C>T variant converts a pre-existing GC site into a strong GT donor site that, together with an upstream cryptic acceptor site (online supplementary figure 6), promotes the inclusion of a pseudoexon between exons 12 and 13. We annotated this new transcript as ▼12A (r.4185\_4186ins4185+3990\_4185+4103), which is predicted to introduce four new amino acids and a stop codon, generating a truncated *BRCA1* protein (p.Gln1395\_Gln1396insSerLys-SerLeu\*) (figure 2C).

*BRCA1* global expression was measured in two independent qPCR assays using probes located in exons 5–6 and exons 23–24, respectively. Results showed a notable reduction (>2-fold) of *BRCA1* expression levels in carrier compared with controls (figure 4). This reduction was similar to a known *BRCA1* pathogenic splicing variant c.302-1G>A, used as positive control.

*BRCA1* reference FL transcript specific assessment was determined by qualitative, semiquantitative and quantitative approaches. Qualitative allele-specific analysis was performed

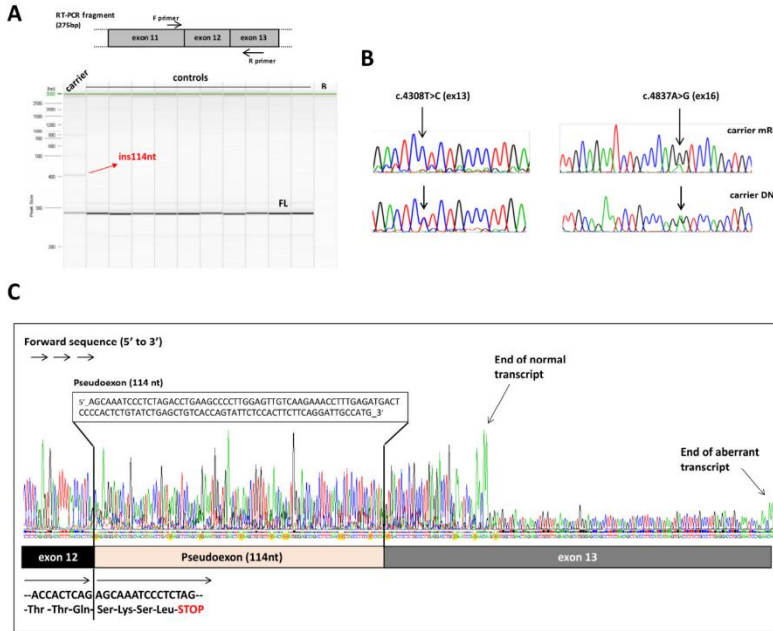
**Table 1** Candidate *BRCA1/2* deep intronic variants predicted to alter splicing and characterisation in patient RNA

GENE	Sample ID	Family phenotype	Variant (HGVS nomenclature*)	Location	dbSNP	Population variant frequencies (gnomAD)	Local splicing effect† (Alman software)	Clinvar review status*	RNA splicing analysis†	Allelic imbalance assessment	
<i>BRCA1</i>	12-144†	BC-35 + DC	c.800>907G>C	intron 2	rs18816980	ALLO.11%—AFR0.40%—	Cryptic donor strongly activated	Not reported	No effect	na	
	14-234	BC-35	c.871>1305delA	intron 2	—	ALLO.005%—AFR0.24%—	Cryptic acceptor weakly activated	Not reported	No effect	na	
	12-238†	TBC + DC	c.1344>1823>T	intron 3	—	—	Cryptic acceptor weakly activated	Not reported	No effect	na	
	15-07†	DC-50	c.1344>1834A>G	intron 3	—	—	New acceptor site	Not reported	No effect	No imbalance	
	14-530	DC-50	c.1335-1582T>A	intron 3	rs52026256	ALLO.12%—AFR0.023%—AMR0.12%—	New acceptor site	Not reported	No patient sample	No patient sample	
	08-213	3 or more BC	c.1351>1534dup	intron 3	rs98428768	WED0.27%—FIN0.057%—	Cryptic acceptor strongly activated	Not reported	Not reported	No patient sample	No patient sample
	15-371	ZBC (TBC<35)	c.135-533C>T	intron 3	rs35071846	FIN0.03%—AFR0.10%—AMR0.24%—	New donor site	Not reported	No effect	No effect	na
	12-238†	DC	c.212>235G>A	intron 5	rs18932191	WED0.39%—FIN0.059%—OTR0.21%—	Cryptic donor strongly activated	Not reported	No effect	No effect	na
	13-265	DC-50	c.441>2340C>A	intron 7	rs56845937	ALLO.17%—AFR0.023%—AMR0.12%—	New acceptor site	Not reported	No effect	No effect	na
	14-106	3 or more BC (TBC<5, DC)	c.671>1200C>T	intron 10	—	ALLO.19%—G0.0.07%—AFR0.027%—	Cryptic acceptor strongly activated	Not reported	No effect	No effect	na
	15-210†	3 or more BC (TBC<5)	c.4185+194C>G	intron 12	rs55068894	—	New acceptor site	Not reported	No effect	No effect	na
	14-431	DC-50	c.4185+1684A>C	intron 12	—	ALLO.11%—AFR0.012%—NFE0.17%—	Cryptic acceptor strongly activated	Not reported	Not reported	No imbalance	No imbalance
	12-309	ZBC	c.4185+2957G>C	intron 12	—	FIN0.14%—OTR0.10%—	Cryptic donor strongly activated	Not reported	Not reported	No effect	No imbalance
	12-144†	BC-35 + DC	c.4185+1105C>T	intron 12	—	—	New donor site	Not reported	Not reported	No effect	No imbalance
	11-018	BC-35	c.4185+4137A>G	intron 12	—	ALLO.002%—NFE0.0067%—	Cryptic donor strongly activated	Not reported	Not reported	No effect	No imbalance
	15-074	BC-36, BC-35+DC-50	c.4186-2320A>G	intron 12	rs87242619	—	New acceptor site	Not reported	Not reported	No effect	No imbalance
	14-006	3 or more BC (DC<10C)	c.4186-2448A>T	intron 12	rs32046884	ALLO.032%—AMR0.12%—NFE0.053%—	Cryptic acceptor weakly activated	Not reported	Not reported	No effect	No imbalance
	11-196	3 or more BC (DC<10C)	c.4338-3856G>A	intron 13	rs19131108	OTR0.10%—	New acceptor site	Not reported	Not reported	No effect	No imbalance
	13-274	3 or more BC	c.4338-2324C>T	intron 13	rs11818895	ALLO.055%—AFR0.012%—NFE0.093%—	New acceptor site	Likely benign (one submission)	Likely benign (one submission)	No effect	na
	15-07†	DC-50	c.4484+4495G>G	intron 14	rs72350860	FIN0.029%—OTR0.10%—	New donor site	New donor site	Benign (reviewed by expert panel)	No patient sample	No patient sample
15-150	3 or more BC	c.4675-113C>G	intron 15	rs18218638	ALLO.023%—AMR0.39%—NFE0.027%—	Cryptic acceptor strongly activated	Not reported	Not reported	No effect	No imbalance	
14-61	3 or more BC	c.4676-381-4678>285delAGTGAACAT	intron 15	—	—	New acceptor site	Not reported	Not reported	No effect	No imbalance	
15-069	BC-35	c.4987-4934A>G	intron 16	rs150756329	ALLO.21%—AFR0.035%—A310.66%—	New donor site	Likely benign (one submission)	Likely benign (one submission)	No effect	No imbalance	
14-31†	BC-35	c.5075-1546G>C	intron 17	rs73126600	WED0.33%—FIN0.064%—OTR0.43%—	Cryptic donor strongly activated	Not reported	Not reported	No effect	No imbalance	
14-368	3 or more BC (TBC<35)	c.5075-1445G>C	intron 17	—	—	Cryptic acceptor strongly activated	Not reported	Not reported	No effect	na	
15-0328†	BC-35	c.5278-3032C>T	intron 20	—	—	Cryptic acceptor strongly activated	Not reported	Not reported	No effect	No imbalance	
14-101†	DC-50	c.5278-1708A>T	intron 20	—	—	New donor site	Not reported	Not reported	No effect	No imbalance	
15-0382†	BC-35	c.5278-4238C>T	intron 20	—	—	New donor site	Not reported	Not reported	No effect	No imbalance	
14-111	BC-35	c.5278-4815>G	intron 20	—	—	New donor site	Not reported	Not reported	No effect	No imbalance	

Continued







**Figure 2** *BRCA1* c.4185+4105C>T characterisation in patient RNA. (A): QIAxcel electrophoresis of RT-PCR assay covering exons 11–13 (275 bp). An extra band of ~400 bp was detected in variant carrier (ins114nt), not present in controls. (B): Sanger electropherogram showing allelic imbalance at polymorphisms c.4308T>C and c.4837A>G. (C): Sanger sequencing confirmed the insertion of 114 nucleotides (nt) between exons 12 and 13, generating a new transcript that we annotated as ▼12A (r.4185\_4186ins4185+3990\_4185+4103). This transcript is predicted to encode for a truncated *BRCA1* protein (p.Gln1395\_Gln1396insSerLysSerLeu\*).

by FL specific amplification and posterior Sanger sequencing of two exonic polymorphisms (rs1060915, rs1799966) known to be heterozygous at DNA level in variant carrier. Visual examination of RT-PCR products by CE detected less amplification of FL transcript in variant carrier, and visual inspection of Sanger peaks at polymorphic positions showed main contribution from only one allele (figure 5A). Semiquantitative measurement of FL transcript using QIAxcel CE data showed a 2-fold reduction of FL levels in carrier sample compared with controls (figure 5B), suggesting that variant allele does not produce normal transcript. Accordingly, specific amplification of FL transcript by qPCR using a probe targeting exons 12–13 junction also showed a significant reduction of FL levels (figure 5C). Consistent with data obtained from qualitative and semiquantitative experiments, these data indicate that the variant allele is not generating normal transcript.

Furthermore, we compared FL levels between *BRCA1* c.4185+4105C>T and the *BRCA2* c.6937+594T>G deep intronic variant reported to alter splicing by Anczuków and colleagues.<sup>16</sup> Semiquantitative CE data from QIAxcel showed lower FL levels in *BRCA1* c.4185+4105C>T carrier (mean=0.27) compared with *BRCA2* c.6937+594T>G carrier (mean=0.74) (figure 5D).

**Variant genotyping in Spanish HBOC families and controls**

A total of 1030 index cases from Spanish HBOC families testing negative for *BRCA1/2* and 327 Spanish controls were genotyped

at *BRCA1* c.4185+4105C>T position. The variant was not identified in any additional family or control, suggesting that this variant is a very rare event (online supplementary figure 7). Moreover, this variant is not reported in 1000 genomes database which includes a set of 165 Spanish controls (77 females and 88 males) and has not been reported in gnomAD which includes whole-genome data from ~15 500 unrelated individuals.

**DISCUSSION**

The aim of this study was to identify novel germline *BRCA1/2* variants in deep intronic regions that could explain hereditary predisposition to BC/OC in high-risk families with uninformative *BRCA1/2* test results. The analysis of 192 high-risk HBOC families by targeted gene sequencing identified 28% of patients carrying rare (MAF <0.005) *BRCA1/2* deep intronic variants located beyond positions +20/-20 from canonical splice sites, with indicative *in silico* predictions of altering splicing. Overall, our results ruled out the presence of predominant splicing alterations occurring in variant carriers, indicating a low specificity for *in silico* tools used in this study. Only *BRCA1* c.4185+4105C>T variant was correctly predicted and its effect was confirmed in patient RNA, producing a novel frameshift transcript ▼12A due to the activation of a cryptic donor site (figure 2A–C). High-resolution CE did not identify this transcript in control samples (blood and normal breast tissue) (figure 3), and it has not been reported in previous RNA studies.<sup>24 25</sup> The

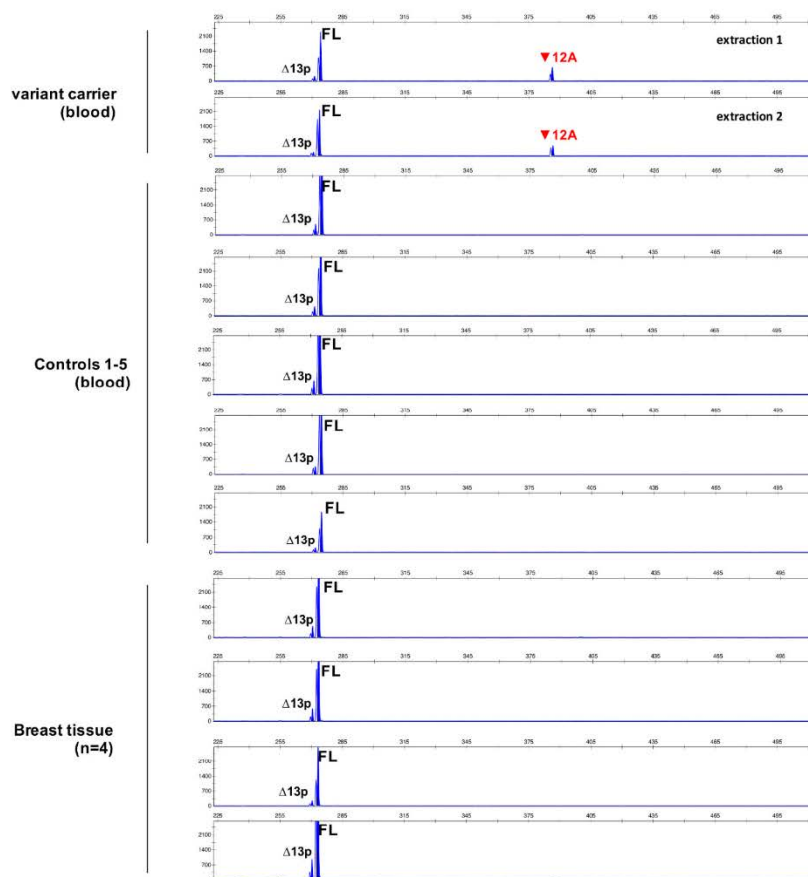
J Med Genet: first published as 10.1136/jmedgenet-2018-105606 on 24 November 2018. Downloaded from <http://jmg.bmj.com/> on 26 November 2018 by guest. Protected by copyright.

## Cancer genetics

presence of this transcript was confirmed in a second RNA extraction, indicating that  $\nabla 12A$  is a true splicing event and it is not a product of illegitimate splicing occurring due to technical artefacts, such as blood processing or blood ageing.<sup>26,27</sup> This transcript introduces a premature stop codon (PTC) 99nt upstream of the next exon-exon junction (pseudoxon-exon13 junction), making the transcript highly likely to be degraded by the NMD system. This mechanism by which PTC transcripts are detected and degraded within cells has been well described in eukaryotes,<sup>28-30</sup> and it is considered a surveillance mechanism to target aberrant mRNAs that would lead to the synthesis of proteins with deleterious effects for the organism. The activity of NMD has been well documented for *BRCA1* PTC transcripts, showing that NMD is triggered by the majority of *BRCA1* PTC mutations, resulting in a 1.5-fold to 5-fold reduction in mRNA abundance.<sup>31</sup> We quantified global *BRCA1* expression in variant

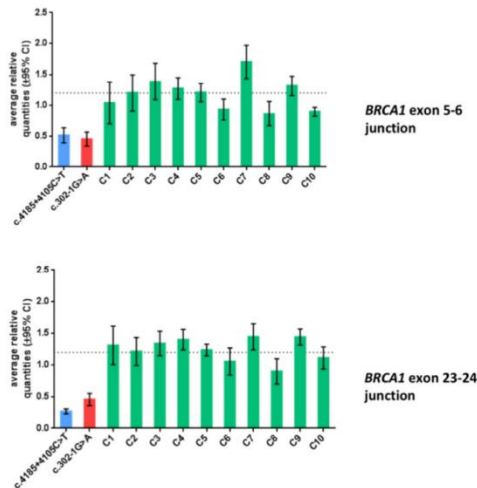
carrier by quantitative PCR, obtaining an indirect measure of NMD activity. We used two different Taqman assays and both showed a >2-fold decrease of *BRCA1* mRNA levels in carrier compared with controls (figure 4), indicating the degradation of transcripts produced by the variant allele and leading to a state of haploinsufficiency. Another approach indicating transcript degradation was the examination of heterozygous *BRCA1* exonic positions in variant carrier, which showed differential allelic expression (figure 2B).

Nucleotide conservation analysis across 100 vertebrate species using PyloP, PhastCons and Multiz Alignment tools, showed less conservation in the pseudoxon from *BRCA1*  $\nabla 12A$  transcript compared with exons 12 and 13 (online supplementary figure 8), suggesting a non-functional role for this region. Furthermore, nucleotide conservation comparison between the pseudoxon included in the aberrant transcript



**Figure 3** Capillary Electrophoresis (CE) assays in carrier and controls. CE of FAM-labelled amplicons from variant carrier, five blood controls and four normal breast tissues. Results from two different RNA extractions in variant carrier are shown. Full-length transcript and the minor alternative isoform  $\Delta 13p$  were detected in all samples, whereas aberrant transcript  $\nabla 12A$  was only detected in variant carrier. FL, full-length.





**Figure 4** *BRCA1* expression in carrier and controls by qPCR. *BRCA1* global expression was measured in two independent qPCR assays, using two probes targeting exons 5–6 and 23–24 junctions, respectively. Both assays show a reduction of *BRCA1* levels in c.4185+4105C>T carrier compared with controls (C1–C10). Grid lines represent *BRCA1* expression levels in controls (mean=1.2). Variant carrier and a pathogenic splicing variant c.302-1G>A show similar *BRCA1* levels. Mean±95% CI are shown.

*BRCA1* ▼12A and the pseudoexon included in the alternative isoform *BRCA1* ▼13A, indicated a higher conservation in the alternative isoform (online supplementary figure 9). The alternative transcript ▼13A has been detected in human control blood tissue and breast tissue,<sup>24,32</sup> whereas ▼12A has not been described previously.

*BRCA1* c.4185+4105C>T variant was identified in co-occurrence with *BRCA1* c.80+909T>C, located in intron 2. A recent study analysing the functional impact of non-coding *BRCA1* variants remarked the importance of this intron because it contains regulatory regions that may affect *BRCA1* promoter activity.<sup>33</sup> In our study, we could not formally exclude that this variant contributes to the allelic imbalance observed in the carrier. However, this variant is located outside the non-coding regulatory sequences from intron 2 (CNS-1 and CNS-2), known to alter transcriptional activity when mutated,<sup>34</sup> supporting a non-functional role for this variant.

A systematic *BRCA1/2* RNA analysis in patients with HBOC in the French population identified the first *BRCA2* deep intronic variant (c.6937+594T>G) causing the inclusion of a pseudoexon by activation of a cryptic donor site.<sup>16</sup> Authors identified this variant in eight additional HBOC families and indicated a pathogenic role for the variant. However, a recent study based on case-control analysis did not observe an association between *BRCA2* c.6937+594T>G and BC risk,<sup>35</sup> and the variant has been classified as benign by the Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) expert panel (<https://enigmaconsortium.org/>). The differences observed in FL levels from *BRCA1* c.4185+4105C>T and *BRCA2* c.6937+594T>G carriers measured by semiquantitative

CE, suggested that *BRCA2* c.6937+594G allele still generates FL transcripts (figure 5D), in agreement with its clinical classification as benign. Although these results need to be confirmed with quantitative approaches and allele-specific assays, semiquantitative measurement of FL transcripts could serve as indicative of variant pathogenicity. The frequency of *BRCA1* c.4185+4105C>T variant in Spanish HBOC cases and control population was also investigated to collect more clues about its pathogenicity, and genotyping analysis did not identify additional carriers in either group. Although variant absence in a larger control group (>1000 individuals) is required to consider evidences of moderate pathogenicity,<sup>36</sup> there is a general assumption that high-penetrance disease-causative variants occur at very low population frequencies. This, combined with the clinical phenotype from variant carrier (BC+OC), supports pathogenicity for *BRCA1* c.4185+4105C>T.

Additionally, the analysis by targeted sequencing of other BC/OC susceptibility genes in *BRCA1* c.4185+4105C>T carrier, including *ATM*, *BRIP1*, *CHEK2*, *EPCAM*, *MLH1*, *MSH2*, *MSH6*, *PALB2*, *PMS2*, *PTEN*, *RAD51C*, *RAD51D*, *STK11* and *TP53*, did not identify any deleterious variant that could explain the family phenotype,<sup>19</sup> supporting also a pathogenic role for the *BRCA1* deep intronic variant.

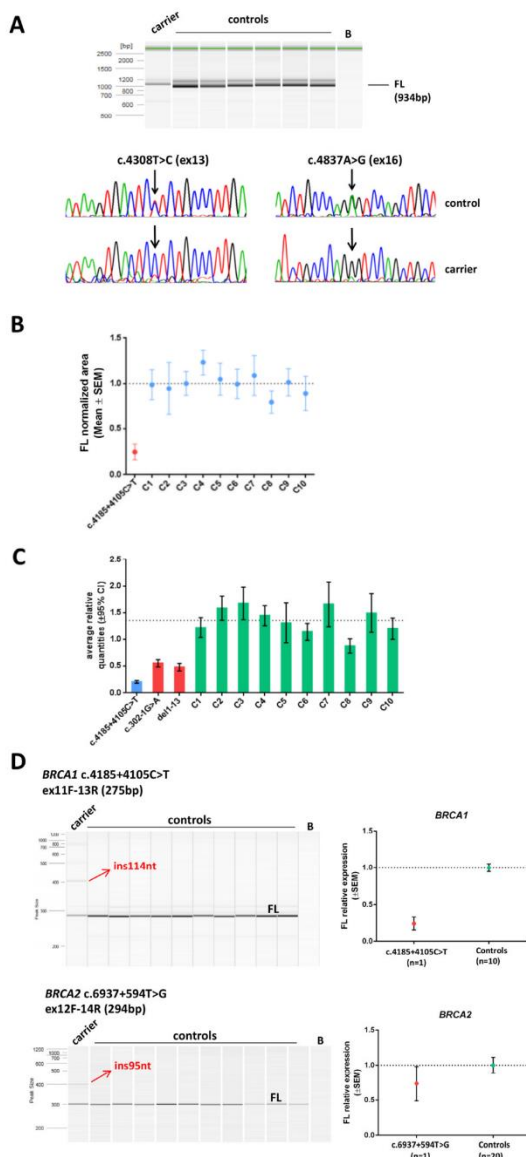
According to the ENIGMA guidelines for *BRCA1/2* variant classification, when the variant allele is assessed by specific transcriptional assays and reveals only the expression of aberrant transcripts, the variant should be classified as pathogenic. In our case, although semiquantitative and quantitative analyses supported a monoallelic contribution to FL transcript expression (figure 5B–C), a residual signal (≈5%–10%) from variant allele was observed at heterozygous sites after the specific amplification of FL transcript (figure 5A). However, whether this is a true contribution from variant allele or a PCR artefact caused by unspecific primer hybridisation cannot be concluded from our results. In any case, the production of functional *BRCA1* protein from the variant allele would be lower than the rescue threshold (≈30%) proposed in de la Hoya work,<sup>37</sup> supporting a likely pathogenic role for our variant. Taking this into account and from an analytical point of view, the variant should be classified as likely pathogenic (Class-4). Further evidence from segregation analysis within family relatives and loss of heterozygosity analysis in tumour samples could help to unequivocally define a pathogenic classification, but unfortunately no samples were available.

In clinical laboratories, deep intronic regions of HBOC genes are generally not screened, and the frequency of pathogenic mutations in these regions could account for a proportion of HBOC cases by either affecting splicing, transcriptional activation or mRNA stability.<sup>38</sup> Pseudoexon insertion events directly related to cancer pathologies caused by the creation of new splicing donor or acceptor sites represent the more frequent occurrence of this type of mutational events.<sup>15</sup> The identification of pseudoexons is particularly interesting for the design of novel therapeutic molecules based on RNA biology.<sup>39</sup> The use of therapies based on antisense oligonucleotides has shown to prevent the inclusion of cryptic exons by blocking the binding of trans-splicing regulatory factors to mutant 5' or 3' splice sites and restore the correct reading frame.<sup>16,40,41</sup>

Here, we report the first deep intronic mutation occurring at *BRCA1* locus that promotes the inclusion of a pseudoexon in mature mRNA and is associated with HBOC risk. Although the frequency of pseudoexon events caused by spliceogenic variants in *BRCA1/2* deep intronic regions appears to be low in our population, our findings highlight the relevance of integrating massive sequencing of whole genomic regions of HBOC genes

Cancer genetics

J Med Genet: first published as 10.1136/jmedgenet-2018-105606 on 24 November 2018. Downloaded from <http://jmg.bmj.com/> on 26 November 2018 by guest. Protected by copyright.



**Figure 5** Specific assessment of *BRCA1* full-length transcript. (A) Specific amplification of FL transcript by RT-PCR shows less amplification in carrier compared with controls. Sanger inspection shows main contribution to FL from only one allele. (B) FL levels measured by semiquantitative CE experiments show a >2-fold reduction in variant carrier. Grid line represents mean of normalised FL levels in control group ( $y=1$ ). Mean $\pm$ SEM are shown. (C) FL transcript measurement by qPCR using a Taqman probe targeting exons 12–13 junction. FL levels show a >2-fold reduction in variant carrier. *BRCA1* pathogenic variants c.302–1G>A and exon1-13 deletion were used as positive controls (in red). Mean $\pm$ 95% CI are shown. (D) RT-PCR evaluation of *BRCA1* and *BRCA2* deep intronic variants and semiquantitative measurement by QIAxcel electrophoresis of FL transcript. Variant carriers were analysed in parallel with non-carrier controls (10 for *BRCA1* and 20 for *BRCA2*). Full-length levels were not drastically reduced in the *BRCA2* carrier, compared with *BRCA1*. Grid line represents mean of normalised FL levels in control group ( $y=1$ ). Mean  $\pm$ SEM are shown. B, blank; FL, full-length transcript.



and RNA analysis to complement genetic diagnosis of familial breast and ovarian cancers.

**Acknowledgements** The authors acknowledge the Cellex Foundation for providing research facilities and equipment. We also thank Leo Judkins for English language editing.

**Contributors** GM, OD and SG-E designed the study. GM conducted RNA experiments and drafted and edited the manuscript. GM, SB, AT, AG-B and VB performed experiments and procedures. AM-F performed bioinformatics analysis. EC, AL-F, NS and JB provided samples and patient data. OD and SG-E supervised experiments. All authors read and reviewed the manuscript.

**Funding** This work was supported by Spanish Instituto de Salud Carlos III (ISCIII) funding, an initiative of the Spanish Ministry of Economy and Innovation partially supported by European Regional Development FEDER Funds: FIS PI12/02585 and PI15/00355 (to O Diez), PI13/01711 and PI16/01218 (to SG-E). SG-E and SB are supported by the Miguel Servet Program (CP16/00034) and Asociación Española Contra el Cáncer (AÉCC) contract, respectively.

**Competing interests** None declared.

**Patient consent** Not required.

**Ethics approval** Clinical Research Ethics Committee (CEIC), Vall d'Hebron Research Institute (VHIR).

**Provenance and peer review** Not commissioned; internally peer reviewed.

## REFERENCES

- Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, Jervis S, van Leeuwen FE, Milne RL, Andrieu N, Goldgar DE, Terry MB, Rookus MA, Easton DF, Antoniou AC, McGuffog L, Evans DG, Barrowdale D, Frost D, Adlard J, Ong KR, Izatt L, Tischkowitz M, Eeles R, Davidson R, Hodgson S, Ellis S, Nogues C, Lasset C, Stoppa-Lyonnet D, Fricke JP, Faivre L, Berthet P, Hoening MJ, van der Kolk LE, Kets CM, Adank MA, John EM, Chung WK, Andrulis IL, Southey M, Daly MB, Buys SS, Osorio A, Engel C, Kast K, Schmutzler RK, Caldes T, Jakubowska A, Simard J, Friedlander ML, McLachlan SA, Machackova E, Foretova L, Tan YY, Singer CF, Olah E, Gerdes AM, Arver B, Olsson H. BRCA1 and BRCA2 Cohort Consortium. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* 2017;317:2402–16.
- Couch FJ, Nathanson KL, Offit K. Two decades after BRCA: setting paradigms in personalized cancer care and prevention. *Science* 2014;343:1466–70.
- Easton DF, Pharoah PD, Antoniou AC, Tischkowitz M, Nathanson KL, Devilee P, Meindl A, Couch FJ, Southey M, Goldgar DE, Evans DG, Chenevix-Trench G, Rahman N, Robson M, Domchek SM, Foulkes WD. Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med* 2015;372:2423–57.
- Valencia OM, Samuel SE, Viscusi RK, Riall TS, Neumayer LA, Aziz H. The role of genetic testing in patients with breast cancer: a review. *JAMA Surg* 2017;152:589–94.
- Erkko H, Xia B, Nikkila J, Schleutker J, Syrjäkoski K, Mannermaa A, Kallioniemi A, Pylkäs K, Karpinen SM, Rappakko K, Miron A, Sheng Q, Li G, Mattila H, Bell DW, Haber DA, Grip M, Reiman M, Jukkola-Vuorinen A, Mustonen A, Kere J, Aaltonen LA, Kosma VM, Kataja V, Soini Y, Drapkin IR, Livingston DM, Winqvist R. A recurrent mutation in PALB2 in Finnish cancer families. *Nature* 2007;446:316–9.
- Meindl A, Hellebrand H, Wieck C, Erven V, Wappenschmidt B, Niederacher D, Freund M, Lichtner P, Hartmann L, Schaaf H, Ramser J, Honisch E, Kubisch C, Wichmann HE, Kast K, Deissler H, Engel C, Müller-Myhok B, Neveling K, Kiechle M, Mathew CG, Schindler D, Schmutzler RK, Haneberg H. Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. *Nat Genet* 2010;42:410–4.
- Loveday C, Turnbull C, Ramsay E, Hughes D, Ruark E, Frankum JR, Bowden G, Kalmyrzaev B, Warren-Perry M, Snape K, Adlard JW, Barwell J, Berg J, Brady AF, Brewer C, Brice G, Chapman C, Cook J, Davidson R, Donaldson A, Douglas F, Greenhalgh L, Henderson A, Izatt L, Kumar A, Lalloo F, Miedzybrodzka Z, Morrison PJ, Paterson J, Porteous M, Rogers MT, Shanley S, Walker L, Eccles D, Evans DG, Renwick A, Seal S, Lord C, Ashworth A, Reis-Filho JS, Antoniou AC, Rahman N. Breast Cancer Susceptibility Collaboration (UK). Germline mutations in RAD51D confer susceptibility to ovarian cancer. *Nat Genet* 2011;43:879–82.
- Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, Jayatilake H, McGuffog L, Hanks S, Evans DG, Eccles D, Easton DF, Stratton MR. Breast Cancer Susceptibility Collaboration (UK). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* 2007;39:165–7.
- Desmond A, Kurian AW, Gabree M, Mills MA, Anderson MJ, Kobayashi Y, Horick N, Yang S, Shannon KM, Tung N, Ford JM, Lincoln SE, Ellisen LW. Clinical actionability of multigene panel testing for hereditary breast and ovarian cancer risk assessment. *JAMA Oncol* 2015;1:943–51.
- Nielsen FC, van Oeveren Hansen T, Sørensen CS. Hereditary breast and ovarian cancer: new genes in confined pathways. *Nat Rev Cancer* 2016;16:599–612.
- Dhir A, Buratti E. Alternative splicing: role of pseudoxons in human disease and potential therapeutic strategies. *Febs J* 2010;277:841–55.
- Pozzoli U, Sironi M. Silencers regulate both constitutive and alternative splicing events in mammals. *Cell Mol Life Sci* 2005;62:1579–604.
- Sironi M, Menozzi G, Riva L, Cagliari R, Comi GP, Bresolin N, Giorda R, Pozzoli U. Silencer elements as possible inhibitors of pseudoxon splicing. *Nucleic Acids Res* 2004;32:1783–91.
- Vaz-Drago R, Custódio N, Carmo-Fonseca M. Deep intronic mutations and human disease. *Hum Genet* 2017;136:1093–111.
- Romano M, Buratti E, Baralle D. Role of pseudoxons and pseudointrons in human cancer. *Int J Cell Biol* 2013;2013:1–16.
- Anczuków O, Buisson M, Léoné M, Coutanson C, Lasset C, Calender A, Sinilnikova OM, Mazoyer S. BRCA2 deep intronic mutation causing activation of a cryptic exon: opening toward a new preventive therapeutic strategy. *Clin Cancer Res* 2012;18:4903–9.
- Gambino G, Tancredi M, Falaschi E, Aretini P, Caligo MA. Characterization of three alternative transcripts of the BRCA1 gene in patients with breast cancer and a family history of breast and/or ovarian cancer who tested negative for pathogenic mutations. *Int J Mol Med* 2015;35:950–6.
- Byers H, Wallis Y, van Veen EM, Lalloo F, Reay K, Smith P, Wallace AJ, Bowers N, Newman WG, Evans DG. Sensitivity of BRCA1/2 testing in high-risk breast/ovarian/male breast cancer families: little contribution of comprehensive RNA/NGS panel testing. *Eur J Hum Genet* 2016;24:1591–7.
- Bonache S, Esteban I, Moles-Fernández A, Tenés A, Duran-Lozano L, Montalban G, Bach V, Carrasco E, Gadea N, López-Fernández A, Torres-Escaius S, Mancuso F, Caratú G, Vivancos A, Tuset N, Balmaña J, Gutiérrez-Enriquez S, Diez O. Multigene panel testing beyond BRCA1/2 in breast/ovarian cancer: Spanish families and clinical actionability of findings. *J Cancer Res Clin Oncol* 2018 [Epub ahead of print 10 Oct 2018].
- Whitley PJ, de la Hoya M, Thomassen M, Becker A, Brandão R, Pedersen IS, Montagna M, Menéndez M, Quiles F, Gutiérrez-Enriquez S, De Leener K, Tenés A, Montalban G, Tserpelis D, Yoshimatsu T, Tirapo C, Raponi M, Caldes T, Blanco A, Santamariña M, Guidugli L, de Garibay GR, Wong M, Tancredi M, Fachel L, Ding YC, Kruse T, Lattimore V, Kwong A, Chan TL, Colombo M, De Vecchi G, Caligo M, Baralle D, Lázaro C, Couch F, Radice P, Southey MC, Neuhausen S, Houdayer C, Fackenthal J, Hansen TV, Vega A, Diez O, Blok R, Claes K, Wappenschmidt B, Walker L, Spurdle AB, Brown MA. ENIGMA consortium. Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. *Clin Chem* 2014;60:341–52.
- de Garibay GR, Acedo A, García-Casado Z, Gutiérrez-Enriquez S, Tosar A, Romero A, Garre P, Lloret G, Thomassen M, Diez O, Pérez-Segura P, Diaz-Rubio E, Velasco EA, Caldes T, de la Hoya M. Capillary electrophoresis analysis of conventional splicing assays: IARC analytical and clinical classification of 31 BRCA2 genetic variants. *Hum Mutat* 2014;35:53–7.
- Gutiérrez-Enriquez S, Coderch Y, Masas M, Balmaña J, Diez O. The variants BRCA1 IVS6-1G>A and BRCA2 IVS15+1G>A lead to aberrant splicing of the transcripts. *Breast Cancer Res Treat* 2009;117:461–5.
- Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paeppe A, Speleman F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002;3:research0034.1.
- Colombo M, Blok MJ, Whitley P, Santamariña M, Gutiérrez-Enriquez S, Romero A, Garre P, Becker A, Smith LD, De Vecchi G, Brandão RD, Tserpelis D, Brown M, Blanco A, Bonache S, Menéndez M, Houdayer C, Foglia C, Fackenthal JD, Baralle D, Wappenschmidt B, Diaz-Rubio E, Caldes T, Walker L, Diez O, Vega A, Spurdle AB, Radice P, De La Hoya M; kConFab Investigators. Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. *Hum Mol Genet* 2014;23:3666–80.
- Lattimore VL, Pearson JF, Currie MJ, Spurdle AB, Robinson BA, Walker LC. kConFab Investigators. Investigation of Experimental Factors That Underlie BRCA1/2 mRNA Isoform Expression Variation: Recommendations for Utilizing Targeted RNA Sequencing to Evaluate Potential Spliceogenic Variants. *Front Oncol* 2018;8:140.
- Ars E, Serra E, de la Luna S, Estivill X, Lázaro C. Cold shock induces the insertion of a cryptic exon in the neurofibromatosis type 1 (NF1) mRNA. *Nucleic Acids Res* 2000;28:1307–12.
- Liu Y, Malaviarachi P, Beggs M, Emanuel PD. PTEN transcript variants caused by illegitimate splicing in 'aged' blood samples and EBV-transformed cell lines. *Hum Genet* 2010;128:609–14.
- Le Hir H, Gatfield D, Izaurralde E, Moore MJ. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *Embo J* 2001;20:4987–97.
- Lykke-Andersen S, Jensen TH. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol* 2015;16:665–77.
- Lindeboom RG, Supek F, Lehner B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat Genet* 2016;48:1112–8.
- Perrin-Vidol L, Sinilnikova OM, Stoppa-Lyonnet D, Lenoir GM, Mazoyer S. The nonsense-mediated mRNA decay pathway triggers degradation of most BRCA1 mRNAs bearing premature termination codons. *Hum Mol Genet* 2002;11:2805–14.

Montalban G, et al. *J Med Genet* 2018;0:1–12. doi:10.1136/jmedgenet-2018-105606

11

## Cancer genetics

- 32 Davy G, Rousselin A, Goardon N, Castéra L, Harter V, Legros A, Muller E, Fouillet R, Brault B, Smirnova AS, Lemoine F, de la Grange P, Guillaud-Bataille M, Caux-Moncouter V, Houdayer C, Bonnet F, Blanc-Fournier C, Gaildrat P, Frebourg T, Martins A, Vaur D, Krieger S. Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer. *Eur J Hum Genet* 2017;25:1147–54.
- 33 Dos Santos ES, Caputo SM, Castera L, Gendrot M, Briaux A, Breault M, Krieger S, Rogan PK, Mucaki EJ, Burke LJ, Bléche I, Houdayer C, Vaur D, Stoppa-Lyonnet D, Brown MA, Lallemand F, Rouleau E. ENIGMA consortium. Assessment of the functional impact of germline BRCA1/2 variants located in non-coding regions in families with breast and/or ovarian cancer predisposition. *Breast Cancer Res Treat* 2018;168:311–25.
- 34 Wardrop SL, Brown MA. kConFab Investigators. Identification of two evolutionarily conserved and functional regulatory elements in intron 2 of the human BRCA1 gene. *Genomics* 2005;86:316–28.
- 35 Dutil J, Godoy L, Rivera-Lugo R, Arroyo N, Albino E, Negrón L, Monteiro AN, Matta JL, Echenique M. No Evidence for the Pathogenicity of the *BRCA2* c.6937 + 594T>G Deep Intronic Variant: A Case-Control Analysis. *Genet Test Mol Biomarkers* 2018;22:85–9.
- 36 Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL. ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–23.
- 37 de la Hoya M, Soukariéh O, López-Perollo I, Vega A, Walker LC, van Ierland Y, Baralle D, Santamariña M, Lattimore V, Wijnen J, Whaley P, Blanco A, Raponi M, Hauke J, Wappenschmidt B, Becker A, Hansen TV, Behar R, Investigators K, Niederacher D, Arnold N, Dworniczak B, Steinemann D, Faust U, Rubinstein W, Hulick PJ, Houdayer C, Caputo SM, Castera L, Pesaran T, Chao E, Brewer C, Southey MC, van Asperen CJ, Singer CF, Sullivan J, Poplawski N, Mai P, Peto J, Johnson N, Burwinkel B, Surowy H, Bojesen SE, Flyger H, Lindblom A, Margolin S, Chang-Claude J, Rudolph A, Radice P, Galastri L, Olson JE, Hallberg E, Giles GG, Milne RL, Andrusis IL, Glendon G, Hall P, Czene K, Blows F, Shah M, Wang Q, Dennis J, Michailidou K, McGuffog L, Bolla MK, Antoniou AC, Easton DF, Couch FJ, Tavtigian S, Vreeswijk MP, Parsons M, Meeks HD, Martins A, Goldgar DE, Spurdle AB. Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. *Hum Mol Genet* 2016;25:2256–68.
- 38 Mucaki EJ, Caminsky NG, Perri AM, Lu R, Laederach A, Halvorsen M, Knoll JH, Rogan PK. A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer. *BMC Med Genomics* 2016;9:19.
- 39 Burnett JC, Rossi JJ. RNA-based therapeutics: current progress and future prospects. *Chem Biol* 2012;19:60–71.
- 40 Hammond SM, Wood MJ. Genetic therapies for RNA mis-splicing diseases. *Trends Genet* 2011;27:196–205.
- 41 Davis RL, Homer VM, George PM, Brennan SO. A deep intronic mutation in FGB creates a consensus exonic splicing enhancer motif that results in afibrinogenemia caused by aberrant mRNA splicing, which can be corrected in vitro with antisense oligonucleotide treatment. *Hum Mutat* 2009;30:221–7.



## Article 7

### Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants.

Cline MS, Babbi G, Bonache S, Cao Y, Casadio R, de la Cruz X, Díez O, Gutiérrez-Enríquez S, Katsonis P, Lai C, Lichtarge O, Martelli PL, Mishne G, Moles-Fernández A, Montalban G, Mooney SD, O'Conner R, Ootes L, Özkan S, Padilla N, Pagel KA, Pejaver V, Radivojac P, Riera C, Savojardo C, Shen Y, Sun Y, Topper S, Parsons MT, Spurdle AB, Goldgar DE; ENIGMA Consortium.

**Hum Mutat.** 2019 Sep;40(9):1546-1556.

doi: 10.1002/humu.23861



## Assessment of blind predictions of the clinical significance of *BRCA1* and *BRCA2* variants

Melissa S. Cline<sup>1</sup>  | Giulia Babbì<sup>2</sup> | Sandra Bonache<sup>3</sup>  | Yue Cao<sup>4</sup> | Rita Casadio<sup>2</sup>  | Xavier de la Cruz<sup>5,6</sup>  | Orland Díez<sup>3,5</sup> | Sara Gutiérrez-Enríquez<sup>3</sup>  | Panagiotis Katsonis<sup>7</sup>  | Carmen Lai<sup>8</sup> | Olivier Lichtarge<sup>7,9,10,11</sup>  | Pier L. Martelli<sup>2</sup>  | Gilad Mishne<sup>8</sup> | Alejandro Moles-Fernández<sup>5</sup>  | Gemma Montalban<sup>5</sup>  | Sean D. Mooney<sup>12</sup>  | Robert O'Conner<sup>13</sup> | Lars Ootes<sup>5</sup> | Selen Özkan<sup>5</sup> | Natalia Padilla<sup>5</sup> | Kymberleigh A. Page<sup>14</sup>  | Vikas Pejaver<sup>12</sup>  | Predrag Radivojac<sup>14,15</sup>  | Casandra Riera<sup>5</sup> | Castrense Savojardo<sup>2</sup>  | Yang Shen<sup>4</sup>  | Yuanfei Sun<sup>4</sup> | Scott Topper<sup>8</sup>  | Michael T. Parsons<sup>16</sup>  | Amanda B. Spurdle<sup>16</sup>  | David E. Goldgar<sup>17</sup>  | The ENIGMA Consortium

<sup>1</sup>Genomics Institute, UC Santa Cruz, Santa Cruz, California

<sup>2</sup>FaBIT Department, Biocomputing Group, University of Bologna, Bologna, Italy

<sup>3</sup>Oncogenetics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain

<sup>4</sup>Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas

<sup>5</sup>Clinical and Translational Bioinformatics Research Unit, Vall d'Hebron Institute of Research (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain

<sup>6</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

<sup>7</sup>Department of Medical and Human Genetics, Baylor College of Medicine, Houston, Texas

<sup>8</sup>Color Genomics, Burlingame, California

<sup>9</sup>Department of Biochemistry & Molecular Biology, Baylor College of Medicine, Houston, Texas

<sup>10</sup>Department of Pharmacology, Baylor College of Medicine, Houston, Texas

<sup>11</sup>Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, Texas

<sup>12</sup>Biomedical Informatics and Medical Education, University of Washington, Seattle, Washington

<sup>13</sup>Sage Bionetworks, Seattle, Washington

<sup>14</sup>Computer Science and Informatics, Indiana University, Bloomington, Indiana

<sup>15</sup>Khoury College of Computer Science, Northeastern University, Boston, Massachusetts

<sup>16</sup>Molecular Cancer Epidemiology, QIMR Berghofer Medical Research Institute, Brisbane, Australia

<sup>17</sup>Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah

### Correspondence

Melissa S. Cline, Genomics Institute, UC Santa Cruz, 1156 High Street, Santa Cruz, CA 95060.  
Email: mcline@ucsc.edu

### Funding information

National Institutes of Health, Grant/Award Numbers: NIH U41 HG007346, NIH R13 HG006650, NIH R35GM124952; National Human Genome Research Institute, Grant/Award Number: U54HG007990; National Institute of General Medical Sciences, Grant/Award Numbers: NIH-GM079656,

### Abstract

Testing for variation in *BRCA1* and *BRCA2* (commonly referred to as *BRCA1/2*), has emerged as a standard clinical practice and is helping countless women better understand and manage their heritable risk of breast and ovarian cancer. Yet the increased rate of *BRCA1/2* testing has led to an increasing number of Variants of Uncertain Significance (VUS), and the rate of VUS discovery currently outpaces the rate of clinical variant interpretation. Computational prediction is a key component of the variant interpretation pipeline. In the CAG15 ENIGMA Challenge, six prediction teams submitted predictions on 326 newly-interpreted variants from the ENIGMA



NIH-GM066099; Newcastle University; NMHRC Senior Research Fellowship, Grant/Award Number: ID 1061778; NIH National Institute on Aging, Grant/Award Number: R01-AG061105

**Consortium.** By evaluating these predictions against the new interpretations, we have gained a number of insights on the state of the art of variant prediction and specific steps to further advance this state of the art.

**KEYWORDS**

BRCA, BRCA1, BRCA2, CAGI, CAGI5, variant interpretation

## 1 | INTRODUCTION

While women have a 12% lifetime risk of breast cancer on average (Howlander et al., 2017), that risk rises to roughly 70% in women with pathogenic variants in *BRCA1/2* (Kuchenbaecker et al., 2017). For ovarian cancer, the average lifetime risk is approximately 1.3% for women in the general population, while the risk is 44% for *BRCA1* carriers and 17% for *BRCA2* carriers. These facts and the decreasing cost of sequencing have led to an upsurge of *BRCA1/2* testing in recent years (Kolor et al., 2017). This increased rate of *BRCA1/2* testing has led to an increasing discovery of new variants, and this rate of variant discovery has outpaced the rate of variant interpretation. Out of 21,695 variants currently listed at BRCA Exchange, the largest public source of *BRCA1/2* variation data (Cline et al., 2018), almost half (9,225) have no clinical interpretation in either ClinVar (Landrum & Kattman, 2018) or the Leiden Open Variation Database (LOVD) (Fokkema et al., 2011). Further, only 7,225 so far have expert interpretations by the ENIGMA Consortium (Spurdle et al., 2012), the ClinGen expert panel for curation of variants in *BRCA1/2*. These numbers underscore the need for both developing robust, high-throughput methods for *BRCA1/2* variant interpretation and gaining a clear understanding of the capabilities of the existing methods.

The CAGI5 ENIGMA Challenge provided an opportunity to evaluate the current state of the art in predicting the clinical significance of *BRCA1/2* variants, leveraging blind prediction. The ENIGMA Consortium provided the CAGI organizers with not-yet-published clinical interpretations for hundreds of *BRCA1/2* variants. Six research teams predicted the clinical significance of these variants, using 14 methods altogether. In this paper, we compare results from these 14 prediction methods, as well as three widely-used reference

methods from the literature, against the expert clinical interpretations, with the goals of evaluating what types of approaches were most effective and identifying areas for further improvement.

## 2 | METHODS

This challenge featured 326 variants that were recently interpreted by the ENIGMA Consortium, as detailed in another paper in this issue (Parsons, Tudini, Li, Goldgar, & Spurdle, 2019). This paper also details the variant classification process used by ENIGMA researchers. Briefly, unclassified variants were prioritized by the ENIGMA Consortium for classification based on the amount of available evidence and/or prior likelihood of pathogenicity based on variant location and predicted effect (Tavtigian, Byrnes, Goldgar, & Thomas, 2008; Vallée et al., 2016). These variants were classified using multifactorial analysis (Goldgar et al., 2008, 2004). While the standard ACMG guidelines evaluate multiple lines of evidence by qualitative rules (Richards et al., 2015), the multifactorial analysis combines evidence types quantitatively in a Bayesian network to estimate the overall likelihood of pathogenicity. Of the 326 variants that were shared with the prediction teams, all were exonic and were either missense variant or in-frame deletions (ENIGMA had provided additional intronic variants that were not shared with the predictors). All of these variants were assessed as Benign, Likely Benign, Likely Pathogenic or Pathogenic at the time of submission to CAGI5. None of the variants had a population frequency of 1% or greater in any reference population studied, and none were predicted truncating variants (Parsons et al., 2019).

Table 1 summarizes the 326 ENIGMA Challenge variants included in the CAGI5 challenge, comprising 318 single-nucleotide variants and

**TABLE 1** Summarizes the variants of the BRCA challenge according to gene, domain region, and clinical significance as interpreted by the ENIGMA Consortium. Note: The Domain column indicates which variants were part of a clinically-significant protein domain, by the criteria of the ENIGMA Consortium (ENIGMA Consortium, 2017). The rows marked "None" indicate variants that are not part of a clinically-significant domain; 318 of the variants were single-nucleotide substitutions, while the remaining eight were in-frame deletions.

Gene	Domain	Benign	Likely benign	Uncertain significance	Likely pathogenic	Pathogenic	Total
BRCA1	BRCT	2	7	1	4	1	15
	None	29	90	2		2	123
	RING	1	4			4	9
BRCA2	DNB	11	27		3	3	44
	None	20	110	2			132
	TR2/RAD5	1	2				3
Total		64	240	5	7	10	326

**TABLE 2** Summary of the BRCA challenge blind prediction teams and methods

Predictor	Method	Brief description
Lichtarge lab	Evolutionary action (EA)	Estimates pathogenicity from evolutionary and phylogenetic information, and substitution likelihood.
	Normalized EA	Normalizes EA predictions with the estimated fraction of BRCA isoforms affected by the mutation.
Mooney-Radivojac 2018	MutPred2	Machine learning predictor with features including estimated changes in structural and functional properties for single-nucleotide variants augmented with an unpublished method for in-frame deletions.
TransBio-Inf	TBI_1	Neural network trained to predict clinical significance from estimated splice site impact and sequence-based features.
	TBI_2	Similar to TBI_1, but with multiple linear regression prediction of functional assay scores.
	TBI_3	Similar to TBI_1, but with no estimated splicing impact.
	TBI_4	Similar to TBI_2, but with no estimated splicing impact.
Bologna bio-computing	SNPs&GO	Machine learning predictor that integrates features extracted from the sequence, sequence profile and GO functional annotation of the input protein.
	Disease Index Matrix	Statistical scale estimating the probability of a variation type to be associated with disease.
AIBI	AIBI	Weakly supervised linear regression using categorized inexact labels from ClinVar, 15 selected features from MutPred2, and designed loss functions.
Color genomics	LEAP 1	Two-class logistic regression using function predictions, splicing predictions, frequency of cancer in individuals and their families, co-occurrence with pathogenic variants, and literature and cancer associations from HGMD.
	LEAP 2	As LEAP 1, but with publicly-available information only.
	LEAP 3	As LEAP 1, but with random forest classifier rather than logistic regression.
	LEAP 4	As LEAP 1, but with three-class logistic regression rather than two-class.

eight in-frame deletions. All *BRCA1* variants reflect the NM\_007294.3 transcript, and all *BRCA2* variants reflect NM\_000059.3. Although the ENIGMA Consortium had prioritized some variants according to the prior likelihood of pathogenicity, most of the variants were either Benign or Likely Benign. This skew is consistent with the actual proportions of the different clinical significance annotations for *BRCA1/2* variants (Cline et al., 2018). During the course of the CAGI experiment, the ENIGMA Consortium reclassified several of these variants with new evidence. Five variants were reclassified from Likely Benign to VUS after ENIGMA received new evidence that conflicted with previous evidence, and these variants were omitted from the assessment. Seventeen variants were reclassified from Likely Benign to Benign, and one was reclassified from Likely Pathogenic to Pathogenic. That is, the majority of the reclassifications increased certainty in the assignment. Since the CAGI5 challenge examined benign and likely benign as one group, and pathogenic and likely pathogenic as another, these reclassifications did not affect the assessment except for removing five variants from the assessment pool.

The fact that 23 variants were reclassified illustrates two things: all interpretations have some uncertainty, the level of which is inherent in the probability of pathogenicity and the class assigned; additional data are helpful to move variants from "likely" categories to outer categories with higher probabilities in favor of a benign or pathogenic classification

## 2.1 | Prediction methods

Six teams submitted blind predictions, using a total of 14 methods. The methods are summarized in Table 2 and summarized below. Most teams have submitted methods papers to the CAGI5 publication set, and we have referenced those methods for further detail. We have also indicated which methods were executed as published; for the others, further details are available in the Supporting Information section.

- The Lichtarge Lab submitted predictions with Evolutionary Action (EA; special issue; Katsonis & Lichtarge, 2019). EA estimates variant pathogenicity through evolutionary information by using an analytic equation. The components of the equation are the likelihood that the reference and alternative amino acids substitute to each other in numerous multiple sequence alignments (MSA), and the sensitivity of the protein function to residue changes calculated by the Evolutionary Trace method (Lichtarge, Bourne, & Cohen, 1996; Mihalek, Res, & Lichtarge, 2004) using MSA and phylogenetic information. The *Normalized EA* predictions had the EA scores adjusted to the fraction of the isoforms affected by the mutation. See Supporting Information for further details.
- The Mooney-Radivojac 2018 team submitted predictions for single-nucleotide variants with *MutPred2* (Pejaver et al., 2017), a machine learning predictor that incorporates contextual features

from protein sequence, conservation, and homology, along with features that encode mutation-induced changes in protein structure and function, as predicted by over 50 built-in machine learning models. The in-frame deletions were scored using *MutPred-indel*, a neural network-based pathogenicity prediction method that incorporates similar features, representative of protein structure, function, and conservation (unpublished).

- The TransBioInf (Vall d'Hebron University Hospital) team submitted four sets of predictions, as detailed in the CAGI5 special issue (Padilla et al., 2019). *TBL\_1* and *TBL\_3* predict clinical significance with neural networks, given features including sequence alignment conservation and biophysical measures of the differences between the reference and alternative amino acids. *TBL\_2* and *TBL\_4* predict functional assay scores with multiple linear regression and a similar set of input features. In addition, *TBL\_1* and *TBL\_2* incorporate estimates of the impact of the mutation on existing splice sites, while *TBL\_3* and *TBL\_4* do not.
- Bologna Biocomputing submitted predictions with SNPs&GO (Calabrese, Capriotti, Fariselli, Martelli, & Casadio, 2009) and the *Disease Index Matrix* (Casadio, Vassura, Tiwari, Fariselli, & Luigi Martelli, 2011), both executed as published. SNPs&GO is a machine learning predictor that estimates pathogenicity from information on the Gene Ontology terms associated with the protein, as well as features describing amino acid conservation, the local sequence environment, and the evolutionary likelihood of the reference and alternative amino acids. The Disease Index Matrix ( $P_d$ ) is a scale that associates each variant type (i.e., pair of wild type and variant residues) with the probability of being related to the disease. The scale has been estimated with a statistical analysis of a large data set of disease-related and neutral variations retrieved from UniProtKB and dbSNP databases.
- *ABI* directly predicted the probability of pathogenicity with weakly supervised linear regression, as detailed in the CAGI5 special issue (Cao et al., 2019) as the exact probabilities are not available for supervised machine learning. They used variants annotated with the class of pathogenicity in ClinVar, selected from MutPred2 15 features about molecular impacts upon variation, and designed parabola-shaped loss functions that penalize the predicted probability of pathogenicity according to its supposed class.
- Color Genomics submitted four sets of predictions with *LEAP* (Lai et al., 2018), a machine learning framework that predicts variant pathogenicity according to features including:
  - population frequencies from gnomAD;
  - function prediction from SnpEFF (Cingolani et al., 2012), SIFT (Ng & Henikoff, 2003), PolyPhen-2 (Adzhubei, Jordan, & Sunyaev, 2013) and MutationTaster2 (Schwarz, Cooper, Schuelke, & Seelow, 2014);
  - splice impact estimation from Alamut (Interactive Biosoftware, Rouen, France) and Skippy (Woolfe, Mullikin, & Elnitski, 2010);
  - indications of publications mentioning the variant and cancer associations from the subscription version of HGMD, indicating whether or not the variant is included in HGMD, whether or not it is associated with one or more articles curated by HGMD, and

whether HGMD associates the variant with cancer (Stenson et al., 2017); and

- aggregate information from individuals who have undergone genetic testing. This information consists of frequencies of cancer in the individuals tested, and within the individuals' families (covering many cancer types, not simply breast and ovarian cancer), and co-occurrence of pathogenic variants in the same individual.

*LEAP 1* estimates pathogenicity with a two-class regularized logistic regression model, *LEAP 2* serves as a control and is equivalent to *LEAP 1* except for omitting any inputs that are not publicly-available (including data from HGMD). *LEAP 3* uses a random forest rather than regularized logistic regression. *LEAP 4* uses a three-class regularized logistic regression model (*Benign*, *VUS*, *Pathogenic*) rather than a two-class model (*Benign*, *Pathogenic*). While the *LEAP* method is not publicly-available at this time, Color Genomics anticipates making the predictions by *LEAP* publicly available during 2019 (Lai et al., 2018). The authors of *LEAP* are preparing a publication on their method, which will be added to the CAGI5 collection upon publication.

For reference, we analyzed the variants with the popular methods SIFT Version 5.2.2 (Ng & Henikoff, 2003), PolyPhen-2 Version 2.2 (Adzhubei et al., 2013) and REVEL (Version December 2018; Ioannidis et al., 2016). SIFT applies substitution matrices to estimate the likelihood that a variant is pathogenic. PolyPhen-2 scores variants based on substitution matrices, evolutionary information, differences in the biophysical properties of the reference and alternative amino acids, functional residue and domain annotations, and predicted secondary structure. REVEL is a meta-predictor that estimates variant pathogenicity on the basis of several individual predictors, including SIFT, PolyPhen-2, MutPred, and MutationTaster. We generated scores for these methods via the Ensembl Variant Effect Predictor (McLaren et al., 2016) in December 2018.

The ENIGMA Consortium incorporates "priors" of variant pathogenicity prediction as part of its variant interpretation process (Parsons et al., 2019). This pathogenicity prediction incorporates splice site impact, protein conservation, and expert knowledge, as detailed in a recent publication (Tavtigian et al., 2008; Vallée et al., 2016). Briefly, the impact of a variant on known splice sites and the likelihood of a variant introducing an ectopic splice site, are assessed by MaxEntScan (Yeo & Burge, 2004). To estimate the impact of missense variants, the variants are binned according to two factors: whether or not the variant is inside a clinically-important protein domain (Tavtigian et al., 2008); and ranges of substitution scores from Align-GVGD (<http://agvgd.iarc.fr/>), which estimates substitution likelihoods from alignments of orthologous protein sequences. The variants within each bin are assigned a probability of pathogenicity which was estimated from previous analyses of disease-causing variation (Easton et al., 2007). This approach share features with some of the predictors in this experiment; the TransBioInf and *LEAP* methods use similar splicing information, and many methods use genomic conservation (which is related to the protein conservation implicit in Align-GVGD).



However, these similarities are minor. Moreover, ENIGMA variant interpretation is based on many lines of evidence beyond pathogenicity prediction, namely several forms of clinical observation. None of the predictors use such lines information except for LEAP, and LEAP uses similar observations that were collected independently. The LEAP predictors reflect individuals who have been tested by Color Genomics (which was founded in 2013), while ENIGMA priors reflect individuals who had been tested before the mid-2000s. These sets of individuals could overlap, but this overlap is likely to be minor given the gap in time. In summary, the ENIGMA priors include similar information to some of the prediction methods, but the risk of bias from this similarity is minimal.

## 2.2 | Assessment methods

The 326 variants that were submitted for prediction analysis in the CAGI5 ENIGMA Challenge had all been interpreted by the ENIGMA Consortium, the ClinGen-approved expert panel for *BRCA1/2* variant interpretation. Table S1 lists these variants along with the ENIGMA interpretation, and the predicted probability of pathogenicity and prediction standard deviation by all 14 prediction methods and the three reference methods (see Table S1).

Of the 326 *BRCA1/2* variants that were shared with the CAGI5 prediction teams, ENIGMA interpreted 64 as Benign (Class 1), 240 as Likely Benign (Class 2), 5 as VUS (Class 3), 7 as Likely Pathogenic (Class 4), and 10 as Pathogenic (Class 5). As described earlier, these were the final interpretations by ENIGMA; the consortium had interpreted these variants when they were submitted to CAGI, and subsequently revised the interpretation of several variants according to new evidence that became available during the CAGI experiment. By IARC classification criteria (Plon et al., 2008), Benign variants include those with posterior estimates of pathogenicity of less than 0.001 in the multifactorial estimation, while the threshold posteriors for Likely Benign variants is 0.049; the threshold for Likely Pathogenic variants is 0.95 while that for Pathogenic variants is 0.99. Two aspects of this statistical modeling are the evidence that the variant is damaging or increases disease risk and the strength of the evidence. For example, suppose two variants have a similar impact on the protein function, but one is observed in a very few individuals while the second is observed much more frequently. The first variant might be classified, as Likely Pathogenic while the second is Pathogenic, because the smaller amount of evidence might not reach the threshold for Pathogenic classification. The amount of evidence on a variant is not relevant to pathogenicity prediction, while the predicted impact on function is. In our assessment, we grouped the Benign and Likely Benign variants together (assigning them a target probability of 0.025), grouped the Pathogenic and Likely Pathogenic variants together (assigning them a target probability of 0.975), omitted the VUS (each of which had been classified previously as Likely Benign but were reclassified based on additional evidence), and evaluated the prediction methods on their accuracy at predicting these target probabilities.

Most predictors submitted numerical predictions of pathogenicity ranging from 0.0 (predicted benign) to 1.0 (predicted pathogenic). One team submitted class labels (Class 1–5), which we

translated to random probabilities selected from within ENIGMA's posterior probability ranges (ENIGMA Consortium, 2017). Most predictors submitted standard deviations to accompany their estimated probabilities, and some submitted comments on their predictions.

We approached the assessment by computing several different summary statistics, as each can offer distinct insights. These included both threshold-dependent and threshold-independent metrics. The threshold-dependent metrics included:

- **Precision:** the ratio of true positives to true and false positives, or variants accurately predicted as pathogenic as related to all pathogenic predictions;
- **Recall:** the ratio of true positives to true positives and false negatives, or variants accurately predicted as pathogenic as related to all pathogenic variants (also known as *sensitivity*);
- **Accuracy:** the ratio of true positive and true negative predictions to all true and false predictions, or the fraction of variants accurately classified as benign or pathogenic relative to the number of variants;
- **F1:** the harmonic mean of precision and recall.

A contrast between Accuracy and F1 is that Accuracy reflects in part the number of True Negatives, benign variants predicted as such, while F1 does not. In cases such as this, with a large skew between the positive and negative sets, F1 is generally considered more meaningful. Accordingly, we leveraged F1 for threshold selection, and empirically selected one threshold for each predictor by sampling candidate thresholds across the prediction range and selecting the threshold that yielded the largest F1. We applied these thresholds in measuring Precision, Recall, and Accuracy. Table S2 lists these thresholds along with these performance metrics.

We applied the following threshold-independent methods:

- **ROC AUC:** area under the ROC curve, which relates sensitivity (recall) to specificity (which in this context represents the fraction of benign variants correctly classified as benign). ROC AUC is a widely-used classification metric, which lends itself easily to probabilistic interpretation.
- **P/R AUC:** area under the Precision-Recall curve. This metric is similar to ROC AUC but is more effective for datasets such as this one with a large skew between positives and negatives.
- **RMSD:** root-mean-squared deviation describes the numerical distance between the prediction and its target value.
- **Pearson correlation:** this is a standard parametric correlation metric. Like RMSD, it tends to reward predictions that are numerically close to the target value.

We also evaluated Kendall correlation but found that for these data, it was redundant with ROC AUC (data not shown).

To evaluate significance in predictor performance, with confidence intervals, we performed 10,000 iterations of bootstrapping. For predictors that supplied standard deviations (as most did), in

each bootstrapping iteration, we added a small amount of noise, sampled at random from a normal distribution with a mean of zero and the standard deviation supplied by the predictor, and measured all summary statistics on these data. We computed the standard deviation of these bootstrapped summary statistics. We considered the difference between two prediction methods to be significant if their summary statistics differed by more than one standard deviation. When a prediction was accompanied by a large standard deviation (which communicates a high degree of uncertainty), the bootstrapping communicated wide confidence intervals around the prediction metrics; small or no standard deviations translated to greater certainty around the summary statistics. Note that the bootstrapping was used only to estimate the error bars around the summary statistics, and the summary statistics themselves were computed on the actual prediction values.

One last component of the assessment was to identify a subset of variants that had proved to be challenging in general and analyze the commonalities of these variants. To identify these difficult variants, we computed the median predicted probability from all prediction methods and selected the pathogenic variants with lower median predictions and the benign variants with higher median predictions.

All of the software used in this assessment is publicly available at <https://github.com/melissacline/CAG15-BRCA-Assessment>. Table S2 provides all of the assessment statistics for each method assessed (see Table S2).

### 3 | RESULTS

We evaluated results from 14 blind prediction methods and three reference methods. With few exceptions, the blind prediction methods reported values for the same variants, so their results can be compared directly. The three reference methods did not report values for many of these variants, and due to the number of missing values, their results should be viewed as only rough approximations of their performance. Figure 1 shows a dendrogram of the predictions and indicates the missing values. As shown, there were

very few missing values. Almost all predictors submitted predictions on the same variants; the results were not confounded by missing values. The dendrogram shows that, unsurprisingly, different methods by the same teams tend to cluster together.

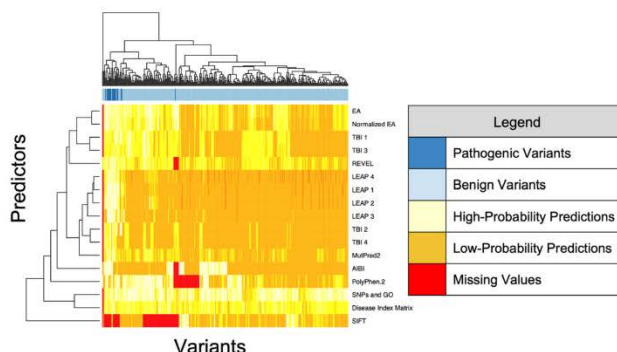
Figure S1 shows the distributions of probabilities estimated by each method and contrasts the probabilities for Benign and Likely Benign variants to those for Pathogenic and Likely Pathogenic variants (Figure S1). Figure S2 breaks this down further by separately showing predictions for the Benign, Likely Benign, Likely Pathogenic, and Pathogenic classes (Classes 1, 2, 4, and 5 respectively; see Figure S2). This figure illustrates that the predictions were not necessarily stronger for Pathogenic versus Likely Pathogenic variants, nor for Benign versus Likely Benign variants. This supports the assertion that the difference between Benign and Likely Benign, and between Pathogenic and Likely Pathogenic, reflects the strength of the clinical evidence rather than the expected functional impact of the variant, and is not relevant to this assessment.

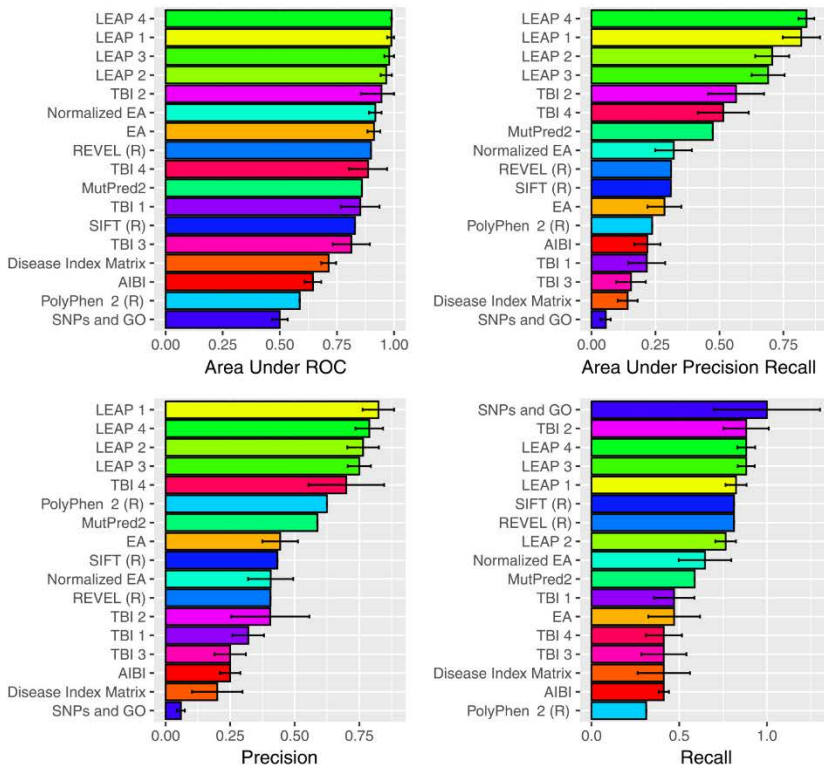
Figure 2 summarizes the performance of the methods in terms of four metrics chosen as most illustrative: ROC AUC, P/R AUC, Precision and Recall. Table S2 lists the complete set of performance metrics (see Table S2). While each metric has nuances, the rank order was largely consistent between the metrics. As a reflection of the overall performance accuracy, the strongest F1 accuracy was achieved by LEAP 4 at 0.83. In other words, on this particular dataset, the state of the art methods were correct in roughly four out of five cases, which illustrates that variant prediction remains a hard problem.

Overall, most methods fared better at predicting pathogenic variants as pathogenic than predicting benign variants as benign, as seen by comparing the Precision and Recall graphs in Figure 2. The LEAP methods were an exception, with strong precision as well as recall at the best empirically-selected threshold.

Interpretability was a design objective for the LEAP methods. LEAP 1 and LEAP 2, which are both regularized logistic regression methods, listed the input features which were most significant for each prediction. These include, scores from LRT, MutationTaster, SIFT, PolyPhen 2, and phastCons 100way vertebrate conservation.

**FIGURE 1** Dendrogram illustrating the predictions on all variants by all prediction methods





**FIGURE 2** Shown is the performance of the 14 blind prediction methods and three reference methods (denoted with R), for four selected performance metrics. The bar lengths and the error bars reflect the mean performance and standard deviation observed in random benchmarks, where each estimated probability was permuted according to standard deviation supplied by the predictor. No benchmarking was performed on methods for which the predictor supplied no standard deviation, or on the reference methods

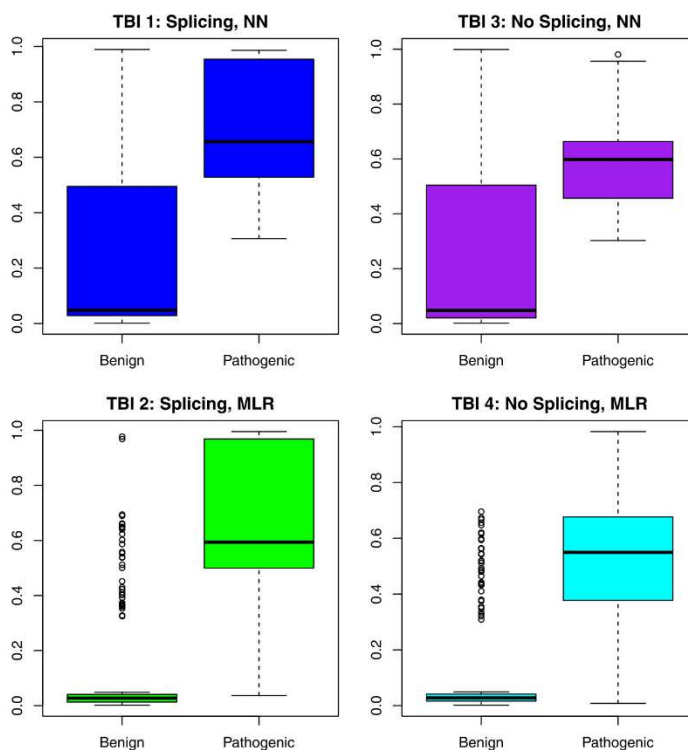
These features are also inputs to the reference method REVEL, which did not score quite as well on these variants. One possible explanation for LEAP's performance advantage concerns differences in what the methods were trained to do. REVEL was trained to predict variants that are pathogenic in disease in general, and there may have been some variation in the methods that were used to interpret the variants in its training set. LEAP was trained to predict pathogenicity in cancer specifically and was trained on variants that were interpreted consistently, according to the ACMG Guidelines, by board-certified medical geneticists. In general, the methods that had been trained to identify disease variants, in general, did not fare as well on this challenge. This includes PolyPhen-2 and Disease Index Matrix. Arguably, LEAP addressed an easier problem by limiting its scope to cancer.

A second explanation is that features shared by LEAP and REVEL were necessary but not sufficient, and LEAP's performance can be attributed to additional features. Important features that were

distinct to LEAP included patient-derived information. Co-occurrence with known pathogenic variants was valuable in the accurate prediction of roughly one-third of the benign variants. Information on individuals who carry the variant and the frequency of cancer in these individuals and their families was a strong predictor for a few difficult pathogenic variants (Lai et al., 2018). Since patient-level information informs clinical variant interpretation, including the ENIGMA variant interpretations, it comes as no surprise that it is also valuable for improving variant pathogenicity prediction above that based on bioinformatic information alone.

Another form of information that benefitted LEAP was population frequencies from gnomAD. Higher minor allele frequencies within a distinct out-bred population is a characteristic of benign variants. While ENIGMA omitted variants with sufficiently high population frequencies to meet the ACMG Guidelines as benign, higher population frequencies still suggested benign variants. Since population frequency repositories are publicly-available, and growing





**FIGURE 3** Of the four methods by the TransBioInf team, two (left) used predicted splicing information while two (right) did not. Further, the methods used two different learning frameworks and objective functions: neural network prediction of clinical significance (top), and multiple linear regression of functional assay scores (bottom). These boxplots show that in both architectures, including the splicing information improved prediction accuracy

in size and quality, their demonstrated value to LEAP's performance suggests that they may be valuable to other methods as well.

While the variants interpreted by ENIGMA had no publicly-available interpretations at the time of the challenge, many of them proved to be in HGMD, where additional information is available to paid subscribers. The information in HGMD includes assigned categories (particularly, the "Disease-causing mutation" or DM category) and the presence of the variant in the literature. This proved to be a strong source of information for the methods that included data from HGMD in their input set (LEAP 1, LEAP 3, LEAP 4, and REVEL). In fact, for pathogenic variants, the data from HGMD was among the more important inputs to LEAP 1: the presence of the variant in the literature was instrumental in accurate prediction 15 of the 17 pathogenic variants, and the HGMD-assigned category of DM was an instrumental inaccurate prediction of 12 of these variants. There were two pathogenic variants for which LEAP 1 did not indicate HGMD features as key inputs (*BRCA2* c.7819A>C and

*BRCA2* c.8975\_9100del), and LEAP 1 mispredicted on these two variants. Since HGMD features papers on pathogenic or damaging variants, it makes sense that the mere fact that the mere mention a variant in HGMD is a strong predictor of pathogenicity. The rules of the CAGI experiment stipulate that each prediction team can use whatever information they have available, including private information. While the merits of subscription databases can be argued elsewhere, the scientific lesson is that the added information in these databases appears to be valuable. The lesson for the larger scientific community is that there exists additional data that could in theory be shared publicly (should its owner so decide), and sharing these data would advance the science of variant interpretation.

While most pathogenic variants were accurately predicted as pathogenic, there were a few that received lower predictions on average, and a review of these variants was instructive. A number of mispredicted pathogenic variants were proximal to splice sites.

Examples include *BRCA1:c.4675G>A* and *BRCA1:c.4484G>C* (which are adjacent to splice sites) and *BRCA1:c.5144G>A* and *BRCA2:c.7819A>C* (which are in close proximity to splice sites). The TransBioInf team further illustrated the impact of splicing information in the construction of their four methods, which themselves were a controlled experiment. Two included the predicted splicing impact, while two did not. In addition, the methods used two different learning frameworks: neural network prediction of clinical significance and multiple linear regression estimation of functional assay scores. As shown in Figure 3, including the predicted splicing impact improved prediction accuracy in both learning frameworks. Splicing-related information was also valuable to LEAP, such as the distance to the nearest splice site and exon length. Exon length is an interesting quantity that other researchers have also found to be valuable in such prediction; there appears to be valuable information encoded in exon length, beyond whether or not this length is of modulo three (Jagadeesh et al., 2019), to indicate information such as if the loss of the exon would introduce a frameshift. This further suggests that future method development may benefit from a greater emphasis on splicing information and might look beyond the splice sites themselves to additional regulatory features.

We reviewed a number of variants that were annotated as Benign or Likely Benign by ENIGMA, yet predicted as pathogenic by most predictors. Many of these variants were in conserved, clinically-important domains, yet in solvent-exposed, loop regions within these domains. Examples include *BRCA1:c.5312C>G* and *BRCA2:c.8764A>G*. Many predictors in this experiment use protein structure information, either directly or indirectly through predictors they incorporate, yet perhaps the protein structure information is being masked by the strong conservation signal.

#### 4 | CONCLUSION

In the CAGI5 ENIGMA Challenge, six teams submitted blind predictions with 14 methods on a set of 326 *BRCA1/2* exonic variants (mostly missense variants plus a few in-frame indels). These variants all had a minor allele frequency of less than 1%, and had recently been assessed for clinical significance by the ENIGMA Consortium using multifactorial likelihood analysis methodology. While this set of variants was skewed to the Benign and Likely Benign category, this skew is representative of the *BRCA1/2* variants encountered in clinical practice. Yet it should be noted that given the small size of the data, and particularly the small number of pathogenic variants, this assessment should not be over-interpreted. For example, a predictor who predicts 100% sensitivity on a set of 17 pathogenic variants can still have a Wilson confidence interval as low as 82% (Wilson, 1927). Predicting the clinical impact of variants remains a hard problem. This experiment showed that the best methods achieved an F1 accuracy of just over 0.8, implying that variant prediction might be the wrong one out of every five variants, at best. Variant prediction is not yet ready for clinical application in the absence of other data. With that said, this assessment may provide useful qualitative information.

A confounding factor in this experiment was that most of the pathogenic variants were in the subscription version HGMD and were predicted as pathogenic ("Disease-causing Mutations") by HGMD. This information was only available to HGMD subscribers. In theory, paid HGMD subscriptions are available to anyone; in practice, the subscription fees are beyond the means of many academic labs and smaller institutions. This information was available to the LEAP methods (minus LEAP 2, which used publicly-available information only), and appears to have been instrumental in many correct pathogenic predictions by LEAP 1. Recognizing this potential bias, the results of this experiment should best be used as a motivation for methods development rather than a guide for direct clinical interpretation. Yet these results present a powerful lesson for the scientific community that there is private data that shows value invariant prediction. By extension, efforts to make such data more broadly-available are likely to advance the science of variant prediction.

Nonetheless, we learned several valuable lessons in this experiment, including the value of population frequency data. The LEAP methods leveraged population frequencies from gnomAD, which were instrumental in many accurate predictions. This is an information source that was not used by most of the variant prediction methods, yet is available now and stands to improve as more population-scale sequencing studies become available (Lek et al., 2016).

While the pathogenic variants were few in number, they presented a clear story on the importance of splicing information. In the LEAP methods, splicing information as instrumental at predicting both pathogenic variants as pathogenic, and benign variants as benign. The results of the TransBioInf team demonstrated that splicing information improved prediction in two distinct architectures. When we assessed the pathogenic variants that were not predicted as pathogenic by many methods, many of them were proximal to splice sites. Our observations suggest that predictive methods should routinely include prediction of splicing impact. As our knowledge of splicing regulation improves, this improved knowledge may translate to further improvements invariant prediction methods.

The LEAP team from Color Genomics was able to draw upon their large database of patient-level clinical results, as well as a subscription to HGMD. They observed that the cancer frequencies of individuals and their families were valuable input for some variants that would otherwise be difficult to classify. We observed that variant co-occurrence information was an important factor in their correctly predicting many of the benign variants as benign. It should come as no surprise that the types of information that are valuable for variant interpretation are also informative for predicting variant pathogenicity. This offers an optimistic note on how data sharing might improve the practice of variant prediction. While individual-level (or case-level) data is difficult to share for privacy reasons, case-derived information such as family history summary statistics and variant co-occurrences can be shared in ways that do not compromise patient privacy. As progress is made to share such information, those who benefit will include the developers and users of variant prediction methods.

## ACKNOWLEDGMENTS

We deeply thank the organizers of the CAGI5 experiment for their hard work and dedication. Blind prediction has had a profound benefit on the field of bioinformatics, yet it relies on the persistence and determination of the organizers, as well as many individuals who keep the infrastructure complete and intact. The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650. M. S. C. is supported by grant U54HG007990 from the National Human Genome Research Institute (genome.gov). O. L. and P. K. were supported by the National Institute of General Medical Sciences (NIGMS) NIH-GM079656 and the NIH-GM066099 grants. MTP is supported by a grant from Newcastle University, UK. A. B. S is supported by NMHRC Senior Research Fellowship ID 1061778. YC, YS, and YS are supported by NIH R35GM124952. K. P. and O. L. are submitted by NIH National Institute on Aging (NIA) R01-AG061105.

## ORCID

Melissa S. Cline  <http://orcid.org/0000-0002-0148-1956>  
 Sandra Bonache  <http://orcid.org/0000-0002-7395-8143>  
 Rita Casadio  <http://orcid.org/0000-0002-7462-7039>  
 Xavier de la Cruz  <http://orcid.org/0000-0002-9738-8472>  
 Sara Gutiérrez-Enríquez  <http://orcid.org/0000-0002-1711-6101>  
 Panagiotis Katsonis  <http://orcid.org/0000-0002-7172-1644>  
 Olivier Lichtarge  <http://orcid.org/0000-0003-4057-7122>  
 Pier L. Martelli  <http://orcid.org/0000-0002-0274-5669>  
 Alejandro Moles-Fernández  <http://orcid.org/0000-0003-0252-6084>  
 Gemma Montalbán  <http://orcid.org/0000-0002-6958-4759>  
 Sean D. Mooney  <http://orcid.org/0000-0003-2654-0833>  
 Kymberleigh A. Pagel  <http://orcid.org/0000-0001-8544-9250>  
 Vikas Pejaver  <http://orcid.org/0000-0002-1943-0284>  
 Predrag Radivojac  <http://orcid.org/0000-0002-6769-0793>  
 Castrese Savojardo  <http://orcid.org/0000-0002-7359-0633>  
 Yang Shen  <http://orcid.org/0000-0002-1703-7796>  
 Scott Topper  <http://orcid.org/0000-0003-1612-7201>  
 Michael T. Parsons  <http://orcid.org/0000-0003-3242-8477>  
 Amanda B. Spurdle  <http://orcid.org/0000-0003-1337-7897>  
 David E. Goldgar  <http://orcid.org/0000-0003-0697-9347>

## REFERENCES

- Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. In Haines, J. L. (Ed.), *Current protocols in human genetics*. John Wiley & Sons Inc.
- Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., & Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation*, 30(8), 1237–1244.
- Cao, Y., Sun, Y., Karimi, M., Chen, H., Moronfoye, O., & Shen, Y. (2019). Predicting pathogenicity of missense variants with weakly supervised regression. *Human Mutation*.
- Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., & Luigi Martelli, P. (2011). Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome. *Human Mutation*, 32(10), 1161–1170.
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2), 80–92.
- Cline, M. S., Liao, R. G., Parsons, M. T., Paten, B., Alquaddoomi, F., Antoniou, A., & Spurdle, A. B. (2018). BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. *PLoS Genetics*, 14(12):e1007752.
- Easton, D. F., Deffenbaugh, A. M., Pruss, D., Frye, C., Wenstrup, R. J., Allen-Brady, K., ... Goldgar, D. E. (2007). A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *American Journal of Human Genetics*, 81(5), 873–883.
- ENIGMA Consortium. (2017, June 29). ENIGMA BRCA1/2 Gene Variant Classification Criteria, Version 2.5.1. Retrieved from [https://enigmaconsortium.org/wp-content/uploads/2018/10/ENIGMA\\_Rules\\_2017-06-29-v2.5.1.pdf](https://enigmaconsortium.org/wp-content/uploads/2018/10/ENIGMA_Rules_2017-06-29-v2.5.1.pdf)
- Fokkema, I. F. A. C., Taschner, P. E. M., Schaafsma, G. C. P., Celli, J., Laros, J. F. J., & den Dunnen, J. T. (2011). LOVD v2.0: The next generation in gene variant databases. *Human Mutation*, 32(5), 557–563.
- Goldgar, D. E., Easton, D. F., Byrnes, G. B., Spurdle, A. B., Iversen, E. S., & Greenblatt, M. S., IARC Unclassified Genetic Variants Working Group. (2008). Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. *Human Mutation*, 29(11), 1265–1272.
- Goldgar, D. E., Easton, D. F., Deffenbaugh, A. M., Monteiro, A. N. A., Tavtigian, S. V., & Couch, F. J. (2004). Integrated evaluation of DNA sequence variants of unknown clinical significance: Application to BRCA1 and BRCA2. *American Journal of Human Genetics*, 75(4), 535–544.
- Howlader, N., Noone, A. M., Krapcho, M., Miller, D., Bishop, K., Kosary, C. L., & Cronin, K. A. (Eds.). (2017, April). SEER Cancer Statistics Review, 1975–2014 (Version based on November 2016 SEER data submission, posted to the SEER web site). SEER, Bethesda, MD: National Cancer Institute. Retrieved from [https://seer.cancer.gov/csr/1975\\_2014/](https://seer.cancer.gov/csr/1975_2014/)
- Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., & Sieh, W. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *American Journal of Human Genetics*, 99(4), 877–885.
- Jagadeesh, K. A., Paggi, J. M., Ye, J. S., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2019). S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nature Genetics*, 51, 755–763. <https://doi.org/10.1038/s41588-019-0348-4>.
- Katsonis, P., & Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness. *Genome Research*, 24(12), 2050–2058.
- Katsonis, P., & Lichtarge, O. (2019). CAGI5: Objective Performance Assessments of Predictions Based on the Evolutionary Action Equation. *Human Mutation*. PRODUCTION: PUBLISHES IN SAME SPECIAL ISSUE.
- Kolor, K., Chen, Z., Grosse, S. D., Rodriguez, J. L., Green, R. F., Dotson, W. D., & Khoury, M. J. (2017). BRCA genetic testing and receipt of preventive interventions among women aged 18–64 years with employer-sponsored health insurance in nonmetropolitan and metropolitan areas - United States, 2009–2014. *Morbidity and Mortality Weekly Report. Surveillance Summaries*, 66(15), 1–11.
- Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K. A., Mooij, T. M., Roos-Blom, M.-J., & Olsson, H. (2017). Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA: The Journal of the American Medical Association*, 317(23), 2402–2416.
- Lai, C., O'Connor, R., Topper, S. J., Ji, J., Stedden, W., Homburger, J., & Mishne, G. (2018, February). *Using Machine Learning to Support Variant Interpretation in a Clinical Setting*. Presented at the Advances in Genome Biology and Technology (AGBT). Retrieved from [https://static.getcolor.com/pdfs/research/Color\\_AGBT\\_PH\\_Poster\\_2018.pdf](https://static.getcolor.com/pdfs/research/Color_AGBT_PH_Poster_2018.pdf)

- Landrum, M. J., & Kattman, B. L. (2018). ClinVar at five years: Delivering on the promise. *Human Mutation*, 39(11), 1623–1630.
- Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., & Fennell, T., Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291.
- Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996). Evolutionarily conserved Galphabeta $\gamma$  binding surfaces support a model of the G protein-receptor complex. *Proceedings of the National Academy of Sciences of the United States of America*, 93(15), 7507–7511.
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., & Cunningham, F. (2016). The ensembl variant effect predictor. *Genome Biology*, 17(1), 122.
- Mihalek, I., Res, I., & Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. *Journal of Molecular Biology*, 336(5), 1265–1282.
- Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13), 3812–3814.
- Padilla, N., Moles-Fernández, A., Riera, C., Montalban, G., Özkan, S., Ootes, L., ... de la Cruz, X. (2019). BRCA1- and BRCA2-specific in silico tools for variant interpretation in the CAGI 5 ENIGMA challenge. *Human Mutation*. PRODUCTION: PUBLISHES IN SAME SPECIAL ISSUE.
- Parsons, M. T., Tudini, E., Li, H., Goldgar, D. E., & Spurdle, A. B. (2019). Largescale multifactorial likelihood analysis of BRCA1 and BRCA2 variants within ENIGMA: A resource to inform qualitative classification criteria. *Human Mutation*. PRODUCTION: PUBLISHES IN SAME SPECIAL ISSUE.
- Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H. -J., & Radivojac, P. (2017). MutPred2: inferring the molecular and phenotypic impact of amino acid variants. *bioRxiv*, <https://doi.org/10.1101/134981>
- Plon, S. E., Eccles, D. M., Easton, D., Foulkes, W. D., Genuardi, M., & Greenblatt, M. S., IARC Unclassified Genetic Variants Working Group. (2008). Sequence variant classification and reporting: Recommendations for improving the interpretation of cancer susceptibility genetic test results. *Human Mutation*, 29(11), 1282–1291.
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., & Gastier-Foster, J., ACMG Laboratory Quality Assurance Committee. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 17(5), 405–424.
- Schwarz, J. M., Cooper, D. N., Schuelke, M., & Seelow, D. (2014). MutationTaster2: Mutation prediction for the deep-sequencing age. *Nature Methods*, 11(4), 361–362.
- Spurdle, A. B., Healey, S., Devereau, A., Hogervorst, F. B. L., Monteiro, A. N. A., Nathanson, K. L., & ENIGMA (2012). ENIGMA—evidence-based network for the interpretation of germline mutant alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Human Mutation*, 33(1), 2–7.
- Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., & Cooper, D. N. (2017). The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, 136(6), 665–677.
- Tavtigian, S. V., Byrnes, G. B., Goldgar, D. E., & Thomas, A. (2008). Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. *Human Mutation*, 29(11), 1342–1354.
- Vallée, M. P., Di Sera, T. L., Nix, D. A., Paquette, A. M., Parsons, M. T., Bell, R., & Tavtigian, S. V. (2016). Adding in silico assessment of potential splice aberration to the integrated evaluation of BRCA gene unclassified variants. *Human Mutation*, 37(7), 627–639.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *J Am Stat Association.*, 22, 209–212.
- Woolfe, A., Mullikin, J. C., & Elnitski, L. (2010). Genomic features defining exonic variants that modulate splicing. *Genome Biology*, 11(2), R20.
- Yeo, G., & Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, 11(2-3), 377–394.

#### SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Cline MS, Babbi G, Bonache S, et al. Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants. *Human Mutation*. 2019;40:1546–1556. <https://doi.org/10.1002/humu.23861>

## Article 8

BRCA1- and BRCA2-specific in silico tools for variant interpretation in the CAGI 5 ENIGMA challenge.

Padilla N, Moles-Fernández A, Riera C, Montalban G, Özkan S, Ootes L, Bonache S, Díez O, Gutiérrez-Enríquez S, de la Cruz X.

**Hum Mutat.** 2019 Sep;40(9):1593-1611.

doi: 10.1002/humu.23802.





Received: 8 January 2019 | Revised: 15 May 2019 | Accepted: 17 May 2019

DOI: 10.1002/humu.23802

## SPECIAL ARTICLE

Human Mutation  WILEY

# BRCA1- and BRCA2-specific in silico tools for variant interpretation in the CAGI 5 ENIGMA challenge

Natàlia Padilla<sup>1</sup>  | Alejandro Moles-Fernández<sup>2</sup>  | Casandra Riera<sup>1</sup> | Gemma Montalban<sup>2</sup>  | Selen Özkan<sup>1</sup>  | Lars Ootes<sup>1</sup>  | Sandra Bonache<sup>2</sup>  | Orland Díez<sup>2,3</sup>  | Sara Gutiérrez-Enríquez<sup>2</sup>  | Xavier de la Cruz<sup>1,4</sup> 

<sup>1</sup>Research Unit in Clinical and Translational Bioinformatics, Vall d'Hebron Institute of Research (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain

<sup>2</sup>Oncogenetics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain

<sup>3</sup>Area of Clinical and Molecular Genetics, University Hospital of Vall d'Hebron, Barcelona, Spain

<sup>4</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

**Correspondence**

Xavier de la Cruz, Vall d'Hebron Institute of Research (VHIR), Passeig de la Vall d'Hebron, 119-129, 08035 Barcelona, Spain.  
Email: xavier.delacruz@vhir.org

Sara Gutiérrez-Enríquez, Vall d'Hebron Institute of Oncology (VHIO), Cellex Center, C/Natzaret, 115-117, 08035 Barcelona, Spain.  
Email: sgutierrez@vhio.net

**Funding information**

Ministerio de Economía y Competitividad, Grant/Award Numbers: BIO2012-40133, SAF2016-80255-R; European Regional Development Fund, Grant/Award Numbers: Interreg program POCTEFA, Pirepre (EFA086/15); Foundation for the National Institutes of Health, Grant/Award Numbers: NIH R13 HG006650, NIH U41 HG007346; Fundación Científica Asociación Española Contra el Cáncer; Instituto de Salud Carlos III, Grant/Award Numbers: PI15/00355, PI13/01711, PI16/01218, Miguel Servet Program [CP16/00034], PI12/02585

**Abstract**

*BRCA1* and *BRCA2* (*BRCA1/2*) germline variants disrupting the DNA protective role of these genes increase the risk of hereditary breast and ovarian cancers. Correct identification of these variants then becomes clinically relevant, because it may increase the survival rates of the carriers. Unfortunately, we are still unable to systematically predict the impact of *BRCA1/2* variants. In this article, we present a family of in silico predictors that address this problem, using a gene-specific approach. For each protein, we have developed two tools, aimed at predicting the impact of a variant at two different levels: Functional and clinical. Testing their performance in different datasets shows that specific information compensates the small number of predictive features and the reduced training sets employed to develop our models. When applied to the variants of the *BRCA1/2* (ENIGMA) challenge in the fifth Critical Assessment of Genome Interpretation (CAGI 5) we find that these methods, particularly those predicting the functional impact of variants, have a good performance, identifying the large compositional bias towards neutral variants in the CAGI sample. This performance is further improved when incorporating to our prediction protocol estimates of the impact on splicing of the target variant.

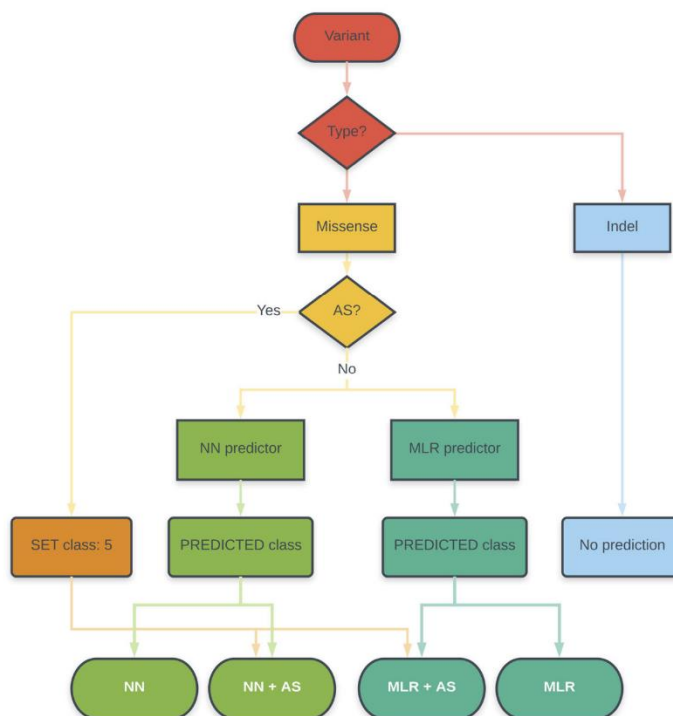
**KEYWORDS**

bioinformatics, breast cancer, functional assays, gene-specific predictor, homology-directed DNA repair (HDR), molecular diagnosis, ovarian cancer, pathogenicity predictions, protein-specific predictor, splicing predictions

## 1 | INTRODUCTION

Germline variants disrupting the DNA protective role of *BRCA1* and *BRCA2* (*BRCA1/2*) result in an increased risk of developing hereditary breast and ovarian cancers (HBOC; Roy, Chun, & Powell, 2012; Venkitaraman, 2014). Identification of the carriers of these variants is clinically relevant because it allows channeling

these individuals to surveillance, prevention programs and targeted therapies (Paluch-Shimon et al., 2016). As a result, these patients increase their survival rates; however, not all of them will benefit equally, because we lack an exact knowledge of the functional impact of *BRCA1/2* variants. In these cases, a straightforward decision can only be taken when the variant is overtly deleterious (insertions, deletions, and substitutions codifying



**FIGURE 1** Prediction protocol. In this article, we present a protocol for the prediction of missense variants that includes assessment of the impact of this variant on splicing and protein function. This protocol has been used to interpret the variants of the ENIGMA challenge in the CAGI 5 community experiment. MLR and NN refer to our two protein-specific predictors, based on a multiple linear regression model and a neural network model, respectively. AS refers to the procedure to predict variants resulting in affected splicing (Moles-Fernández et al., 2018). Abbreviation: CAGI 5, fifth Critical Assessment of Genome Interpretation

truncated proteins). When the variant has an uncertain effect on protein function (e.g., missense, synonymous, intronic, and 5'-untranslated region [5'-UTR] or 3' UTR variants) the best course of action becomes unclear. Solving this problem is not easy because experimentally measuring the impact of these variants on the activity of BRCA1 and BRCA2 (BRCA1/2), requires complex cell-based assays (reviewed in Guidugli et al., 2013; Millot et al., 2012) that are technically challenging for a systematic application (Starita et al., 2015).

In these circumstances, *in silico* pathogenicity predictors of missense substitutions—Align-GVGD (Tavtigian et al., 2006), PolyPhen-2 (Adzhubei et al., 2010), SIFT (Kumar, Henikoff, & Ng, 2009), PON-P2 (Niroula, Urolagin, & Vihinen, 2015) and so on—are employed as an inexpensive, easy-to-use alternative. The predictions obtained are applied for prioritizing variants for experimental evaluation and as a contribution to decision models that integrate different sources of evidence (Karbassi et al., 2016; Lindor et al.,

2012; Moghadasi, Eccles, Devilee, Vreeswijk, & van Asperen, 2016; Vallée et al., 2016). However, the moderate success rate of these tools is an obstacle for their extended use in a clinical environment (Riera, Lois, & de la Cruz, 2014). In the specific case of BRCA1/2, Ernst et al. (2018) suggest, after testing the performance of Align-GVGD, SIFT, PolyPhen-2, and MutationTaster2 on a set of 236 BRCA1/2 variants of known effect, that *in silico* results cannot be used as stand-alone evidence for diagnosis. In terms of molecular effect, two independent, massive functional assays of BRCA1 variants (Findlay et al., 2018; Starita et al., 2015) show that *in silico* predictors provide only a limited view of the functional impact of these variants. In summary, we need to improve the predictive power of these tools, if we want to increase their usage in the clinical setting and augment their value for healthcare stakeholders.

The slow progression in performance displayed by pathogenicity predictors along time shows that ameliorating them is a difficult task (Riera et al., 2014). In this scenario, the use of rigorous performance

estimates becomes an important factor, since improvements are expected to be small and hard to establish. Generally, these estimates are obtained using a standard N-fold cross-validation procedure (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000; Riera et al., 2014; Vihinen, 2012). However, given the increasing availability of variant data, independent testing of predictors is emerging as a valuable option to complement cross-validated performance estimates. Sometimes this testing is done in specific systems for which new variants with impact annotations become available, either at specific/general databases or through experimental testing of their function. For example, Riera et al. (2015) cross-validate their Fabry-specific predictor with a set of 332 pathogenic and 48 neutral variants, and provide independent validation, using a set of 65 pathogenic variants obtained from an update of the Fabry-specific database. Wei and Dunbrack (2013) test five *in silico* predictors using an independent set of 204 variants (79 deleterious, 125 neutral) of the human cystathionine beta-synthase whose impact they establish with an *in vitro* assay. Large variant sets, including data from different genes, are also frequently used to assess and compare the performance of several predictors simultaneously (reviewed in Niroula & Vihinen, 2016). While relevant, the value of these approaches to validation is limited by different factors, such as the fact that the standard of performance evaluation may vary between works, the manuscripts may not always be easy to find, and so on. In this situation, Critical Assessment of Genome Interpretation (CAGI) (Hoskins et al., 2017), a community experiment where developers can assess the performance of their methods in specific challenges, offers an excellent opportunity to obtain an independent view on their work. For users, it allows having an idea on the state of the art for a protein or disease of their interest.

In this manuscript, we present: (a) A novel family of pathogenicity predictors for scoring *BRCA1* and *BRCA2* missense variants; and (b) their performance in the recently held CAGI 5 community experiment.

The four tools described in this work (two for *BRCA1* and two for *BRCA2*) are protein-specific (Crockett et al., 2012; Ferrer-Costa, Orozco, & de la Cruz, 2004; Pons et al., 2016; Riera, Padilla, & de la Cruz, 2016), that is, only variants for a given protein are used to train its two predictors. These two predictors differ on their objective: One is trained to estimate the molecular-level impact of variants and the other their clinical impact (neutral/pathogenic). Technically, for the first predictor we employed a standard multiple linear regression (MLR) approach and for the second, a neural network (NN) model with no hidden layers.

Once obtained, these predictors were applied to the variants constituting the *BRCA1/2* (ENIGMA) challenge in CAGI 5. This was done following a protocol that combined predictions of AS and protein impact and was the same for both proteins (Figure 1). Evaluating these two effects of genetic variants (on splicing and protein function) is routine in general diagnostic procedures (Richards et al., 2015) and there are specific tools in the case of *BRCA1/2* variants (Vallée et al., 2016; <http://priors.hci.utah.edu/PRIORS/>). In our protocol, given an unknown variant, it was first

tested for its effect on the splicing pattern, using a recently developed approach (Moles-Fernández et al., 2018). If the variant had no detectable effect, it was subsequently tested for its impact on protein function, using the predictors here presented. Our results show that all our protein-specific predictors can discriminate (with different degrees of success) between neutral and pathogenic variants, both for *BRCA1* and *BRCA2*. For this binary discrimination problem (neutral/pathogenic) their performances are comparable with, or better than, those of general predictors (CADD, PolyPhen-2, PON-P2, PMut, and SIFT). When applied to the variants of the CAGI challenge, where the goal is to classify them in one of the IARC 5-tier classes (or a reduced version with three classes) we see the same trend. In spite of a decrease in performance, our methods are able to predict the biased composition of the dataset, mainly our predictors trained using data from the homology-directed DNA repair (HDR) assay. Most of the neutral variants are correctly identified by these predictors and, for pathogenic variants, *in silico* prediction of AS enhances the final success rate.<sup>1</sup>

## 2 | MATERIALS AND METHODS

In this work, we present: (a) The development of a family of predictors for *BRCA1/2* missense variants, and (b) the use of these tools to predict the pathogenicity of the ENIGMA variants in the CAGI challenge. We first describe the overall prediction protocol (Figure 1), which integrates predictors of splicing and protein impact, and then focus on the description of the specific predictors.<sup>2</sup>

### 2.1 | Overall prediction protocol

In this section and in Figure 1, we describe the protocol followed in our contribution to the CAGI 5 experiment, an experiment that presents participants with different challenges revolving around a central theme (Hoskins et al., 2017): The prediction of variant pathogenicity and its applications. We focused our efforts on the set of *BRCA1* and *BRCA2* variants provided by the ENIGMA consortium (Spurdle et al., 2012), and we submitted four sets of predictions per protein (Table S1). These four sets correspond to different combinations of our approaches for the prediction of variants leading to affected splicing (AS; one method; Moles-Fernández et al., 2018) or affecting protein function/structure (two methods: MLR and NN). They are the following:

<sup>1</sup>Note on terminology: We have italicized the gene symbols (*BRCA1* and *BRCA2*) and not the protein symbols (*BRCA1* and *BRCA2*). In general, because we are presenting protein-specific predictors, when referring to them, to the training variants, and so on, we have utilized the non-italicized version. However, we are aware that at some points it is unclear which option is preferable and our decision may be arbitrary.

<sup>2</sup>When referring to a variant regarding its impact on protein function, we will speak of "functional", "intermediate", or "non-functional" variants, as those that result in a protein that preserves its function, has lost part of it or has lost all of it, respectively. We will preserve the terms "neutral", "unknown" (or "uncertain"), and "pathogenic" to refer to the clinical phenotype of the variant.



1. Set MLR+AS: AS impact+protein impact with MLR
2. Set NN+AS: AS impact+protein impact with NN
3. Set MLR+nAS: Predict protein impact with MLR, no AS predictions used
4. Set NN+nAS: Predict protein impact with NN, no AS predictions used

The submission format was the same for each set and was provided by the organizers. It comprised the following information per variant: Three fields for the identification (DNA variant; Gene; protein variant); three fields for the prediction (predicted IARC 5-tier class;<sup>3</sup> probability of the variant being pathogenic, which we call "p"; confidence of each prediction probability, which we call "sd"); and one field for "Comments."

For the sets MLR+AS and NN+AS, any variant predicted as "pathogenic" by the AS predictor was arbitrarily assigned values of  $p = 1$  and  $sd = 0$ , and the ENIGMA class "5". Otherwise, the variant was annotated using our protein impact predictors, which were obtained as explained below. That is, the protein impact was estimated only if the variant had no predicted effect on AS. One can distinguish these situations by the text in the "Comments" column: (a) "Splicing," which means that the variant is annotated with the AS predictor; (b) "protein," which means that the variant is annotated with the protein-based predictors (MLR or NN); (c) "arbitrary," which is only used for variants for which we have not a predictor (annotation is arbitrarily set to the following: ENIGMA class = 5;  $p = .5$ ; and  $sd = 0.5$ ).

For the sets MLR+nAS and NN+nAS we did not use AS predictors. All the variants are annotated using our protein impact predictors (obtained as explained below). As before, these situations are distinguished in the "Comments" field with the labels "protein," if the variant is annotated with the protein-based predictors (MLR or NN).

## 2.2 | Prediction of AS variants

To score the effect on splicing of the CAGI variants from the ENIGMA challenge, we have used the results of our recent work (Moles-Fernández et al., 2018) where we identified the best combination of in silico tools for predicting splice site alterations, among those predictors available in the package Alamut Visual v2.10. More precisely, we showed that the HSF+SSF-like combination (with  $\Delta$ -2% and  $\Delta$ -5% as thresholds, respectively) for donor sites and the SSF-like ( $\Delta$ -5%) for acceptor sites, exhibited an optimal performance in a benchmark combining RNA in vitro testing and a dataset of variants retrieved from public databases and reported in the literature. For the CAGI challenge (Figure 1), a variant predicted to produce splice site alterations was arbitrarily assigned Class 5,  $p = 1$

<sup>3</sup>The five ENIGMA classes used correspond to the IARC 5-tier classification system (Goldgar et al., 2008; Plon et al., 2008): 1 = "Not pathogenic," 2 = "Likely not pathogenic," 3 = "Uncertain," 4 = "Likely pathogenic," 5 = "Pathogenic" and were taken from CAGI's website for the BRCA1 and BRCA2 challenge ([https://genomeinterpretation.org/content/BRCA1\\_BRCA2](https://genomeinterpretation.org/content/BRCA1_BRCA2)).

and  $sd = 0$ ; in the comments column it was identified as "splicing". Variants giving no signal for splice site alterations were directly channeled to the protein predictors.

## 2.3 | Protein-based predictors

We have developed two methods for predicting the impact of protein sequence variants of BRCA1 and BRCA2. One is based on a NN and is trained to produce a binary output reflecting the pathogenic nature—cancer risk (high/low)—of a cancer variant. The other method is based on an MLR and is trained to predict the values of the HDR assay for a variant. Both methods are protein-specific: There is a version of MLR for BRCA1 and another for BRCA2, and the same for NN. We describe them below; we start with the NN because it employs more predictive features (6) than the MLR, which only uses a subset of these (3).

### 2.3.1 | The NN method

We have followed our approach to produce protein-specific predictors (Riera et al., 2016), which comprises the three steps described below: (a) Obtention of a variant dataset true to the prediction goal; (b) labeling of variants with chosen features; and (c) obtention of the NN model.

#### Obtention of BRCA1 and BRCA2 variants

Missense variants in this dataset were selected with clinical impact in mind. This was done by manually reviewing several gene-specific databases that collect BRCA1 and BRCA2 variants along with published literature: Leiden Open Variation Database (LOVD) describing functional studies of specific BRCA1 and BRCA2 variants (<http://databases.lovd.nl/shared/genes/BRCA1>; <http://databases.lovd.nl/shared/genes/BRCA2>), LOVD-IARC dedicated to variants that have been clinically reclassified using an integrated evaluation ([http://hci-exlovd.hci.utah.edu/home.php?select\\_db=BRCA1](http://hci-exlovd.hci.utah.edu/home.php?select_db=BRCA1)), BRCA Share™ (formerly Universal Mutation Database UMD-BRCA mutations database <http://www.umd.be/BRCA1/>; <http://www.umd.be/BRCA2/>), CLINVAR, that provides clinical relevance of genetic variants (<https://www.ncbi.nlm.nih.gov/clinvar/>), and BRCA1 CIRCOS which compiles and displays functional data on all documented BRCA1 variants (<https://research.nhgri.nih.gov/bic/circos/>). Finally, each variant was validated by combining different sources of evidence.

Variants for which the pathogenic role was attributable to splice site alterations (assessed using Alamut Visual biosoftware 2.6, from Interactive Biosoftware) were eliminated. This was done to ensure, as far as possible, that our model was trained using variants whose damaging/neutral nature was a consequence of their impact in protein function/structure only.

The final datasets (Table S1) were constituted by Table 1: (a) BRCA1: 77 "pathogenic" and 149 "neutral" variants; and (b) BRCA2: 36 "pathogenic" and 105 "neutral" variants.

### Features

We used a total of six features to label the variants for the predictor training. We have previously used them for the development of protein-specific predictors (Riera et al., 2016). We describe them below for the benefit of the reader.

Two features are based on the use of multiple sequence alignments (MSA): Shannon's entropy and position-specific scoring matrix element. Shannon's entropy is equal to  $-\sum_{i,j} p_{ij} \log(p_{ij})$ , where the index  $i$  runs over all the amino acids at the variant's MSA column. Position-specific scoring matrix element for the native amino acid ( $pssm_{nat}$ ) is equal to  $\log(f_{nat,i}/f_{nat,MSA})$ , where  $f_{nat,i}$  is the frequency of the native amino acid at the locus  $i$  of the variant and  $f_{nat,MSA}$  is the frequency of the same amino acid in the whole MSA. We used two different MSA, psMSA, and oMSA, which resulted in two versions of the NN predictor. psMSA were obtained using the same protocol utilized for the protein-specific predictors (Riera et al., 2015, 2016) which, briefly, consists of two steps: (a) Recovery of BRCA1/2 homologs using a query search of UniRef100; (b) elimination of remote homologs (<40% sequence identity); alignment of the remaining sequences with muscle (Edgar, 2004). The resulting MSA is available on demand from the authors. The oMSA, available from the group of Sean Tavtigian (Tavtigian, Greenblatt, Lesueur, & Byrnes, 2008), comprise only orthologs of BRCA1 and BRCA2, and are publicly available at the web of the Huntsman Cancer Institute, University of Utah (<http://agvvd.hci.utah.edu/alignments.php>). The NN predictions submitted to CAGI were those obtained with the method developed using the psMSA, although results for the second predictor are mentioned below.

Three features, each measuring the difference between native and mutant amino acids for a single physicochemical property: van der Waals volume (Bondi, 1964), hydrophobicity scale (estimated from water/octanol transfer free energy measurements) (Fauchere & Pliska, 1983), and the element of the Blossum62 matrix (Henikoff & Henikoff, 1992) corresponding to the amino acid replacement.

Finally, a sixth feature, that is, binary (1/0) and summarizes the information available on the functional/structural role of the native residue at the UniProt database. It is set to "1" when the native residue has a functional annotation on that database, and "0" if this is not the case.

### NN predictor

The NN predictor was built using WEKA (v3.6.8; Hall et al., 2009). After our experience in the development of protein-specific predictors with small datasets (Riera et al., 2016), we employed the simplest NN model: a single-layer perceptron. Sample imbalances in the training set were corrected with SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

The NN model gives two outputs: (a) a binary prediction for the variant, either pathogenic or neutral; (b) a continuous score, comprised between 0 and 1, that reflects the probability of pathogenicity.

A Leave-one-out cross-validation (LOOCV) of the model was done also using the WEKA (v3.6.8; Hall et al., 2009) package.

### CAGI output

As mentioned above, the CAGI submission requires three pieces of information for each variant prediction: The predicted IARC 5-tier class,  $p$  (probability of pathogenicity), and  $sd$  (reliability). We took as " $p$ " the output from the NN: It varies between 0 (minimal probability of pathogenicity) and 1 (maximal probability of pathogenicity). For the  $sd$  value, we used the following formula (Ferrer-Costa et al., 2004):  $sd = 0.5 - |0.5 - p|$ . It goes from 0 (maximal reliability) to 0.5 (minimal reliability). Finally, the predicted IARC 5-tier class was obtained from the  $p$ , using the ENIGMA conversion table at the CAGI site (Class 5:  $p > .99$ ; Class 4:  $.95 < p < .99$ ; Class 3:  $.05 < p < .949$ ; Class 2:  $.001 < p < .049$ ; Class 1:  $p < .001$ ).

## 2.3.2 | The MLR method

This method aims to predict the values of the HDR assay for a given variant, which is a measure of the impact of this variant on BRCA1/2 molecular function. Because the output of the HDR assay is a continuous value, we opted for using an MLR as a modeling tool, as implemented in the python package Scikit-learn (Pedregosa et al., 2011). The LOOCV of the model was done with the same package. For a given variant, the output of our model is  $HDR_{pred}$ , the predicted value of the HDR assay.<sup>4</sup>

To develop our method we used experimental HDR results available from the literature: 44 variants for BRCA1 (Starita et al., 2015) and 185 variants for BRCA2 (Guidugli et al., 2013, 2018) proteins. However, to reinforce the strength of the signal, relative to experimental noise, we did not employ the full datasets. The BRCA1 training dataset was constituted by those variants used to build the NN predictor (see the previous section) for which HDR values were available; for BRCA2 we followed the same approach. The final number of HDR values was 28 for BRCA1. For BRCA2, we worked with 92 HDR values that corresponded to 56 variants (some had been tested twice; Guidugli et al., 2013, 2018).

Given the small size of these variant datasets, to try to minimize overfitting problems, we used only three of the previous features (see Section 3.1.2, Shannon's entropy,  $pssm_{nat}$ , and Blossum62 element) as independent variables in the regression model. Like for NN methods, the MSA-based features were computed with the psMSA and the oMSA, thus leading to two versions of the MLR. Only the predictions for the oMSA-based MLR were submitted to CAGI; however, the results for the second predictor are also provided in this manuscript.

### CAGI output

To adapt the MLR predictions to the CAGI format, we used the following steps:

<sup>4</sup>When obtaining the HDR predicted values using this method, in a few cases the result was a slightly negative number. In these cases, the predicted value was set to 0, because the output of the HDR experiment is always a positive number.

1. Obtain  $HDR_{pred}$ , the MLR predictions for the variants in the BRCA1 and BRCA2 training datasets.
2. Separately for BRCA1 and BRCA2, compute the mean and standard deviations of the HDR values of the known "pathogenic" and "neutral" variants. At this point, we have four values for each protein:  $m_P, sd_P, m_N, sd_N$ .
3. After the "pathogenicity" assignment, we computed CAGI's "p" as follows:  $N(x; m_P, sd_P) / (N(x; m_P, sd_P) + N(x; m_N, sd_N))$ , where  $N(x; m, sd)$  represents a normal probability distribution of mean  $m$  and standard deviation  $sd$ . The resulting value is comprised between 0 ("neutral") and 1 ("pathogenicity") and reflects the probability of a variant being "pathogenic", according to our model.
4. The  $sd$  value was obtained, as for the NN methods, using the following formula (Ferrer-Costa et al., 2004):  $sd = 0.5 - |0.5 - p|$ .

## 2.4 | Performance assessment

As mentioned before, during the development process predictor performance was estimated using a standard LOOCV procedure for each predictor (Riera et al., 2016), regardless of whether it was MLR or NN.

The parameters used to measure the success rate of the predictors vary depending on the number of classes predicted. During the development process, the NN method predicted only two classes: Pathogenic and neutral; in subsequent validations, including that of the CAGI submissions, three and five classes were considered. We describe below the performance parameters employed in each case.

### 2.4.1 | Binary performance assessment

Here success rate was measured with four commonly employed parameters for binary predictions (Baldi et al., 2000; Vihinen, 2013): sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC). They are computed as follows:

.-Sensitivity:

$$\frac{TP}{TP + FN}$$

.-Specificity:

$$\frac{TN}{TN + FP}$$

.-Accuracy:

$$\frac{TP + TN}{TP + FP + TN + FN}$$

.-MCC:

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN) \cdot (TN + FP) \cdot (TP + FP) \cdot (TN + FN)}}$$

where TP and FN are the numbers of correctly and incorrectly predicted pathological variants; TN and FP are the numbers of correctly and incorrectly predicted neutral variants, respectively.

### 2.4.2 | Multiclass performance assessment

In our case, we need to evaluate the performance of our methods when their score is transformed into a five or three class prediction; for example, this happens when assessing the CAGI submission (we predict five classes) and the application of our MLR to the recently published exhaustive, functional assay of BRCA1 variants (Findlay et al., 2018), where we predict three classes. For multiclass problems, the number of options available is smaller than for binary problems (Baldi et al., 2000; Vihinen, 2013). Here we have utilized the following: The confusion matrices, the accuracies per class, the overall accuracy, and the multiclass MCC (Gorodkin, 2004; Jurman, Riccadonna, & Furlanello, 2012).

For a multiclass problem with  $M$  classes the confusion matrix,  $C = (c_{ij})$ , is an  $(M \times M)$  matrix where  $c_{ij}$  is the number of times a class  $i$  input is predicted as class  $j$ . The sum of the  $c_{ij}$  corresponds to the sample size  $N$ , which in our case is the total number of variants predicted. This matrix provides the simplest description of the performance of a predictor; its diagonal and off-diagonal elements correspond to the predictor's successes and failures, respectively. If we normalize each diagonal element by its row total ( $c_{ii} / \sum_j c_{ij}$ , where  $j = 1, M$ ) we obtain the accuracy of the predictor for class  $i$ . If we add all the diagonal elements and divide the result by  $N$  ( $\sum_i c_{ii} / N$ , where  $i = 1, M$ ), we obtain the overall accuracy.

The multiclass MCC (Gorodkin, 2004; Jurman et al., 2012) was obtained using the implementation in the python package Scikit-learn (Pedregosa et al., 2011).

## 3 | RESULTS

In this article, we describe the obtention of a novel family of pathogenicity predictors specific for BRCA1/2 proteins (MLR and NN) and their application to the variants in the CAGI challenge, within a protocol that also includes AS predictions (Figure 1). Sections 3.2–3.5 correspond to the first part, and Section 3.6 corresponds to the second part.

As we have seen in Section 2, we have considered the use of different MSA (psMSA and oMSA) to develop our predictors. However, we center our descriptions on the versions employed for the CAGI challenge: MLR based on oMSA and NN based on psMSA. For completeness, we also provide the performance of our methods when developed using psMSA (for MLR) and oMSA (for NN).

### 3.1 | The variant datasets

In Table 1a we give the size of the datasets employed in this work. In Table 1b, we report the overlap between the CAGI and the remaining



**TABLE 1a** Size of the datasets used in this work (CAGI: missense +AS)

	NN	MLR	CAGI	SGENaN
BRCA1	226 (P = 77/N = 149)	28	144	1,837
BRCA2	141 (P = 36/N = 105)	56	174	-

Abbreviations: AS, affected splicing; CAGI, Critical Assessment of Genome Interpretation; MLR, multiple linear regression; N, neutral; NN, neural network; P, pathogenic; SGE, saturation genome editing.  
<sup>a</sup>Dataset extracted from Findlay et al. (2018)

**TABLE 1b** Overlap between datasets (CAGI: missense+AS)

	NN-CAGI	MLR-CAGI	MLR-SGENaN
BRCA1	18 (P = 7/N = 11)	2	28
BRCA2	5 (P = 2/N = 3)	4	-

Abbreviations: AS, affected splicing; CAGI, Critical Assessment of Genome Interpretation; MLR, multiple linear regression; N, neutral; NN, neural network; P, pathogenic; SGE, saturation genome editing.  
<sup>a</sup>Dataset extracted from Findlay et al. (2018)

datasets. Note that the CAGI class information on each variant was made public only after the challenge was closed.

**3.1.1 | Training datasets for NN and MLR**

The number of missense variants in the NN training sets (BRCA1, 226; BRCA2, 141) is comparable with that used for developing protein-specific predictors with the same NN model and variant features (Riera et al., 2016). The situation is different for the MLR training sets, which were small (BRCA1, 28; BRCA2, 56), thus imposing a severe limitation in the number of features that can be used in the model (see Section 2).

**3.1.2 | Validation dataset for BRCA1 MLR**

This set is obtained from the results of a recently published (Findlay et al., 2018) experiment for BRCA1. The authors

functionally score a large number of single-nucleotide variants; we retrieved the 1,837 cases corresponding to missense variants. We refer to this dataset as SGE (from "saturation genome editing"). We used SGE to further test the performance of our BRCA1 MLR because Findlay et al. (2018) find that there is a correspondence between their functional score and the score of the HDR assay.

**3.1.3 | CAGI datasets**

Their size (BRCA1, 144; BRCA2, 174) is of the same magnitude as that of the NN training datasets. In Table 2 we provide two partitions of these datasets, corresponding to: (a) the original, 5-class ENIGMA partition; and (b) a reduced, 3-class partition. For the latter, the "Pathogenic" and "Likely pathogenic" classes have been unified into a single "Pathogenic class" and the "Likely not pathogenic" and "Not pathogenic" classes have been unified into a single "Neutral class". The "Uncertain class" (or "Unknown") has been left untouched. It must be noted the high compositional imbalance of the CAGI dataset, with the total of classes 1 and 2 being 10 and 25 times higher than that of the remaining classes, for BRCA1 and BRCA2, respectively. In particular, the absolute numbers of variants for classes 3, 4, and 5 are so low that they can hardly lead to reliable estimates for class-dependent parameters. For example, there are only two variants of class 3 for both BRCA1 and BRCA2; two and three variants for classes 4 and 5, respectively, in BRCA2; and four and seven variants for classes 4 and 5, respectively, in BRCA1.

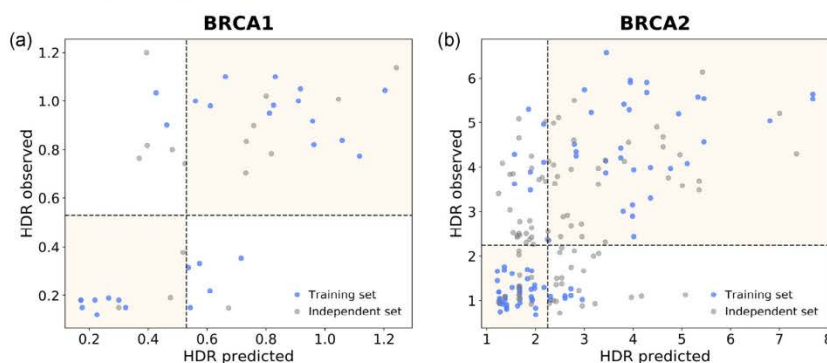
**3.2 | Predicting the functional impact of BRCA1/2 variants: The MLR predictor**

We have developed two MLR methods, one per protein. The goal of these methods is to predict the impact of a given variant on protein function, as measured by the HDR experiment. To this end, they were trained with a set of variants with known experimental values for the HDR assay and the features chosen

**TABLE 2** Composition of the ENIGMA dataset in the CAGI 5 challenge

<b>(A) BRCA1</b>					
IARC 5-tier class	1 (<0.001)	2 (0.001-0.049)	3 (0.05-0.949)	4 (0.95-0.99)	5 (>0.99)
CAGI	31	100	2	4	7
Three Class <sup>a</sup>	Neutral		Unknown	Pathogenic	
CAGI	131		2	11	
<b>(B) BRCA2</b>					
IARC 5-tier class	1 (<0.001)	2 (0.001-0.049)	3 (0.05-0.949)	4 (0.95-0.99)	5 (>0.99)
CAGI	31	136	2	2	3
Three Class <sup>a</sup>	Neutral		Unknown	Pathogenic	
CAGI	167		2	5	

Abbreviations: CAGI, Critical Assessment of Genome Interpretation; IARC, International Agency for Research on Cancer.  
<sup>a</sup>This classification is a simplified version of the IARC 5-tier scheme (see manuscript) where the Neutral class corresponds to IARC classes 1 and 2, the Pathogenic class corresponds to IARC classes 4 and 5, and Unknown corresponds to IARC class 3.



**FIGURE 2** Observed versus predicted HDR values for (a) BRCA1 and (b) BRCA2. In blue, we show the variants used for the training/testing of our MLR method (the version trained with oMSA, used to generate CAGI predictions). The HDR predicted values are cross-validated (LOOCV, see Section 2). For completeness, we show in gray the points from the original HDR experiments that were excluded from the training process after applying our filtering procedure (see Section 2). CAGI, Critical Assessment of Genome Interpretation; HDR, homology-directed DNA repair; LOOCV, leave-one-out cross-validation; MLR, multiple linear regression; MSA, multiple sequence alignment

are related to the effect variants can have on protein structure, protein-protein interactions, and so on. (Ferrer-Costa, Orozco, & de la Cruz, 2002; Riera et al., 2014). In Figure 2, we see that there is a statistically significant correlation between observed versus predicted (LOOCV) HDR values (BRCA1, 0.72;  $p = 1.5 \times 10^{-5}$ ; BRCA2, 0.73;  $p = 3.3 \times 10^{-17}$ ). Visual inspection reveals that the variants tend to group into two clusters, showing that MLR predictions approximately reproduce the bimodal pattern of HDR assays (Guidugli et al., 2013; Starita et al., 2015). We also show (gray color), the predictions for the variants which were left outside the training set, after applying the pathogenicity condition (see Section 2); they are more scattered than those forming the training set, illustrating how the filtering worked.

We explored how good this level of accuracy is for a standard two-class (pathogenic/neutral) prediction of the variant's pathogenicity. To this end we discretized the predictions applying a decision boundary: A variant was called pathogenic or neutral when its predicted HDR score was below or above a given threshold, respectively. These thresholds, taken from the experimental papers, where: 0.53 for BRCA1 (Starita et al., 2015) and 2.25 for BRCA2 (Guidugli et al., 2013). In Table 3 we give the parameters measuring the success rate of the discretized MLR methods. Their accuracies, 0.75 for BRCA1 and 0.86 for BRCA2, fall within the 0.79–0.99 accuracy range for protein-specific predictors (Riera et al., 2016); the same happens for the MCC, 0.50 for BRCA1 and 0.71 for BRCA2. We detect that specificity (0.85) and sensitivity (0.86) are closer for BRCA2 than for BRCA1 (spec, 0.87; sens, 0.62). Actually, for BRCA1 sensitivity tends to be small when compared with that of protein-specific predictors (Riera et al., 2016). Overall, these results indicate that the continuous HDR predictions of our MLR model can be transformed into binary predictions preserving a non-random

prediction power, comparable with that of predictors trained with binary encodings (pathogenic/neutral) of the variant impact.

### 3.3 | Validation of the BRCA1 MLR predictor with functional data

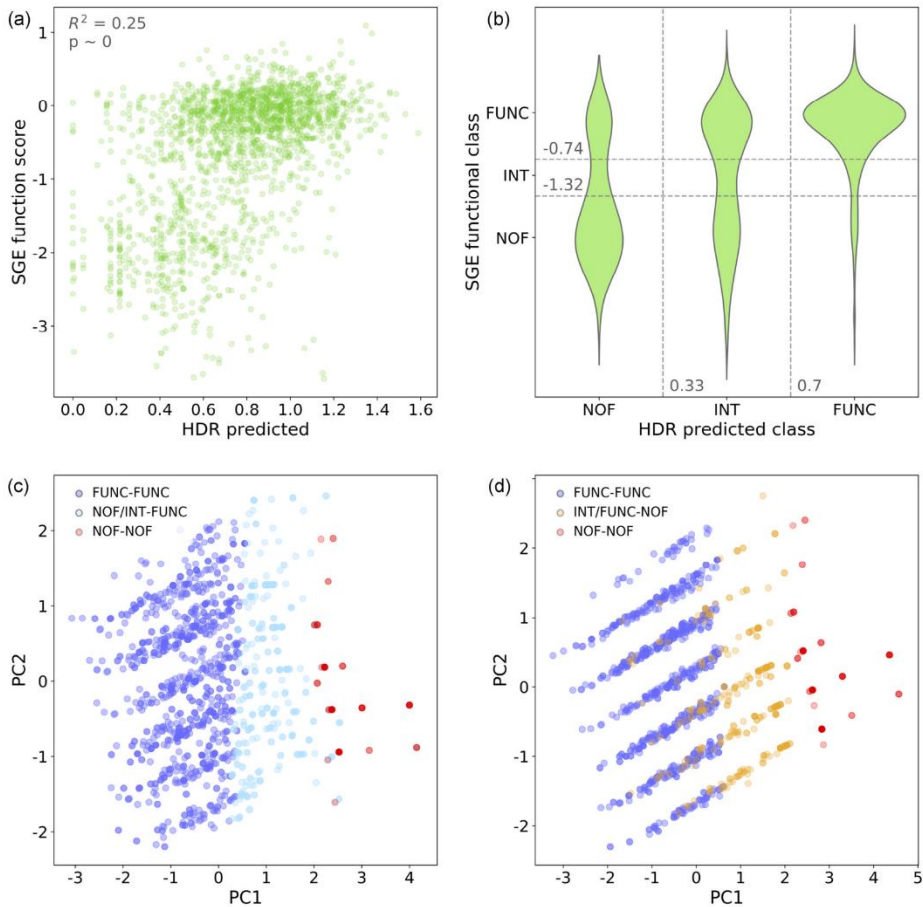
The recent publication (Findlay et al., 2018) of a massive functional assay of BRCA1 variants has given us the opportunity to check the performance of our MLR model on a set of 1,837 variants. The output of this experiment is a continuous value measuring the impact of sequence variants on BRCA1 function. When we represent these values against our HDR predictions (Figure 3a) we observe two clusters of points (below and above  $SGE = -1$ ) that reflect the bimodal behavior of both assays, with a statistically significant rank correlation (Spearman's  $\rho = 0.47$ ;  $p < 0$ ). This overall coincidence is

**TABLE 3** Two-class (binary) performance of our predictors

Protein	Method	SN	SP	ACC	MCC
BRCA1	MLR (psMSA)	0.692	0.933	0.821	0.651
	MLR-CAGI (oMSA)	0.615	0.867	0.75	0.502
	NN (oMSA)	0.922	0.852	0.876	0.746
BRCA2	NN-CAGI (psMSA)	0.857	0.718	0.765	0.546
	MLR (psMSA)	0.828	0.741	0.786	0.571
	MLR-CAGI (oMSA)	0.862	0.852	0.857	0.714
	NN (oMSA)	0.75	0.867	0.837	0.592
	NN-CAGI (psMSA)	0.75	0.771	0.766	0.473

Note: "CAGI" identifies the predictors used for this challenge. Abbreviations: ACC, accuracy; CAGI, Critical Assessment of Genome Interpretation; MCC, Matthews correlation coefficient; MLR, multiple linear regression; MSA, multiple sequence alignment; NN, neural network; SN, sensitivity; SP, specificity.





**FIGURE 3** Prediction of the “saturation genome editing” (SGE) experiment in *BRCA1*. We use our impact prediction to check the correspondence between our HDR predictions and the results of the SGE experiment (Findlay et al., 2018). (a) Scatterplot representing SGE values versus HDR predictions for the 1,837 missense variants from (Findlay et al., 2018; Spearman’s  $\rho = 0.47$ ;  $p \sim 0$ ). (b) Violin plot showing the distribution of variants for the different combinations of SGE and HDR functional categories: “functional” (FUNC), “intermediate” (INT), and “non-functional”(NOF). Points in the off-diagonal quadrants correspond to outliers: Points whose SGE (observed) and HDR (predicted) functional classes do not coincide. (c) Principal component analysis of three variant populations (HDR-SGE classes): FUNC-FUNC (dark blue), NOF-NOF (red) and the outliers NOF-FUNC plus INT-FUNC (light blue). (d) Principal component analysis of three variant populations (HDR-SGE classes): FUNC-FUNC (dark blue), NOF-NOF (red) and the outliers INT-NOF plus FUNC-NOF (yellow). PC1 and PC2 refer to the first two principal components (those which accumulate the highest variance). HDR, homology-directed DNA repair

limited by a substantial scatter. Part of it may be due to technical/biological (intraexon normalization procedures, the impact of RNA levels, etc.) differences between the SGE and HDR experiments that introduce some dispersion in the comparison between both experiments (see Figure 9m from Extended Data Section in Findlay et al., 2018). Another part of the scatter is due to limitations of our model. To better understand these, we divided the SGE-HDR plane into nine

regions, corresponding to the  $3 \times 3$  combinations of SGE (“functional,” “intermediate,” and “non-functional”; Findlay et al., 2018) and HDR (“High,” “Int,” and “Low”; Starita et al., 2015) equivalent, functional classes. The main blocks of outliers correspond to the two top-left and the two bottom-right regions. We separately used the variants inside each block for a principal component analysis (PCA), using as variables the three features in our model (Shannon’s

entropy,  $\text{psm}_{\text{nat}}$ , and Blossum62 element). As a reference, for each PCA we also included the variants from the upper ("functional") and lower ("non-functional") diagonal regions. In the plane of the first two principal components (PC1 and PC2 in Figure 3c,d) the chosen variants adopt a three-layered disposition, where we successively find the "functional," the outliers and the "non-functional" ones. This disposition reflects the contrast between the bimodal nature of the SGE experiment and the smoother nature of our model.

In fact, in Figure S1 we can see that those outlier variants indeed tend to have intermediate values (comprised between those of the "functional" and "non-functional" populations) for the features in our model. This suggests that for these variants we need to improve our representation of protein impact with new properties, to reproduce more accurately the results of the SGE experiment. However, it may also indicate the need to consider the effect of variants on other aspects of gene function, like RNA levels (Findlay et al., 2018).

### 3.4 | Predicting the clinical impact of BRCA1/2 variants: The NN predictors

We have developed two NN methods, one per protein. These methods were trained with the idea of predicting the clinical impact of a given variant. To this end, during the training process, each variant was labeled with a binary version of this clinical impact: Pathogenic/neutral. Here, the larger amount of data (Table 1aa) allowed us to work with three additional features, fully adhering to our protocol for the obtention of protein-specific predictors (Riera et al., 2016). As for the MLR predictors, the results obtained (Table 3) are comparable to those of other protein-specific predictors. Their accuracies, 0.77 for both BRCA1 and BRCA2, are almost within the 0.79–0.99 accuracy range for protein-specific predictors; the same happens for the MCC, 0.55 for BRCA1 and 0.47 for BRCA2. The sensitivities and specificities are more balanced for both BRCA1 (spec, 0.72; sens, 0.86) and BRCA2 (spec, 0.77; sens, 0.75) when compared with what happened for the MLR predictors.

Overall, as in the case of MLR, the results indicate that the more clinically flavored NN predictors have a prediction power comparable to that of other protein-specific predictors (Riera et al., 2016).

### 3.5 | Comparison with general pathogenicity predictors

To put in context the results of our protein-specific predictors, we give the performance, on our training datasets, of a representative set of general predictors: CADD (Kircher et al., 2014), PolyPhen-2 (Adzhubei et al., 2010), SIFT (Kumar et al., 2009), PON-P2 (Niroula et al., 2015), and PMut (López-Ferrando, Gazzo, De La Cruz, Orozco, & Gelpí, 2017). Care must be exercised when considering the results of this comparison, because the variants in our datasets can be found in databases, like UniProt (Bateman et al., 2017), commonly used to develop pathogenicity predictors (Riera et al., 2014). Therefore, it is likely that some of these variants have been used in the training of the general methods, thus leading to optimistic estimates of their

performance. An additional limitation of the comparison is the small sample size involved, for example, training of BRCA1 MLR was done using only 28 variants.

In general, we observe (Figure 4) that our specific methods have success rates comparable with those of general methods. For MCC, our methods are only surpassed by PMut. For BRCA2, our NN is slightly surpassed by PON-P2 (MCC of 0.47 vs. 0.49), but our MLR surpasses PON-P2 (MCC of 0.71 vs. 0). The sensitivities and specificities of our methods are generally smaller and larger, respectively than those of other methods. However, our methods have an equilibrated performance for pathogenic and neutral variants (Figure 4e,f), because they display the smallest differences between sensitivity and specificity, 0.14 (BRCA1) and 0.021 (BRCA2) for NN, respectively, and 0.25 (BRCA1) and 0.01 (BRCA2) for MLR. Only PMut has closer values for the MLR training set of BRCA1, 0.06.

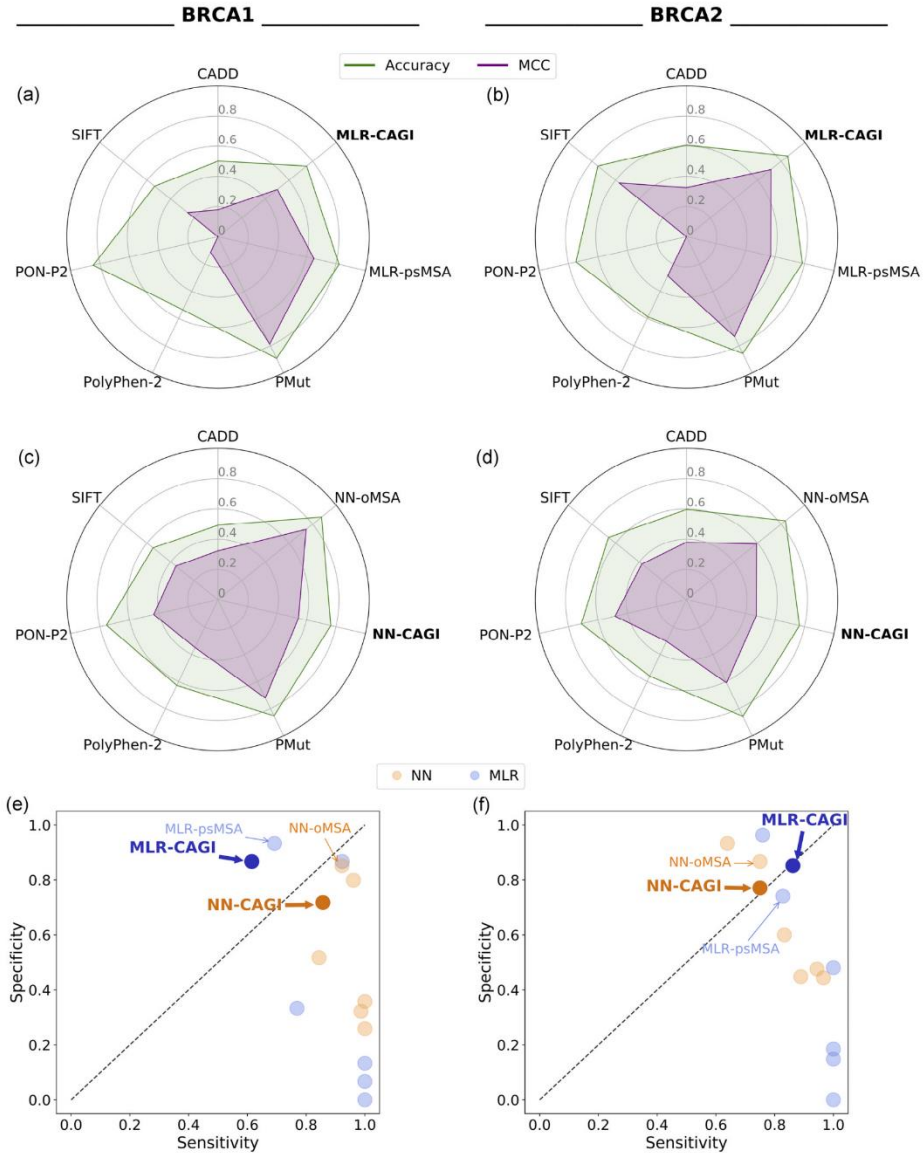
### 3.6 | Results of the predictors in the CAGI experiment

In this section, we present the results of applying our prediction protocol (Figure 1) to the CAGI variants. For each protein, we submitted to the CAGI challenge the results of four versions of this protocol (Figure 1): MLR+AS, NN+AS, MLR, and NN. For simplicity, we will restrict our analysis to the complete protocols (MLR+AS, NN+AS), mentioning protein predictions (MLR, NN) only for discussing the contribution of the AS predictors. The performance was assessed using the class assignments provided by the CAGI organizers after the challenge was closed. More precisely, we computed the ability of our protocols to correctly assign a variant to its class in two different classification schemes. One is the IARC 5-tier classification system (Goldgar et al., 2008; Plon et al., 2008), which was the one requested by the organizers; the other is a 3-class version of this system (see Section 2).

The fact that we must consider the performance for more than two classes makes the evaluation problem more difficult: In multi-class problems confusion matrices retain their explanatory power, but summary measures are not easy to generalize, nor to interpret (Baldi et al., 2000; Vihinen, 2012). In our case, the severity of this problem is augmented by the compositional imbalance in the CAGI dataset (Table 2). For these reasons, we focus our analysis mainly on the confusion matrices (represented as heatmaps) because they provide the basal information in any prediction process and allow a direct interpretation. More concretely, we consider: (a) The diagonal elements to see how good our predictions are; and (b) the off-diagonal elements to see how incorrect predictions distribute among classes. We treat separately BRCA1 and BRCA2 cases because the performance of specific and general pathogenicity predictors is gene-dependent (Riera et al., 2016).

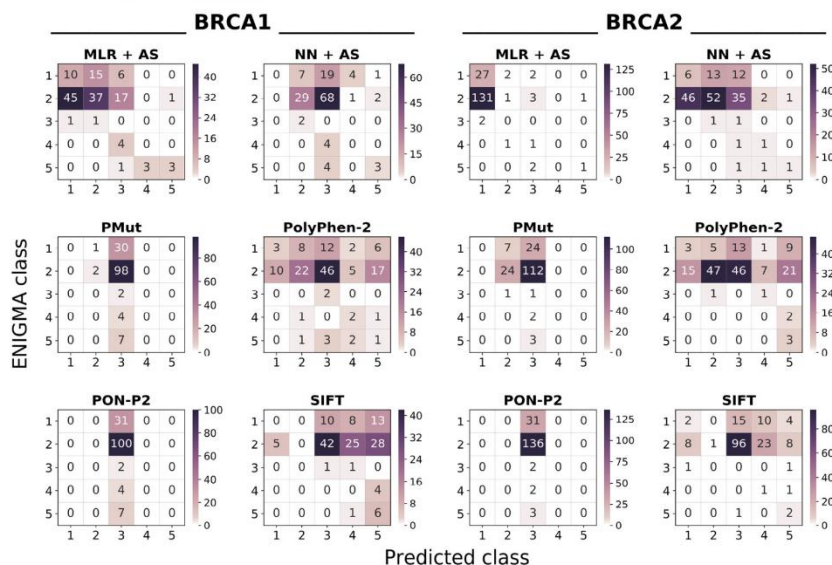
#### 3.6.1 | BRCA1 variants

Looking at the diagonals of their confusion matrices (Figure 5), we observe that MLR+AS and NN+AS can recognize, with varying



**FIGURE 4** Binary, cross-validated performance of the predictors. We represent the performance of our MLR and NN methods, as well as that of general predictors (CADD, PolyPhen-2, PMut, PON-P2, and SIFT), using four parameters: accuracy and MCC (radar plots [a], [b], [c], and [d]) and sensitivity and specificity (scatterplots, [e] and [f]). The methods labeled MLR-CAGI and NN-CAGI are those used to generate our CAGI predictions; for completeness, we give the performance of the other versions: MLR-psMSA (entropy and  $psm_{nat}$  values were obtained from psMSA-based parameters) and NN-oMSA (entropy and  $psm_{nat}$  values were obtained from oMSA-based parameters). In (e) and (f) points are colored according to the set in which sensitivity and specificity were estimated: blue and orange for the MLR and NN sets, respectively. CAGI, Critical Assessment of Genome Interpretation; MCC, Matthews correlation coefficient; MLR, multiple linear regression; MSA, multiple sequence alignment; NN, neural network





**FIGURE 5** Heatmap of the predictor performances on the CAGI datasets. Each heatmap represents the confusion matrix of a predictor. We provide six heatmaps per protein, two for our predictors (MLR+AS and NN+AS) and four for the general predictors (PolyPhen-2, PON-P2, PMut, and SIFT). In all the plots, the vertical and horizontal axes correspond to the observed (provided by CAGI organizers) and predicted IARC 5-tier classes, respectively. Diagonal and off-diagonal elements correspond to successful and failed predictions, respectively. NOTE: given the range differences in the predictions, each plot has its color scale. AS, affected splicing; CAGI, Critical Assessment of Genome Interpretation; IARC, International Agency for Research on Cancer; MLR, multiple linear regression; NN, neural network

accuracies, members from three (1,2,5) and two classes (2,5), respectively. This overall trend is reflected in the class accuracies, which are higher for MLR-based protocols than for NN-based ones (Table 4). If AS predictions are not included, the two methods also fail to recognize class 5 variants (Table 4). In fact, for MLR+AS and NN+AS protocols AS predictions are responsible for the accuracy of class 5, which is 0.43 (3 out of 7 correctly predicted variants) in both cases; AS predictions lead to a single failure, for a class 2 variant.

To understand the distribution of incorrect predictions among classes, we consider the off-diagonal elements of the confusion matrices (Figure 5). For MLR+AS, incorrect predictions mostly group at positions adjacent to the diagonal, with only 9 out of 144 variants breaking this trend. For NN+AS this number grows to 31 and predictions (both correct and incorrect) seem to cluster around the class 3 column.

If we analyze the predictions within the unified 3-class framework, we find that the class accuracies increase for MLR+AS: 0.82 and 0.56 for "Neutral" and "Pathogenic," respectively. For NN+AS, this is not the case, due to the previously mentioned clustering of predictions around class 3. Accuracy for the "Unknown" class is the same as that for IARC 5-tier class 3 because the classes are the same.

Finally, we compare the performance of our predictors with that for the general predictors for which the output directly corresponded

to a probability of pathogenicity (we only excluded CADD, because the score has another scale; Figure 5). For the chosen predictors (PMut, PolyPhen-2, PON-P2, and SIFT) their score is a probability of pathogenicity that can be transformed into an equivalent of the IARC 5-tier classes, using the ENIGMA conversion table (see Section 2). Focusing on the most frequent CAGI variants (31 from class 1; 100 from class 2), we see that MLR+AS performs better than general methods; for class 5, all general methods, except SIFT, identify fewer correct variants. The case of SIFT is of interest since some of the class 5 variants appear to be splicing variants according to our AS predictions: At this point, and without further evidence, it is unclear which is the correct view, the amino acid view provided by SIFT or the nucleotide view provided by AS predictions. For classes 3 and 4, the size of the sample, two and four variants, respectively, limits the value of the results, which are: For the two variants of class 3, MLR+AS performs worse than general methods; for the four variants of class 4, only PolyPhen-2 correctly identifies two of them. A remarkable feature of MLR+AS, relative to general methods, is that its predictions form a band around the diagonal, while general methods either scatter their predictions (PolyPhen-2, SIFT) or cluster them around class 3 (PON-P2 and PMut). Comparison of NN+AS with general methods (Figure 5) shows similarities with PON-P2 and PMut, and a failure to identify members of class 1 that is shared with

**TABLE 4** Class accuracies for the CAGI variants (IARC 5-tier and 3-class unified classes)

<b>(A) BRCA1</b>					
IARC 5-tier	1 (<0.001)	2 (0.001-0.049)	3 (0.05-0.949)	4 (0.95-0.99)	5 (>0.99)
MLR	0.323	0.37	0	0	0
MLR +AS	0.323	0.37	0	0	0.429
NN	0	0.29	0	0	0
NN + AS	0	0.29	0	0	0.429
Three Class	Neutral		Unknown	Pathogenic	
MLR	0.817		0	0.273	
MLR +AS	0.817		0	0.545	
NN	0.275		0	0	
NN + AS	0.275		0	0.273	

<b>(B) BRCA2</b>					
IARC 5-tier	1 (<0.001)	2 (0.001-0.049)	3 (0.05-0.949)	4 (0.95-0.99)	5 (>0.99)
MLR	0.871	0.007	0	0	0
MLR +AS	0.871	0.007	0	0	0.333
NN	0.194	0.382	0.5	0.5	0
NN + AS	0.194	0.382	0.5	0.5	0.333
Three Class	Neutral		Unknown	Pathogenic	
MLR	0.97		0	0	
MLR +AS	0.964		0	0.2	
NN	0.701		0.5	0.4	
NN + AS	0.701		0.5	0.6	

Note: The color shading reflects the correspondence between the two class systems.

Abbreviations: AS, affected splicing; CAGI, Critical Assessment of Genome Interpretation; IARC, International Agency for Research on Cancer; MLR, multiple linear regression; NN, neural network.

all general methods, except PolyPhen-2; again, AS predictions favor our method for class 5, except in the case of SIFT.

The comparison within the three-class framework (Figure S2) confirms the previous trends, with MLR+AS having the largest class accuracy for "Neutral," 0.82, well over that of general methods (0.33

for PolyPhen-2; 0.04 for SIFT; 0.02 for PMut; and 0 for PON-P2). MLR+AS displays the second best accuracy for "Pathogenic," together with PolyPhen-2 and behind SIFT. NN+AS again shows a performance below that of these two general methods, but above that of PON-P2 and PMut.

### 3.6.2 | BRCA2 variants

For *BRCA2*, the situation is somewhat different. The diagonal elements of the confusion matrix (Figure 5) show that NN+AS can recognize variants from the five classes, with varying accuracies (Table 4), while MLR+AS recognizes only variants from classes 1, 2, and 5. In addition, for the most frequent classes (1, 2) NN+AS is more balanced than MLR+AS (Figure 5; Table 4): 0.19 (1) and 0.38 (2) vs. 0.87 (1) and 0.01 (2), respectively. Inspection of the off-diagonal elements shows that wrong predictions are more spread for NN+AS than for MLR+AS. For example, for MLR+AS, essentially all (97%) the incorrect predictions of Class 2 go to Class 1, while this figure drops to 55% for NN+AS. As before, the tiny number of variants in the remaining classes reveals no clear trends. The AS predictions result in one correctly identified member of Class 5 for the two versions of our protocol; AS predictions lead to a single failure, for a Class 2 variant.

As for *BRCA1*, reduction of the five IARC 5-tier classes to a 3-class system reveals a reversion in the previous trend, with a high-class accuracy for "Neutral," higher for MLR+AS (0.96) than for NN+AS (0.70). Accuracy for the "Unknown" class is the same as that for IARC 5-tier Class 3 because the classes are the same. For the "Pathogenic" class, NN+AS still performs better than MLR+AS (Figure 5; Table 4).

Finally, we compare the performance of our predictors with that for the general predictors for which the output directly corresponded to a probability of pathogenicity (we only excluded CADD, because the score has another scale; Figure 5). Focusing on the most frequent CAGI variants (31 from Class 1; 136 from class 2), we see that NN+AS performs better than general methods; MLR+AS is only better for Class 1; for Class 2 its accuracy is low, the same as SIFT, and below that of PolyPhen-2 and PMut. For Classes 3, 4, and 5, the sample size is smaller than that of *BRCA1* (2, 4, 7 vs. 2, 2, 3 variants for *BRCA1* and *BRCA2*, respectively); for this reason, we believe that for these variants it is preferable to wait for next rounds of the CAGI challenge to assess the performance of the different *in silico* tools, including ours.

The comparison within the three-class framework (Figure S2) confirms the previous trends, showing that for the "Neutral" class (167 out of 174 CAGI variants) both MLR+AS and NN+AS surpass general methods (Figure S2). For the "Pathogenic" class (5 variants), PolyPhen-2, and SIFT have the best performances, while our methods rank third (MLR+AS) and fourth (NN+AS).

## 4 | DISCUSSION

Obtaining good estimates of the functional impact and cancer risk of *BRCA1* and *BRCA2* sequence variants plays a vital role in the diagnosis and management of inherited breast and ovarian cancers (Eccles et al., 2015; Findlay et al., 2018; Guidugli et al., 2018; Moreno et al., 2016; Paluch-Shimon et al., 2016). *A priori*, *in silico* tools, can be used to obtain these estimates; however, their moderate success

rate restricts their applicability (Ernst et al., 2018). In this work, we have addressed this issue focusing on the problem of predicting the pathogenicity of *BRCA1/2* missense variants using protein-specific information (Riera et al., 2014). This approach has been validated in different proteins (Crockett et al., 2012; Riera et al., 2016); recent results (Hart et al., 2019) show that it can improve the identification *BRCA1/2* pathogenic variants. Here, we present a new family of *BRCA1*- and *BRCA2*-specific tools that we validate in two different ways: (a) In isolation, using manually curated sets of functionally and clinically annotated variants; and (b) in combination with predictors of splicing impact (Figure 1), to interpret the variants from the ENIGMA challenge of the CAGI 5 experiment.

### 4.1 | The performance of *BRCA1*- and *BRCA2*-specific predictors in isolation

When tested in isolation, we find that our two methods (MLR and NN) are competitive when compared with general methods (Section 3.5; Table 3; Figure 4), for both *BRCA1* and *BRCA2*. In particular, their specificities are among the best, a property desirable from the point of view of HBOC diagnosis requirements (Ernst et al., 2018); they also have the best balances between specificity and sensitivity, with the only exception of PMut in *BRCA1*, which has slightly better figures for the MLR training set. General methods also show good success rates in our training sets (Figure 4), in contrast with the usually lower performance estimates cited in the literature. For example, the last version of PMut displays an MCC of 0.31 for both *BRCA1* (63 variants) and *BRCA2* (104 variants; López-Ferrando et al., 2017). In the same work, we find MCC values for other tools, computed on the same dataset: For *BRCA1* they vary between 0.17 (PROVEAN) and 0.38 (LRT); for *BRCA2* they vary between 0.01 (PROVEAN) and 0.19 (Mutation Assessor). In a previous study, using a small dataset of *BRCA2* variants, Karchin, Agarwal, Sali, Couch, and Beattie (2008) find that general tools display good sensitivities but low specificities. A similar trend has been recently reported by Ernst et al. (2018), after testing PolyPhen-2, SIFT, Align-GVGD, and MutationTaster2 in a set of 236 *BRCA1/2* variants. These authors express concern about the moderate performance observed, particularly about the low specificities observed relative to HBOC diagnosis requirements (e.g., PolyPhen-2: 0.67 and 0.72 for *BRCA1* and *BRCA2*, respectively). We believe that our higher estimates for general predictors (Table 3; Figure 4), relative to those in the literature, may partly result from the overlap between their training sets and our manually curated dataset.

Presently, stand-alone use of *in silico* methods for HBOC diagnosis is discouraged (Ernst et al., 2018). Nonetheless, it is considered that these methods can be fruitfully combined with the results of functional assays, to provide an alternative to multifactorial models in the absence of family information (Guidugli et al., 2018). The tools presented in this work are amenable to this type of approach because of their extreme simplicity and interpretability. This is a consequence of the small number of features utilized (3 and 6 for MLR and NN, respectively) and of the low complexity of our



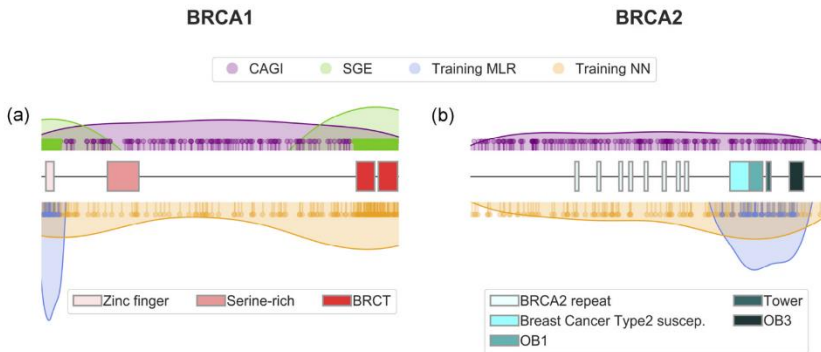
models (Riera et al., 2014). In addition, our MLR models allow a direct interpretation of a variant's impact at the molecular level, because they produce estimates of the HDR assay for the target variant. In this sense, the MLR approach resembles that of Starita et al. (2015) who estimate HDR values using the results of other functional assays (E3 ligase scores and BARD1-binding scores). In our case, we use instead of a few sequence-based features, with two conservation measures (Shannon's entropy and psssm<sub>naa</sub>) standing among them given their recognized predictive power (Ferrer-Costa et al., 2004). Conceptually, this makes MLR methods an implementation of the idea of addressing pathogenicity prediction problems focusing on endophenotypes, rather than on clinical phenotypes. Endophenotypes are quantitative measures of intermediate phenotypes with clinical relevance (Masica & Karchin, 2016); they are closer to the genotype and, for this reason, may result in predictors with high success rates, given the small contribution of genetic background and environmental effects to the outcome of the variant. In general, this is the case when looking at the clinical performance (Table 3; Figure 4). However, for BRCA1, the sensitivity (0.62) is low compared with specificity (0.87); while this may be a consequence of the discretization of the HDR prediction, it may also be a consequence of the extreme simplicity of our model. When testing the MLR model with SGE data we observe a significant correlation (Spearman's  $\rho = 0.47$ ;  $p \sim 0$ ), comparable with that of Align-GVGD ( $\rho = 0.46$ ) and better than that of CADD ( $\rho = 0.40$ ), PhyloP ( $\rho = 0.36$ ), SIFT ( $\rho = 0.36$ ), and PolyPhen-2 ( $\rho = 0.28$ ; values obtained from Figure 9 in Extended Data Section in Findlay et al., 2018). However, visual inspection shows the presence of substantial deviations from a monotonic relationship (Figure 3a,b). If we analyze the population of outliers using PCA and value distributions of the features in our model (Figure S1) we see that, generally, they have an intermediate behavior between "functional" and "non-functional" variants for all

features. This points to an aspect of the variant's impact that is poorly represented by our present set of features, like the effect of the mutation in RNA levels.

Finally, it is worth mentioning that our MLR predictors have been trained with small sets of variants that are concentrated in a reduced region of BRCA1 and BRCA2 (Figure 6). This is in contrast with the broader range of positions covered by the NN and the CAGI datasets. The fact that, in spite of this situation, the MLR tools are competitive suggests that they capture some general effect of variants on protein function/structure, like impact on stability (Yue, Li, & Moul, 2005).

#### 4.2 | The performance of BRCA1- and BRCA2-specific predictors in the CAGI 5 experiment

The ENIGMA challenge within the CAGI experiment provides a good opportunity to independently validate the performance of pathogenicity predictors for BRCA1/2. Two aspects are specific to the ENIGMA challenge. First, if some of the target variants are pathogenic, the participants do not know what molecular effect originates their pathogenicity: It can be the impact on protein function, but it can also be the impact on splicing (Eccles et al., 2015). For this reason, we decided to combine predictions for these two effects in our protocol (Figure 1). A second, distinctive aspect of the challenge is that the submissions had to provide the predicted IARC 5-tier class for each variant (see Section 2.1). This is relevant since this classification is strongly related to the clinical actions associated to each class (Goldgar et al., 2008; Moghadasi et al., 2016; Plon et al., 2008) which are in turn related to factors such as the impact on the counselee or cost to the healthcare system. Collective consideration of these factors crystallizes into five decision regions (Plon et al., 2008) that are applied to the posterior probability of pathogenicity, a probability obtained after integrating different sources of clinical/



**FIGURE 6** Distribution of the variants along the BRCA1 and BRCA2 sequences. Each variant dataset used in this work is represented with a set of pins (indicating the location of each variant) and a colored surface that provides a general, smoothed view of the distribution. The different functional domains in each structure are represented with boxes; for representation purposes, BRCA1 (1863 aa) and BRCA2 (3418 aa) are displayed with the same length. The color codes for the different sets are: CAGI (lilac), SGE (green), MLR training (blue), and NN training (orange). CAGI, Critical Assessment of Genome Interpretation; MLR, multiple linear regression; NN, neural network; SGE, saturation genome editing

**TABLE 5** Overall accuracies (ACC) and MCC for our two methods (MLR and NN, with and without splicing) and the general methods (PMut, PolyPhen-2, PON-P2, and SIFT) in the CAGI dataset

(A) BRCA1								
IARC 5-tier	MLR	MLR+AS	NN	NN+AS	PMut	PolyPhen-2	PON-P2	SIFT
ACC	0.326	0.347	0.201	0.222	0.028	0.208	0.014	0.049
MCC	-0.041	0.006	0.015	0.056	-0.002	0.031	0	0.021
Three Class	MLR	MLR+AS	NN	NN+AS	PMut	PolyPhen-2	PON-P2	SIFT
ACC	0.764	0.785	0.25	0.271	0.035	0.354	0.014	0.118
MCC	-0.237	0.354	-0.012	0.055	0.026	0.136	0	0.123
(B) BRCA2								
IARC 5-tier	MLR	MLR+AS	NN	NN+AS	PMut	PolyPhen-2	PON-P2	SIFT
ACC	0.161	0.167	0.345	0.351	0.144	0.305	0.011	0.034
MCC	-0.109	-0.068	-0.017	-0.006	-0.029	0.078	0	0.017
Three Class	MLR	MLR+AS	NN	NN+AS	PMut	PolyPhen-2	PON-P2	SIFT
ACC	0.931	0.931	0.69	0.695	0.184	0.431	0.011	0.086
MCC	0.18	0.277	0.185	0.213	-0.013	0.125	0	0.022

Abbreviations: ACC, accuracy; AS, affected splicing; IARC, International Agency for Research on Cancer; MCC, Matthews correlation coefficient; MLR, multiple linear regression; NN, neural network.

biomedical evidence. In our case, this probability was estimated using only molecular information; nonetheless, to adapt our output to the CAGI requirements we directly applied the ENIGMA boundaries (Sections 2.3.1 and 2.3.2, "CAGI output"). We computed our performances on the basis of this assignment; however, we also obtained the performances for a simplified version of the ENIGMA classification, separately collapsing its neutral and pathogenic classes (Table 2).

Assessment of the results obtained (Figure 5; Figure S2; Tables 4 and 5) shows some clear trends. For the 5-class problem, all the methods (both ours and the general methods) have poor per class performances; however, our methods are more successful at reproducing the compositional bias of the sample and outperform general methods for the most abundant classes (1 and 2) in BRCA1/2, with only one exception, for Class 2 in BRCA2, both PolyPhen-2 and PMut surpass MLR+AS; our methods also have a better distribution of wrong predictions among classes, because they tend to cluster nearby the correct class. These trends are reinforced when reducing the number of classes from five to three. Overall, the results for the CAGI challenge show that our methods can identify low-risk variants

with an accuracy higher than that of general methods, a desirable property for HBOC diagnosis (Ernst et al., 2018). Part of this improved performance could be attributed to an unequal effect of applying the ENIGMA decision boundaries to the posterior probability generated by general methods. We believe that this mapping procedure may play a role, but not a determining one since comparison of the original, binary predictions of the general methods with those of the binary versions of our tools (MLR scores binarized as explained in Section 3.2) gives a similar result (Table 6) again. MLR+AS has the top specificities for BRCA1/2 and high sensitivities; NN+AS has the same sensitivities but lower specificities, nonetheless these are only surpassed by PMut.

In summary, we have applied the protein-specific approach to building a pathogenicity predictor for BRCA1/2 variants, using either clinical phenotypes or endophenotypes. The results obtained from our methods indicate that this approach can contribute to improving our ability to discriminate between high- and low-risk variants for BRCA1/2. Of particular interest is the MLR+AS tool, because it gives an estimate of the molecular impact of a sequence replacement that is easy to interpret because it corresponds to an *in silico* version

**TABLE 6** Binary performances (sensitivities and specificities) for our predictors and the general predictors (PMut, PolyPhen-2, PON-P2, SIFT)

(A) BRCA1							
	MLR+AS	NN+AS	CADD	PMut	PolyPhen-2	PON-P2	SIFT
Sensitivity (P = 11)	0.909	0.909	1	0.818	0.727	1	1
Specificity (N = 131)	0.977	0.718	0.456	0.817	0.557	0.188	0.435
(B) BRCA2							
	MLR+AS	NN+AS	CADD	PMut	PolyPhen-2	PON-P2	SIFT
Sensitivity (P = 5)	0.8	0.8	1	0.6	1	1	0.8
Specificity (N = 167)	0.97	0.886	0.533	0.958	0.653	0.625	0.731

Abbreviations: AS, affected splicing; MLR, multiple linear regression; N, neutral; NN, neural network; P, pathogenic.



of the HDR assay. Participation in the CAGI experiment has allowed us to obtain independent estimates of the performance of our predictors, to compare them with other predictors and to help us clarify the classification level at which in silico tools could be useful for HBOC diagnosis. This participation has also underlined the role that splicing predictions can play in the correct annotation of BRCA1/2 variants, particularly when integrated into protocols that combine different views of a variant's impact.

#### ACKNOWLEDGMENTS

This work was supported by grants BIO2012-40133 and SAF2016-80255-R from the Spanish Ministry of Economy and Competitiveness (MINECO) and Pirepred (EFA086/15) from the Interreg program POCTEFA, supported by the European Regional Development Fund (ERDF) to X. d.C.; FIS P12/02585 and P15/00355 to O. D.; P13/01711 and P16/01218 to S. G-E. from the Spanish Instituto de Salud Carlos III (ISCIII) funding, an initiative of the Spanish Ministry of Economy and Innovation partially supported by European Regional Development FEDER Funds. S. B. and S. G-E. are supported by the Asociación Española contra el Cáncer [AECC] and the Miguel Servet Program [CP16/00034], respectively. The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650.

#### ORCID

Natàlia Padilla  <http://orcid.org/0000-0002-7872-2279>  
 Alejandro Moles-Fernández  <http://orcid.org/0000-0003-0252-6084>  
 Gemma Montalban  <http://orcid.org/0000-0002-6958-4759>  
 Selen Özkan  <http://orcid.org/0000-0002-7398-1351>  
 Lars Ootes  <http://orcid.org/0000-0001-6954-1620>  
 Sandra Bonache  <http://orcid.org/0000-0002-7395-8143>  
 Orland Diez  <http://orcid.org/0000-0001-7339-0570>  
 Sara Gutiérrez-Enríquez  <http://orcid.org/0000-0002-1711-6101>  
 Xavier de la Cruz  <http://orcid.org/0000-0002-9738-8472>

#### REFERENCES

- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), 248–249. <https://doi.org/10.1038/nmeth0410-248>.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 6(5), 412–424. <https://doi.org/10.1093/bioinformatics/16.5.412>
- Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., ... Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(Database issue), D158–D169. <https://doi.org/10.1093/nar/gkw1099>.
- Bondi, A. (1964). van der Waals volumes and radii. *Journal of Physical Chemistry*, 68, 441–451. <https://doi.org/10.1021/j100785a001>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Crockett, D. K., Lyon, E., Williams, M. S., Narus, S. P., Facelli, J. C., & Mitchell, J. A. (2012). Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants. *Journal of the American Medical Informatics Association*, 19, 207–211. <https://doi.org/10.1136/amiajnl-2011-000309>
- Eccles, E. B., Mitchell, G., Monteiro, A. N. A., Schmutzler, R., Couch, F. J., Spurdle, A. B., ... Goldgar, D. (2015). BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance. *Annals of Oncology*, 26, 2057–2065. <https://doi.org/10.1093/annonc/mdv278>
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Ernst, C., Hahnen, E., Engel, C., Nothnagel, M., Weber, J., Schmutzler, R. K., & Hauke, J. (2018). Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Medical Genomics*, 11(1), 35. <https://doi.org/10.1186/s12920-018-0353-y>
- Fauchere, J., & Pliska, V. (1983). Hydrophobic parameters of amino acid side-chains from the partitioning of N-acetyl-amino-acid amides. *European Journal of Medicinal Chemistry*, 18, 369–375
- Ferrer-Costa, C., Orozco, M., & de la Cruz, X. (2002). Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *Journal of Molecular Biology*, 315(4), 771–786. [https://doi.org/10.1006/jmbi.2001.5255/n5002283601952556\[pilj\]](https://doi.org/10.1006/jmbi.2001.5255/n5002283601952556[pilj])
- Ferrer-Costa, C., Orozco, M., & de la Cruz, X. (2004). Sequence-based prediction of pathological mutations. *Proteins*, 57, 811–819. <https://doi.org/10.1002/prot.20252>
- Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., ... Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, 562, 217–222. <https://doi.org/10.1038/s41586-018-0461-z>
- Goldgar, D. E., Easton, D. F., Byrnes, G. B., Spurdle, A. B., Iversen, E. S., & Greenblatt, M. S. (2008). Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. *Human Mutation*, 29(11), 1265–1272. <https://doi.org/10.1002/humu.20897>
- Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28, 367–374. <https://doi.org/10.1016/j.compbiolchem.2004.09.006>
- Guidugli, L., Pankratz, V. S., Singh, N., Thompson, J., Erding, C. A., Engel, C., ... Couch, F. J. (2013). A classification model for BRCA2 DNA binding domain missense variants based on homology-directed repair activity. *Cancer Research*, 73(1), 265–275. <https://doi.org/10.1158/0008-5472.CAN-12-2081>
- Guidugli, L., Shimelis, H., Masica, D. L., Pankratz, V. S., Lipton, G. B., Singh, N., ... Couch, F. J. (2018). Assessment of the clinical relevance of BRCA2 missense variants by functional and computational approaches. *American Journal of Human Genetics*, 102(2), 233–248. <https://doi.org/10.1016/j.ajhg.2017.12.013>
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, 11, 10–18. <https://doi.org/10.1088/1751-8113/44/8/085201>
- Hart, S. N., Hoskin, T., Shimelis, H., Moore, R. M., Feng, B., Thomas, A., ... Couch, F. J. (2019). Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models. *Genetics in Medicine*, 21(1), 71–80. <https://doi.org/10.1038/s41436-018-0018-4>
- Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of*

- the United States of America, 89(22), 10915–10919. <https://doi.org/10.1073/pnas.89.22.10915>
- Hoskins, R. A., Repo, S., Barsky, D., Andreoletti, G., Moul, J., & Brenner, S. E. (2017). Reports from CAGI: The Critical Assessment of Genome Interpretation. *Human Mutation*, 38(9), 1039–1041. <https://doi.org/10.1002/humu.23290>
- Jurman, G., Riccadonna, S., & Furlanello, C. (2012). A comparison of MCC and CEN error measures in multi-class prediction. *PLoS One*, 7(8), e41882. <https://doi.org/10.1371/journal.pone.0041882>
- Karbsani, I., Maston, G. A., Love, A., Divincenzo, C., Braastad, C. D., Elzinga, C. D., ... Higgins, J. J. (2016). A standardized DNA variant scoring system for pathogenicity assessments in mendelian disorders. *Human Mutation*, 37(1), 127–134. <https://doi.org/10.1002/humu.22918>
- Karchin, R., Agarwal, M., Sali, A., Couch, F., & Beattie, M. S. (2008). Classifying variants of undetermined significance in BRCA2 with protein likelihood ratios. *Cancer Informatics*, 6, 203–216. <https://doi.org/10.4137/CIN.S618>
- Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3), 310–315. <https://doi.org/10.1038/ng.2892>
- Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7), 1073–1081. <https://doi.org/10.1038/nprot.2009.86>
- Lindor, N. M., Guidugli, L., Wang, X., Vallée, M. P., Monteiro, A. N. A., Tavtigian, S., ... Couch, F. J. (2012). A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). *Human Mutation*, 33(1), 8–21. <https://doi.org/10.1177/104398629200800406>
- López-Ferrando, V., Gazzo, A., De laCruz, X., Orozco, M., & Gelpí, J. L. (2017). PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Research*, 45(W1), W222–W228. <https://doi.org/10.1093/nar/gkx313>
- Masica, D. L., & Karchin, R. (2016). Towards increasing the clinical relevance of in silico methods to predict pathogenic missense variants. *PLoS Computational Biology*, 12(5), e1004725
- Millot, G. A., Carvalho, M. A., Caputo, S. M., Vreeswijk, M. P. G., Brown, M. A., Webb, M., ... Monteiro, A. N. A. (2012). A guide for functional analysis of BRCA1 variants of uncertain significance. *Human Mutation*, 33(11), 1526–1537. <https://doi.org/10.1002/humu.22150>
- Moghadas, S., Eccles, D. M., Devilee, P., Vreeswijk, M. P. G., & vanAsperen, C. J. (2016). Classification and clinical management of variants of uncertain significance in high penetrance cancer predisposition genes. *Human Mutation*, 37(4), 331–336. <https://doi.org/10.1002/humu.22956>
- Moles-Fernández, A., Duran-Lozano, L., Montalban, G., Bonache, S., López-Perolio, I., Menéndez, M., ... Gutiérrez-Enríquez, S. (2018). Computational tools for splicing defect prediction in breast/ovarian cancer genes: How efficient are they at predicting RNA alterations? *Frontiers in Genetics*, 9, 366. <https://doi.org/10.3389/fgene.2018.00366>
- Moreno, L., Linossi, C., Esteban, I., Gadea, N., Carrasco, E., Bonache, S., ... Balmaña, J. (2016). Germline BRCA testing is moving from cancer risk assessment to a predictive biomarker for targeting cancer therapeutics. *Clinical & Translational Oncology*, 18, 981–987. <https://doi.org/10.1007/s12094-015-1470-0>
- Niroula, A., Urolagin, S., & Vihinen, M. (2015). PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS One*, 10(2), e0117380. <https://doi.org/10.1371/journal.pone.0117380>
- Niroula, A., & Vihinen, M. (2016). Variation interpretation predictors: Principles, types, performance, and choice. *Human Mutation*, 37(6), 579–597. <https://doi.org/10.1002/humu.22987>
- Paluch-Shimon, S., Cardoso, F., Sessa, C., Balmana, J., Cardoso, M. J., Gilbert, F., ... ESMO Guidelines Committee (2016). Prevention and screening in BRCA mutation carriers and other breast/ovarian hereditary cancer syndromes: ESMO clinical practice guidelines for cancer prevention and screening. *Annals of Oncology*, 27(5), v103–v110. <https://doi.org/10.1093/annonc/mdw327>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://doi.org/10.1016/j.molcel.2012.08.019>
- Plon, S. E., Eccles, D. M., Easton, D., Foulkes, W. D., Genuardi, M., Greenblatt, M. S., ... Tavtigian, S. V. (2008). Sequence variant classification and reporting: Recommendations for improving the interpretation of cancer susceptibility genetic test results. *Human Mutation*, 29(11), 1282–1291. <https://doi.org/10.1002/humu.20880>
- Pons, T., Vazquez, M., Matey-Hernandez, M. L., Brunak, S., Valencia, A., & Izarzugaza, J. M. G. (2016). KinMutRF: A random forest classifier of sequence variants in the human protein kinase superfamily. *BMC Genomics*, 17(Suppl. 2), 396. <https://doi.org/10.1186/s12864-016-2723-1>
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., ... Reh, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5), 405–424. <https://doi.org/10.1038/gim.2015.30>
- Riera, C., Lois, S., & de laCruz, X. (2014). Prediction of pathological mutations in proteins: The challenge of integrating sequence conservation and structure stability principles. *WIREs Computational Molecular Science*, 4, 249–268
- Riera, C., Lois, S., Domínguez, C., Fernández-Cadenas, I., Montaner, J., Rodríguez-Sureda, V., & de laCruz, X. (2015). Molecular damage in fabry disease: Characterization and prediction of alpha-galactosidase A pathological mutations. *Proteins*, 83(1), 91–104. <https://doi.org/10.1002/prot.24708>
- Riera, C., Padilla, N., & de laCruz, X. (2016). The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. *Human Mutation*, 37(10), 1013–1024. <https://doi.org/10.1002/humu.23048>
- Roy, R., Chun, J., & Powell, S. N. (2012). BRCA1 and BRCA2: Different roles in a common pathway of genome protection. *Nature Reviews Cancer*, 12(1), 68–78. <https://doi.org/10.1038/nrc3181>
- Spurdle, A. B., Healey, S., Devereau, A., Hogervorst, F. B. L., Monteiro, A. N. A., Nathanson, K. L., ... Goldgar, D. E. (2012). ENIGMA-evidence-based network for the interpretation of germline mutant alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Human Mutation*, 33(1), 2–7. <https://doi.org/10.1002/humu.21628>
- Starita, L. M., Young, D. L., Islam, M., Kitzman, J. O., Gullingsrud, J., Hause, R. J., ... Fields, S. (2015). Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics*, 200(2), 413–422. <https://doi.org/10.1534/genetics.115.175802>
- Tavtigian, S. V., Deffenbaugh, A. M., Yin, L., Judkins, T., Scholl, T., Samollow, P. B., ... Thomas, A. (2006). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *Journal of Medical Genetics*, 43(4), 295–305. <https://doi.org/10.1136/jmg.2005.033878>
- Tavtigian, S. V., Greenblatt, M. S., Lesueur, F., & Byrnes, G. B. (2008). In silico analysis of missense substitutions using sequence-alignment based methods. *Human Mutation*, 29, 1329–1336. <https://doi.org/10.1002/humu.20892>
- Vallée, M. P., Di Sera, T. L., Nix, D. A., Paquette, A. M., Parsons, M. T., Bell, R., ... Tavtigian, S. V. (2016). Adding in silico assessment of potential splice aberration to the integrated evaluation of BRCA gene unclassified variants. *Human Mutation*, 37(7), 627–639. <https://doi.org/10.1002/humu.22973>
- Venkitaraman, A. R. (2014). Cancer suppression by the chromosome custodians, BRCA1 and BRCA2. *Science*, 34(6178), 1470–1475. <https://doi.org/10.1126/science.1252230>

- Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*, 13(Suppl. 4), S2. <https://doi.org/10.1186/1471-2164-13-S4-S2>
- Vihinen, M. (2013). Guidelines for reporting and using prediction tools for genetic variation analysis. *Human Mutation*, 34, 275–282. <https://doi.org/10.1002/humu.22253>
- Wei, Q., & Dunbrack, R. L. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One*, 8(7), e67863. <https://doi.org/10.1371/journal.pone.0067863>
- Yue, P., Li, Z., & Moul, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology*, 353, 459–473. <https://doi.org/10.1016/j.jmb.2005.08.020>

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Padilla N, Moles-Fernández A, Riera C, et al. *BRCA1- and BRCA2-specific in silico tools for variant interpretation in the CAGI 5 ENIGMA challenge*. *Human Mutation*. 2019;40:1593–1611. <https://doi.org/10.1002/humu.23802>

## Article 9

### Multigene panel testing beyond BRCA1/2 in breast/ovarian cancer Spanish families and clinical actionability of findings

Bonache S, Esteban I, Moles-Fernández A, Tenés A, Duran-Lozano L, Montalban G, Bach V, Carrasco E, Gadea N, López-Fernández A, Torres-Esquiús S, Mancuso F, Caratú G, Vivancos A, Tuset N, Balmaña J, Gutiérrez-Enrriquez S, Diez O.

**J Cancer Res Clin Oncol.** 2018 Dec;144(12):2495-2513. doi: 10.1007/s00432-018-2763-9.



increased risk of OC (Nakonechny and Gilks 2016). Finally, published genome-wide association studies provide evidence for approximately a 100 of common variants with low penetrance, conferring breast cancer risks below 1.5 times the risk in the general population (Easton et al. 2015; Fachal and Dunning 2015). Yet the contribution of variants in high, moderate and common low penetrance account for only 40–50% of the familial relative risk (Fachal and Dunning 2015). Therefore, it is very likely that there are other genes associated with HBOC.

Compared to the US (Susswein et al. 2016; Buys et al. 2017; Couch et al. 2017; Kurian et al. 2017), smaller European HBOC cohorts have been tested using massively parallel sequencing to estimate the prevalence of pathogenic variants in high- and moderate-risk genes (Castéra et al. 2014; Schroeder et al. 2015; Lhota et al. 2016; Eliade et al. 2017; Feliubadaló et al. 2017; Kraus et al. 2017; Tavera-Tapia et al. 2017; Tedaldi et al. 2017). Our aim was to identify deleterious variants in high and moderate cancer penetrance genes and describe their clinical actionability, as well as, to determine the genetic profile of potentially associated genes (the so-called candidate genes), in a cohort of HBOC *BRCA1/2* negative Spanish families.

## Patients and methods

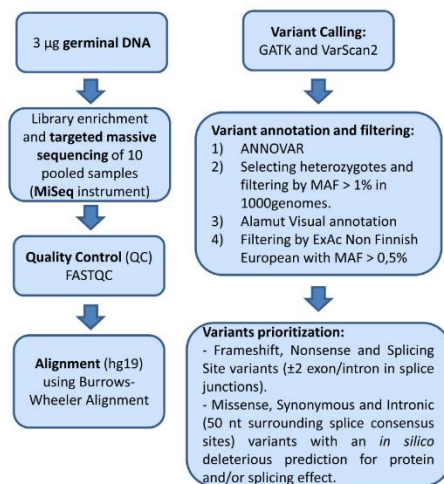
### Patients

The study included a total of 192 unrelated index cases from breast cancer high-risk Spanish families ascertained through the unit of familial cancer of Vall d'Hebron Hospital from Barcelona: 77 (40%) had BC at a young age (<36 years), 60 (31%) had BC and belonged to a family with two or more relatives with BC or OC, 38 probands (20%) had OC (8 of them also had BC or endometrial cancer), seven (4%) had BC after 36 with one first/second degree relative with BC, OC or pancreatic cancer, six (3%) had two BC (bilateral or ipsilateral) regardless of BC family history and four (2%) had colon, endometrium, sarcoma or stomach cancer and a BC/OC family history. All index cases were previously screened for single nucleotide variants and large rearrangements in *BRCA1* and *BRCA2* genes and no pathogenic variant was identified. All were afterwards enrolled for panel testing between January 2013 and December 2015 and received genetic counselling and signed informed consent for the research study, approved by the Clinical Research Ethics Committee of the Hospital Vall d'Hebron. Clinical data including personal and family cancer histories, tumour histology and receptor status of breast tumours were collected through medical chart review. Confirmation of cancer among first- and/or second-degree relatives was obtained whenever possible.

### Massively parallel sequencing and variant classification

Ninety-seven genes were included in our research panel. Thirty-four genes were well-known high/moderate-risk cancer genes (16 related to BC/OC and 18 associated to other cancers), and 63 were candidate genes with only initial evidence of cancer risk association and/or related to DNA repair (Supplementary Table 1 lists all the sequenced genes, also providing reference to publications that point to the potential link between each candidate gene and the familial breast and ovarian cancer predisposition). The protocols and methodology for DNA extraction, capture design, library preparation, sequencing, data alignment, variant calling and variant prioritization are described in detail in Supplementary Methods and summarized in Fig. 1.

The frameshift, nonsense and exonic or intronic variants with a RNA analysis data that indicated a protein impact, were classified as pathogenic for the risk-associated genes but as deleterious for candidate genes. For a detailed RNA evaluation protocol, see qualitative and quantitative cDNA analysis section in Supplementary Methods. Missense variants with well-known reported clinical effect were also classified as pathogenic for the risk-associated genes. Variants were classified as variants of uncertain significance (VUS) if no functional data were available or the risk was not clearly established according literature or gene databases. These



**Fig. 1** Sequencing platform, bioinformatics and variant prioritization pipeline

in Table 1 and complete in silico predictions as well as other relevant annotations are in Supplementary Table 4. The highest number of pathogenic variants, excluding heterozygous *MUTYH* variants, was in *PALB2* (four variants, three of them novel) and *ATM* (three variants, one not previously described). These were all identified in families with BC and no OC cases. One out of *PALB2* variants (c.3201+5G>T) alters the splicing process through an imbalanced expression of natural RNA isoforms (Table 1). Results obtained from RNA analysis confirmed a splicing alteration consisting of an imbalanced expression of several *PALB2* alternative RNA isoforms (Duran-Lozano et al. 2018). The variant up-regulates isoforms  $\Delta 11, 12$  (in-frame) and  $\Delta 11$  (frameshift), and down-regulates isoform  $\Delta 12$  (frameshift). All transcripts are predicted to encode for non-functional proteins. Isoform  $\Delta 11, 12$  presumably contributes to variant pathogenicity by encoding a *PALB2* protein lacking 79aa of the WD40 domain that mediates direct interactions between *PALB2* and key proteins involved in homologous recombination. Semi-quantitative and quantitative analysis of *PALB2* full-length transcript indicated haploinsufficiency in carriers (Duran-Lozano et al. 2018). One stop gain variant in *PTEN* was found in a patient with BC diagnosed at 46, a suggestive Cowden syndrome and a family history of BC/OC. Two *TP53* pathogenic variants (c.587G>C, p.Arg196Pro, and c.783-1G>A) were found in probands with early onset BC (before age 30) and the absence of Li–Fraumeni family history. The missense variant is predicted to be deleterious by three bioinformatics in silico tools (Supplementary Table 4) and it is placed at the DNA-binding domain of TP53 ([https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?seqinpu=NP\\_000537.3](https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?seqinpu=NP_000537.3)). The effect of the splicing variant in *TP53* was confirmed by RNA analysis, showing the retention of intron 7 (r.782+1\_783-1ins) and deletion of the first 24 nucleotides of exon 8 (r.783\_806del) as aberrant transcripts, which would encode a truncated and in frame proteins, respectively (Table 1; Fig. 3a). Semi-quantitative QIAxcel CE data showed > 0.5 reduction of FL transcript levels in the carrier compared to controls, suggesting that the variant allele is unable to produce FL transcript (Fig. 4b). Splicing fraction (SF) estimation showed that the isoform retaining intron 7 was the most expressed (54.4%), whereas the isoform lacking the first 24 nucleotides of exon 8 was present in a 14.5% (Fig. 4c). Both *TP53* variants (c.587G>C, p.Arg196Pro, and c.783-1G>A) are listed in the IARC *TP53* database (<http://p53.iarc.fr/TP53GeneVariations.aspx>, R18) (Bouaoun et al. 2016) reported in Li–Fraumeni and Li–Fraumeni like families, and in ClinVar as pathogenic or likely pathogenic.

Regarding OC-associated genes, two *RAD51D* variants were identified in two OC patients (c.94\_95delGT, c.694C>T), and one pathogenic variant (c.1702\_1703delAA)

in *BRIP1* was identified in a young onset (32y) BC patient without OC family history.

The *PMS2* variant c.989-2A>G was identified in a proband with personal and family history of BC who did not meet Amsterdam or Bethesda criteria. The splicing effect, confirmed by RNA analysis (Table 1; Fig. 3b), results in an in-frame exon 10 skipping (r.989\_1144del), with the loss of part of the *PMS2* dimerisation domain. Semi-quantitative analysis showed a 0.5 reduction of the FL transcript in the carrier (Fig. 4b) compared to controls. Splicing fraction estimation showed that the  $\Delta 10$  isoform was higher expressed (58%) than the FL (41.8%) in the carrier allele (Fig. 4c). The same splicing alteration was also obtained by Borràs et al. (Borràs et al. 2013) and it appears as likely pathogenic in InSIGHT (The International Society for Gastrointestinal Hereditary Tumours) database. The same patient carries a frameshift variant (c.580\_581del) in *BARD1* (Supplementary Fig. 1), a BC candidate gene whose protein interacts with *BRCA1*. The proband's mother, diagnosed with bilateral BC, is an obligate carrier of both *BARD1* and *PMS2* variants (data not shown).

Three monoallelic pathogenic variants in *MUTYH* were found in seven patients, the novel c.1101dup, and the recurrent c.536A>G (p.Tyr179Cys) and c.1187G>A (p.Gly396Asp), also known as p.Tyr165Cys and p.Gly382Asp (Lipton and Tomlinson 2004), respectively. The expected carrier rate for *MUTYH* monoallelic pathogenic variants in healthy controls of 1.5–2% (Nielsen et al. 2011) is lower than that observed in our series of 192 HBOC patients (3.6%).

The *APC* moderate risk variant for colorectal cancer in people of Ashkenazi Jewish decent c.3920T>A (p.Ile1307Lys) (Liang et al. 2013) was found in two BC patients. This variant is present in 0.3% of the Spanish population (CIBERER Spanish Variant Server, <http://csvs.babelomics.org/>) and has recently been associated with BC (Leshno et al. 2016). The *CHEK2* variant c.470T>C (p.Ile157Thr), a founder variant in Northern European populations and considered a low penetrance variant for BC (Han et al. 2013), was found in one BC family originally from East Germany.

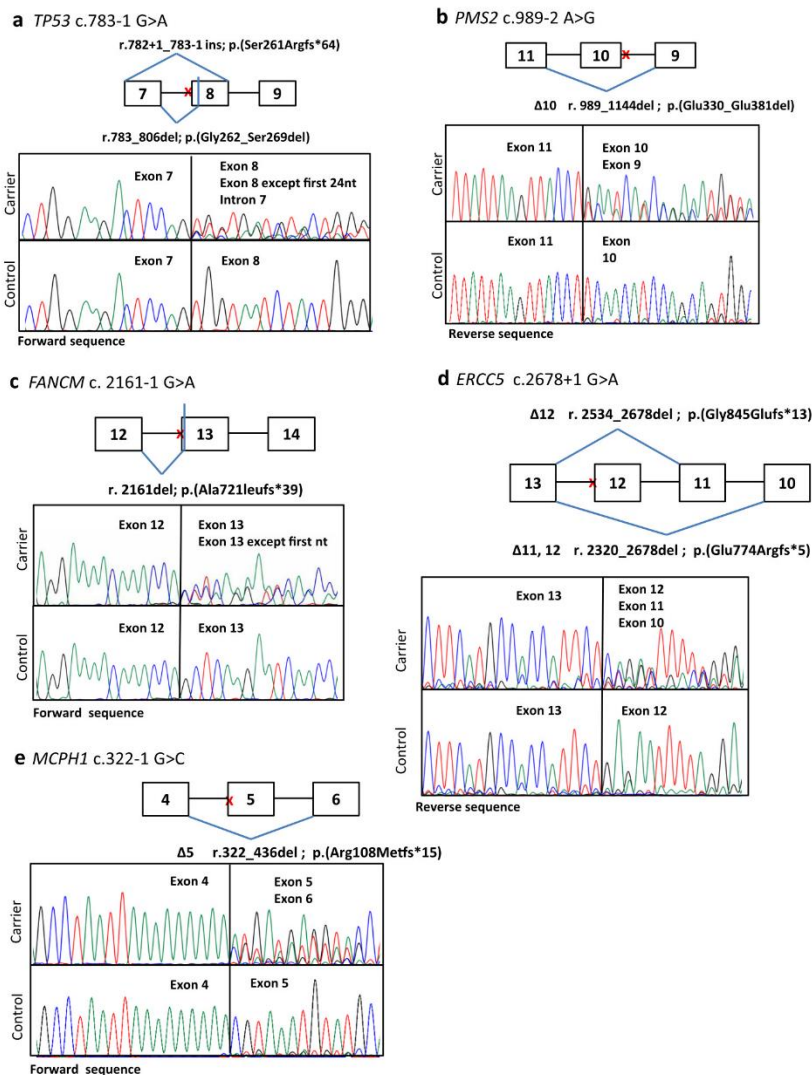
In the 34 risk-associated cancer genes, from a total of 427 unique variants we categorized 383 as VUS (Fig. 2). After an in silico analysis, literature and database revision, only 8% (35/427) were prioritized as deleterious (Table 2, Supplementary Table 5). The genetic characteristics, familial phenotype, as well as published data for these prioritized VUS are described in Supplementary Data. The remaining 82% in silico non-prioritized VUS (348/427) may be considered as either simply VUS or likely benign variants (Supplementary Data and Supplementary Table 6). The number of VUS prioritized and non-prioritized for each known-predisposition cancer gene is shown in Fig. 5, where the genes

Table 1 (continued)

Gene	Patient ID	Effect	cNomen	pNomen	Transcripts	rsId	rMAF/EA/ NFE/MAF/ ESP/EA/MAF/FEq/Spansish	Proband diagnosis	Family history
<i>RAD51D</i>	15-312	n	c.694C>T	p.(Arg232*)	NM_002878.3	rs587780104	-0.000019/-/-	OC P5 60	mt c BC42
Lynch syndrome									
<i>PM52</i>	11-204	spI	c.989-2A>G <sup>d</sup>	p.(Gln330_Gln381del)	NM_000535.5	rs587779347	-/-/-/-	BC IDC G2 ER+PR+27, BC IDC TN 42	m BC 44, BC 44, mt u SI 47, mt gm ENDO 40, BC 90
MUTHY monallelic									
<i>MUTYH</i>	11-018	m	c.536A>G	p.(Tyr179Cys)	NM_001128425.1	rs34612342	0.0002/0.002/0.003/0.003	BC IDC G3 ER+PR+33	s BC 48
<i>MUTYH</i>	15-499	m	c.536A>G	p.(Tyr179Cys)	NM_001128425.1	rs34612342	0.0002/0.002/0.003/0.003	BC IDC G2/3 ER+PR+HER2-40, BC G2 ER+PR+HER2+47, THY follicular 55	f PANC 71
<i>MUTYH</i>	14-155	fr	c.1101dupC	p.(Arg368Glnfs*164)	NM_001128425.1	rs708130289	-/-/-/-	OC P5 41	m BC 64, mt gm OC 69
<i>MUTYH</i>	15-292	fr	c.1101dupC	p.(Arg368Glnfs*164)	NM_001128425.1	rs708130289	-/-/-/-	BC IDC ER+PR+HER2+34	-
<i>MUTYH</i>	08-048	m	c.1187G>A	p.(Gly396Asp)	NM_001128425.1	rs36053993	0.002/0.004/0.005/0.011	OC CC 41	b CO 51, CO 52, pt a, BC 80, pt a, CO 59, pt a, CO 63, pt c, OC 59, CO 61, pt c, ENDO 47, pt c, CO 46
<i>MUTYH</i>	15-0334	m	c.1187G>A	p.(Gly396Asp)	NM_001128425.1	rs36053993	0.002/0.004/0.005/0.011	BC IDC 49, BC DCIS 49, BC IDC 63, BC DCIS 63	d BC 40, b PR 59, b PR 70, pt a BC 80, pt c BC 59, pt c BC 28
Moderate colorectal cancer risk									
<i>APC</i>	13-148	m	c.1187G>A	p.(Gly396Asp)	NM_001128425.1	rs36053993	0.002/0.004/0.005/0.011	BC IDC TN 32, BC IDC ER+PR+HER2+52	mt gm BC 49, mt c BC 38
<i>APC</i>	12-070	m	c.3920T>A	p.(Ile1307Lys)	NM_001127510.2	rs1801155	0.0011/0.003/0.016/0.003	BC IDC G3 ER+PR+HER2-33	f 5 polyps, pt a BC 50
<i>APC</i>	15-123	m	c.3920T>A	p.(Ile1307Lys)	NM_001127510.2	rs1801155	0.0011/0.003/0.016/0.003	BC LLC ER+PR+HER2+67, BC IDC ER+PR+HER2+67	s BC 40, pt a BC

MAF minor allele frequency, EA: Exome Aggregation Consortium, NFE Non-Finnish European, ESP exome sequencing project, EA European American, FEq. Spanish MAF from CIBERER Spanish Variant Server, <http://cvs.babelomics.org/>, OV ovarian cancer, CC clear cell, BC breast cancer, PANC pancreatic cancer, PS papillary serous, CO colon, CNS central nervous system, SI small intestine, TC transitional cell, mt maternal, pt paternal, m mother, f father, s sister, b brother, d daughter, gm grandmother, gf grandfather, c cousin, a aunt, n nephew/niece, ga great aunt, fr frameshift, n nonsense, m missense, spI splicing





**Fig. 3** Results of in vitro mRNA analysis of five variants leading to aberrant splicing. Sequencing results are shown for the patient carrying the variant and a negative control. A red X illustrates the position of the variant. Nucleotide and exon numbering for cDNA

is based on NCBI entries NM\_000546.4 (*TP53*), NM\_000535.5 (*PMS2*), NM\_020937.3 (*FANCM*), NM\_000123.3 (*ERCC5*) and NM\_024596.3 (*MCPH1*)



**Table 2** Variants of unknown significance prioritized by in silico analysis and the database or literature revision, in risk-associated cancer genes

Gene	Patient ID	Effect	c Nomen	p Nomen	Transcript	rsid	rMAF/ExAc; NFE MAF/ESP; EA; MAF/PEq; Spanish	Proband diagnosis	Proband diagnosis	Family history
<b>High breast cancer risk</b>										
<i>CDH1</i>	I6-03887 <sup>a</sup>	m	c.1406C>T	p.(Ser469Phe)	NM_004360.3	-	-/-/-	OC 38, ENDO 38	OC PS 38, ENDO 38	s OC 51
<i>CDH1</i>	15-165	m	c.2558C>T	p.(Ser853Leu)	NM_004360.3	rs569928380	0.0002/0.000015/-/0.003	OC 28	OC PS 28	-
<i>CDH1</i>	I6-03887 <sup>a</sup>	m	c.2590G>A	p.(Glu864Lys)	NM_004360.3	rs142927667	-/0.000015/-/-	OC 38, ENDO 38	OC PS 38, ENDO 38	s OC 51
<i>CDH1</i>	14-266	m	c.2644G>A	p.(Asp882Asn)	NM_004360.3	rs200104963	-/-/-	BC 29	BC DCIS 29	mt c STO 52, (DGC) m THY 45
<i>PALB2</i>	15-356	m	c.101G>A	p.(Arg34His)	NM_024675.3	rs144944814	-/0.000015/-/-	BC 30	BC IDC ER+PR-HER2-30	m BC 44, BC 44, mt u SI 47, mt gm ENDO 40, BC 90
<i>PALB2</i>	11-204	m	c.2989G>T	p.(Asp997Tyr)	NM_024675.3	-	-/-/-	BC 27, 42	BC IDC G2 ER+PR-27, BC IDC TN 42	m BC 60, b PR 60, n PANC 45
<i>TP53</i>	13-519	m	c.467G>A	p.(Arg156His)	NM_000546.4	rs371524413	-/3e-05/0.000116/-	CO 64	CO 64	m BC 60, mt a BC 59, mt gm BC 75
<i>TP53</i>	15-027	m	c.869G>A	p.(Arg290His)	NM_000546.4	rs55819519	0.0002/0.0002399/0.000349/-	OC 53	OC PS 53	-
<i>STK11</i>	15-03325	m	c.662C>T	p.(Pro221Leu)	NM_000455.4	-	-/-/-	BC 35	BC IDC TN 35	m BC 60, mt a BC 59, mt gm BC 75
<b>Moderate breast cancer risk</b>										
<i>ATM</i>	15-242	m	c.2932T>C	p.(Ser978Pro)	NM_000051.3	rs139552233	0.001/0.001/0.000465/-	BC 52	BC IDC ER+PR+HER2-52	s BC 48, m BC 53, mt c BC 45
<i>ATM</i>	14-324	pot spl	c.3153+4A>G	p.?	NM_000051.3	-	-/-/-	BC 33	BC IDC ER+PR+HER2-33	-
<i>ATM</i>	15-03582	pot spl	c.3993+5G>T	p.?	NM_000051.3	rs3092842	0.003/0.000015/-/-	BC 24, 29	BC IDC TN 24, BC IDC TN 29	-
<i>ATM</i>	14-325	m	c.4148C>T	p.(Ser1383Leu)	NM_000051.3	rs141087784	-/0.00006/0.000233/-	BC 60	BC ILC ER+PR+HER2-60	s BC 45, m BC 45

**Table 2** (continued)

Gene	Patient ID	Effect	c Nomen	p Nomen	Transcript	rsId	rsMAF/ExAc EA, MAF/TrEq, Spanish	Proband diagnosis	Proband diagnosis	Family history
<i>POLE</i>	15-251	m	c.139G>A	p.(Val437Met)	NM_006231.3	rs115047349	-	OC 50	OC PS 50	mt in PANC 68
<i>POLE</i>	15-292	m	c.2021A>C	p.(Glu674Ala)	NM_006231.3	-	-/-/-	BC 34	BC IDC ER+PR+HER2+34	-
Other syndromic cancer genes										
<i>NF1</i>	14-101	syn	c.3498C>T	p.=	NM_001042492.2	rs2066733	0.004/0.00006/0.000116/-	OC 42	OC PS 42	pt a BC 60
<i>NF1</i>	12-225	pot spl	c.4578-13T>G	p.?	NM_001042492.2	-	-/-/-	BC 48	BC ILC ER+PR+HER2+48	m BC 39,77
<i>NF1</i>	13-310	m	c.8144C>T	p.(Ala2715Val)	NM_001042492.2	-	-/-/-	BC32	BC IDC ER+PR+HER2-32	-
<i>NF2</i>	14-443	m	c.1252C>T	p.(Arg418Cys)	NM_000268.3	rs765540111	-/0.000018/-/-	BC 50, OC 60	BC IDC ER+PR+HER2-50, OC PS 60	-
<i>RET</i>	14-024	m	c.1780C>T	p.(His594Tyr)	NM_020975.4	rs778622905	-/-/-	BC 58	BC IDC ER+PR+58	s BC 56, s BC 74, b PR 49, s LYM, pt a MEL, 67, BL71, pt c BC 45, pt c BC 52, pt c ENDO 51
<i>WT1</i>	13-381	m	c.1048T>C <sup>a</sup>	p.(Cys350Arg)	NM_024426.4	rs142059681	-/0.000660/0.000814/0.001	BC 33	BC IDC TN 33	s CNS 30
<i>WT1</i>	14-161	m	c.1048T>C <sup>a</sup>	p.(Cys350Arg)	NM_024426.4	rs142059681	-/0.000660/0.000814/0.001	BC 34	BC TN 34	pt a BC 50
<i>WT1</i>	15-134	m	c.830G>A	p.(Cys277Tyr)	NM_024426.4	-	-/-/0.001	BC 39	BC IDC ER+PR+HER2+39	m BC 54, mt a BC 63

MAF minor allele frequency, ExAc Exome Aggregation Consortium, NFE Non-Finnish European, ESP exome sequencing project, EA European American, FrEq. Spanish, MAF from CIBERER Spanish Variant Server, <http://cavs.babelomics.org/>, DGC diffuse gastric cancer, OV ovarian cancer, CC clear cell, BC breast cancer, A41, PANC pancreatic cancer, PS papillary serous, CO colon, CNS central nervous system, SI small intestine, TC transitional cell, mt maternal, pt paternal, m mother, f father, s sister, b brother, d daughter, gm grandmother, gf grandfather, c cousin, u uncle

<sup>a</sup>This patient carries three different VUS

<sup>b</sup>Variant previously described in Spanish BC/OV families (Tavera-Tapia et al. 2017)

<sup>c</sup>Same variant is present in two different patients

**Table 3** Overview of surveillance and/or prevention options carried out in the patients with high and moderate pathogenic variants

Gene	Patient ID	Proband actionability	Cascade testing in relatives/at risk (53/71)
High breast cancer risk			
<i>PALB2</i>	14-424	BC screening (MRI)	Yes 3 out of 4
<i>PALB2</i>	13-352	BC screening (MRI)+RRM	Yes 11 out of 14
<i>PALB2</i>	12-336	BC screening (MRI)	Yes 5 out of 8
<i>PALB2</i>	13-051	NA (deceased)	Yes 5 out of 5
<i>PTEN</i>	14-006	RRM + hysterectomy	No
<i>TP53</i>	13-412	LF surveillance + reproductive decision-making	Yes 3 out of 4
<i>TP53</i>	13-331	LF surveillance	Yes 1 out of 3
Moderate breast cancer risk			
<i>ATM</i>	14-172	NA (previous bilateral mastectomy)	Yes 1 out of 1
<i>ATM</i>	14-171	BC screening (MRI)	Yes 3 out of 5
<i>ATM</i>	14-086	BC screening (MRI)	None out of 2
Ovarian cancer risk			
<i>BRIP1</i>	14-430	OC risk assessment	Yes 5 out of 6
<i>RAD51D</i>	14-530	NA (previous oophorectomy for prior OC)	Yes 4 out of 7
<i>RAD51D</i>	15-312	NA (previous oophorectomy for prior OC)	Yes 4 out of 4
Lynch syndrome			
<i>PMS2</i>	11-204	LS surveillance	Yes 8 out of 8
Moderate colorectal cancer risk			
<i>APC</i>	12-070	CRC screening	No relatives at risk
<i>APC</i>	15-123	CRC screening	No relatives at risk
MUTYH monoallelic			
<i>MUTYH</i>	11-018	Pending	NA
<i>MUTYH</i>	15-499	Pending	NA
<i>MUTYH</i>	14-155	Pending	NA
<i>MUTYH</i>	15-292	Pending	NA
<i>MUTYH</i>	08-048	Pending	NA
<i>MUTYH</i>	15-03334	Pending	NA
<i>MUTYH</i>	13-148	Pending	NA
Low risk cancer			
<i>CHEK2</i>	08-147	NA	NA

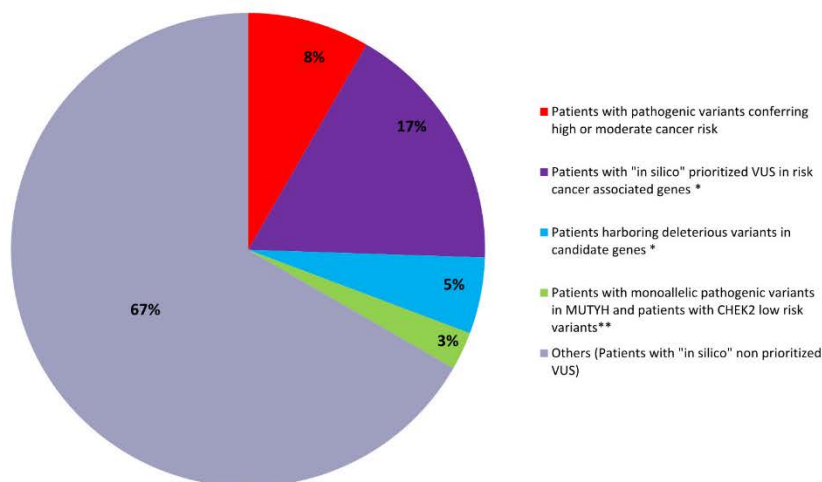
The number of predictive tests is also shown

BC breast cancer, MRI magnetic image resonance, RRM risk reducing mastectomy, NA not applicable, LF Li–Fraumeni, OC ovarian cancer, LS Lynch syndrome, CRC colorectal cancer

variants was found in the *PALB2* and *ATM* genes, both previously associated with BC, and reinforces the role of these two genes as essentially BC risk genes (Easton et al. 2015; Tung et al. 2016). Our results add clinical evidence of the benefit of sequencing through panel testing of several BC/OC susceptibility genes compared to the strategy of sequential testing. Our study also demonstrates the utility of multigene panels in patients who previously underwent non-informative *BRCA1/2* genetic screening. The panel approach provides a high number of variants of unknown significance that require the use of a prioritization system to select those with the highest probability of being associated with risk. In our study, we used in silico and database information to prioritize 35 variants in known cancer genes that merit

additional studies to unequivocally define them as pathogenic. Excluding the two patients who carried concurrently a VUS and a pathogenic variant in high- or moderate-risk cancer genes, 17% of the patients (33/192) harboured a prioritized VUS (Fig. 6).

Identification of deleterious variants in *ATM*, *BRIP1*, *PALB2*, *PMS2*, *PTEN*, *RAD51D*, *TP53* and *APC* genes had a clinical impact, resulting in a change in the medical management of the probands and/or cascade testing in relatives. Overall, 12 out of 16 (75%) of the variants identified in high and moderate penetrance genes were clinically actionable. Other studies have reported an actionability of pathogenic variants in new genes between 69 and 91%, depending on the criteria used to define clinical actionability (Eliade et al.



**Fig. 6** Patient distribution according to the variant classification. \*Not included patients with pathogenic variants. \*\*Not included patients with potentially pathogenic VUS in risk cancer-associated genes

2017; Frey et al. 2017). In this regard, we did not consider actionable the *CHEK2* c.470T>C low penetrance variant, and the heterozygous *MUTYH* variants had not been disclosed at the time of submission. There has been much debate over whether *MUTYH* heterozygotes also have an increased risk of developing colorectal cancer (CRC) or other types of tumours such as BC (Nielsen et al. 2011; Win et al. 2016). Starting colonoscopy at age 40 every 5 years has been proposed and there are no recommendations for BC screening.

Some of the variants we detected would have not been identified without the multiplex testing. The *PMS2* c.989-2A>G variant was found in a family with mainly BC, who did not meet Amsterdam or Bethesda criteria. Recent studies have found pathogenic variants in *PMS2* in BC families undergoing panel testing, suggesting that the use of this methodology might expand the *PMS2*-associated cancer risks (Ten Broeke et al. 2015; Eliade et al. 2017; Espenschied et al. 2017). A patient with early onset BC was found to have the *BRIP1* c.1702\_1703delAA variant. Large studies have established *BRIP1* as being associated with a moderately increased risk of epithelial OC, but association with BC risk is not robust (Ramus et al. 2015; Easton et al. 2016; Couch et al. 2017). However, the *BRIP1* c.1702\_1703delAA variant had previously been associated with significant risk of both OC and BC in Spanish patients and it was also identified in one individual with lung cancer (LC) out of 2,758 Spanish individuals with other cancer types (Rafnar



et al. 2011). Additionally, this variant was also identified in one out of 40 unrelated Spanish CRC patients with strong CRC familial aggregation (Esteban-Jurado et al. 2016). In our study, this variant was identified in the proband's father diagnosed with LC and one paternal aunt with CRC. Overall, c.1702\_1703delAA might be a Spanish founder allele that deserves further research on its association with different cancer types. Neither of the two families with *TP53* mutations fulfilled the 2009 Chompret criteria (Tinat et al. 2009). When these criteria were revised in 2015 it was suggested to consider women diagnosed with BC before the age of 31 to be eligible for *TP53* testing. Our results reinforce the application of these criteria, feasible through panel testing, especially if it impacts the patient's medical management. The identification of secondary findings is a challenge for health care professionals in two different ways: one is the difficulty in determining the best screening for a patient with a pathogenic variant in the absence of the classical phenotype (Rana et al. 2018), and the other one is helping the patient understand and adapt to the implications of this findings during the pre-test and post-test genetic counselling. New prospective studies are warranted to update the cancer spectrum and cancer risk of mutation carriers identified in settings that do not resemble the classical phenotype, as well as the psychological impact associated with these findings. In addition, the health care professional needs to discuss the possibility of finding VUS or moderate penetrance variants, which have been associated with increased uncertainty and



- Cardoso M, Paulo P, Maia S, Teixeira MR (2016) Truncating and missense PPM1D mutations in early-onset and/or familial/hereditary prostate cancer patients. *Genes Chromosom Cancer* 55:954–961. <https://doi.org/10.1002/gcc.22393>
- Castéra L, Krieger S, Rousselin A et al (2014) Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. *Eur J Hum Genet* 22:1305–1313. <https://doi.org/10.1038/ejhg.2014.16>
- Chinnadurai G (2006) CtIP, a candidate tumor susceptibility gene is a team player with luminaries. *Biochim Biophys Acta Rev Cancer* 1765:67–73. <https://doi.org/10.1016/j.bbcan.2005.09.002>
- Couch FJ, Shimelis H, Hu C et al (2017) Associations between cancer predisposition testing panel genes and breast cancer. *JAMA Oncol* 3:1190–1196. <https://doi.org/10.1001/jamaoncol.2017.0424>
- Duran-Lozano L, Montalban G, Bonache S et al (2018) Alternative transcript imbalance underlying breast cancer susceptibility in a family carrying PALB2 c.3201+5G>T. *Breast Cancer Res Treat* (under second revision)
- Easton DF, Pharoah P, Antoniou AC et al (2015) Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med* 372:2243–2257. <https://doi.org/10.1038/nbt.3121>
- Easton DF, Lesueur F, Decker B et al (2016) No evidence that protein truncating variants in BRIP1 are associated with breast cancer risk: implications for gene panel testing. *J Med Genet* 53:298–309. <https://doi.org/10.1136/jmedgenet-2015-103529>
- Eliade M, Skrzypski J, Baurand A et al (2017) The transfer of multigene panel testing for hereditary breast and ovarian cancer to healthcare: what are the implications for the management of patients and families? *Oncotarget* 8:1957–1971. <https://doi.org/10.18632/oncotarget.12699>
- Espenschied CR, LaDuca H, Li S et al (2017) Multigene panel testing provides a new perspective on lynch syndrome. *J Clin Oncol* 35:2568–2575. <https://doi.org/10.1200/JCO.2016.71.9260>
- Esteban I, Vilaró M, Adrover E et al (2018) Psychological impact of multigene cancer panel testing in patients with a clinical suspicion of hereditary cancer across Spain. *Psychooncology* 27(6):1530–1537. <https://doi.org/10.1002/pon.4686>
- Esteban-Jurado C, Franch-Exposito S, Muñoz J et al (2016) The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer. *Eur J Hum Genet* 24:1501–1505. <https://doi.org/10.1038/ejhg.2016.44>
- Fachal L, Dunning AM (2015) From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. *Curr Opin Genet Dev* 30:32–41. <https://doi.org/10.1016/j.gde.2015.01.004>
- Feliubadaló L, Tonda R, Gausachs M et al (2017) Benchmarking of whole exome sequencing and ad hoc designed panels for genetic testing of hereditary cancer. *Sci Rep* 7:37984. <https://doi.org/10.1038/srep37984>
- Frey MK, Sandler G, Sobolev R et al (2017) Multigene panels in Ashkenazi Jewish patients yield high rates of actionable mutations in multiple non-BRCA cancer-associated genes. *Gynecol Oncol* 146:123–128. <https://doi.org/10.1016/j.ygyno.2017.04.009>
- Fu W, Ligabue A, Rogers KJ et al (2017) Human RECQ helicase pathogenic variants, population variation and “Missing” diseases. *Hum Mutat* 38:193–203. <https://doi.org/10.1002/humu.23148>
- Gutiérrez-Enríquez S, Bonache S, Ruíz De Garibay G et al (2014) About 1% of the breast and ovarian Spanish families testing negative for BRCA1 and BRCA2 are carriers of RAD51D pathogenic variants. *Int J Cancer* 134:2088–2097. <https://doi.org/10.1002/ijc.28540>
- Han FF, Guo CL, Liu LH (2013) The effect of CHEK2 variant I157T on cancer susceptibility: evidence from a meta-analysis. *DNA Cell Biol* 32:329–335. <https://doi.org/10.1089/dna.2013.1970>
- Kraus C, Hoyer J, Vasilieou G et al (2017) Gene panel sequencing in familial breast/ovarian cancer patients identifies multiple novel mutations also in genes others than BRCA1/2. *Int J Cancer* 140:95–102. <https://doi.org/10.1002/ijc.30428>
- Kurian AW, Hughes E, Handorf EA et al (2017) Breast and ovarian cancer penetrance estimates derived from germline multiple-gene sequencing results in women. *JCO Precis Oncol* 1–12. <https://doi.org/10.1200/PO.16.00066>
- Leshno A, Shapira S, Liberman E et al (2016) The APC I1307K allele conveys a significant increased risk for cancer. *Int J Cancer* 138:1361–1367. <https://doi.org/10.1002/ijc.29876>
- Lhota F, Zemankova P, Kleiblova P et al (2016) Hereditary truncating mutations of DNA repair and other genes in BRCA1/BRCA2/PALB2-negatively tested breast cancer patients. *Clin Genet* 90:324–333. <https://doi.org/10.1111/cge.12748>
- Liang J, Lin C, Hu F et al (2013) APC polymorphisms and the risk of colorectal neoplasia: a huge review and meta-analysis. *Am J Epidemiol* 177:1169–1179. <https://doi.org/10.1093/aje/kws382>
- Lipton L, Tomlinson I (2004) The multiple colorectal adenoma phenotype and MYH, an excision repair gene. *Clin Gastroenterol Hepatol* 2:633–638. [https://doi.org/10.1016/S1542-3565\(04\)00286-1](https://doi.org/10.1016/S1542-3565(04)00286-1)
- Llort G, Chirivella I, Morales R et al (2015) SEOM clinical guidelines in Hereditary Breast and ovarian cancer. *Clin Transl Oncol* 17:956–961. <https://doi.org/10.1007/s12094-015-1435-3>
- Mantere T, Winqvist R, Kauppila S et al (2016) Targeted next-generation sequencing identifies a recurrent mutation in MCPHI associating with hereditary breast cancer susceptibility. *PLoS Genet* 12:1–14. <https://doi.org/10.1371/journal.pgen.1005816>
- Nakonechny QB, Gilks CB (2016) Ovarian cancer in hereditary cancer susceptibility syndromes. *Surg Pathol Clin* 9:189–199. <https://doi.org/10.1016/j.path.2016.01.003>
- Nielsen M, Morreau H, Vasen HF, Hes FJ (2011) MUTYH-associated polyposis (MAP). *Crit Rev Oncol Hematol* 79:1–16. <https://doi.org/10.1016/j.critrevonc.2010.05.011>
- Paluch-Shimon S, Cardoso F, Sessa C et al (2016) Prevention and screening in BRCA mutation carriers and other breast/ovarian hereditary cancer syndromes: ESMO clinical practice guidelines for cancer prevention and screening. *Ann Oncol* 27:v103–v110. <https://doi.org/10.1093/annonc/mdw327>
- Pharoah PDP, Song H, Dicks E et al (2016) PPM1D mosaic truncating variants in ovarian cancer cases may be treatment-related somatic mutations. *J Natl Cancer Inst* 108:1–5. <https://doi.org/10.1093/jnci/djv347>
- Rafnar T, Gudbjartsson DF, Sulem P et al (2011) Mutations in BRIP1 confer high risk of ovarian cancer. *Nat Genet* 43:1104–1107. <https://doi.org/10.1038/ng.955>
- Ramus SJ, Song H, Dicks E et al (2015) Germline mutations in the BRIP1, BARD1, PALB2, and NBN genes in women with ovarian cancer. *J Natl Cancer Inst* 107:1–8. <https://doi.org/10.1093/jnci/djv214>
- Rana HQ, Gelman R, LaDuca H et al (2018) Differences in TP53 mutation carrier phenotypes emerge from panel-based testing. *JNCI J Natl Cancer Inst* 110:1–8. <https://doi.org/10.1093/jnci/djy001>
- Ruark E, Snape K, Humburg P et al (2013) Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature* 493:406–410. <https://doi.org/10.1038/nature11725>
- Schroeder C, Faust U, Sturm M et al (2015) HBOC multi-gene panel testing: comparison of two sequencing centers. *Breast Cancer Res Treat* 152:129–136. <https://doi.org/10.1007/s10549-015-3429-9>
- Slavin TP, Maxwell KN, Lilyquist J et al (2017) The contribution of pathogenic variants in breast cancer susceptibility genes to familial breast cancer risk. *NPJ Breast Cancer* 9:22. <https://doi.org/10.1038/s41523-017-0024-8>
- Suhasini AN, Brosh RMJ (2013) DNA helicases associated with genetic instability, cancer, and aging. *Adv Exp Med Biol* 767:123–144. <https://doi.org/10.1007/978-1-4614-5037-5>
- Susswein LR, Marshall ML, Nusbaum R et al (2016) Pathogenic and likely pathogenic variant prevalence among the first 10,000

- patients referred for next-generation cancer panel testing. *Genet Med* 18:823–832. <https://doi.org/10.1038/gim.2015.166>
- Tavera-Tapia A, Pérez-Cabornero L, Macías JA et al (2017) Almost 2% of Spanish breast cancer families are associated to germline pathogenic mutations in the ATM gene. *Breast Cancer Res Treat* 161:597–604. <https://doi.org/10.1007/s10549-016-4058-7>
- Tedaldí G, Tebaldí M, Zampiga V et al (2017) Multiple-gene panel analysis in a case series of 255 women with hereditary breast and ovarian cancer. *Oncotarget* 8:47064–47075. <https://doi.org/10.18632/oncotarget.16791>
- Ten Broeke SW, Brohet RM, Tops CM et al (2015) Lynch syndrome caused by germline PMS2 mutations: delineating the cancer risk. *J Clin Oncol* 33:319–325. <https://doi.org/10.1200/JCO.2014.57.8088>
- Tinat J, Bougeard G, Baert-Desurmont S et al (2009) 2009 version of the Chompret criteria for Li Fraumeni Syndrome. *J Clin Oncol* 27:108–109. <https://doi.org/10.1200/JCO.2009.22.7967>
- Tung N, Domchek SM, Stadler Z et al (2016) Counselling framework for moderate-penetrance cancer-susceptibility mutations. *Nat Rev Clin Oncol* 13:581–588. <https://doi.org/10.1038/nrclinonc.2016.90>
- Villani A, Shore A, Wasserman JD et al (2016) Biochemical and imaging surveillance in germline TP53 mutation carriers with Li–Fraumeni syndrome: 11 year follow-up of a prospective observational study. *Lancet Oncol* 17:1295–1305. [https://doi.org/10.1016/S1473-2045\(16\)30249-2](https://doi.org/10.1016/S1473-2045(16)30249-2)
- Win AK, Reece JC, Dowty JG et al (2016) Risk of extracolonic cancers for people with biallelic and monoallelic mutations in MUTYH. *Int J Cancer* 139:1557–1563. <https://doi.org/10.1002/ijc.30197>

## Affiliations

Sandra Bonache<sup>1</sup> · Irene Esteban<sup>2,3</sup> · Alejandro Moles-Fernández<sup>1</sup> · Anna Tenés<sup>4</sup> · Laura Duran-Lozano<sup>1</sup> · Gemma Montalban<sup>1</sup> · Vanessa Bach<sup>1</sup> · Estela Carrasco<sup>2</sup> · Neus Gadea<sup>2,5</sup> · Adrià López-Fernández<sup>2</sup> · Sara Torres-Esquius<sup>2</sup> · Francesco Mancuso<sup>6</sup> · Ginevra Caratú<sup>6</sup> · Ana Vivancos<sup>6</sup> · Noemí Tuset<sup>7</sup> · Judith Balmaña<sup>2,5</sup> · Sara Gutiérrez-Enríquez<sup>1</sup>  · Orland Diez<sup>1,4</sup> 

<sup>1</sup> Oncogenetics Group, Vall d'Hebron Institute of Oncology-VHIO, Lab 2.02A, CELLEX CENTER, c/ Natzaret, 115-117, 08035 Barcelona, Catalonia, Spain

<sup>2</sup> High Risk and Cancer Prevention Group, VHIO, Barcelona, Spain

<sup>3</sup> Genetics and Microbiology Department, Universitat Autònoma de Barcelona, Campus UAB, Bellaterra, Spain

<sup>4</sup> Area of Clinical and Molecular Genetics, University Hospital of Vall d'Hebron, Barcelona, Spain

<sup>5</sup> Medical Oncology Department, University Hospital of Vall d'Hebron, Barcelona, Spain

<sup>6</sup> Cancer Genomics Group, Vall d'Hebron Institute of Oncology, VHIO, Barcelona, Spain

<sup>7</sup> Medical Oncology Department, Hospital Universitari Arnau de Vilanova, Lleida, Spain







54