



UNIVERSITAT DE
BARCELONA

Exploring protein structure changes due to somatic mutations in cancer

Andrea Diéguez Docampo

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tdx.cat) i a través del Dipòsit Digital de la UB (diposit.ub.edu) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tdx.cat) y a través del Repositorio Digital de la UB (diposit.ub.edu) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tdx.cat) service and by the UB Digital Repository (diposit.ub.edu) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

DOCTORAL PROGRAM OF BIOMEDICINE

Exploring protein structure changes due to somatic mutations in cancer

Andrea Diéguez Docampo

Director:

Ivo Glynne Gut

Centro Nacional de Análisis Genómico (CNAG-CRG)

Tutor:

Modesto Orozco López

Universidad de Barcelona

Exploring protein structure changes due to somatic mutations in cancer

Thesis presented to opt for the degree of doctor by the

Universidad de Barcelona

Doctoral program of BIOMEDICINE

Research area: Bioinformatics

Author: Andrea Diéguez Docampo

Director: Ivo Glynne Gut

Tutor: Modesto Orozco

Barcelona, 2022

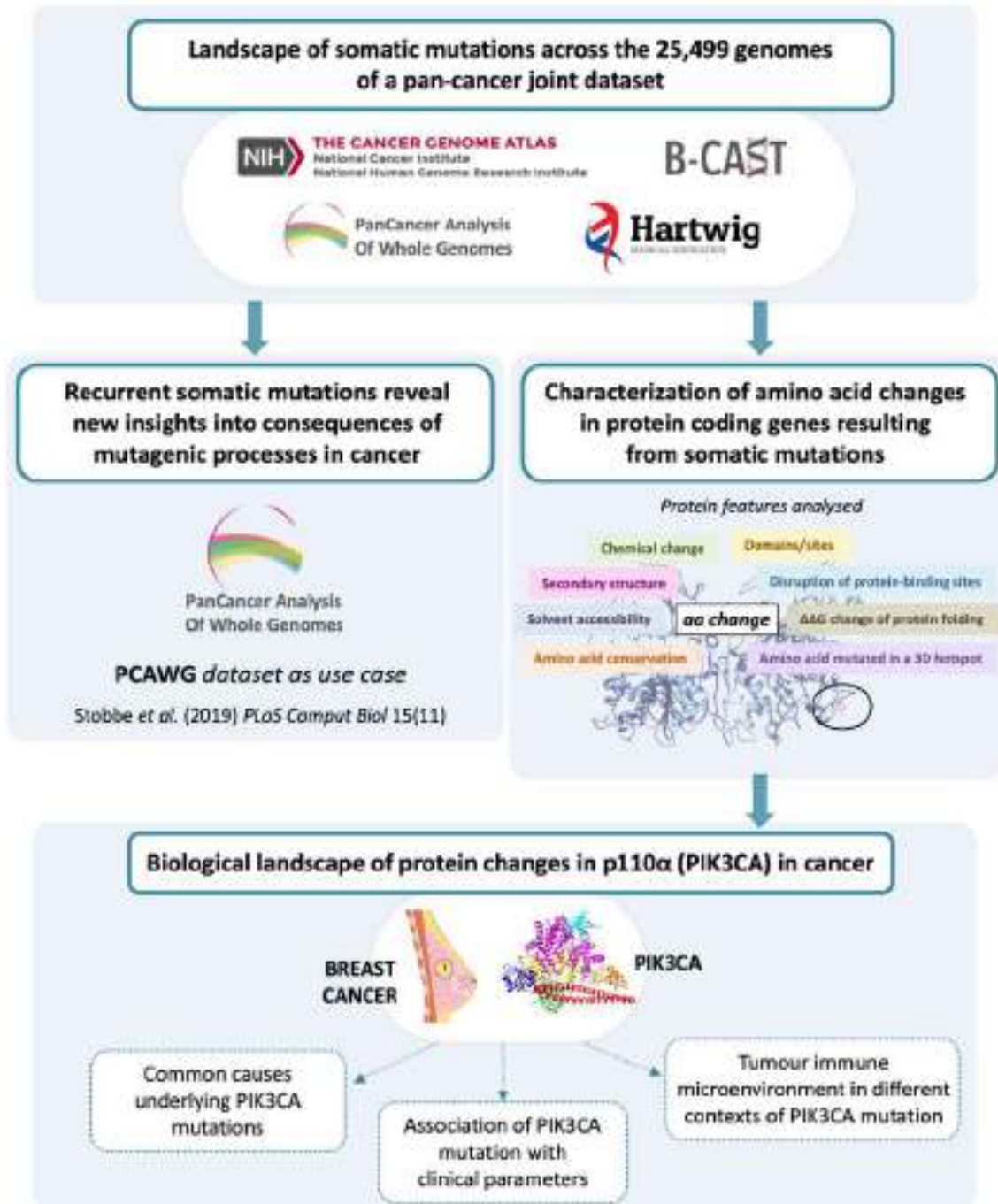


UNIVERSITAT DE
BARCELONA

Exploring protein structure changes due to somatic mutations in cancer

GRAPHICAL CONTENT

EXPLORING PROTEIN STRUCTURE CHANGES DUE TO SOMATIC MUTATIONS IN CANCER



ABSTRACT

Cancer is one of the most common diseases worldwide. Despite that a lot of time and resources have already been spent into resolving cancer, there is still a long way to go to be able to cure every patient and improve their quality of life. To contribute to these efforts, we integrate and study a joint dataset of whole genome, whole exome and panel sequencing data from primary and metastatic tumours from 25,499 donors with different cancer types. This dataset consists out of four cohorts: the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset, the Hartwig Medical Foundation (HMF) dataset, The Cancer Genomes Atlas (TCGA) dataset and the Breast-Cancer Stratification study (B-CAST) dataset. By describing mutations found in the individual cohorts and the joint dataset, we provide an overview of the genomic landscape across various cancer types. We also assess the landscape of mutational signatures in primary and metastatic tumours focused on breast, colorectal and uterus cancer and identify groups based on the dominant mutational signatures. We observe groups with the same dominant signature across all three cancer types, as well as differences between primary and metastatic tumours. To illustrate the importance of studying the genomic landscape we take the PCAWG dataset as a use case and compute 42 genomic features based on either all or only the recurrent mutations. Using these features, we are able to divide the dataset into biologically relevant clusters. Studying recurrent mutations also reveals susceptible sequence motifs, including TT[C>A]TTT and AAC[T>G]T for the Pol ϵ and 'gastric-acid exposure' clusters, respectively.

To go beyond the genomic landscape, we focus on the mutations that results in an amino acid change in the protein and characterize these protein changes with a combination of amino acid, evolutionary and structural properties. We provide an overview of the amino acid changes observed within breast cancer specifically. In our joint dataset, one of the most frequently mutated genes in breast cancer is PIK3CA, which is also frequently mutated in colorectal and uterus cancer. The comparison of the protein changes in p110 α protein, encoded by PIK3CA, and their protein features across these cancer types elucidates differences in the proportion of mutations across the different protein domains. Deciphering the underlying causes of this could provide information on the mechanisms playing a role in the three cancer types. Our results show that

mutational processes such as hypermutation activity of polymerase epsilon (Pol ϵ) or defective DNA damage repair in uterus cancer could be causing the mutations in the ABD domain. For uterus cancer, patients with a PIK3CA mutation have a higher survival rate than those without. In breast cancer we show that there is an association between the ER-positive status of the tumour and having a PIK3CA mutation. Breast cancer is the most diagnosed cancer and characterized by a high heterogeneity. Therefore, improving the stratification of patients is key to tailoring the treatment strategy and to improve the management of this disease. We assess the composition of the tumour microenvironment and demonstrate that its composition is different in PIK3CA mutated breast tumours compared to those without. We also find differences within the group of PIK3CA mutated tumours. For example, tumours with a mutation in the linker ABD-RBD region present an exhausted profile in T cells characterized by a significantly higher expression of LAG3.

In conclusion, the analysis of somatic mutations and corresponding protein changes combined with the evaluation of clinical data and the tumour immune microenvironment (TIME) across and within cancer types is useful to stratify cancer patients and identify groups for whom a specific treatment strategy, such as immunotherapy, might be beneficial.

ABBREVIATIONS

ABD: Adaptor Binding Domain
ASA: Accessible Surface Area
B-CAST: Breast CAncer STRatification
CAF: Cancer Associated Fibroblast
CD: Cluster of Differentiation
CN: Copy Number variations
COSMIC: Catalogue of Somatic Mutations in Cancer
DBS: Double Base Signature
DCs: Dendritic Cells
DE: Differential Expression
EMA: European Medicines Agency
EMT: epithelial-to-mesenchymal transition
ER: Oestrogen Receptor
FAMD: Factor Analysis of Mixed Data
FDA: USA Food and Drug Administration
GSEA: Gene Set Enrichment Analysis
HCPC: Hierarchical Clustering of Principal Components
HER2: Human Epidermal growth factor Receptor 2
HLA: Human Leukocyte Antigen
HMF: Hartwig Medical Foundation
HR: Homologous Recombination
IARC: International Agency for Research on Cancer
ID: Indel (insertion/deletion)
LAM: Lipid Associated Macrophages
MMR: Mismatch Repair
MSI: Microsatellite instability
NK: Natural Killer
PCA: Principal Component Analysis
PCAWG: Pan-Cancer Analysis of Whole Genomes
PDB: Protein Data Bank
PR: Progesterone Receptor
RBD: Ras Binding Domain
SBS: Somatic Single-base Mutations
SIMs: Somatic Insertion/deletion Mutations
SSMs: Somatic Single-base Mutations
TAMs: Tumour-Associated Macrophages
TILs: Tumour-Infiltrating Lymphocytes
TIME: Tumour Immune MicroEnvironment
TME: Tumour MicroEnvironment
TCGA: The Cancer Genomes Atlas
TNBC: Triple Negative Breast Cancer
WES: Whole Exome Sequencing
WGS: Whole Genome Sequencing

LIST OF FIGURES

- Figure 1. Examples of mutational processes and their corresponding SBS mutational signatures.
- Figure 2. Main consequences of single-base somatic substitutions at protein level.
- Figure 3. Amino acid classification according to the charge of their side chain.
- Figure 4. The PI3K-AKT-mTOR pathway and drug targets.
- Figure 5. Structure of p110 α coloured by protein domains.
- Figure 6. Summary of characteristics of each of the main breast cancer subtypes.
- Figure 7. Immune cell lineages involved in the two main immune responses.
- Figure 8. (a) Major immune populations in the tumour microenvironment of breast cancer. (b) Molecules involved in the crosstalk between cancer cells and the tumour immune microenvironment in breast cancer.
- Figure 9. The cancer-immunity cycle.
- Figure 10. List of cancer type abbreviations and complete names in PCAWG and TCGA.
- Figure 11. Scheme of procedure followed by SigProfilerSingleSample.
- Figure 12. Categories among the cases of amino acid changes.
- Figure 13. Protein secondary structures: (a) α -helix, (b) β -strand and (c) turn or loop.
- Figure 14. Amino acid conservation scores computed by ConSurf.
- Figure 15. Schematic example of the use of mutation3D to find clusters in the 3D structure of a protein in our workflow.
- Figure 16. FoldX computation to obtain the change in free energy degrees for protein folding and the categories depending on threshold: destabilizing, stabilizing or not affecting stability.
- Figure 17. Number of donors per cancer type in the PCAWG dataset.
- Figure 18. Distribution of (a) total number of mutations across cancer types, (b) number of SSMs and (c) number of SIMs across cancer types in the PCAWG dataset.
- Figure 19. Distribution of mutation types across cancer types in the PCAWG dataset. Cancer types are ordered by the percentage of C>A mutations.
- Figure 20. Number of donors per location of the primary tumour in the HMF dataset.
- Figure 21. Distribution of (a) total number of mutations across the primary location of the metastatic tumours, (b) number of SSMs and (c) number of SIMs across the primary location of the metastatic tumours in the HMF dataset.
- Figure 22. Distribution of mutation types across primary tumour locations in the HMF dataset. Cohorts are ordered by the percentage of C>A mutations.
- Figure 23. Number of donors per cancer type in the TCGA dataset.
- Figure 24. Distribution of (a) total number of mutations across cancer types, (b) number of SSMs and (c) number of SIMs across cancer types in the TCGA dataset.
- Figure 25. Distribution of mutation types across cancer types in the TCGA dataset. Cancer types are ordered by the percentage of C>A mutations.
- Figure 26. Number of mutations per mutation type and top mutated genes in each case in the B-CAST dataset.
- Figure 27. Breast cancer subtype distribution in the individual datasets.
- Figure 28. Percentage of SBS signatures found in breast tumours in the TCGA, PCAWG and HMF dataset.
- Figure 29. Percentage of SBS signatures found in colorectal tumours in the TCGA, PCAWG and HMF dataset.

Figure 30. Percentage of SBS signatures found in uterus tumours in the TCGA, PCAWG and HMF dataset.

Figure 31. Percentage of SBS signatures across the different breast cancer molecular subtypes in the PCAWG, TCGA and HMF dataset.

Figure 32. Recurrence within each tumour type in absolute numbers and percentages.

Figure 33. Spearman's rank correlation between the 42 mutational features.

Figure 34. Workflow of the recurrence-based approach to group cancer genomes.

Figure 35. Key characteristics of the 16 clusters.

Figure 36. Overview of the 42 features and their association with each cluster.

Figure 37. Enriched sequence motifs.

Figure 38. Summary of take-home messages: Factors impacting on recurrence in the context of the clusters.

Figure 39. Workflow of the analysis of amino acid changes in breast cancer.

Figure 40. Distribution of amino acid changes found in breast tumours across the categories of 'Chemical change'.

Figure 41. Percentage of the different amino acid changes found in breast cancer tumours.

Figure 42. Distribution of the amino acid changes found in breast cancer tumours across (a) location in secondary structure, (b) solvent accessibility and (c) location in a functional site.

Figure 43. Distribution of the amino acid changes found in breast cancer tumours across protein domains.

Figure 44. PIK3CA mutations in different datasets.

Figure 45. Mutation types in PIK3CA gene.

Figure 46. Proportion of the different categories of amino acid changes in p110 α (PIK3CA) protein in breast, colorectal and uterus cancer cohorts in the PCAWG, HMF, TCGA and B-CAST dataset.

Figure 47. Proportion of the different categories of amino acid size change for the amino acid changes found in p110 α (PIK3CA) protein in breast, colorectal and uterus cancer cohorts in the PCAWG, HMF, TCGA and B-CAST dataset.

Figure 48. Proportion of type of secondary structure hit by amino acid changes found in p110 α (PIK3CA) protein in breast, colorectal and uterus cancer cohorts in the PCAWG, HMF, TCGA and B-CAST dataset.

Figure 49. Proportion of amino acid changes found in p110 α (PIK3CA) protein happening in an exposed or buried amino acid in breast, colorectal and uterus cancer cohorts in the PCAWG, HMF, TCGA and B-CAST dataset.

Figure 50. Proportion of amino acid conservation categories for the amino acid changes found in p110 α (PIK3CA) protein in different cohorts in the PCAWG, HMF, TCGA dataset.

Figure 51. 3D-Hotspots of mutations on p110 α protein structure found by mutation3D per dataset (PCAWG, HMF, TCGA and B-CAST).

Figure 52. . 3D-Cluster in the linker ABD-RBD including the amino acids 106, 107, 108 and 111.

Figure 53. Proportion of mutations in each of the main domains of p110 α (PIK3CA) in breast, colorectal and uterus cancer in the PCAWG, HMF and TCGA dataset.

Figure 54. Change in the free energy of protein folding ($\Delta\Delta G$ in kcal/mol) upon amino acid changes in p110 α (PIK3CA) protein in C2 PI3K-type, helical and kinase domain.

Figure 55. Number of missense mutations per position in the p110 α protein in three different cancer types: (a) breast, (b) colorectal and (c) uterus cancer.

Figure 56. Number of non-missense coding mutations per position in the p110 α protein in three different cancer types: breast, colorectal and uterus cancer.

Figure 57. Mutational signatures PIK3CA-mutated breast tumours divided in single mutation vs. multiple mutations in p110 α (PIK3CA) protein.

Figure 58. Mutational signatures PIK3CA-mutated uterus tumours divided in single mutation vs. multiple mutations in p110 α (PIK3CA) protein.

Figure 59. Mutational signatures PIK3CA-mutated colorectal tumours divided in single mutation vs. multiple mutations in p110 α (PIK3CA) protein.

Figure 60. Principal Component Analysis (PCA) followed by hierarchical clustering of principal components of the mutational signatures describing breast and uterus cancer genomes from the PCAWG, TCGA and HMF dataset.

Figure 61. Fisher's Exact test to test for associations of PIK3CA mutations with different mutational processes or conditions across cancer genomes from all cancer types in PCAWG and TCGA datasets.

Figure 62. Methylation values of 19 positions in the PIK3CA gene in breast and uterus tissue.

Figure 63. Kaplan-Meier curves for survival in PIK3CA-mutated versus non-mutated tumours in uterus cancer.

Figure 64. Deconvolution of TME in PIK3CA mutated and non-mutated breast cancer.

Figure 65. Deconvolution of TME in PIK3CA mutated tumours depending on the domain mutated.

Figure 66. Differential expression of immune gene signatures linker ABD against all other domains.

Figure 67. Differential expression of immune gene signatures: all possible pairwise comparisons of C2, helical and kinase domain mutated tumours.

Figure 68. Differential expression of immune gene signatures: ABD domain mutated tumours against all other domains.

LIST OF TABLES

Table 1. Summary of the main characteristics of the four individual datasets.

Table 2. HMF donors excluded.

Table 3. Clinical data available across the different datasets.

Table 4. Amino acid residues weights.

Table 5. Populations in the first level of annotation of the single-cell RNA-Seq reference.

Table 6. Populations in the second level of annotation of the single-cell RNA-Seq reference.

Table 7. Main SBS mutational signatures identified in primary and metastatic tumours from breast, colon and uterus and the mutational processes to which they are related to (if known).

Table 8. Gene signatures related to immune function: signature name, list of genes included in the signature and general description of the function in which they are involved in.

Table 9. Summary results of the significant differentially expressed genes along the different immune gene signatures between PIK3CA mutated and PIK3CA non-mutated tumours.

INDEX

GRAPHICAL CONTENT	7
ABSTRACT	9
ABBREVIATIONS	11
LIST OF FIGURES	12
LIST OF TABLES	14
1. INTRODUCTION	19
1.2. Mutation rates and mutation subtypes across cancer types.....	20
1.3. Mutational signatures: a general overview	20
1.4. From somatic mutations in the genome to protein changes.....	22
1.4.1. Evaluation of the effect of protein changes	23
1.4.2. Protein structure availability	26
1.5. PIK3CA gene encodes p110α (PIK3CA) protein	27
1.5.1. Structural insights of p110 α (PIK3CA) protein.....	29
1.5.2. PIK3CA plays a central role in cancer	29
1.5.3. PIK3CA mutations association to clinicopathological parameters	30
1.5.4. Treatment to target p110 α over-activation: PI3K inhibitors (PI3Ki).....	31
1.5.5. Emerging therapy strategies: immunotherapy	32
1.6. PIK3CA: the most common genomic aberration in breast cancer	33
1.6.1. Morphological characteristics of breast tumours	34
1.6.2. Molecular characteristics: breast cancer subtypes	35
1.6.3. Prognosis and survival	36
1.6.4. Targeted therapies in breast cancer.....	36
1.6.5. Breast cancer and its Tumour Immune Microenvironment	38
2. HYPOTHESIS AND OBJECTIVE.....	43
2.1. Hypothesis.....	43
2.2. General objective	43
2.3. Specific objectives	43
3. MATERIALS	45
3.1. Mutational data.....	45
3.2. RNA-Seq, methylation and clinical data	49
3.3. Protein annotation and protein structures.....	50
3.4. Data availability.....	50
4. METHODS	51
4.1. CHAPTER 1. Genomics landscape of 25,499 cancer genomes.....	51

4.1.1.	Plots.....	51
4.1.2.	Mutational signatures	51
4.2.	CHAPTER 2. Use case in PCAWG dataset. Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer	52
4.2.1.	PCAWG cohort – quality control	52
4.2.2.	PCAWG cohort – mutation calls	52
4.2.3.	Features describing each cancer genome	53
4.2.4.	Principal Component Analysis and hierarchical clustering on Principal Components	54
4.2.5.	Detection and enrichment of motifs	54
4.2.6.	Statistical tests.....	56
4.2.7.	Plots.....	56
4.3.	CHAPTER 3. Characterization of amino acid changes due to somatic mutations in protein coding genes.....	57
4.3.1.	Workflow for the automatic evaluation of missense mutations.....	57
4.3.2.	Statistical methods	64
4.4.	CHAPTER 4. Landscape of protein changes in p110α (PIK3CA) in cancer	65
4.4.1.	Data visualization	65
4.4.2.	Statistics: Chi-squared test / Fisher’s Exact Test	65
4.4.3.	Survival analysis and associations	65
4.4.4.	RNA-Seq analyses: Differential expression analysis and Gene Set Enrichment Analysis	66
4.4.5.	Deconvolution of bulk RNA-Seq samples using SPOTlight	66
5.	RESULTS	69
5.1.	CHAPTER 1. GENOMIC LANDSCAPE OF 25,499 CANCER GENOMES	69
5.1.1.	Genomic description of the individual datasets.....	69
5.1.2.	Consensus of the joint dataset and largest cohort.....	79
5.1.3.	Description of mutational signatures across cancer types.....	80
5.2.	CHAPTER 2. USE CASE IN PCAWG DATASET: RECURRENT SOMATIC MUTATIONS REVEAL NEW INSIGHTS INTO CONSEQUENCES OF MUTAGENIC PROCESSES IN CANCER	85
5.2.1.	Recurrence is higher than expected by chance.....	85
5.2.2.	Number of samples does not always correspond to the level of recurrence	86
5.2.3.	General mutational characteristics versus recurrence.....	88
5.2.4.	Recurrence characteristics divide the cohort.....	88

5.2.5.	High levels of recurrent SSMs and low levels of recurrent SIMs characterize exposure to UV light	92
5.2.6.	High levels of recurrent SSMs characterize deregulated activity of Pol ϵ	95
5.2.7.	High levels of recurrent SIMs characterize microsatellite instability .	97
5.2.8.	Positive association with recurrence of SSMs and SIMs: Gastric-acid exposure and hypermutation of immunoglobulin genes	97
5.2.9.	Negative association with recurrence: Tobacco-smoke exposure, alcohol use and increased activity of cytidine deaminases	99
5.2.10.	The added value of the recurrence-related features	101
5.3.	CHAPTER 3. CHARACTERIZATION OF AMINO ACID CHANGES RESULTING FROM SOMATIC MUTATIONS IN PROTEIN CODING GENES.....	103
5.3.1.	Distribution of the amino acid changes found in breast tumours across the categories established for eight protein features	106
5.3.2.	Proof of concept. Reduction method and clustering to find groups of relevant mutations	112
5.4.	CHAPTER 4. LANDSCAPE OF PROTEIN CHANGES IN p110α (PIK3CA) IN CANCER	115
5.4.1.	Pan-cancer description of PIK3CA mutations.....	116
5.4.2.	Underlying causes of PIK3CA mutations.....	129
5.4.3.	Relationship between p110 α (PIK3CA) mutation and survival or other clinical parameters in cancer.....	138
5.4.4.	Gene set enrichment analysis (GSEA) of p110 α mutated domains..	140
5.4.5.	Assessment of the tumour immune microenvironment in p110 α mutated and non-mutated tumours in breast cancer.....	141
6.	DISCUSSION	155
7.	CONCLUSIONS.....	164
8.	REFERENCES.....	167
9.	SUPPLEMENTARY INFORMATION	181

1. INTRODUCTION

Cancer is one of the most common diseases worldwide, with an estimated 19.3 million new cases and causing almost 10 million deaths in 2020, according to the statistics recorded by the International Agency for Research on Cancer (IARC) [1]. Despite all the effort put in by many researchers all over the world and for many years, there is still a lot to study, find out and solve regarding this disease. Many cancers could be cured if detected early and treated effectively. Making progress in any of these aspects by, for example, discovering molecular causes of cancer initiation, progression and metastasis or biomarkers that make the treatment selection easier, would help fight the disease and improve the patient's quality of life, therefore being of great importance.

1.1. Cancer origin: mutations in the genome and mutational processes

Cancer is caused by the accumulation of mutations in the genome. In the process of a cell becoming a cancer cell, the cellular division speeds up and there is an accumulation of mutations because of a combination of mistakes during the DNA replication and the lack of repair. The mutations that are identified in the tumour but not in the normal tissue are called "somatic mutations". Somatic mutations are not inherited by offspring, in contrast to germline mutations that occur in sperm, eggs and their progenitor cells, and therefore are present in all tissues of the individual. Critical somatic mutations can affect a wide variety of pathways and functions in the cell. This deregulation makes cells grow without control, invading adjacent parts in the body or spreading further to other organs through the blood and lymphatic system (metastasis). Every cell type in every tissue and organ can undergo this malignant transformation, resulting in a large variety of cancer types that can affect any part of the body, from a common lung cancer to rare cancers such as the Kaposi sarcoma that originates in the cells lining lymph or blood vessels. As each organ has a different function and is exposed to different mutagens, for example skin is the most exposed to UV light [2] and lung to smoking [3], different mutational processes and defective DNA repair processes can be involved in the occurrence of mutations, which result in different mutational imprints left on the genome [4]. In addition, the epigenetics, the regulation and the transcription of the genome are different across tissues affecting the mutational patterns observed.

1.2. Mutation rates and mutation subtypes across cancer types

Different cancer types can be clustered according to their different mutational patterns as the result of the influence of the different mutational and/or defective repair processes. Main differences can be the mutation rate and mutation type. The mutation rate is influenced by replication time [5], is linked to epigenomic features (chromatin accessibility) [6], shows a periodic pattern around nucleosomes [7] and can depend strongly on the 5' and 3' flanking base as shown in mutational signatures for several mutational processes [8]. Considering Somatic Single-base Mutations (SSMs), the cancer type with the highest mutation rate is skin cancer [9]. In contrast, haematological and pediatric tumours have a low mutation rate [9]. Considering Somatic Insertion/deletion Mutations (SIMs), they are high in certain cancer types such as renal cell carcinomas [10]. Regarding mutation type, which type of mutation takes place is linked to the mutational process behind it. Generally, the most frequent substitution is C>T, followed by C>A [11], although this ranking can change depending on the cancer type. For example, in lung cancers, C>A (or in particular here G>T) is the most frequent substitution, since this transversion is a typical mutation as a consequence of tobacco smoke carcinogens such as polycyclic aromatic hydrocarbons (PAH) [12].

1.3. Mutational signatures: a general overview

Mutational signatures are the result of the endogenous and exogenous mutational processes affecting the DNA, which provides the individual history of the tumour. The mutational signatures differ in the number, type and distribution of SSMs along the 96 different trinucleotide contexts considering the six conventional mutations in the centre of the trinucleotide. We consider as “conventional mutations” the one indicating the mutation happening in the pyrimidine base (C or T) first. In the version v3.3 of COSMIC (June 2022 - <https://cancer.sanger.ac.uk/signatures/>), signatures are described at four levels: single-base substitutions (SBS), small insertions and deletions (ID), double-base substitutions (DBS) and copy number variation (CN) signatures. There are 94 different SBS signatures and 18 ID signatures described, however the aetiology for many of them is not known. Examples of well-known signatures are shown in **Figure 1**.

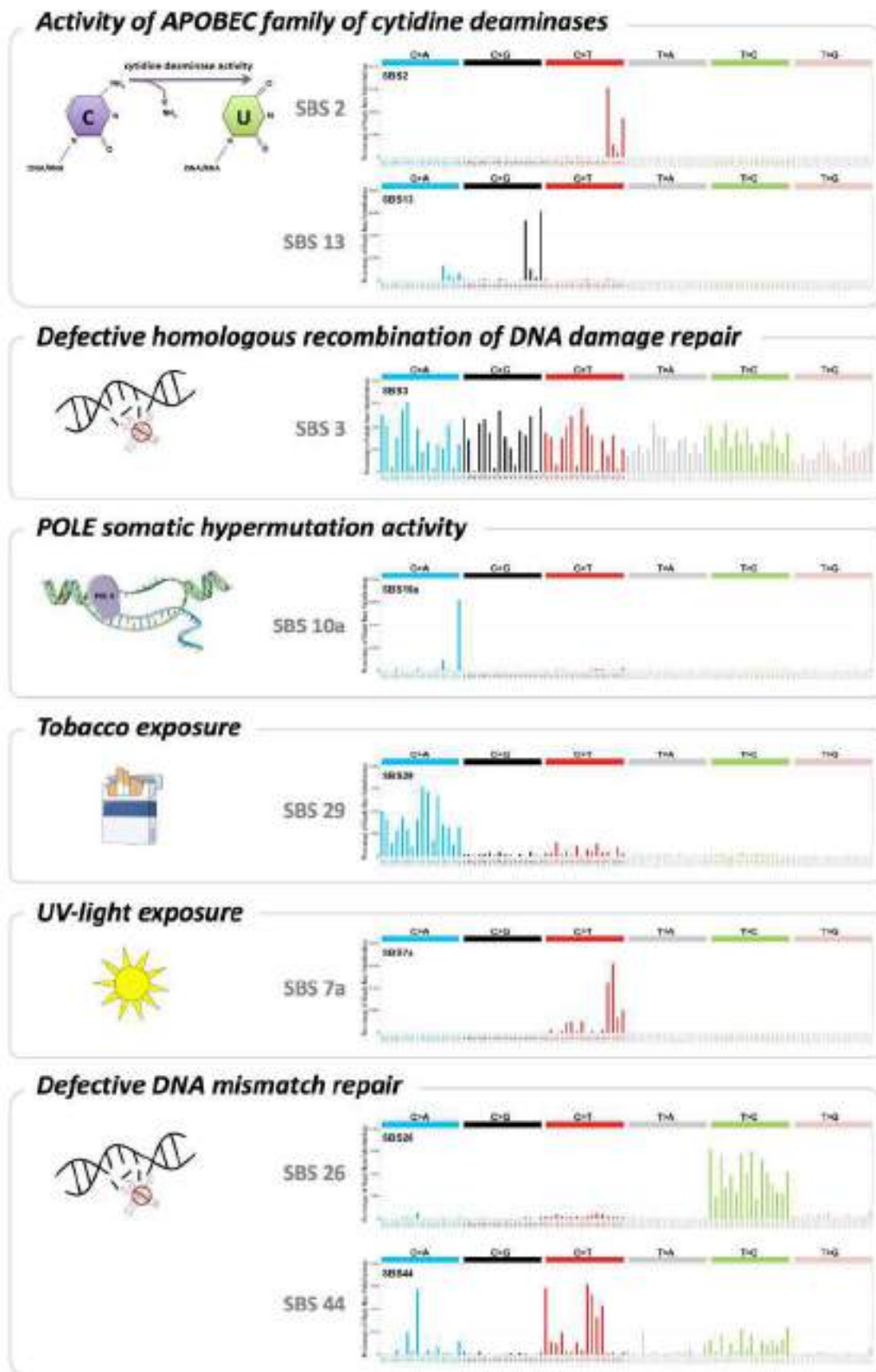


Figure 1. Examples of mutational processes and their corresponding SBS mutational signatures. All signature plots were obtained from COSMIC v3.3 - June 2022 (<https://cancer.sanger.ac.uk/signatures/>).

1.4. From somatic mutations in the genome to protein changes

The vast majority of somatic mutations happen in the non-coding sequence of the genome, since it is 99% of the whole genome against the 1% that corresponds to the coding sequence. Despite the limited size of coding sequence, driver mutations are commonly found in this region. We refer to 'driver' as somatic mutations with the ability to drive tumourigenesis after conferring to the cell certain advantages that are positively selected with respect to its neighbouring cells [13]. Although the impact of non-coding mutations is more difficult to establish, it is known that a small proportion of them are also potential cancer drivers involved in the initiation of the tumour, or can contribute to cancer progression once initiated, such as those affecting regulatory elements (e.g. mutations in the TERT promoter [14][15]). Recurrence plays an important role to find these cases since it is difficult to detect the functional effect of mutations in non-coding regions. Sequencing and mapping artefacts, incomplete annotation of regulatory regions, inaccurate estimation of the background mutation rate and poorly understood localized hypermutations processes [16][17][18] are some aspects that add to the challenge of non-coding driver identification [19]. In the case of mutations affecting the coding region of the genome, their impact can be studied when translating them into the changes that they make in the corresponding protein. Proteins are structural and motor elements, serving as catalysts in virtually every biochemical reaction in our cells. Their folded conformation depends directly on their linear amino acid sequence [20]. Changes in this sequence caused by mutations in the gene encoding the protein could affect their structure and, consequently, its function. Point changes, *i.e.*, substitutions of a single nucleotide, in the protein-coding region of the genome can be divided into synonymous, which do not change the amino acid sequence of the protein, and non-synonymous, which do cause a change (**Figure 2**). The latter change can be missense, non-sense or non-stop mutations if the consequence is an amino acid change, the appearance of an early stop codon or the deletion of the expected stop codon, respectively. Aside from point changes, there are also insertions or deletions (indels) of nucleotides, which in coding regions can be divided into frameshift and in-frame mutations, if they cause a shift in the reading frame of the original protein or not, respectively.

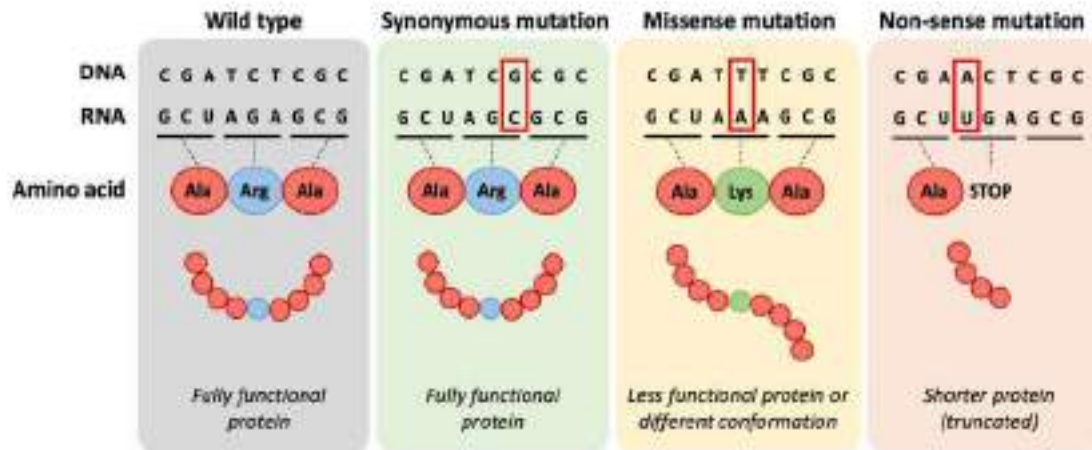


Figure 2. Main consequences of single-base somatic substitutions at protein level. Adapted from Martínez & Quiñones (2018) Chapter in book “ADME Processes in Pharmaceutical Sciences”.

1.4.1. Evaluation of the effect of protein changes

Understanding the effect of the coding mutations that the tumours harbour is extremely important to be able to find targets to develop new cancer treatments. All coding mutations are potential contributors to cancer development and progression. There are drastic mutations that can produce a big change in the protein and therefore have a clear deleterious effect, such as large insertions or deletions, as well as nonsense mutation at the beginning of the protein. In contrast, other mutations that produce just a subtle change in the protein, such as missense mutations, are more uncertain with respect to their pathogenicity and more in-depth study is required to clarify their possible involvement in cancer in the absence of more drastic changes to the protein structure. In addition, understanding the effect of these subtle mutations at molecular level is interesting for the study of drug treatment responses across individuals [21]. Current efforts in this field are, therefore, aimed at predicting how much they affect the protein and whether or not these mutations are deleterious [22]. The features used to make such predictions are many, but can be classified in three main categories according to what they are taking into account: a) amino acid properties, b) evolutionary properties and c) structural properties [22].

a) Amino acid properties

Although there are over 500 amino acids found in nature, the human genetic code only directly encodes 20 amino acids [23]. These 20 amino acids differ in size, shape, solubility and ionization properties of its side chain [24]. Amino acids can be classified in different ways according to the different characteristics mentioned. Focusing on the characteristics of the side chain, they can be classified as nonpolar (divided into alkyl or aromatic group), polar uncharged, acidic polar (negatively charged) or basic polar (positively charged) (**Figure 3**). Missense mutations that produce amino acid changes that result in small differences in properties between the amino acids are expected to affect less the protein function than those that result in more drastic changes, such as the appearance of an amino acid with a different charge on its side chain.

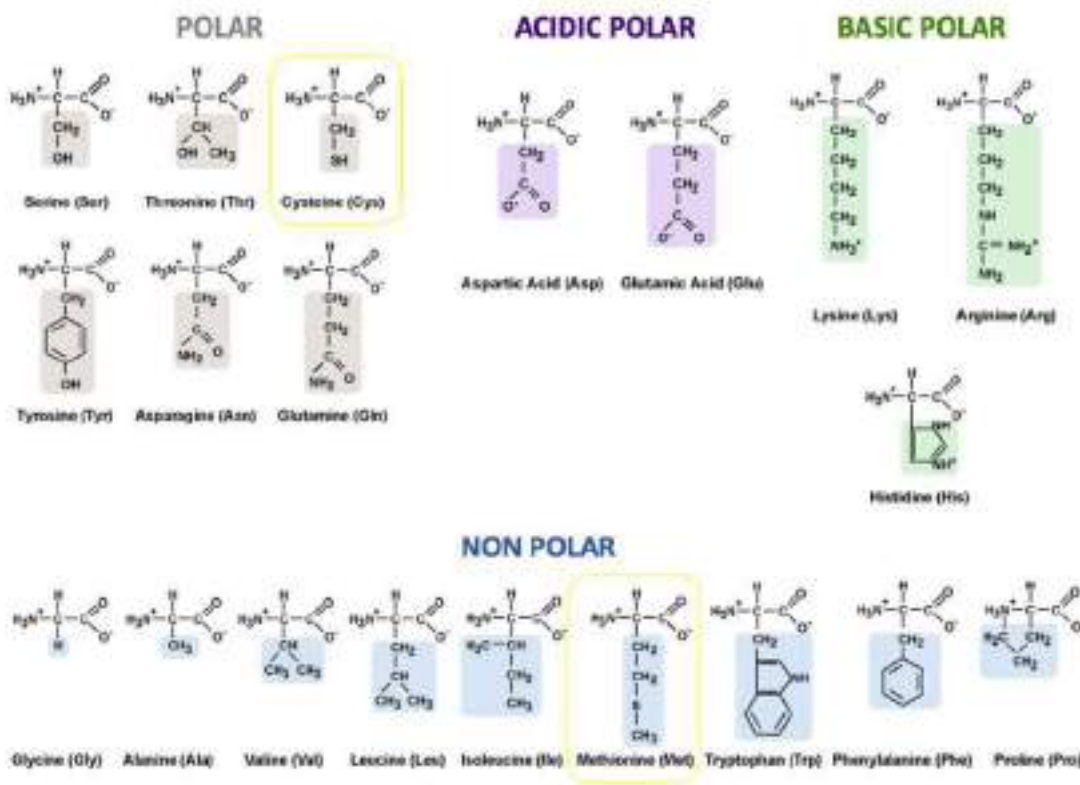


Figure 3. Amino acid classification according to the charge of their side chain. Adapted from Karki (2018) [24].

b) Evolutionary properties

The evolutionary properties of the amino acid mutated are also important when predicting its effect. The conservation of amino acid in a specific protein is measured by carrying out a multiple sequence alignment with the sequences of the same protein in different species. This allows determining whether a specific amino acid has remained the same across the different species or if it is variable. Highly conserved positions in multiple sequence alignments can help to identify functional sites [25], since amino acids conserved across species during evolution of the protein tend to be critical for the function of the protein or in maintaining its structural integrity [21].

c) Structural properties

The localization of the altered amino acid in the protein structure provides insights into its potential effect on the protein. Residues involved in intramolecular interactions, such as cysteine bridges, hydrogen bonds or zinc fingers can affect the structure if when mutated the protein can no longer maintain an important interaction. Something similar can be said for residues forming catalytic and regulatory sites, a mutation could prevent the normal function of the protein. The interpretation of the effect of the mutation could also be different if we know if the mutated residue is buried or exposed. In addition, having structural information gives us the opportunity to study protein folding. Protein folding is the process by which the linear protein sequence is shaped to build the final 3-dimensional (3D) structure [26]. Mutations can change the free energy that a protein needs to fold and therefore affect protein stability.

There are also other advantages of using 3D structures. Amino acids that are far apart in the linear sequence can be close in structure so there is possibility to find a 3D cluster of mutations. Interestingly, for example, less frequent mutations can be located close to a hotspot mutation and despite their rarity they could have the same effect as the hotspot. In this way, mutations clustered in the 3D structure may indicate relevant regions for protein function, reflecting sites that when mutated in the protein they would have a particular effect on protein performance and could be selected for in cancer. There are different softwares that aim to find 3D clusters of mutations in protein structures, such as HotSpot3D [27], HotMaps [28], CLUMPS [29] or Mutation3D [30].

All these three categories combined have been very informative in delineating the effects of pathogenic mutations and understanding the underlying mechanisms of cancer [31]. For example, mutations in sites buried in the protein molecule or involved in macromolecular interactions are frequently pathogenic if the mutation results in a drastic change of the amino acid physicochemical properties, *i.e.* a change from negative to positive charge would prevent interactions from being established [31]. This is the case for the sodium iodide symporter (NIS) gene, which encodes an iodide transporter. It has been shown that different missense mutations that hit amino acids with charged side chains can affect the electrostatic interactions in the transmembrane domains of the NIS protein. These affected interactions have been proposed to affect the protein functionality and therefore associated with a iodide transport defect [32][33][34].

1.4.2. Protein structure availability

A bottleneck in using the protein structure to predict the impact of missense mutations is the availability of reliable structures. Protein structures can be obtained experimentally mainly from X-ray crystallography, NMR spectroscopy and electron microscopy. An alternative to obtaining the actual protein structure is the computation of homology models from other structures which share a similar sequence, since it is postulated that when the sequence similarity is sufficiently high the protein structure would also be similar [35]. A 35% or higher sequence identity is thought to be enough for ensuring the structural similarity of two proteins, while with a sequence identity of 20–35%, often referred to as ‘twilight zone’, structural similarity is less common [35]. Recently, a new source of protein structures has come available with AlphaFold, an artificial intelligence system developed by Deepmind that predicts the 3D protein structures from its amino acid sequence [36]. The AlphaFold Protein Structure Database (AlphaFold DB, <https://alphafold.ebi.ac.uk>) provides open access to their results, which account for over 200 million protein structure predictions. Despite that the number of proteins for which we have the actual protein structure available is limited, adding these predictions provides us with a good subset to work with on the characterization of mutations.

The availability of a complete structure containing from the first to last amino acid of the protein sequence is not common. Some exceptions are the cases of proteins that are frequently involved in cancer, such as p110 α protein. The structure of p110 α protein is almost complete allowing us to study the variety of mutations found in different regions of the protein.

1.5. PIK3CA gene encodes p110 α (PIK3CA) protein

A gene that is frequently mutated in several cancer types is PIK3CA, which encodes the p110 α protein. This protein corresponds to the catalytic subunit of a heterodimeric enzyme called phosphatidylinositol 3-kinase (PI3K). This enzyme belongs to the phosphoinositide 3-kinase (PI3K) family, a group of lipid kinases that act as signal transducers in various signalling pathways. They regulate a wide range of signalling, membrane trafficking and metabolic processes by phosphorylating the inositol ring of phosphoinositides in nearly all membranes in the cell [37]. Different isoforms of the catalytic and regulatory subunits combine and form different complexes that have their specific function or target. One ubiquitous complex is the p110 α -p85 α complex. Here, the catalytic subunit (p110 α) is encoded by PIK3CA, a gene located in chromosome 3 (3q26.3), while the regulatory subunit (p85 α) is encoded by PIK3R1, a gene located in chromosome 5 (5q13.1). One of the key pathways in which this complex is involved is the PI3K/Akt/mTOR signalling pathway, which regulates diverse cellular processes including protein synthesis, cell proliferation and survival, glucose metabolism, apoptosis, DNA repair and genome stability (**Figure 4**) [38].

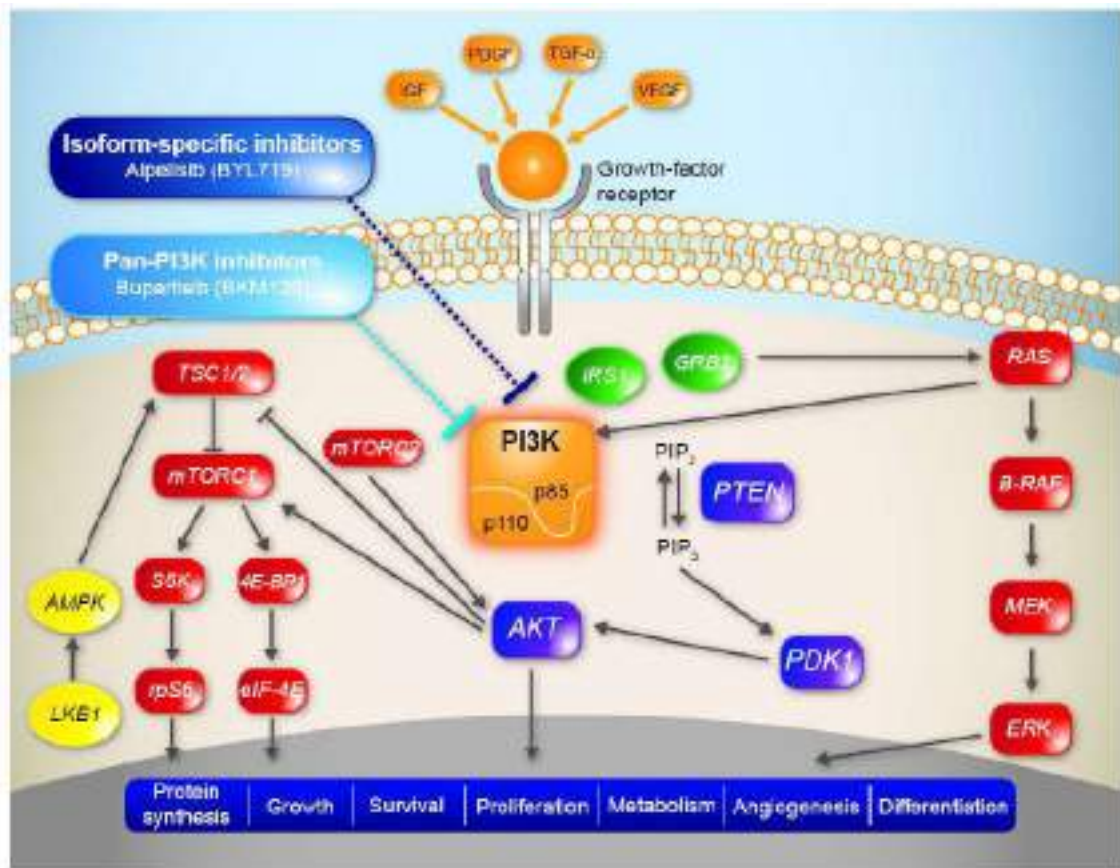


Figure 4. The PI3K-AKT-mTOR pathway and drug targets. From Massaccesi *et al.* (2016) [39].

When the catalytic and regulatory subunit are together in a complex, the protein remains in an inactive, cytosolic state. The enzyme is activated when the complex is recruited to the plasma membrane by the binding of the SH2 domains of p85 to tyrosine-phosphorylated proteins, such as receptor tyrosine kinases, or other membrane-bound proteins, such as the insulin receptor substrate proteins [38]. This results in the disinhibition (by detachment) of the p85-p110 complex and the association of p110 with its lipid substrates in the membrane [40], the phosphatidylinositol 4,5-bisphosphate (PtdIns(4,5)P₂ or PIP₂). These lipids are phosphorylated into phosphatidylinositol-3,4,5-triphosphate (PtdIns(3,4,5)P₃; also known as PIP₃), which acts as second messengers for the recruitment of many effector proteins with PIP₃-binding domains, such as protein kinases (i.e. AKT, PDK1, BTK), Ras super-family guanine nucleotide exchange factors (GEFs), GTPase-activating proteins (GAP) and adaptor proteins [41].

1.5.1. Structural insights of p110 α (PIK3CA) protein

There are five domains described in p110 α : an adaptor-binding domain (ABD), a Ras-binding domain (RBD), a C2 homology type (C2 PI3K-type) domain, a helical domain and a kinase domain (**Figure 5**). The parts of the protein that are not assigned to any domain we call the 'linker' regions between domains. For example, the sequence of amino acids between the ABD and RBD will be referred to as 'linker ABD-RBD'. The regulatory subunit (p85 α) that forms the complex with p110 α contains six domains: a Src homology 3 (SH3) domain, a GAP domain, two Src homology 2 (SH2) domains, the N-terminal and C-terminal SH2 domains (nSH2 and cSH2), which are separated by a coiled-coil domain known as the inter-SH2 linker (iSH2)[41].

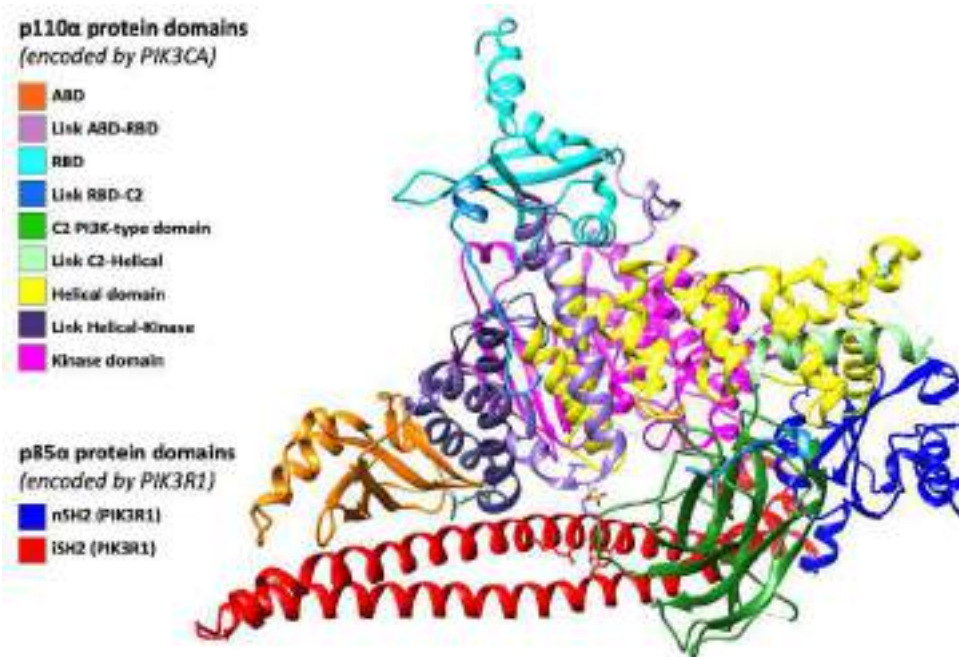


Figure 5. Structure of p110 α coloured by protein domains. Only two domains are shown for the regulatory subunit (p85 α). Figure obtained from Chimera software visualization of the PDB structure 4L23.

1.5.2. PIK3CA plays a central role in cancer

The PI3K/Akt/mTOR signalling pathway (**Figure 4**) is one of the most frequently deregulated pathways in cancer. It can be aberrantly activated through multiple mechanisms, including genomic alterations in PIK3CA, which are common not only in

one but in several cancer types. PIK3CA somatic mutations are particularly frequent in breast, colorectal and endometrial cancer [42].

Point mutations in PIK3CA can increase the enzymatic activity of the protein and thereby contribute to tumourigenesis through increased cell proliferation, decreased apoptosis and autophagy, loss of contact inhibition, induction of angiogenesis, and increased tumour invasion [43]. These mutations mimic and enhance dynamic events that take place in the natural activation of PIK3CA, as described by Burke *et al.* [44], who examined the activation of the wild-type p110 α -p85 α complex and a spectrum of oncogenic mutants. Examples of these dynamic events are: **(1)** the movement of the ABD domain and linker with respect to the rest of the catalytic subunit, **(2)** breaking the C2-iSH2 interface, **(3)** breaking the nSH2-helical domain contact caused by phosphotyrosine containing peptides binding to the enzyme and **(4)** interaction of the C lobe of the kinase domain with the membrane. Examples of mutations inducing each of these dynamic events are, respectively: **(1)** mutations in the linker between the ABD and RBD domain (G106V and G118D), **(2)** mutations in the C2 domain (N345K and C420R), **(3)** E545K mutation in the helical domain and **(4)** specific mutations in the kinase domain (*e.g.* H1047R) [44]. An increase in activity can also be achieved by, for example, mutations in the C2 domain, which are thought to facilitate p110 α localizing to the plasma membrane by increasing the positive surface charge of this domain [45].

1.5.3. PIK3CA mutations association to clinicopathological parameters

Alqahtani *et al.* [46] reviewed the relations between PIK3CA mutation and clinicopathological parameters in 2020 concluding that for some associations there was agreement in the literature, while there are discrepancies for others. Some associations for which we can find agreement are that PIK3CA mutations are positively associated to HR-positive breast tumours while negatively associated to triple-negative breast cancer tumours [47]. There is also agreement on that so far no associations have been found when considering the age at diagnosis, tumour grade or the presence of lymph node metastasis and PIK3CA mutations [47]. On the contrary, there is no agreement on the effect of PIK3CA mutations on prognosis and survival, some studies reported associations with poor survival [48], while others with better prognosis [48], and even

no association at all has been reported [47]. There is also controversy regarding the response to therapy.

1.5.4. Treatment to target p110 α over-activation: PI3K inhibitors (PI3Ki)

PI3K has been recognized as an attractive molecular target [49] because of the frequent involvement of the PI3K pathway in many cancer types. Different inhibitors have been developed and tested in clinical trials over the past decades focused on both solid and haematological malignancies [50]. Some pan-PI3K inhibitors as well as isoform specific ones have already been approved for treatment [51], such as copanlisib or idelalisib. The first progress on isoform specific ones was made in haematological malignancies. PI3K δ -specific inhibitor, idelalisib, was approved in 2014 as treatment for follicular B-cell non-Hodgkin lymphoma (FL) and small lymphocytic lymphoma (SLL) as monotherapy, as well as for chronic lymphocytic leukaemia in combination with rituximab[52]. Isoform specific inhibitors have also been developed for p110 α (PIK3CA). The first and, for now, only α -specific PI3K inhibitor approved is alpelisib (BYL719), which is being used to treat advanced breast cancer. The USA Food and Drug Administration (FDA) and the European Medicines Agency (EMA) approved this drug in 2018. As these drugs are meant to target the over-activated p110 α protein, establishing PIK3CA mutation status in cancer patients is informative for treatment choice. After the completion of the SOLAR-1 trial, the first phase 3 trial leading to an approval specifically for advanced breast cancer patients with PIK3CA mutation, a list of PIK3CA mutations was determined to select the patients that would likely benefit and have a progression-free survival after being treated with alpelisib [53].

Several companies have successfully developed panels to test for PIK3CA mutations. The Therascreen[®] PIK3CA RGQ PCR Kit from Qiagen was the first one approved by the FDA to aid in the selection of breast cancer patients that could potentially benefit from treatment with alpelisib in combination with fulvestrant. The panel of this kit allows the identification of 11 somatic PIK3CA mutations (C420R, E542K, E545A, E545D (c.1635G>T), E545G, E545K, Q546E, Q546R, H1047L, H1047R, H1047Y). Other kits include higher number of mutations, such as the cobas[®] PIK3CA Mutation Test CE-IVD

from Roche, which tests for 17 different mutations in exons 2, 5, 8, 10 and 21. The mutations added are R88Q, N345K, Q546K, Q546L, G1049R and M1043I. The ClearSEEK™ PIK3CA Panel from Agena Bioscience® tests for 20 clinically actionable PIK3CA mutations in breast cancer and has the added advantage of lowering the variant allele frequency needed for mutation detection. The mutations included in this panel that are not in the Therascreen® are E542Q, E545D (c.1635G>C), E545Q, E545V, Q546K, Q546L, Q546P, H1047N and H1047P. However, the detection of the mutations listed until now might not be enough to ensure treatment efficacy in every patient. Results from various clinical studies have demonstrated that not all patients with a PIK3CA mutation benefitted from the treatment combining alpelisib and fulvestrant [53].

In some cases, there is improvement and even cancer remission, while in other patients the disease gets worse. Due to the variable response to treatment across patients it is clear more investigation is needed regarding PIK3CA mutation contexts and precision medicine would be needed, to do a better selection of the group or individual patients that could be treated successfully with this strategy. For patients with PIK3CA mutated tumours that are not eligible for alpelisib treatment or that did not respond to the treatment, it is necessary to develop new strategies. An emerging therapy strategy with less side effects and that it is showing a high efficacy in some cancer types [54] is immunotherapy.

1.5.5. Emerging therapy strategies: immunotherapy

Cancer immunotherapies, treatments that harness the immune system's natural ability to recognise and eliminate tumour cells [55], look promising. Knowledge about the tumour microenvironment of solid tumours is needed to be able to apply this kind of therapies. For example, the presence of tumour-infiltrating lymphocytes (TILs) is a biomarker for considering the use of immunotherapy [56]. Also, tumour-associated macrophages (TAMs) are often associated with poor prognosis and are recognized as important emerging targets for cancer immunotherapy [57].

The relationship between the mutations in the tumour and the response to immunotherapies are also being studied [58]. The same as normal cells, cancer cells also need to break and recycle their proteins. Since many proteins are mutated in cancer,

from their degradation novel peptides are released harbouring these mutations, called neoantigens [10]. These neoantigens are placed in the human leukocyte antigen (HLA), which could be perceived as foreign by the immune system. T cells can naturally recognize the neoantigens coming from the mutated proteins that are unique to cancer cells, with the advantage of targeting the tumour without affecting the healthy cells [59]. An important point to consider is that mutations in cancer are largely unique to each patient, except for mutations in driver genes that are recurrent across patients. These mutations are therefore a good target for immunotherapy since it would allow the same treatment to be applied to a high number of patients. Therefore, these treatments are being developed for the most mutated genes. In particular, Chandran *et al.* (2022) and colleagues focused on neoantigens derived from driver mutations in PIK3CA [59].

1.6. PIK3CA: the most common genomic aberration in breast cancer

Breast cancer is a very heterogeneous cancer type, both within the same tumour, due to the diversity in cell populations that can be found, and across tumours from different patients. Heterogeneity within a tumour increases its ability to adapt constantly changing constraints, which affects negatively a patient's prognosis, therapy response and clinical outcome [60][55] due to the difficulties to correctly fight it. Breast cancer is not only characterized by this intra-tumoral heterogeneity, but also inter-tumoral heterogeneity, since tumours from different patients can highly differ at both morphological and molecular level [60]. Morphologically, breast cancer heterogeneity comes from differences in, for example, the size of the tumour, lymph node involvement, stage and grade [60]. Molecularly, the heterogeneity can be seen already starting from the different subtypes that are defined [60]. Genome sequencing as well as other omics like expression and methylation profiling have also provided insights into heterogeneity between tumours even from the same pathological subtype [60].

1.6.1. Morphological characteristics of breast tumours

Morphologically, breast tumours can differ in their histological type, stage and grade, which are evaluated in clinical practice.

Histological types

There are more than 20 histological types described [61]. For example, a breast tumour can be ductal or lobular, depending on if it originated in the ducts or in the milk-producing glands, respectively [62], or a mix of the two. Other less common morphological types are tubuloductal, comedo, medullary, mucinous and Paget types [62].

Staging

The most widely used system for staging breast carcinoma is the TNM classification, published by the American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC) [63]. The stage is derived from the extent of cancer at the primary site (T), at the regional lymph nodes (N) and spread to distant metastatic sites (M) [63] [64]. These three measurements are combined to create five stages (stage 0 to IV). Stage 0 indicates that the disease is only in the ducts of the breast tissue without having spread to the surrounding tissue, what it is known as non-invasive or in situ cancer [65]. The other extreme, stage IV, indicates that the cancer is metastatic.

Grade

The assessment of histological grade is based on three tumour features: the proportion of cancer cells that are in tubule formation, the variation of nuclear size and shape between the cells (anisokaryosis) and the number of cell divisions (mitotic counts) [64][63]. Each feature is scored with a three-tier system and summed up, resulting in a final grade (G1, G2 or G3) [63]. This grade represents the potential aggressiveness of the cancer and is therefore a strong prognostic factor [63].

1.6.2. Molecular characteristics: breast cancer subtypes

Breast cancer can be clinically classified into four main molecular subtypes based on gene expression profiling using the PAM50 gene signature and/or immunophenotypic characteristics [66][67]. These subtypes are luminal A, luminal B, human epidermal growth factor receptor 2 (HER2) enriched and Triple-Negative (TNBC, also known as Basal-like). A fifth subtype that sometimes is included is Normal-like (or unclassified). Finally, a sixth subtype that has been reported is called “claudin-low”[68]. Claudin-low subtype expresses specifically markers of epithelial-to-mesenchymal transition (EMT) and stemness, as well as stromal and other immune-related signatures [60]. The different subtypes vary in their biological properties, frequency, prognosis and outcome [66], as summarised in **Figure 6**. Luminal A and Luminal B which are both HER2-negative can be differentiated checking the expression of the nuclear antigen Ki-67, low or negative in luminal A (Ki67-), while positive in luminal B (Ki67+). The breast cancer subtype Normal-like closely resembles luminal A, since it is also Oestrogen Receptor (ER)-positive, Progesterone receptor (PR)-positive, HER2-negative and Ki67-. It is reported with different grades (from 1 to 3) and an intermediate outcome.

Molecular subtype	Luminal A	Luminal B	HER2	TNBC
ER/PR	+		-	
HER2	-	+	-	
Frequency ^a	50-60%	30%	10%	10-20%
Grade ^b	Low		High	
Prognosis ^c	Good		Poor	
5-y survival rate ^d	94.3%	90.5%	84.0%	76.9%

Figure 6. Summary of characteristics of each of the main breast cancer subtypes. From Burguin et al. (2021) [64]. ER: oestrogen receptor; PR: progesterone receptor; HER2: human epidermal growth factor receptor 2; TNBC: triple-negative breast cancer. a. Frequency derived from Al-thoubaity et al. [46] and Hergueta-Redondo et al. [69]. b. Grade derived from Engstrom et al. [70]. c. Prognosis derived from Hennigs et al. [71] and Fragomeni et al. [72]. d. The 5-year survival rate derived from the latest survival statistics of SEER [73].

1.6.3. Prognosis and survival

Depending on the characteristics previously described the survival rate is highly variable. Major prognostic factors are the lymph node metastasis, distant metastasis, tumour size, locally advanced disease, lymphovascular invasion and inflammatory carcinoma. For example, in the first case, if there are no nodes involved the ten years survival rate is 70-80%, if the number of nodes involved is between 1 and 3, the ten years survival goes down to 35-40% and if there are more than 10 nodes involved then the ten years survival is not expected to be more than 10-15%. Other minor prognostic factors are the histologic grade, the expression of the ER, PR and HER2, the proliferative rate and the response to neoadjuvant chemotherapy. TNBC is the cancer subtype with the worse prognosis followed by HER2-enriched subtype, while Luminal A is the least aggressive [63]. Also, Luminal A subtype is less common to metastasize [63]. The spread of breast carcinoma can be directly to skin, including the nipple and areola, or the chest wall; can be through lymphatics (axillary, internal mammary and supraclavicular) or through the blood to mainly the lungs, liver, brain and bone [74].

1.6.4. Targeted therapies in breast cancer

Due to the variety of morphologic, molecular and clinical manifestations of breast cancer, its therapy is still nowadays of high complexity [75] and is continuously evolving. Breast cancer heterogeneity also results in a range of responses to treatment [76]. Ideally, the treatment needs to be tailored to every tumour and every patient, with the main challenges of dealing with treatment resistance, recurrence and metastasis [64]. Therefore, the treatment strategy selected will vary depending on the tumour features, mainly the molecular subtype, grade and stage of the tumour [63][64]. For example, different strategies are needed when targeting early stages of breast cancer compared to advanced stages. The management of the disease can be divided into *localised strategies*, such as surgery or radiotherapy, and *systemic therapy approaches*. Some examples of systemic treatments are: endocrine therapy (suggested for HR-positive cases), anti-HER2 therapy (suggested for HER2-positive cases), chemotherapy and bone stabilizing agents [64]. Most of the previously mentioned therapies can have severe

adverse effects and patients can develop resistance to the treatments [64]. Other therapies for specific cases of breast cancer have been developed, for example, PARP inhibitors (PARPi) such as olaparib, talazoparib, veliparib or rucaparib [64]. This treatment is directed to patients with BRCA1 or BRCA2 mutations, which are mainly found in cases of TNBC. BRCA1 and BRCA2 genes are translated into proteins that are involved in DNA repair. The PARP (poly-(ADP-ribose) polymerase) proteins are also involved in the DNA damage response. They recruit DNA repair proteins, such as these BRCA1 and BRCA2, to different damaged sites in the DNA to perform the repair [77]. PARPi inhibit PARP proteins and, consequently, cells defective in BRCA functions are not recruited to repair DNA damage [78]. Other therapies that are emerging can be divided according to the molecular subtype of breast cancer to which they are directed to [64]:

- **Emerging therapies for HR-positive breast cancer**

For this subtype of breast cancer, there are inhibitors targeting the mTOR/PI3K/Akt signalling pathway, such as Pan-PI3K, isoform specific PI3K, mTORC1, Akt and CDK4/6 inhibitors.

- **Emerging therapies for HER2-positive breast cancer**

The previous therapies mentioned (mTOR/PI3K inhibitors and CDK4/6 inhibitors) can be also included here, as well as new antibodies, such as antibody drug conjugates (ADCs) (*e.g.* trastuzumab-emtansine or T-DM1), chimeric antibodies or bio-specific antibodies; HER2-derived peptide vaccines and new Tyrosine Kinase inhibitors (TKIs).

- **Emerging therapies for Triple Negative breast cancer**

Antibody drug conjugates and targeted antibodies are also being explored for this breast cancer subtype, as well as vaccines and other forms of immunotherapy.

The recent emergence of immunotherapy and the heterogeneity across breast cancer subtypes, makes it necessary to extend the analysis of the tumour immune microenvironment across subtypes to know if there are mechanisms that could allow to target patients and therefore make them eligible for this kind of therapy, which has not been the most common therapy for this cancer type until now.

1.6.5. Breast cancer and its Tumour Immune Microenvironment

Breast cancer develops in a context where the most abundant cell type is the cancer-associated fibroblast, but the tumour microenvironment (TME) also includes the surrounding blood vessels, either pre-existing or newly formed, immune cells and components of the extracellular matrix [63][61]. The immune part of the TME it is known as Tumour Immune Microenvironment (TIME) and refers to the different subpopulations of the immune system that are found in the tumour niche. In general, in the immune system we can differentiate a group of immune cells that derive from a common myeloid progenitor (monocytes, macrophages and dendritic cells), which are responsible of the innate immune response, and a group of immune cell that derive from a common lymphoid progenitor, which are responsible of the adaptive immune response (B cells, T cells and NK cells) (**Figure 7**) [79]. Natural Killer (NK) cells are an exception to this, they come from a lymphoid progenitor that also forms the T and B cells, but they share several similarities in function with the myeloid cells (**Figure 7**) [79]. The major players in the TIME of breast cancer can be divided into immunosuppressive (pro-tumoral: M2-like macrophages, myeloid-derived suppressor cells and regulatory T cells), and immunostimulating cells (anti-tumoral: dendritic cells, CD4/CD8 cytotoxic T cells and NK cells) (**Figure 8**) [80].

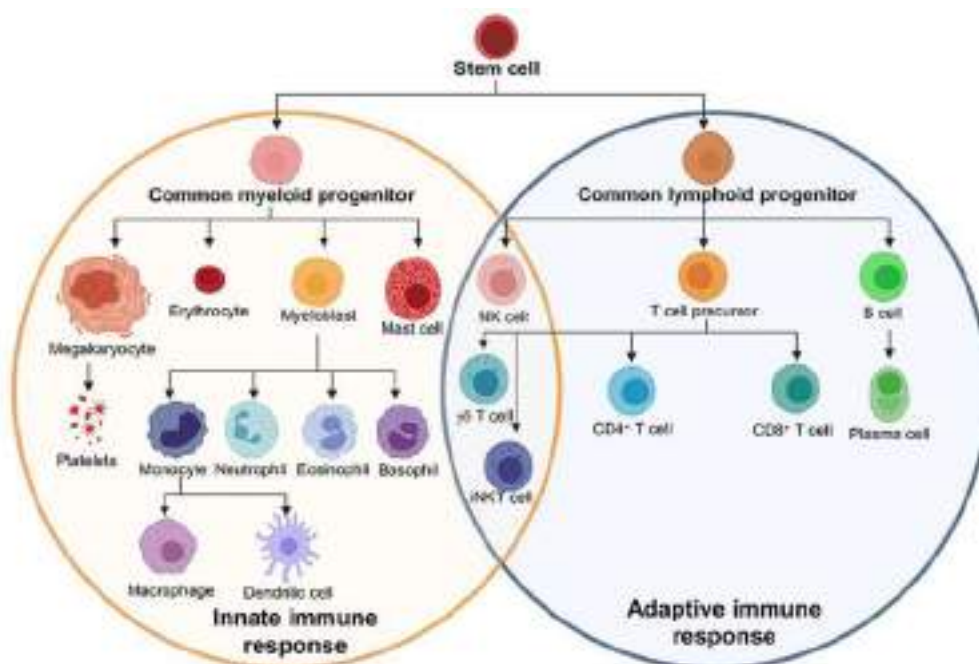


Figure 7. Immune cell lineages involved in the two main immune responses. NK cells derive from a lymphoid progenitor but share functions with the cells derived from a common myeloid progenitor. Source: Charles D. Murin, *Frontiers in Immunology* (2020) [79].

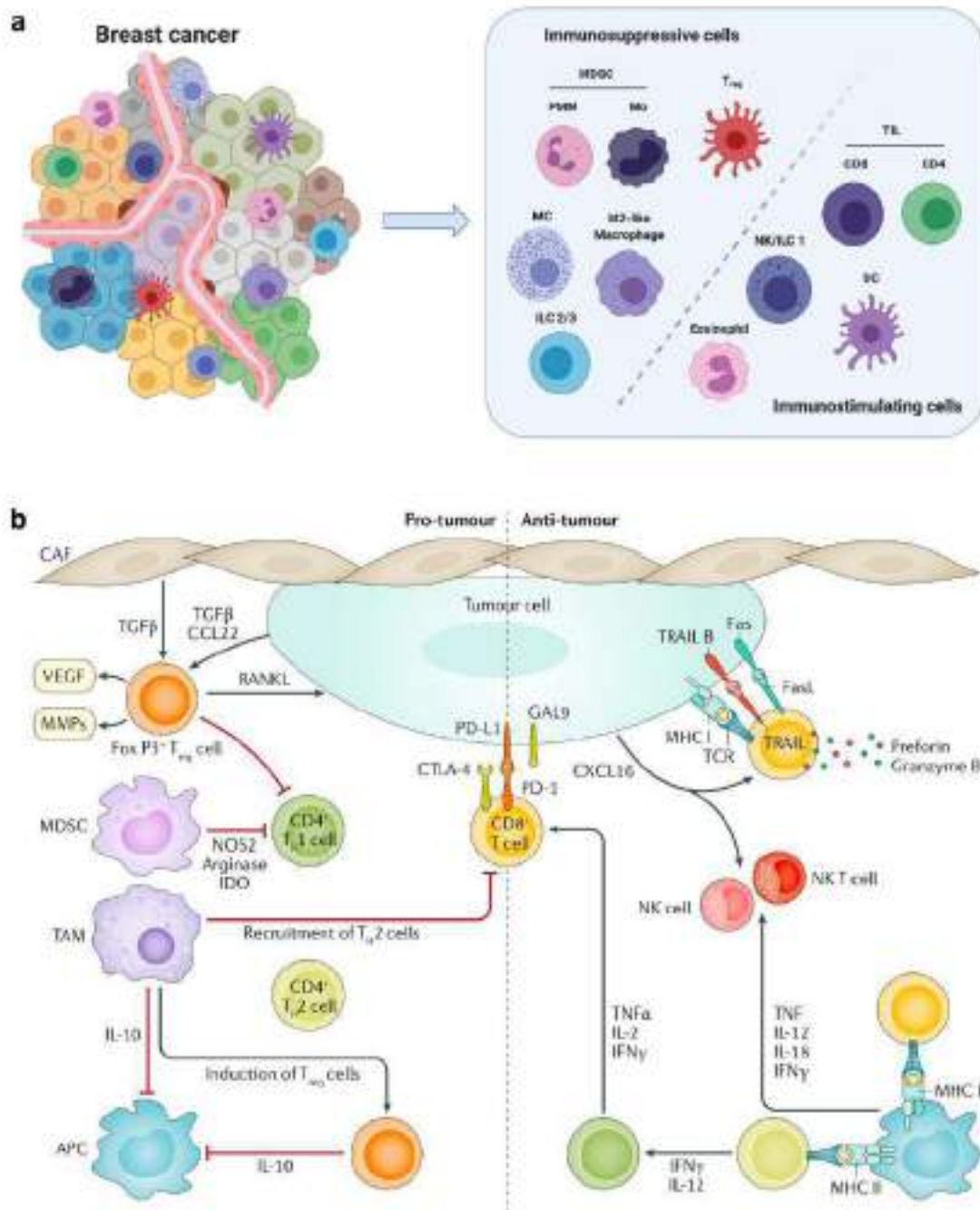


Figure 8. (a) Major immune populations in the tumour microenvironment of breast cancer. Adapted from Salemme et al. (2021) [80]. Populations are dividing depending on if they are involved in an immunosuppressive (pro-tumour) context or in an inflammatory (anti-tumour) context. PMN: PolyMorphoNuclear. Mo: monocytic. MDSC: Myeloid-Derived Suppressor Cell. MC: Mast Cell. ILC: Innate Lymphoid Cell. TIL: Tumour Infiltrating Lymphocytes. NK/ILC-1: Natural Killer/Innate Lymphoid Cell Type 1. DC: Dendritic Cell. **(b) Molecules involved in the crosstalk between cancer cells and the tumour immune microenvironment in breast cancer.** Pro-tumour (left) and anti-tumour (right) context. From Harbeck et al. (2019) [63].

The continuous and dynamic interaction between the tumour and its microenvironment can either promote or hinder cancer progression (**Figure 8b**) [80]. Tumour infiltrating immune cells protect from tumour progression by eliminating immunogenic neoplastic cells (**Figure 9**), while at the same time, once the tumour becomes invasive, they can contribute to tumour resistance to therapies, shaping tumour immunogenicity and selecting resistant tumour clones able to escape the immune response [81][80][63]. One example of a mechanism that induces resistance is the expression of PD-L1 on tumour cells, which can bind PD1 expressed in T cells CD8+ and trigger inhibitory effects on these cells [82]. With this or other mechanisms, TIME can influence the outcome of immunotherapy and of many other anti-cancer therapies [80][61].

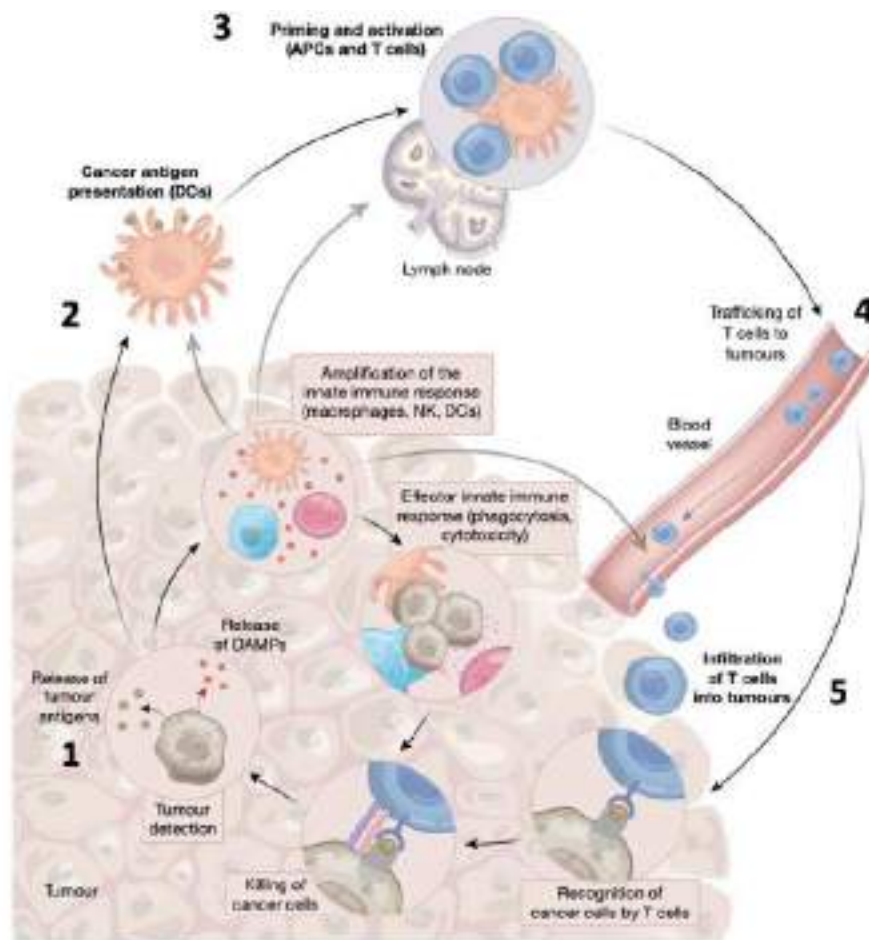


Figure 9. The cancer-immunity cycle. The cycle of immunity against the tumour starts with the presentation of cancer antigens that are liberated from dying cells (1). Tissue-resident Dendritic Cells (DCs) or DCs in draining lymph nodes sense and capture these cancer antigens (2) and initiate an immune response by presenting them to naïve T cells in lymphoid tissues (3). Naïve T cells CD8+ differentiate into cytotoxic T lymphocytes. T cells migrate through blood and lymphatic vessels (4) and can infiltrate through both to reach the tumour (5). Once inside the tumour, T cell can recognize the cancer cells and initiate the process to kill them. Killing of malignant cells can lead to the antigen release and DC activation (endogenous vaccination), thereby closing the cycle. Figure adapted from Demaria *et al.* (2019)[83] and image description (next) adapted from Palucka *et al.* (2016)[55].

Previously, breast cancer was considered a poor immunogenic cancer with a low response to immunotherapies, but the introduction of these therapies in the clinic have been reported to improve the outcome for many breast cancer patients [80]. The immunogenicity of breast cancer depends on the molecular subtype. TNBC and HER2-enriched are the highest immunogenic subtypes, while luminal A and B are the lowest [63]. It has been shown that the amount of tumour-infiltrating lymphocytes (TILs) influences positively the response to neoadjuvant treatment and the prognosis of breast cancer of TNBC or HER2-enriched subtypes [75][63], while the involvement of TILs in luminal subtypes is still not clear and there is still a lot of variability in response efficacy when it is applied [84]. This shows the need of further studies in this field.

2. HYPOTHESIS AND OBJECTIVE

2.1. Hypothesis

Exploring somatic mutations in a pan-cancer dataset at genome level could help to decipher consequences of mutational processes and cluster the cancer genomes into biologically relevant groups. Also, at protein level, the assessment of characteristics of protein coding changes underlying these somatic mutations could uncover relevant patterns within or across cancer types. All together would help to stratify patients in biologically relevant groups to personalise specific strategies of treatments.

2.2. General objective

To decipher consequences of mutational processes and cluster cancer genomes into biologically relevant groups exploring somatic mutation and their corresponding protein changes.

2.3. Specific objectives

Objective 1. To describe the landscape of somatic mutations of 25,499 cancer genomes.

Objective 2. To provide insights into the consequences of mutational processes in cancer based on the recurrent mutations in a pan-cancer dataset and to cluster cancer genomes according to the characteristics measured using 42 different genomic features.

Objective 3. To characterize the amino acid changes resulting from somatic mutations in a pan-cancer dataset considering different amino acid, evolutionary and structural properties.

Objective 4. To describe the protein changes of a highly mutated gene across cancer types and to study the association of different mutations with clinical and immunological characteristics.

The *Methods* (Section 4) and the *Results* (Section 5) are divided into four chapters that correspond to these four specific objectives (1-4) that were developed in this PhD thesis.

3. MATERIALS

3.1. Mutational data

We analysed a joint dataset of 25,499 cancer genomes covering >40 cancer types at the level of somatic mutations (substitutions and insertions/deletions) and their corresponding protein changes for the ones affecting coding. We combined the following four cohorts: (a) the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset, (b) The Cancer Genome Atlas (TCGA) dataset, (c) the Hartwig Medical Foundation (HMF) dataset and (d) the Breast Cancer Stratification (B-CAST) dataset. **Table 1** provides a basic description of these datasets with the type of specimen and sequencing, number of donors and number of different cancer types. The list of cancer types that are included in PCAWG and TCGA dataset are shown in **Figure 10**, together with their corresponding abbreviations and the number of donors.

Table 1. Summary of the main characteristics of the four individual datasets.

Dataset	Specimen Type	Sequencing	Number of donors	Number of different tumour types
PCAWG	Primary	WGS	2,583	37
TCGA	Primary	WES	9,104	32
B-CAST	Primary	Panel (323 genes)	9,255	1
HMF	Metastatic ¹	WGS	4,557	38

¹ All donors have metastatic disease, but for 100 donors the biopsy was taken from the primary tumour.

A description of the individual datasets at the level of genomic mutations is presented in *Results - 5.1 (Chapter 1)*. For the HMF dataset we had initially 4,901 samples for 4,570 donors. We excluded ten donors because according to the metadata they had multiple primary tumours in different organs (**Table 2-A**). It was not clear whether this was truly the case or that the primary location was revised throughout the treatment of the patient. An example of the latter is possibly a donor that was listed as having primary tumours in the stomach, oesophagus and gastroesophageal junction. Another three donors were excluded because they had primary tumours of different subtypes (**Table 2-B**). For another 292 donors there was more than one sample available, in which case we selected one sample per donor to not have multiple measurements for the same donor in our data. To select a single sample per donor, we gave priority to the samples taken from the metastatic tumour and not the primary tumour. This reduced the

number of samples to one for just three donors. For one donor it reduced it to two samples. The next criterion we applied was that we selected the sample with the highest maximum tumour purity, under the condition that the range of the lowest estimated purity to the maximum did not overlap with any of the samples of the same donor. This was sufficient for 208 donors. For 28 donors we selected the sample that had the highest maximal purity and added the criterion of having RNA-Seq data available. For 35 donors we added the criterion of selecting the sample with the earliest biopsy date to reduce the number of treatments the donor had undergone. For 17 donors there was no RNA-Seq data available and we therefore selected the sample with the earliest biopsy date. Finally, for one donor we selected the sample with the highest maximal purity, despite the range of the lowest estimated purity to the maximum purity overlapped with other samples, as for this donor no RNA-Seq data was available and the biopsy dates were missing.

Table 2. HMF donors excluded. Donors excluded from the HMF dataset because of potentially conflicting metadata regarding the primary tumour location or primary tumour subtype.

A	Donor ID	Primary tumour location
	HMF001168	Uterus and Bone/Soft tissue
	HMF001668	Urothelial tract and Uterus
	HMF000726	Gastroesophageal and Stomach and Esophagus
	HMF003321	Vagina and Uterus
	HMF000963	Gallbladder and Bile duct
	HMF001663	Esophagus and Stomach
	HMF003533	Colorectum and Breast
	HMF001184	Anus and Colorectum
	HMF002723	Lymphoid tissue and Skin
HMF002243	Skin and Kidney	
B	Donor ID	Primary tumour subtype
	HMF002878	ER-positive/HER2-negative and Adenocarcinoma ¹
	HMF001187	Small cell carcinoma and Non-small cell carcinoma
	HMF002363	ER-positive/HER2-negative and Triple negative

¹ The sample with the adenocarcinoma annotation had a later biopsy date and thereby more precise information on the subtype seems to have been revised.



Liver-HCC: hepatocellular carcinoma	BRCA: Breast Invasive carcinoma
Panc-AdenoCA: pancreatic adenocarcinoma	LUAD: Lung adenocarcinoma
Prost-AdenoCA: prostate adenocarcinoma	LGG: Brain Lower Grade Glioma
Breast-AdenoCA: breast adenocarcinoma	HNSC: Head and Neck squamous cell carcinoma
Kidney-RCC: renal cell carcinoma	PRAD: Prostate adenocarcinoma
CNS-Medullo: central nervous system - medulloblastoma	THCA: Thyroid carcinoma
Ovary-AdenoCA: ovary adenocarcinoma	LUSC: Lung squamous cell carcinoma
Lymph-BNHL: B-cell non-Hodgkin lymphoma	SKCM: Skin Cutaneous Melanoma
Skin-Melanoma: skin melanoma	UCEC: Uterine Corpus Endometrial Carcinoma
Eso-AdenoCA: esophagus adenocarcinoma	STAD: Stomach adenocarcinoma
Lymph-CLL: chronic lymphocytic leukaemia	BLCA: Bladder Urothelial Carcinoma
CNS-PiloAstro: central nervous system – pilocytic astrocytoma	KIRC: Kidney renal clear cell carcinoma
Panc-Endocrine: pancreatic endocrine neoplasm	LHC: Liver hepatocellular carcinoma
Stomach-AdenoCA: stomach adenocarcinoma	GBM: Glioblastoma multiforme
Head-SCC: head/neck squamous cell carcinoma	COAD: Colon adenocarcinoma
ColoRect-AdenoCA: colorectal adenocarcinoma	CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma
Thy-AdenoCA: thyroid adenocarcinoma	KIRP: Kidney renal papillary cell carcinoma
Lung-SCC: lung squamous cell carcinoma	SARC: Sarcoma
Uterus-AdenoCA: uterus adenocarcinoma	ESCA: Esophageal carcinoma
Kidney-ChRCC: chromophobe renal cell carcinoma	PCPG: Pheochromocytoma and Paraganglioma
CNS-GBM: central nervous system – glioblastoma multiforme	PAAD: Pancreatic adenocarcinoma
Lung-AdenoCA: lung adenocarcinoma	TGCT: Testicular Germ Cell Tumours
Bone-Osteosarc: bone osteosarcoma	THYM: Thymoma
Biliary-AdenoCA: biliary adenocarcinoma	ACC: Adrenocortical carcinoma
Bladder-TCC: bladder transitional cell carcinoma	READ: Rectum adenocarcinoma
Myeloid-MPN: myeloproliferative neoplasm	MESO: Mesothelioma
SoftTissue-Liposarc: soft tissue liposarcoma	UVM: Uveal Melanoma
Cervix-SCC: cervix squamous cell carcinoma	KICH: Kidney Chromophobe
CNS-Oligo: central nervous system - oligodendroglioma	OV: Ovarian serous cystadenocarcinoma
Bone-Benign: benign neoplasm of the bone	UCS: Uterine Carcinosarcoma
SoftTissue-Lelomyo: soft tissue leiomyosarcoma	DLBC: Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
Breast-LobularCA: breast lobular carcinoma	CHOL: Cholangiocarcinoma
Myeloid-AML: acute myeloid leukaemia	
Bone-Epith: epithelial neoplasm of the bone	
Breast-DCIS: breast ductal carcinoma in situ	
Cervix-AdenoCA: cervix adenocarcinoma	
Myeloid-MDS: myelodysplastic syndromes	

Figure 10. List of cancer type abbreviations and complete names in PCAWG and TCGA. Ordered from highest to lowest number of samples.

Definition of mutations

For SSMs there are 16 possible subtypes. However, we can neither detect substitutions with a base of the same type (*e.g.*, A>A) nor do we usually know on which strand the (pre-)mutagenic event happened first (*e.g.*, A>C is equivalent to T>G on the other strand). Therefore, we combined the substitutions that are each other's reverse complement and refer to them by the pyrimidine of the mutated base pair: C>A, C>G, C>T, T>A, T>C and T>G. Analogously to SSMs, for 1 bp SIMs, these are the four subtypes A/T deletions, C/G deletions, A/T insertions and C/G insertions.

The four datasets differ in how they deal with multiple substitutions close to each other in the sequence. In the case of PCAWG all were considered as single-base substitutions. A consensus of four mutation callers (*see Methods: 4.2.2. PCAWG cohort – mutation calls*) was used and in several cases the individual callers only supported one single-base event, and only the consensus resulted in a multiple base substitution call. We regarded substitutions directly next to each other (median number across samples: 25) as separate single-base events since, aside from the very limited numbers, in several cases the individual callers only supported one single-base event, and only the consensus resulted in a multiple base substitution call. For the other three data sets there are multiple base substitutions. In addition, only HMF also considers the following type of cases as a single event: ATA>CTC. We left all mutation calls as provided and we only used the mutations that were marked as 'PASS'.

Overlap between datasets

For the analyses in which we combined the data from all datasets we had to consider the following in terms of overlap:

- PCAWG includes a subset of the TCGA donors. When we worked with TCGA dataset alone, we worked with the complete set of donors, but for the analysis for which we used TCGA and PCAWG together, we excluded the donors they have in common with the TCGA dataset (653 donors).

- One cancer genome in HMF has a percentage of overlap on the level of mutations higher than expected with a cancer genome in PCAWG and was therefore excluded.

3.2. RNA-Seq, methylation and clinical data

RNA-Seq data was available for a subset of donors in PCAWG, TCGA and HMF. PCAWG and TCGA used STAR as aligner and to obtain the counting data. HMF used STAR as their aligner and provided the results of Isofox [85], a tool they developed for counting fragment support for identifying and counting gene and transcript features using genome aligned RNA-Seq data. We explored the option to combine the three datasets to work with all the RNA-Seq samples together. We performed a principal component analysis (PCA) and observed that the samples split according to the cohort they belong to, with less split between PCAWG and TCGA. One possible explanation for this is the use of different pipelines. However, as the HMF dataset contains mostly metastatic samples it was not to be excluded that this also may explain some of the differences. Therefore, I decided to work with the datasets separately across the different analyses.

Methylation data was available in the PCAWG dataset and in TCGA dataset. For both datasets Sesame was used to compute the beta values.

The clinical data available for the different datasets is summarised in **Table 3**, although in some cases it was not available for all the samples in that dataset.

Table 3. Clinical data available across the different datasets.

	PCAWG	TCGA	B-CAST	HMF
Tumour grade	✓	✓	✗	-
Tumour stage	✓	✓	✗	-
Cancer subtype	✓	✓	✓	✓
Survival	✓	✓	✗	✓
Age of the patient	✓	✓	✗	✓

3.3. Protein annotation and protein structures

Protein structures were downloaded from the Protein Data Bank (PDB) [86] using the European Portal (Protein Data Bank in Europe or PDBe). For the proteins with a low-quality structure or without any structure we downloaded the protein model (if any) from the Swiss-Model Repository [87]. For some specific analyses we needed to use structures of high quality and we selected those crystal structures with a resolution < 2 Å. For example, to compute the change in the free energy of protein folding upon mutation, we used FoldX, which requires accurate structures to be able to predict the potential changes successfully.

3.4. Data availability

PCAWG data was downloaded from the ICGC Data Portal at the section “DCC Data releases - PCAWG” that can be accessed at <https://dcc.icgc.org/releases/PCAWG>. The mutational, methylation and clinical data from TCGA included in this study is all public and was downloaded from <https://gdc.cancer.gov/about-data/publications/pancanatlas>. The RNA-Seq data (counts format) per cancer type was downloaded through the ‘TCGABiolinks’ R package. The information related to the Immune Landscape of Cancer in TCGA dataset is available at: <https://gdc.cancer.gov/about-data/publications/panimmune>. HMF data was available upon request at <https://www.hartwigmedicalfoundation.nl/en/data/data-access-request/>. At the moment of the deposition of this PhD Thesis, B-CAST data was under embargo which will not be lifted until the main paper has been published, after which one will be able to apply for access through the European Genome-Phenome Archive (EGA). I had access for this data as partner in the B-CAST project.

4. METHODS

Next, we describe the methods employed per chapter.

4.1. CHAPTER 1. Genomics landscape of 25,499 cancer genomes

4.1.1. Plots

All plots were done with the 'ggplot2' R package [88] under R version 3.6.0.

4.1.2. Mutational signatures

Mutational signatures were used as a proxy for the mutational processes that are predicted to be active in each sample. We were provided with the mutational signatures for PCAWG and TCGA. For HMF we generated the mutational signatures using *SigProfiler-SingleSample* [89] [90], which attributes a known set of mutational signatures to an individual sample. The inputs of the tool were the somatic mutations in the sample (VCF file) and the set of known signatures that we wanted to be assigned. We used as reference the COSMIC signatures v3.3. First, *SigProfilerMatrixGenerator* creates mutational matrices for all types of somatic mutations in the file. Next, the mutational matrices are fitted to the COSMIC matrices and the attribution of signatures is done. Per sample we obtain the relative percentage of the signatures that had been assigned to each sample.



Figure 11. Scheme of procedure followed by SigProfilerSingleSample. From an input of a file with the somatic mutations in the sample, the tool makes use of SigProfiler-MatrixGenerator, SigProfiler-Attribution and Sigprofiler-Plotting to do a final attribution of the mutations in the sample to a known set of COSMIC mutational signatures. 'Chromosomes' and 'signatures' images have been taken from Ashiquil et al. (2022) [91].

4.2. CHAPTER 2. Use case in PCAWG dataset. Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer

4.2.1. PCAWG cohort – quality control

We used the cohort of cancer genomes assembled by the PCAWG project [92] of the ICGC and TCGA. For every donor, whole-genome sequencing data was available for a normal-tumour pair and all samples were analysed uniformly. A detailed description of the quality control is provided in the PCAWG marker paper [92]. In short, 176 samples were excluded for various reasons as part of the quality control, most commonly because of contamination with RNA. Samples of another 75 donors were of borderline quality for various reasons, including a high percentage of paired reads mapping to different chromosomes [92] [93]. We decided not to include the samples of those donors, which left us with genomic data of 2,583 donors covering 37 tumour types (Appendix 1 - S1 Table). The distribution of the samples across the tumour types is also indicated in Appendix 1 - S1 Table. In case there were multiple tumour samples for the same donor, we selected a single sample following the decision made within the consortium. To make the decision five criteria were used as described by the PCAWG Drivers and Functional Interpretation Group [94]. In order of importance, they prioritized the sample: 1) of a primary tumour over metastatic and recurrent ones; 2) with a OxoG score over 40, which indicates low levels of oxidative damage artefacts [95]; 3) with the highest quality according to the star rating system [93]; 4) with RNA-Seq data available; 5) with the lowest level of contamination with foreign DNA. If none of these criteria led to the selection of a single sample, a random selection was made.

4.2.2. PCAWG cohort – mutation calls

The description of the procedure for the mutation calls is provided in the marker paper of the PCAWG consortium [92]. In brief, the sequenced reads of the respective normal and tumour sample pairs were aligned with BWA-MEM to the GRCh37/h19 genome. Four mutation calling pipelines were run on the resulting BAM-files for each normal/tumour sample pair. The pipelines used for calling SSMs were MuSE [96] and

three in-house pipelines developed at the Deutsches Krebsforschungszentrum (DKFZ) in collaboration with the European Molecular Biology Laboratory (EMBL), Wellcome Sanger Institute and Broad Institute, respectively. A consensus set was built by keeping those calls on which two or more callers agreed. SIMs were called by SMuFIN [97] and three pipelines developed by the same institutes as mentioned for SSMs. The consensus was determined by stacked logistic regression instead, as the level of agreement between the callers was lower than for SSMs. Furthermore, the SIM calls were left-aligned to make them comparable across samples. Several filters were applied to both the SSM and SIM calls to remove, among other things, calls due to oxidative damage artefacts [95] and germline variants. Great care was taken by the consortium to reduce the number of false positive mutation calls, resulting in a reliable dataset that is believed to be a conservative representation of the true set of mutations.

4.2.3. Features describing each cancer genome

We computed 29 general features and 13 related to recurrence (Table A in Appendix 1 - S1 File) to characterize different aspects of the somatic mutations in a cancer genome. We used the `vcfR` package in R to read in the VCF files [98]. The general features comprised the number of SSMs and SIMs (two features), the percentage of SIMs with respect to the total number of mutations (one feature), the distribution of SSMs and SIMs across the different subtypes (six and four features, respectively), and the homopolymer context of 1 bp SIMs for each of the four subtypes (four times four features). We used the `BCFtools` (version 1.5) to compute recurrence using the VCF files as input. Recurrence was captured by the overall percentage of recurrent SSMs and SIMs (two features), percentage of recurrent mutations of type SIM (one feature) and recurrence per SSM and SIM subtype (six and four features, respectively). The homopolymer context is not included in the recurrence features, as the number of recurrent SIMs is too low to stratify into 16 additional features. Except for the number of SSMs and SIMs, all other 40 features were in percentages.

4.2.4. Principal Component Analysis and hierarchical clustering on Principal Components

The R package FactoMineR (v1.41) was used for the PCA [99]. All input features for the PCA were scaled to zero mean and unit variance to account for the differences between the ranges of the features, especially with respect to the two features in absolute terms versus the ones in terms of percentages. The first 18 PCs explained together over 80% of the variance of the data. The remaining components were assumed to mostly represent noise in the data. The PCs were used as input to the ‘hierarchical clustering on principal components’ (HCPC) function from the FactoMineR package. The Euclidean distance was used as a measure of dissimilarity and the Ward criterion for linkage. We cut the hierarchical clustering tree at various heights to see a more global down to a more specific division of the samples. The HCPC function includes a consolidation step in the form of k-means clustering [100], which uses the centroids of the hierarchical clustering as a starting point. This consolidation step was repeated a maximum of 10 times. The k-means clustering increased the variance between clusters from 17.5 to 18.9. Other advantages of this hybrid approach are that it reduces the sensitivity of k-means clustering to outliers and the initial centroids are selected in an informed way instead of at random. As a consequence of this step, some samples were finally assigned to a different cluster than after the hierarchical clustering. We decided a division into 16 clusters that were named alphabetically (the details about this decision will be explained in *Results - Section 5.3.4*). A ‘v test’, included in the FactoMineR package, was used to determine which features were significantly associated with each cluster. This test compares the mean of a particular feature in a cluster to the overall mean in the dataset. We corrected the p-values of all ‘v tests’ for multiple testing using the Benjamini-Yekutieli method. A feature is considered to be significantly associated to a cluster if the adjusted p-value < 0.05.

4.2.5. Detection and enrichment of motifs

We collected for clusters A, E, G, H, L and M all SSMs of the subtype that is the most characteristic. This is C>A for clusters A and H, C>G for cluster E and M, C>T for cluster G and T>G for cluster L. In addition, we looked at T>G SSMs in cluster H to compare them

to cluster L. Next, we extracted from the reference genome (GRCh37/h19) the ten adjacent bases in 5' and 3' direction of the mutation using the *Rsamtools* package in R. We used the extracted sequence context as input to construct two sequence logos per cluster: one for the mutations that are recurrent within the cluster and one for those that are not. We include each recurrent mutation only once to avoid giving extra weight to highly recurrent mutations. As a measure of information content, we used the relative entropy [101] [102], which is defined for position i by:

$$RE_i = \sum_{b \in \{A,C,G,T\}} f(b_i) \log_2 \frac{f(b_i)}{P(b)}$$

Here, $f(b_i)$ stands for the frequency of base b (A, C, G or T) in position i and $P(b)$ stands for the prior probability of base b as determined by the frequency in the human genome (GRCh37/h19). The height of each base in the sequence plot is proportional to $f(b_i) \log_2 \frac{f(b_i)}{P(b)}$. A positive value corresponds to an enrichment of the base with respect to the prior probability and a negative value to a depletion. The relative entropy (RE_i) is zero, if all four bases are observed with the same frequency as the prior in position i . We set 0.25 as a threshold for RE_i to define the enriched motif. Furthermore, we computed per cluster the percentages of all, non-recurrent and recurrent SSMs that were in the sequence context that was found to be enriched in the recurrent SSMs. To estimate the percentage of the respective motifs in the human genome, we first slid a window of the same size (k) as the motif across the genome with a shift equal to the length of the motif and counted all possible k -mers. Next, we added to this the counts retrieved in the same way for the reverse complement of the reference sequence (corresponding to the opposing strand), since we also combined the reverse complements for each of the SSM subtypes. From this we computed the percentage of the enriched motif with respect to all k -mers and to the k -mer with the base that is mutated in the enriched motif at the same position.

4.2.6. Statistical tests

The correlation between every possible pair of the 42 features was measured by the Spearman's rank correlation coefficient using the R package Hmisc (v4.1–1). Multiple testing correction of the p-values of all correlation tests (including those in Appendix 1 - S2 Text) was done by the Benjamini-Yekutieli method. For the other correlations mentioned we also used the Spearman's rank correlation coefficient.

We used the Wilcoxon rank-sum test with continuity correction as the test of significance for differences in features observed between clusters.

The different proportions of sequence motifs between recurrent and non-recurrent SSMs were assessed by using χ^2 tests.

4.2.7. Plots

Figures 32, 34, 36 and 37, the pie charts in **Figure 35** and the plots in Appendix 1, except for S1, were made using the R package ggplot2 (v3.0.0). Figure 37, S3 Fig (Appendix 1) and S4 Fig (Appendix 1) additionally required ggseqlogo (v0.1) [103] and **Figure 33** was made with the use of the R package corrplot (v0.84). **Figure 38** was made using Microsoft PowerPoint and we also included images from the Servier Medical Art website (<http://smart.servier.com/>). The 'clustering tree' in S1 Fig (Appendix 1) was made using the clustree R package [104]. We have manually replaced the nodes in the tree with the pie diagram showing the distribution of tumour types in each cluster. For the colours of the different tumour types, we have made use of the script provided by the PCAWG consortium, available at: <https://github.com/ICGC-TCGA-PanCancer/pcawg-colour-palette>.

4.3. CHAPTER 3. Characterization of amino acid changes due to somatic mutations in protein coding genes

4.3.1. Workflow for the automatic evaluation of missense mutations

Missense mutations selection

We focused on the analysis of all the missense mutations in our joint dataset, irrespective whether they were the result of a single-base substitution or a multiple base substitution. Starting for the genomic mutation, for the genes with more than one transcript we selected the mutation in the canonical transcript, since the same mutation in a different transcript could translate to a different amino acid. As TCGA mutation data already is given for a list of canonical transcripts, for the genes that were in common with PCAWG or HMF we used the one selected in TCGA. For the genes that were just mutated in PCAWG or HMF we selected the canonical according to UniProt annotation. For the cases without a canonical transcript, we followed the UniProt rule of selecting the longest one.

Protein features defined

Eight features were selected for the evaluation of the amino acid changes, which we considered to provide interesting information for elucidating their potential relevance. These eight features are described next.

1) Chemical change

The side chain properties such as volume, polarity, acidity, basicity, conformational flexibility and the ability to form, for example, a hydrogen bond or a salt bridge, vary across the different groups of amino acids [105]. These characteristics could play a crucial role in protein folding, stability, interaction of protein-protein complexes and protein function. Therefore, a mutation that results in a different amino acid with a different biochemical group usually involves a significant alteration. We considered the classification of the amino acids according to the charge of their side chain (polar, non-polar, acidic polar and basic polar amino acids, **Figure 3**) and established nine categories

depending on which was the change (**Figure 12**). For example, “same category” indicates that the original amino acid and the amino acid resulting from the mutation belong to the same group (e.g., both acidic polar amino acids). The category “gain of polarity” indicates that the original amino acid was non-polar and the one after the mutation is polar.

	POLAR	ACIDIC POLAR	BASIC POLAR	NON POLAR
POLAR	same category	Gain (-) charge	Gain (+) charge	loss of polarity
ACIDIC POLAR	Loss of charge	same category	Change of charge	loss of polarity
BASIC POLAR	Loss of charge	Change of charge	same category	loss of polarity
NON POLAR	Gain polarity	Gain (-) from non polar	Gain (+) from non polar	same category

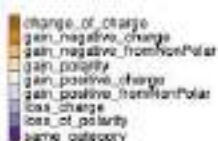


Figure 12. Categories among the cases of amino acid changes.

We also annotated if the original amino acid was replaced with an amino acid of a similar or different size. We considered 4 categories depending of the weight of the residues (**Table 4**): ‘big’ (tryptophan (W), tyrosine (Y), arginine (R) and phenylalanine (F)), ‘medium’ (histidine (H), methionine (M), glutamic acid (E), lysine (K) and glutamine (Q)), ‘small’ (aspartic acid (D), asparagine (N), isoleucine (I), leucine (L), cysteine (C), threonine (T), valine (V), proline (P)) and, finally, ‘tiny’ (serine (S), alanine (A), glycine (G)). We gave a score from 0 to 3, ‘0’ if the two amino acids (before and after mutation) were in the same size category, ‘1’ if the two amino acids were one position away from each other (e.g., from a ‘small’ to a ‘medium’ amino acid), ‘2’ if the two amino acids were two positions away in the size category (e.g., from a ‘small’ to a ‘big’ amino acid) and ‘3’ if the two amino acids were three positions away in the size categories (e.g., ‘tiny’ amino acid mutated to a ‘big’ amino acid or vice versa).

Table 4. Amino acid residues weights. Source: Bio-Synthesis Inc [106].

Amino acid (AA) name	AA (3 letters code)	AA (1 letter code)	Residue weight (Da)
Tryptophan	Trp	W	186.22
Tyrosine	Tyr	Y	163.18
Arginine	Arg	R	156.19
Phenylalanine	Phe	F	147.18
Histidine	His	H	137.14
Methionine	Met	M	131.20
Glutamic acid	Glu	E	129.12
Lysine	Lys	K	128.18
Glutamine	Gln	Q	128.13
Aspartic Acid	Asp	D	115.09
Asparagine	Asn	N	114.11
Isoleucine	Ile	I	113.16
Leucine	Leu	L	113.16
Cysteine	Cys	C	103.15
Threonine	Thr	T	101.11
Valine	Val	V	99.13
Proline	Pro	P	97.12
Serine	Ser	S	87.08
Alanine	Ala	A	71.08
Glycine	Gly	G	57.05

2) Solvent accessibility

The solvent accessibility of an amino acid or Accessible Surface Area (ASA) of the amino acid in the protein structure refers to the degree to which the amino acid is exposed to the solvent in which the protein is contained or if it is facing the inner core of the protein [107]. Considering a threshold, it can be established if an amino acid is exposed to the solvent in which the protein is contained or if it is in the inner core. In the first case, the amino acids are more likely involved in interactions with other proteins or substrates, while in the second case the amino acids would be buried in the structure and more relevant for maintaining the core of the protein. Changes in the solvent accessibility after mutation has been suggested to provide hints about the maintenance or change of protein packaging [108]. It has been suggested that pathogenicity is more frequently associated to the buried residues than to the exposed ones [108].

We used ASAquick (<http://mamiris.com/ASAquick/>) to obtain the ASA of the amino acid that was being mutated and computed the relative ASA, to classify the amino acid as buried or exposed (>20% is considered exposed) [108].

3) Secondary structure

Protein secondary structures, which are considered as the linkages between primary and tertiary structures, are defined as local structures that form the backbone of the protein and are stabilized by hydrogen bonds [109]. The main four secondary structures are an α -helix, β -strand and turn/loop (**Figure 13**) [110]. We annotated in which of these secondary structures the mutated amino acid is located the amino acid mutated. We retrieved this information from the PDB file of the protein in question if there was a protein structure available, or we took the PSIPRED (<http://bioinf.cs.ucl.ac.uk/psipred/>) predictions collected in DescribePROT [111] when there was no structure available.

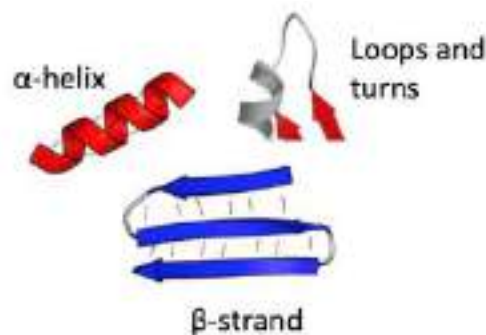


Figure 13. Protein secondary structures: (a) α -helix, (b) β -strand and (c) turn or loop. Adapted from Shafee, T. (2020).

4) Domain

Domains are functional or structural units defined in the proteins [112]. They are normally responsible for a particular function or interaction that contributes to the overall role of the protein [112]. For all the protein coding genes for which we had mutations, we did the crosslink from the Ensembl Transcript ID (the canonical transcripts) to the UniProt ID using the R package *biomaRt* [113]. With the UniProt ID, we downloaded the corresponding 'xml files' that contain all the information for each protein. Parsing the file with an 'in-house' script, we collected the domain information as defined by the InterPro [114] resources PROSITE [115], Pfam [116] and SMART [117]. We annotated for each amino acid mutated whether it is part of a protein domain, if any is defined for the protein in question.

5) Disruption of specific site (active, metal, protein-protein binding sites)

We also retrieve from the UniProt annotation ('xml file') whether the amino acid mutated is in any relevant site, for example an active site, binding site, site, zinc finger or DNA binding site. In the case of PIK3CA, apart from the annotations available in the UniProt file, we annotated the amino acids involved in the interactions between p110 α (PIK3CA) and p85 α (PIK3R1) using the Chimera software [118].

6) Amino acid conservation

The amino acid conservation is based on the estimation of evolutionary rate of the amino acid in the protein sequence or structure [119]. This indicates how well an amino acid is conserved across species. Extracting conservation scores from a multiple sequence alignment of homologous proteins can provide interesting information, since highly conserved residues are generally considered to be critical for protein function [119]. We obtained the pre-calculated evolutionary conservation scores from ConSurf-DB (information obtained after for now their last update: November 4th, 2019). The conservation scores go from 1 to 9 (Figure 14), where '1' is lowly conserved or a more variable amino acid and '9' highly conserved or a not variable amino acid. ConSurf [119] obtains the score per amino acid by doing a multiple sequence alignment (**Figure 14**) with homolog sequences from different species and considers how variable each amino acid position is.

For the PIK3CA analysis, we obtained the results of amino acids conservation from the ConSurf server (<https://consurfdb.tau.ac.il>), to be able to download the different files that the tool provides such as the scripts to display the structure with the corresponding colours according to the conservation score computed per residue (**Figure 14**).

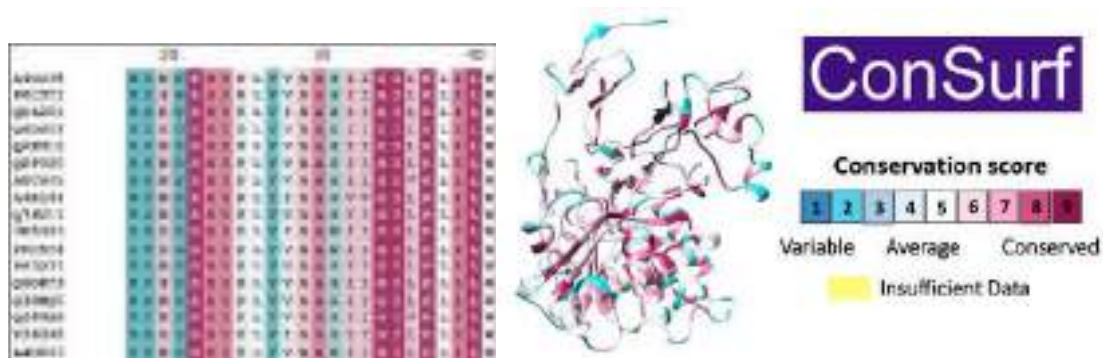


Figure 14. Amino acid conservation scores computed by ConSurf [119]. From the left, piece of a multiple sequence alignment of homologous proteins from different species, each amino acid position is coloured according to the conservation score computed. Next, protein structure with the conservation scores indicated and legend with the meaning of the different colours, towards blue less conserved.

7) Amino acid mutated in a hotspot in the 3D structure

Finding 3D clusters of amino acids that are mutated inside the protein structure can be indicative of relevant sites for the correct function of the protein [120]. Also, the identification of clusters in which there are mutations that are already considered hotspots in cancer may help us find less frequent mutations that could have the same implications as the hotspot and, therefore, are of interest to study. We used mutation3D [30] to look for potential clusters among the mutations in our data (**Figure 15**). We also included the annotation of amino acids, if any was described as involved in any particular function, to point out cluster including these amino acids since they could be relevant in terms of having an effect on protein function.

The computation of the statistical significance of the clusters found by mutation3D, as explained by the authors, is done in the next way: “*mutation3D performs an iterative bootstrapping method to calculate a background distribution of cluster sizes arising from a random placement of an equivalent number of substitutions in a given protein structure. By default, mutation3D will randomly rearrange all amino acid substitutions 15,000 times in a given structure and calculate the minimum complete linkage (CL) distance at which a cluster of size n (where n is all cluster sizes found in the original data) is observed in the randomized data. For each cluster in the original data, P values are computed empirically as the percentile rank of its CL distance among all CL distances for randomized clusters containing the same number of amino acid substitutions*” [30].

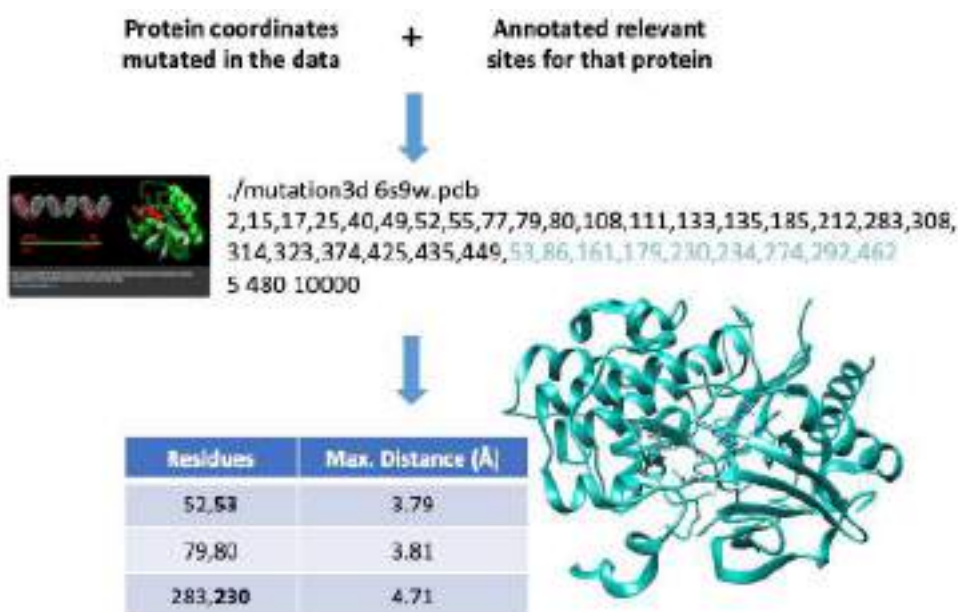


Figure 15. Schematic example of the use of mutation3D to find clusters in the 3D structure of a protein in our workflow.

8) Change in the free energy of protein folding

Protein stability is a fundamental property affecting proteins function, activity, and regulation [121]. The final shape adopted by the protein is the most energetically favourable one. Computing the change in the free energy of protein folding upon mutation is indicative of how the protein structure has been affected. We made use of FoldX [122] to find mutations affecting the stability of the protein (destabilizing or stabilizing mutations) or not affecting. The thresholds used to determine if an amino acid change was destabilizing, stabilizing or not affecting the stability of the protein are shown in **Figure 16**.

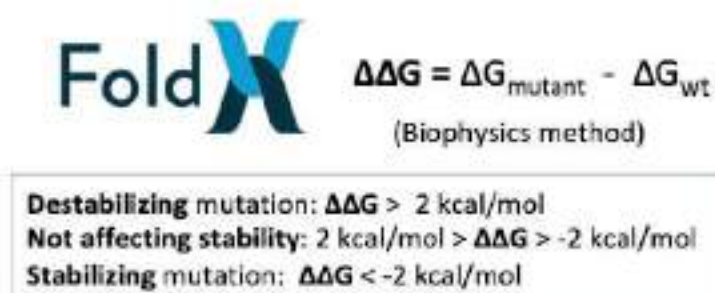


Figure 16. FoldX computation to obtain the change in free energy degrees for protein folding and the categories depending on threshold: destabilizing, stabilizing or not affecting stability.

4.3.2. Statistical methods

Dimensionality reduction method: Factor Analysis of Mixed Data

The Factor Analysis of Mixed Data (FAMD) is a principal component method that allows the analysis of a dataset containing both quantitative and qualitative variables. This method allows to analyse the similarity between individuals by taking into account the mixed types of variables [123]. We applied this method, available in the *FactomineR* package [99], to our set of mutations characterized by the eight protein features to visualize how similar or different the mutations are.

Hierarchical Clustering of Principal Components (HCPC)

As a next step to the FAMD, we performed a hierarchical clustering of principal components (HCPC) [124] with the aim of finding groups of mutations sharing features. This method is also available in the *FactomineR* package.

4.4. CHAPTER 4. Landscape of protein changes in p110 α (PIK3CA) in cancer

4.4.1. Data visualization

All plots were done with *ggplot2* R package under R version 3.6.0.

4.4.2. Statistics: Chi-squared test / Fisher's Exact Test

We applied the independence test Chi-square Test or Fisher's Exact Test (depending on sample size) to determine if there was a significant relationship between two categorical variables regarding different mutational signatures or conditions between (a) 'tumours with PIK3CA mutated' vs. 'tumours without a mutation in PIK3CA' and (b) 'tumours with a protein domain mutated' vs. 'tumours without that particular protein domain mutated'. With previous tests we tested whether the odds ratio was equal to 1 (alternative: two-sided). If the value of the odds ratio is 1 or close to 1 it means that there are no differences between the two conditions compared. These tests were done in R (v 3.6.0).

4.4.3. Survival analysis and associations

In breast, uterus and colorectal cancer cohorts, we investigated if there were differences in survival rate in tumours with PIK3CA mutated vs. not mutated, as well as considering the tumours mutated in the different PIK3CA protein domains. For the survival analysis, we added the parameters age, tumour grade and tumour stage as they may impact on survival. In the case of breast cancer, we also added the breast cancer subtype to the model for this reason.

We carried out univariable and multivariable survival analyses using Cox proportional hazards model. We used Kaplan-Meier curves for the visualization of the results where applicable. For both methods we used the *survminer* and *survival* R packages. All statistical tests were two-sided and we considered results to be statistically significant if the p-value is below 0.05.

4.4.4. RNA-Seq analyses: Differential expression analysis and Gene Set Enrichment Analysis

We performed a differential expression (DE) analysis using the *DESeq2* R package [125] in breast, uterus and colorectal cancer cohorts grouping the samples by PIK3CA mutational status. Next, we performed the DE analysis testing between different domains mutated. We considered a gene to be differentially expressed if the p-value was below 0.05. In addition to looking at the individual genes that were differentially expressed, we also performed a Gene Set Enrichment Analysis (GSEA) [126][127]. In this analysis you consider sets of genes together that individually might not be significantly differentially expressed. We performed the analysis using two different lists of gene sets downloaded from MSigDB (<https://www.gsea-msigdb.org/gsea/msigdb>): HALLMARKS (“h.all.v7.2.symbols”) and KEGG PATHWAYS (“c2.cp.kegg.v7.5.1.symbols.gmt”).

4.4.5. Deconvolution of bulk RNA-Seq samples using SPOTlight

Using SPOTlight [85] and a single-cell RNA-Seq reference for breast cancer, we performed a deconvolution of the bulk RNA-Seq samples from primary breast tumours in TCGA to study the tumour microenvironment (TME). We obtained the cellular composition estimated inside of each sample. The reference allowed us to obtain the different distribution of cancer and normal cells from the breast, stromal cell populations and immune cell populations.

Single-cell RNA-Seq reference

The single-cell RNA-Seq (scRNA-Seq) reference for breast cancer was obtained from Wu *et al.* (2021) [56], which included 26 primary tumours from the three major clinical subtypes of breast cancer: 11 ER-positive, 5 HER2-negative and 10 TNBC.

Deconvolution steps

As input for performing the deconvolution, SPOTlight computed the marker genes that characterize each of the populations in our reference in which we are going to deconvolute the bulk RNA-Seq samples. From these marker genes, we filtered out ribosomal and mitochondrial genes which correspond to bad quality and dead cells. We considered a log₂-fold change cut off of 0.5 (absolute number). We kept a list of genes manually curated specifically related to T cells to be able to separate them, irrespective of the log₂-fold change. The genes kept were: CCR7, CD274, CD3D, CD3E, CD4, CD40LG, CD8A, CD8B, CTLA4, EOMES, FOXP3, GZMA, GZMH, ICOS, IFNG, IL2RA, IL7R, ITGB1, KLRB1, LAG3, LEF1, NKG7, PASK, PDCD1, PDCD1LG2, PTPRC, RORA, SELL, TBX21, TCF7, TIGIT, TOX, TRAC, TRBC1 and TRBC2. In addition, SPOTlight selected the 3,000 highest variable genes in the whole data. The different cell populations that we considered in the different deconvolutions performed are shown in Table 5 and Table 6.

Table 5. Populations in the first level of annotation of the single-cell RNA-Seq reference [56]. With the first deconvolution of the bulk RNA-Seq data we obtained the relative proportion of each of these populations in each sample.

	Cell category	Cell population type
Deconvolution 1	Cancer and normal cells from breast tissue	Cancer SC
		Cancer Cycling
		Luminal Progenitors
		Mature Luminal
		Myoepithelial
	Stroma cells	Endothelial
		Endothelial Lymphatic LYVE1
		CAFs MSC iCAF-like
		CAFs myCAF-like
	Immune cells	DCs
		Macrophage
		Monocyte
		B cells
		Plasmablasts
		T cells + NK cells + NKT cells
		Cycling

Table 6. Populations in the second level of annotation of the single-cell RNA-Seq reference [56]. In the second deconvolution of the bulk RNA-Seq data we focused on macrophages, in the third on CD4 T helper cells and, in the fourth, on CD8 and NK cytotoxic cells. The different subpopulations inside these groups are shown.

	Cell population type	Cell population subtype
Deconvolution 2	Macrophages	Macro_CXCL10
		Macro_EGR1
		Macro_LAM1_FABP5
		Macro_LAM2_APOE
		Macro_SIGLEC1
Deconvolution 3	CD4 T Helper cells	T_cells_CD4+_CCR7
		T_cells_CD4+_IL7R
		T_cells_CD4+_T-regs_FOXP3
		T_cells_CD4+_Tfh_CXCL13
Deconvolution 4	CD8 and NK Cytotoxic cells	T_cells_CD8+_ZFP36
		T_cells_CD8+_GZMK
		T_cells_CD8+_IFIT1
		T_cells_CD8+_IFNG
		T_cells_CD8+_LAG3
		T_cells_NK_cells_AREG
		T_cells_NKT_cells_FCGR3A

5. RESULTS

5.1. CHAPTER 1. GENOMIC LANDSCAPE OF 25,499 CANCER GENOMES

At the basis of studying protein changes are the genomic mutations that caused them. Here we describe the genomic landscape of the four cohorts, individually, and when combined. We will also zoom into breast, colorectal and uterus cancer because of their relevance in the context of studying the PIK3CA gene (Chapter 4).

5.1.1. Genomic description of the individual datasets

Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset

The PCAWG dataset consists off whole genome sequencing of mostly primary tumours from 2,583 donors and 37 different cancer types (**Figure 17**). The largest cohort is liver cancer followed by pancreatic, prostate and breast adenocarcinoma. For breast adenocarcinoma we also have the subtype information available for 91 out of the 211 samples (**Figure 27a**). The most mutated cancer type is skin melanoma considering Somatic Single-base Mutations (SSMs) and Somatic Insertion/deletion Mutations (SIMs) combined (**Figure 18a**). If we only consider SSMs then it is still the most mutated cancer type (**Figure 18b**). The cancer type with the highest median of SIMs is lung squamous cell carcinoma (Lung-SCC) (**Figure 18c**). The mutation subtypes distribute differently depending on the cancer type (**Figure 19**). For example, percentage wise, C>A mutations are the most prevalent in the two forms of lung cancer, C>T mutations in skin cancer and C>G mutations in bladder cancer. Further details on the genomic landscape of the PCAWG dataset are provided in Appendix 1.

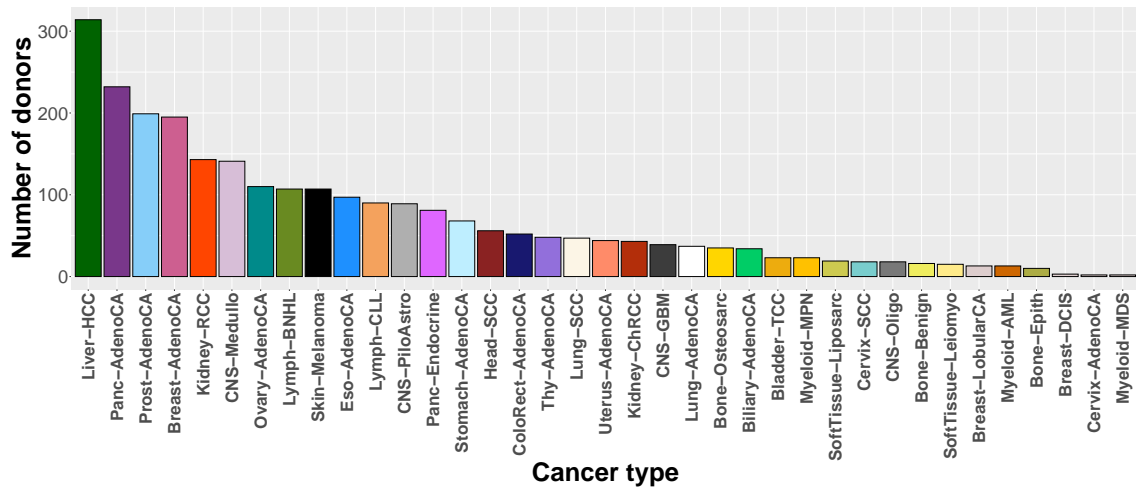


Figure 17. Number of donors per cancer type in the PCAWG dataset.

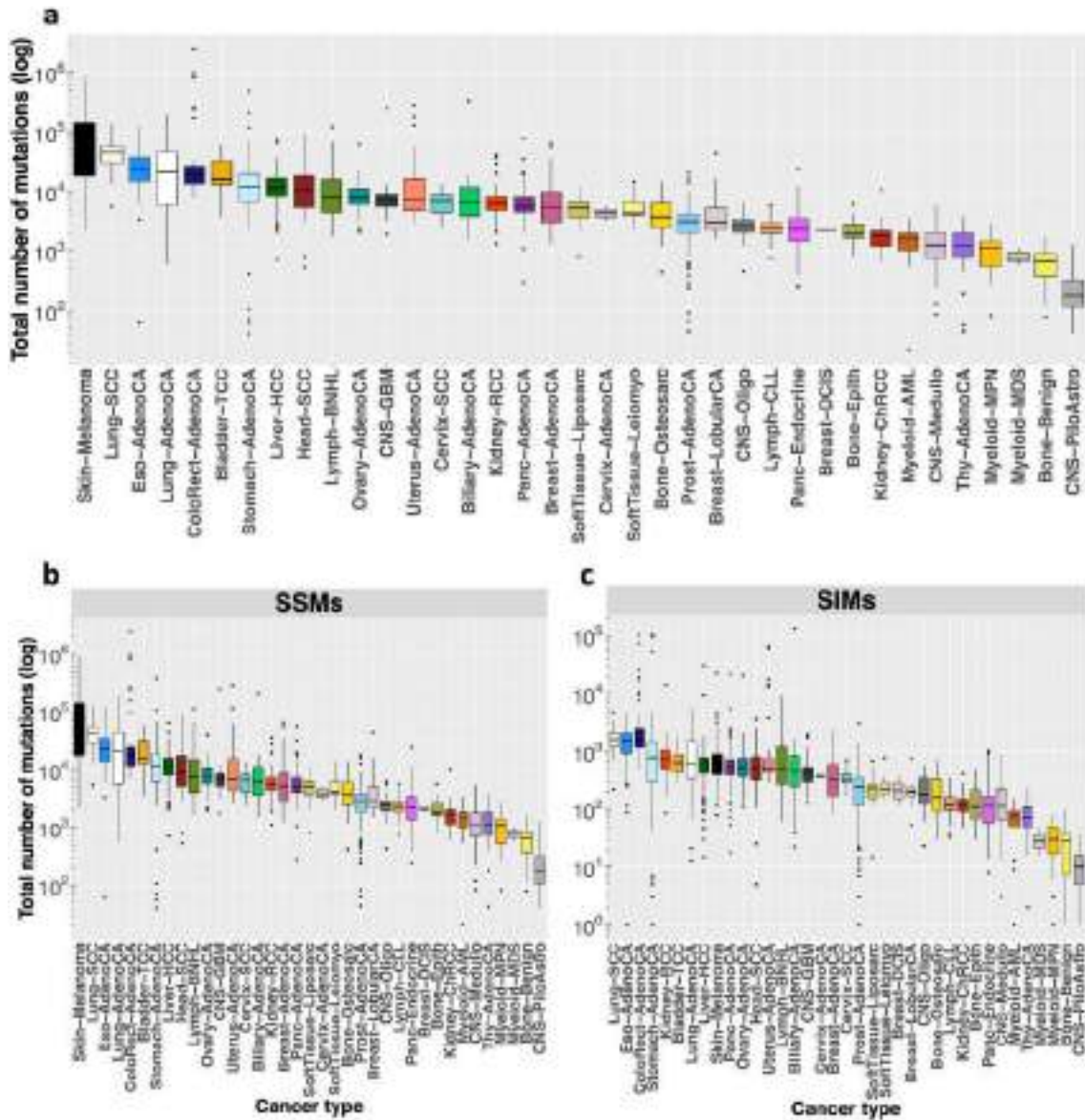


Figure 18. Distribution of (a) total number of mutations across cancer types, (b) number of SSMs and (c) number of SIMs across cancer types in the PCAWG dataset.

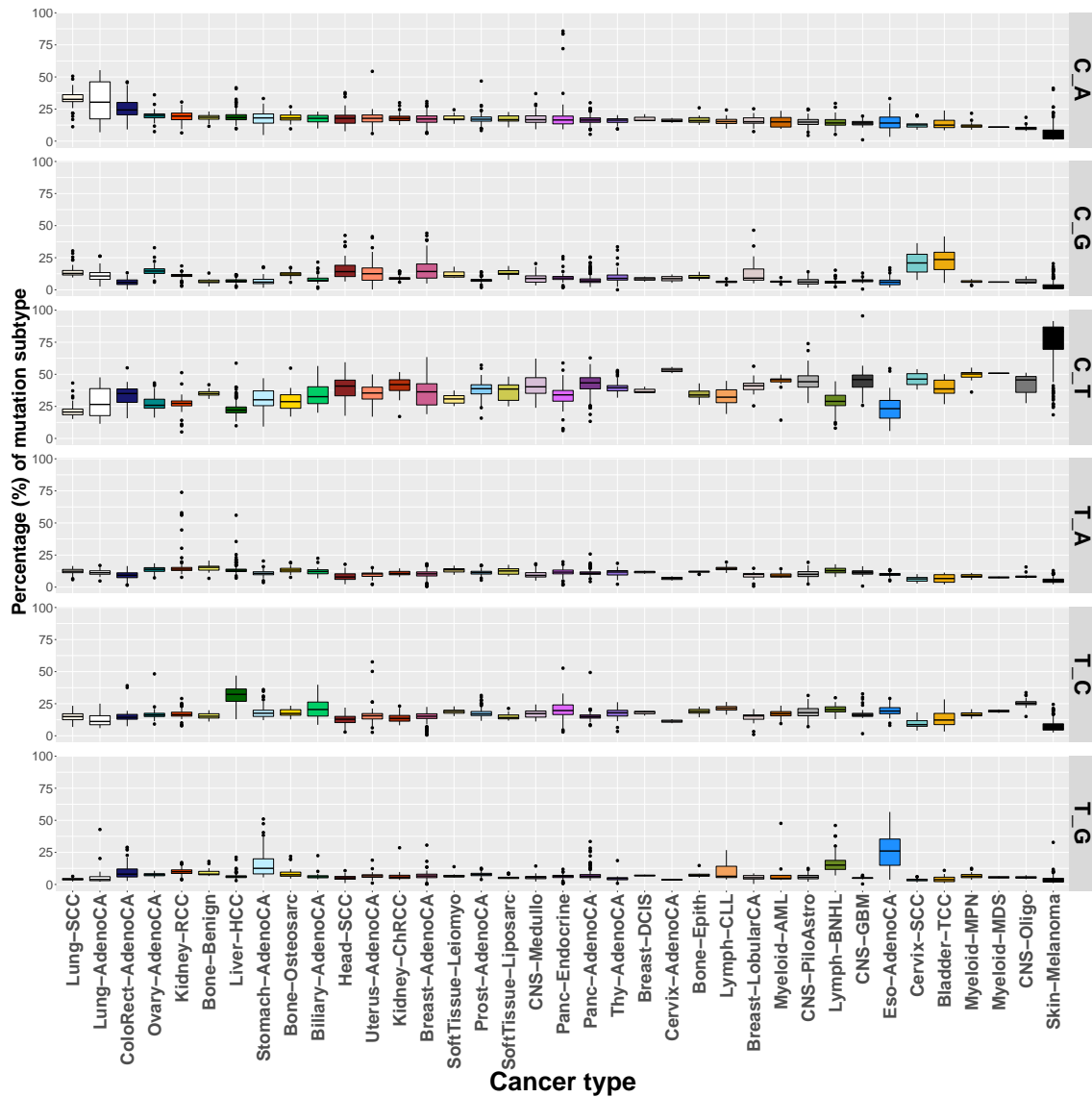


Figure 19. Distribution of mutation types across cancer types in the PCAWG dataset. Cancer types are ordered by the percentage of C>A mutations.

Hartwig Medical Foundation (HMF) dataset

The HMF dataset consists of whole genome sequencing of 4,557 donors with metastatic disease from 38 different primary tumour locations. For 100 donors the biopsy was taken from the primary tumour (44 breast, 25 nervous system, 8 oesophagus, 3 ovary, 3 pancreas, 13 prostate and 4 stomach cancer). For 122 tumours we do not know the primary location. **Figure 20** shows the number of donors available per location of the primary tumour of the corresponding metastatic tumour. The biggest cohort is formed by donors with breast cancer as primary cancer type followed by colorectal, lung and prostate cancer. For 752 out of the 787 breast cancer samples we know the subtype (**Figure 27b**). The most mutated cancer type is skin cancer, considering SSMS and SIMs combined (**Figure 21a**), or only SSMS (**Figure 21b**). The metastatic cancer with the highest median of SIMs is oesophagus cancer (**Figure 9c**). The mutation subtypes distribute differently depending on the primary tumour location (**Figure 22**), but they follow a trend similar to what we see in primary tumours (**Figure 19** and **Figure 25**). For example, the highest percentage of C>A mutations is in lung cancer, C>T mutations in skin cancer and C>G mutations in urothelial tract cancer.

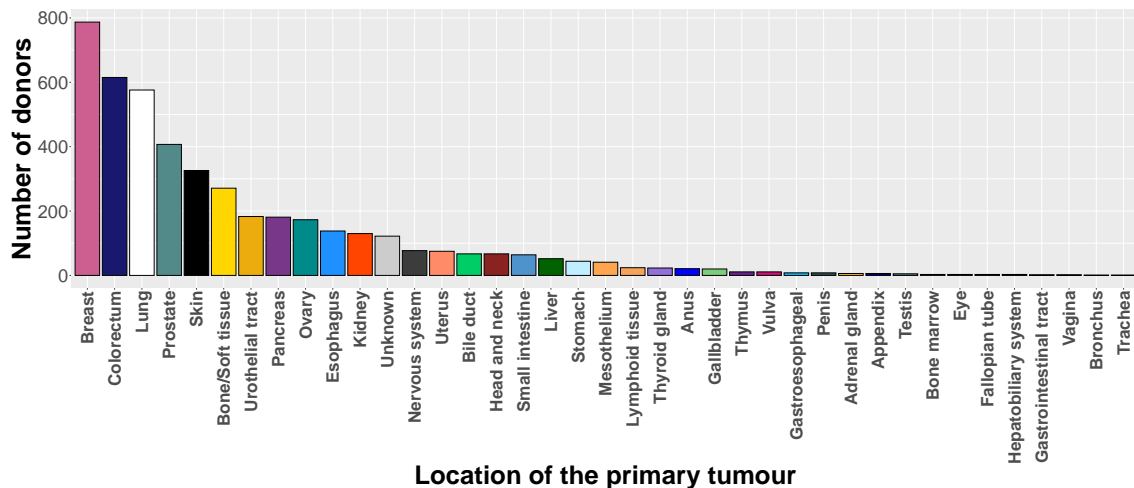


Figure 20. Number of donors per location of the primary tumour in the HMF dataset.

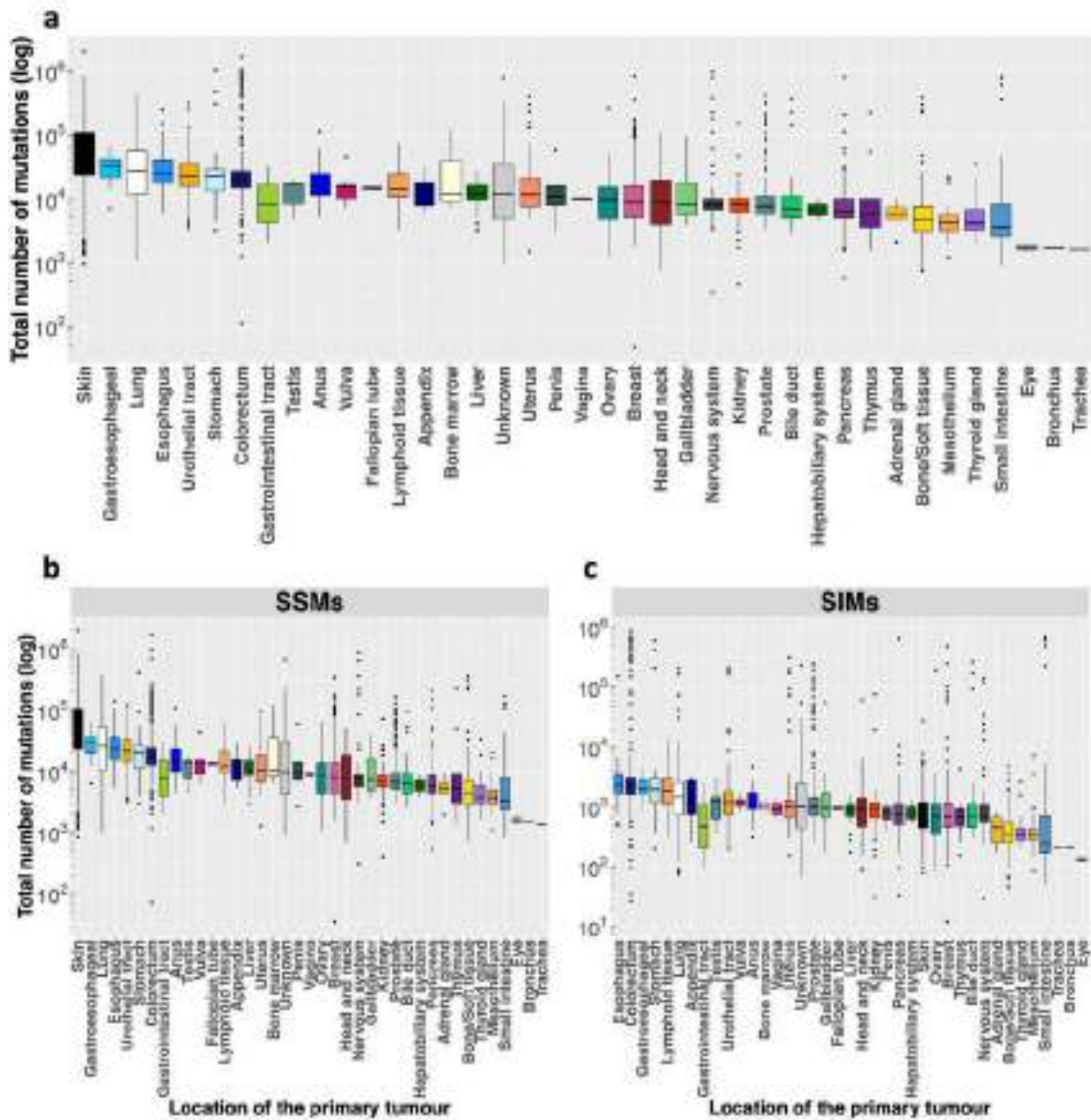


Figure 21. Distribution of (a) total number of mutations across the primary location of the metastatic tumours, (b) number of SSMs and (c) number of SIMs across the primary location of the metastatic tumours in the HMF dataset.

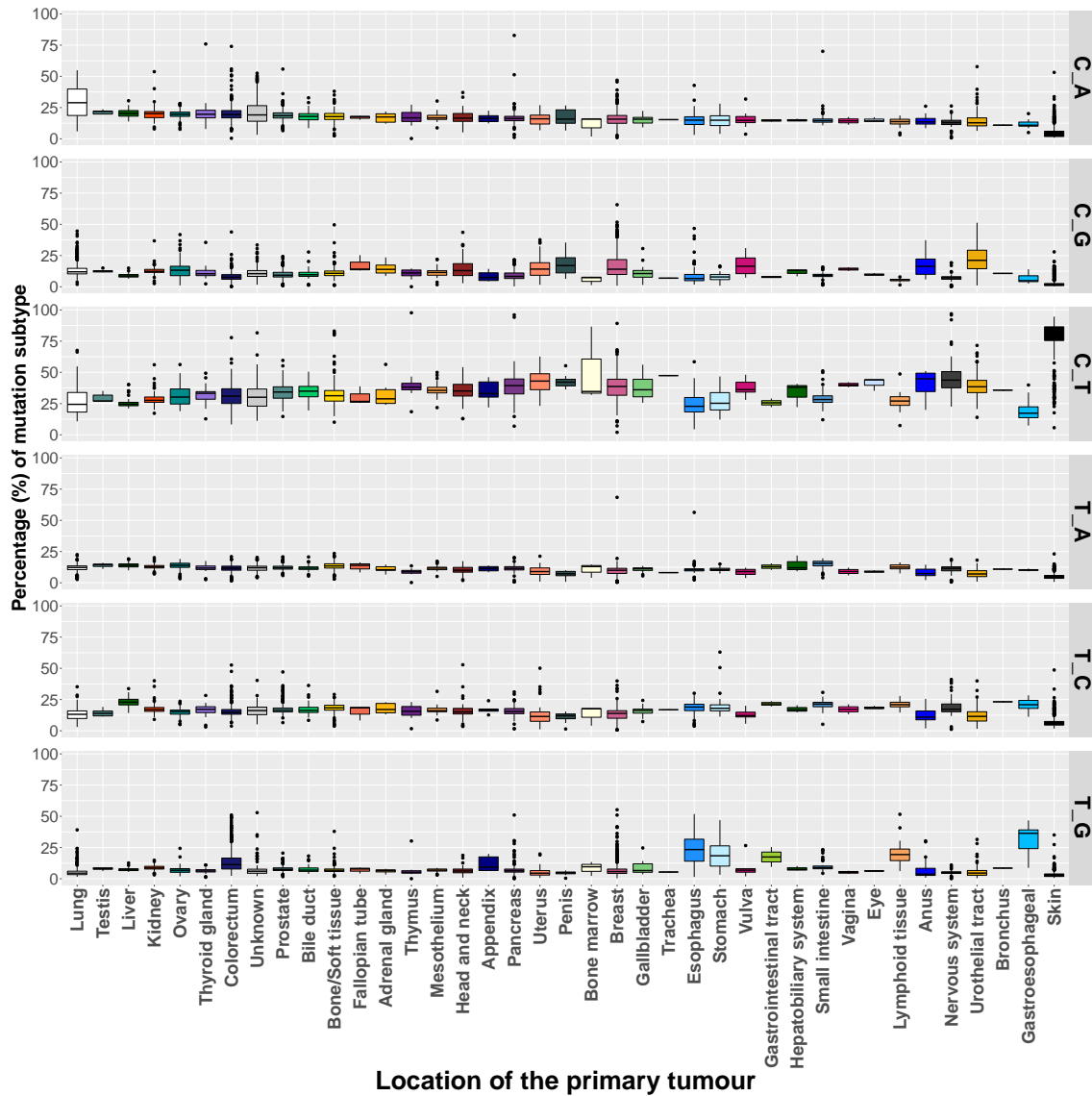


Figure 22. Distribution of mutation types across primary tumour locations in the HMF dataset. Cohorts are ordered by the percentage of C>A mutations.

The Cancer Genomes Atlas (TCGA) dataset

The TCGA dataset consists of whole exome sequencing of mostly primary tumours from 9,104 donors and 32 different cancer types (**Figure 23**). The biggest cohort is breast adenocarcinoma (BRCA), followed by lung adenocarcinoma (LUAD), Brain Lower Grade Glioma (LGG) and Head and Neck Squamous Cell carcinoma (HNSC). The most mutated cancer type is skin cancer (SKCM), when considering SSMs and SIMs (**Figure 24a**), or only SSMs (**Figure 24b**). The cancer type with the highest median of SIMs is lung squamous cell carcinoma (LUSC) (**Figure 24c**). The mutation subtypes distribute differently depending on the cancer type (**Figure 25**), but again we see the same cancer types at the top as in PCAWG. For example, the percentage of C>A mutations is the highest in the lung cancer cohorts, C>T in skin melanoma and C>G in bladder cancer.

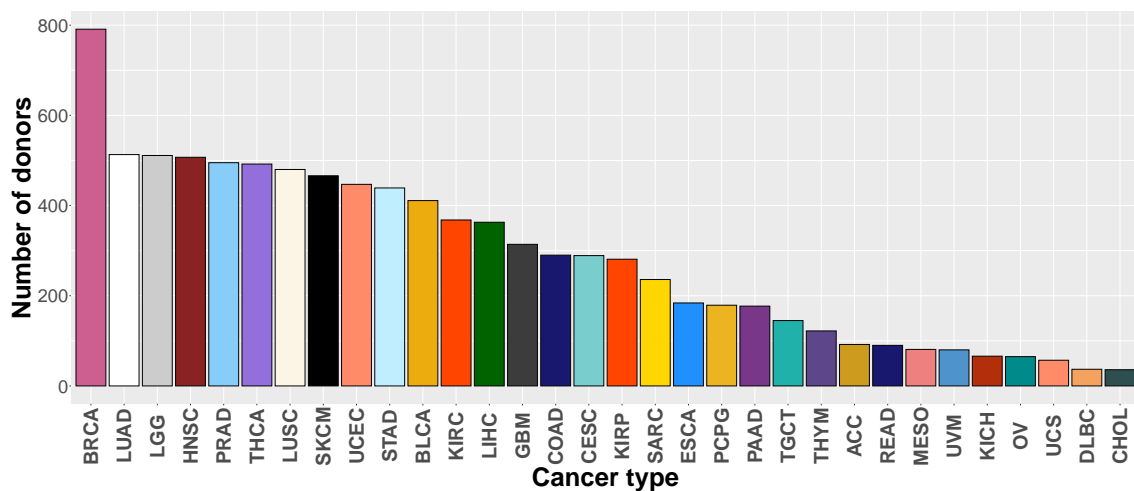


Figure 23. Number of donors per cancer type in the TCGA dataset.

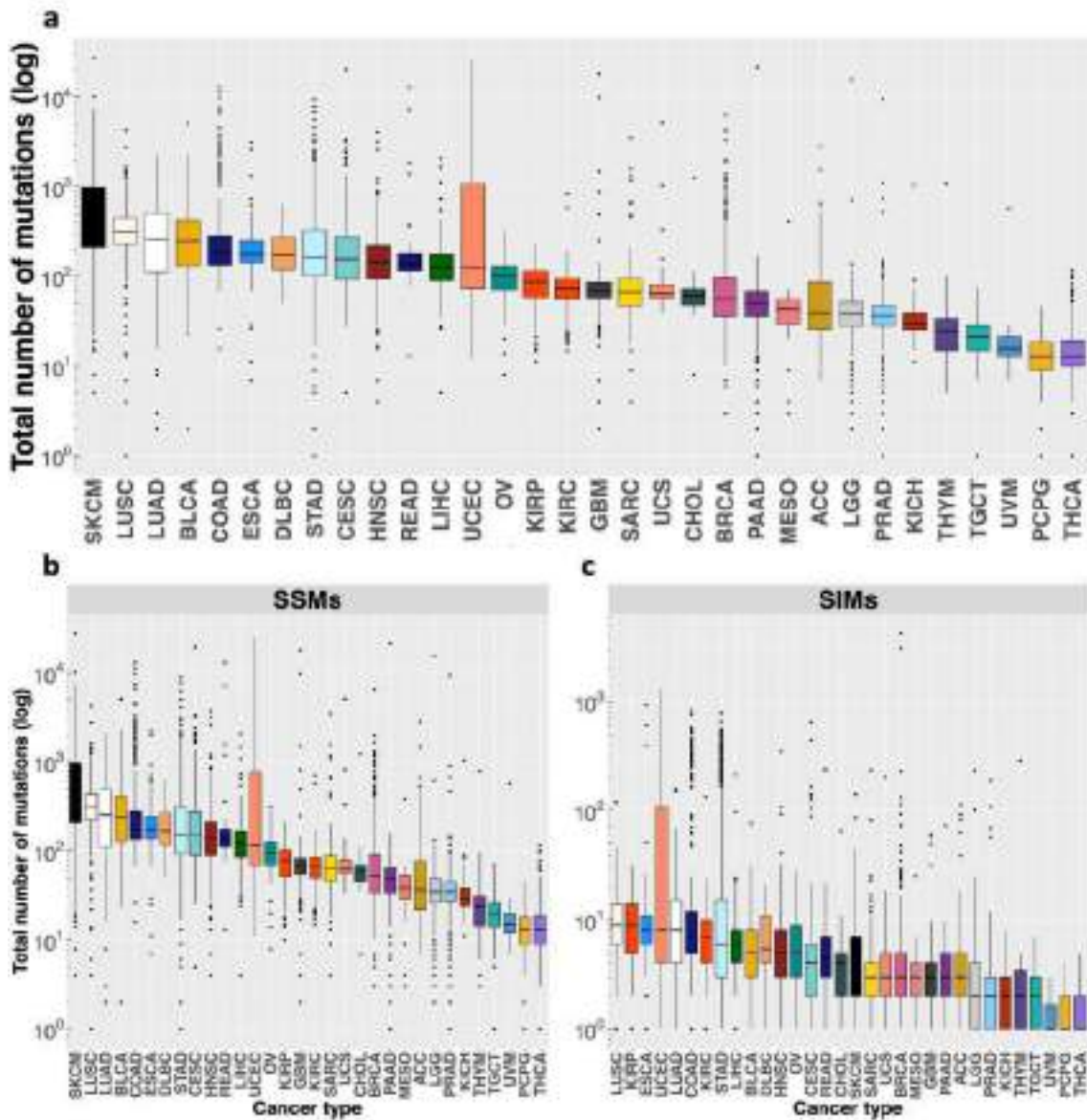


Figure 24. Distribution of (a) total number of mutations across cancer types, (b) number of SSMs and (c) number of SIMs across cancer types in the TCGA dataset.

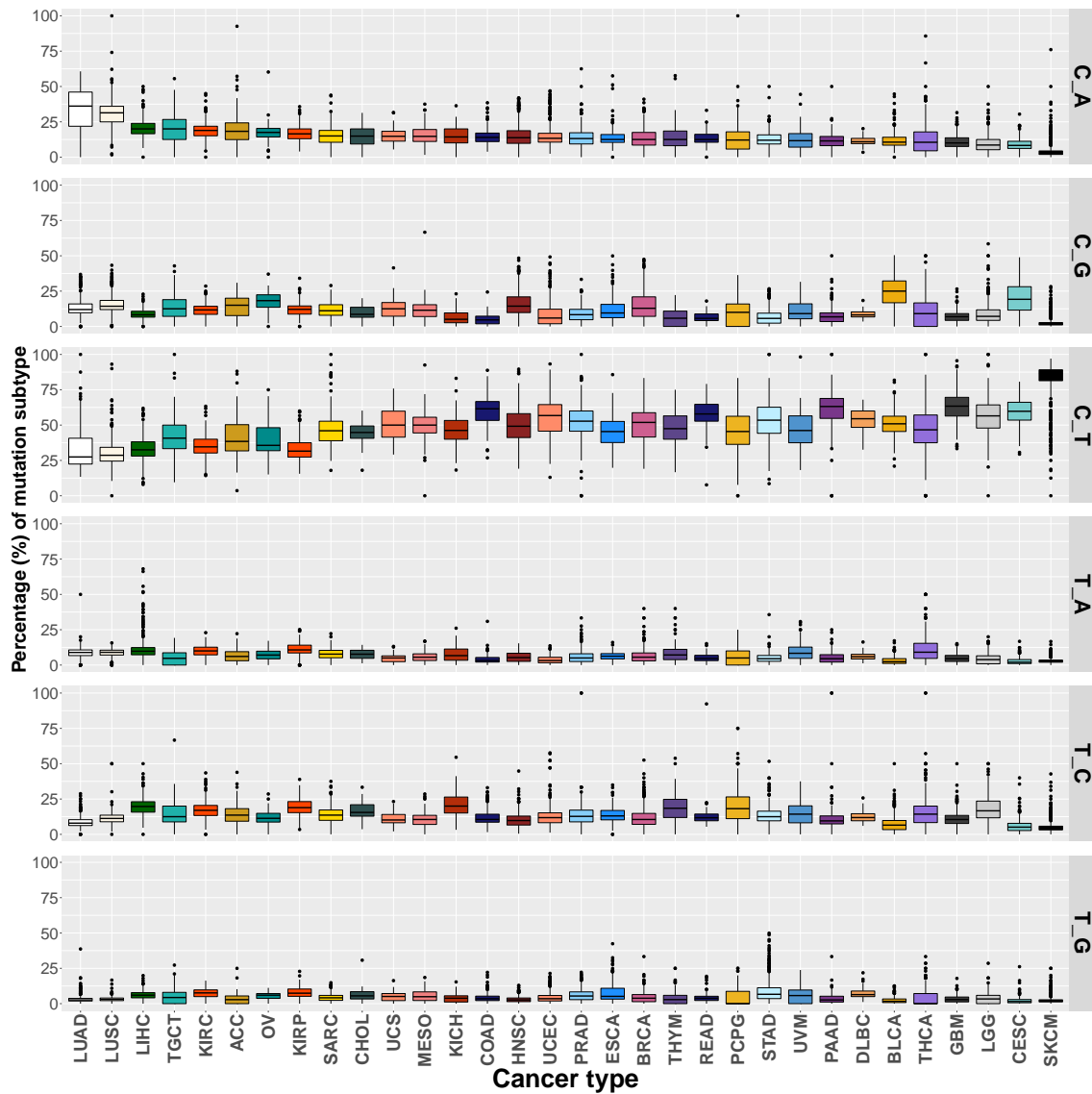


Figure 25. Distribution of mutation types across cancer types in the TCGA dataset. Cancer types are ordered by the percentage of C>A mutations.

Breast CAncer STRatification (B-CAST) study dataset

The B-CAST dataset consists of 9,255 donors with breast primary tumours for which a panel of 323 genes was sequenced. This subset of genes was selected because of their known relevance to breast cancer. From the total number of samples, 345 did not have a mutation in any of the genes in the panel. The highest percentage of tumours are Luminal A (59%), followed by 12% Triple-negative, 11% Luminal B and 5 % HER2 enriched (**Figure 27d**). For 13% of the tumours the molecular subtype is unknown.

Considering only the mutations affecting coding, there are 33,142 somatic mutations across 8,520 donors, considering deletions, insertions, multi- and single-nucleotide substitutions (**Figure 26**). Per mutation type, **Figure 26** includes the top mutated genes either in absolute counts (A.C.) or correcting the number of mutations by the length of the gene (C.C), to account for the fact that larger genes would have more chance of accumulating more mutations. PIK3CA is in the top 10 for single-base substitutions. I will focus on PIK3CA in *Results - Chapter 4*.

Variant type	Frequency	Variant classification	Freq	Top mutated genes
DEL	3,881	Frame-Shift Del	3,137	(A.C) MAP3K1, TP53, CDH1, TBX3, GATA3, KMT2C (C.C) TP53, CBF, CDH1, GATA3, MAP3K1, CDKN1B n=140
		In-Frame Del	652	
		Splice site	49	
		Splice region	29	
		Nonsense mutation	9	
		Translation start site	4	
INS	2,210	Frame-Shift Ins	2,119	(A.C) GATA3, MAP3K1, CDH1, TBX3, TP53, RUNX1 (C.C) GATA3, TP53, CBF, CDH1, MAP3K1, TBX3 n=157
		In-Frame Ins	190	
		Nonsense mutation	8	
		Splice region	3	
MNV	133	Misense mutation	88	(A.C) TP53, CBF, TBX3, KMT2C (C.C) CBF, TP53, TBX3, FOXA1 n=67
		Nonsense mutation	18	
		Splice site	0	
		Silent	1	
		Translation start site	1	
SNV	26,318	Misense mutation	13,301	(A.C) PIK3CA, TP53, MUC16, KMT2C, CDH1, MAP3K1 (C.C) TP53, PIK3CA, CBF, AKT1, FOXA1, MAP2K4 n=322
		Silent	3,976	
		Nonsense mutation	3,469	
		Splice region	199	
		Translation start site	22	
		Nonsense mutation	22	

(A.C) = Absolute count

(C.C) = Count corrected by gene length (CDS)

Figure 26. Number of mutations per mutation type and top mutated genes in each case in the B-CAST dataset. n: number of genes with mutations of that mutation type. CDS: coding sequence. A.C: absolute count. C.C: count corrected by gene length (CDS).

5.1.2. Consensus of the joint dataset and largest cohort

By joining the four datasets described in the previous section we have a total of 24,845 donors, considering that we had to remove 653 patients from the TCGA dataset that overlap with the PCAWG dataset and also one donor from HMF overlapping also with PCAWG. The largest cohort in the joint dataset is breast cancer. For this cancer we have available information of the breast cancer subtype in some cases. In PCAWG dataset, the subtype is known for less than half of the tumours, but the most frequent among the ones known is basal or triple-negative breast cancer (**Figure 27a**). In TCGA, B-CAST and HMF breast cancer donors the most frequent subtype is Luminal A (**Figure 27b-d**).

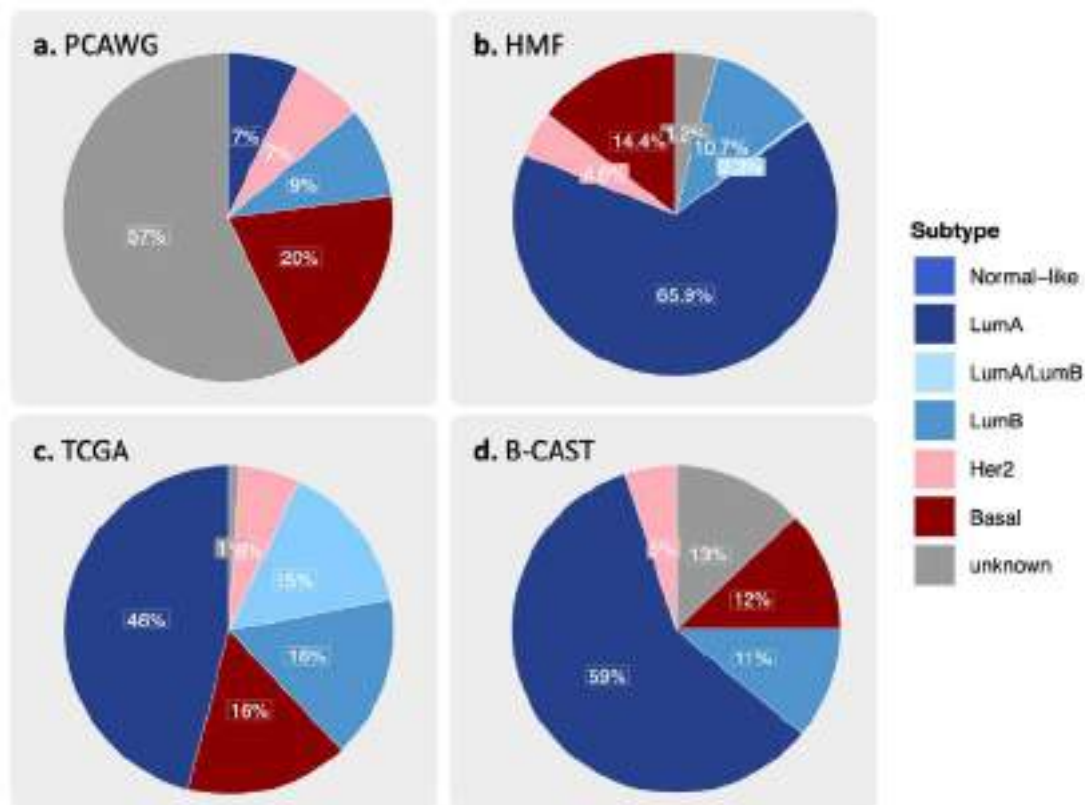


Figure 27. Breast cancer subtype distribution in the individual datasets.
LumA: Luminal A. LumB: Luminal B. Her2: HER2-enriched.

5.1.3. Description of mutational signatures across cancer types

Focusing on breast, colorectal and uterus cancer, there were a total of 64 known mutational signatures present across these tumours (**Table 7**). The dominating signature varied across the three cancer types and there was also stratification of tumour genomes within the same cancer type. We observed the different mutational signatures and their proportions in breast, colorectal and uterus tumours from primary (TCGA, PCAWG) and metastatic (HMF). Primary breast tumours showed a different profile with respect to uterus and colorectal primary tumours, while the metastatic breast cancer had a similar profile to metastatic uterus cancer. In primary breast cancer (**Figure 28**, TCGA and PCAWG) we differentiated three main patterns, a group of samples characterized by defective homologous recombination DNA damage repair (signature SBS3), a second group characterized by the activity of APOBEC family of cytidine deaminases (signatures SBS2 and SBS13) and, third, a group of tumours that has mainly signature SBS5, which is a clock-signature, normally related to the age of the patient. We observed a kind of mutual exclusivity between SBS3 and SBS2/13. When one of the previous mutational signatures is dominating in a sample, the other is practically inexistent. In primary colorectal cancer (**Figure 29**, TCGA and PCAWG) we also observed a mutual exclusivity but different than in breast. Donors with hypermutation activity of polymerase epsilon (Pol ϵ) (SBS10a, SBS10b) did not present at the same time the mutational signature SBS40 nor damage by reactive oxygen species (SBS18). We noted that mutational signatures SBS10a and SBS10b that were associated with mutations in the exonuclease domain of Pol ϵ were often present together with SBS28. SBS28 is a signature of unknown aetiology but has already been related to samples with SBS10a/b signatures [128]. It has been proposed that SBS28 could be the third minor component of the original SBS10 signature related to T>G transversions. The original SBS10 signature would therefore be captured by these three signatures now: a highest prevalence of the C>A component captured by SBS10a, followed by C>T (SBS10b) and T>G (SBS28) [128]. APOBEC signature was largely absent in colorectal tumours, with only isolated cases in TCGA. Primary uterus cancer (**Figure 30**, TCGA and PCAWG) showed a group of samples with SBS40 as dominating signature, other group with defective DNA damage repair (SBS44, SBS15, SBS21), few cases of APOBEC signature (SBS2/13),

samples that had mainly the clock-signature SBS5 and, finally, a clear group dominated by the mutational signatures related to mutations in the exonuclease domain of Pol ϵ (SBS10a, SBS10b). In the three metastatic tumours (**Figure 28, 29** and **30**, HMF) we had a high proportion of samples with SBS40 as dominating signature, for which the aetiology is unknown, but it has been correlated with patients' age in some studies [129]. In the case of metastatic colorectal cancer (**Figure 29**), this SBS40 signature was the dominating signature across almost all donors. In the case of metastatic breast and uterus (**Figure 28** and **30**) we also had a group of tumours that showed the activity of APOBEC cytidine deaminases with SBS2 and SBS13 as dominating signatures. There were no differences in mutational signatures when divided the breast cancer donors according to their molecular subtypes (**Figure 31**).

Table 7. Main SBS mutational signatures identified in primary and metastatic tumours from breast, colon and uterus and the mutational processes to which they are related to (if known). Descriptions were taken from COSMIC v3.3.

Signature	Description
SBS 1	Spontaneous deamination of 5'-methylcytosine (clock-like signature)
SBS 2	Activity of APOBEC family of cytidine deaminases
SBS 13	
SBS 3	Defective homologous recombination DNA damage repair
SBS 5	Clock-like signature
SBS 9	Polymerase eta somatic hypermutation activity
SBS 10a	Polymerase epsilon exonuclease domain mutations
SBS 10b	
SBS 28	Unknown. Often found in samples with SBS10a/SBS10b signatures
SBS 14	Concurrent polymerase epsilon mutation and defective DNA mismatch repair
SBS 6	Defective DNA mismatch repair
SBS 15	
SBS 21	
SBS 26	
SBS 44	
SBS 20	Concurrent POLD1 mutations and defective DNA mismatch repair
SBS 30	Defective DNA base excision repair due to NTHL1 mutations
SBS 36	Defective DNA base excision repair due to MUTYH mutations
SBS 40	Unknown. Correlated with patients' ages for some types of cancer
SBS 17a/b	Unknown, but SBS17b has been associated to fluorouracil chemotherapy treatment and to damage inflicted by reactive oxygen species (in some studies)
SBS 18	Damage by reactive oxygen species
SBS 31	Platinum chemotherapy treatment
SBS 35	

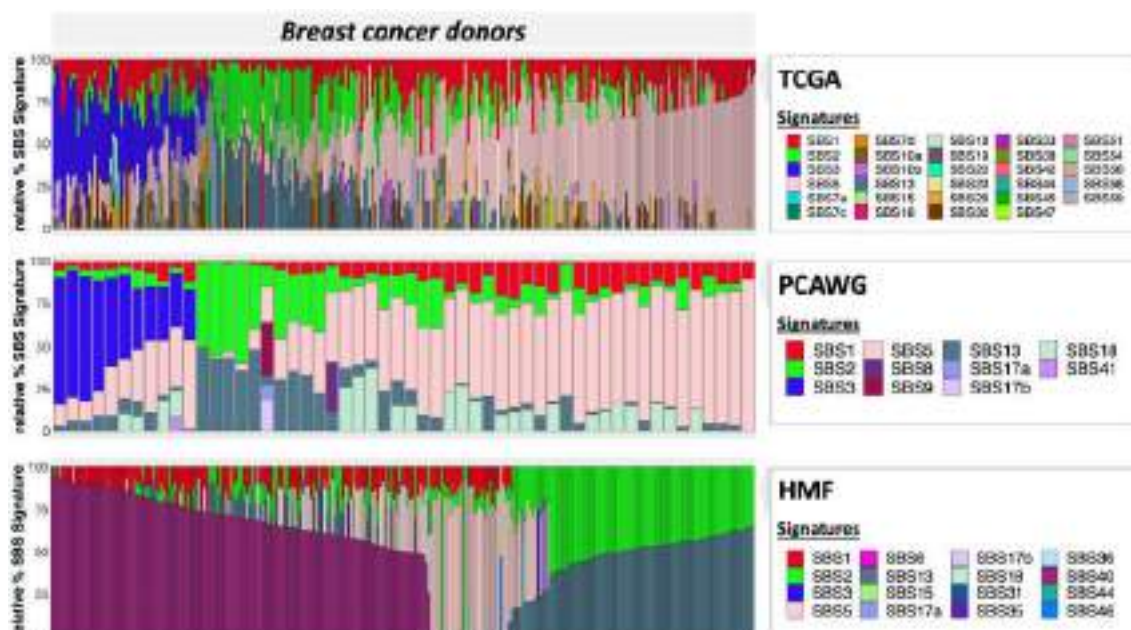


Figure 28. Percentage of SBS signatures found in breast tumours in the TCGA, PCAWG and HMF dataset. Each vertical bar in the plots refers to a tumour genome from a donor and the colours correspond to the SBS signatures found in their different proportions.

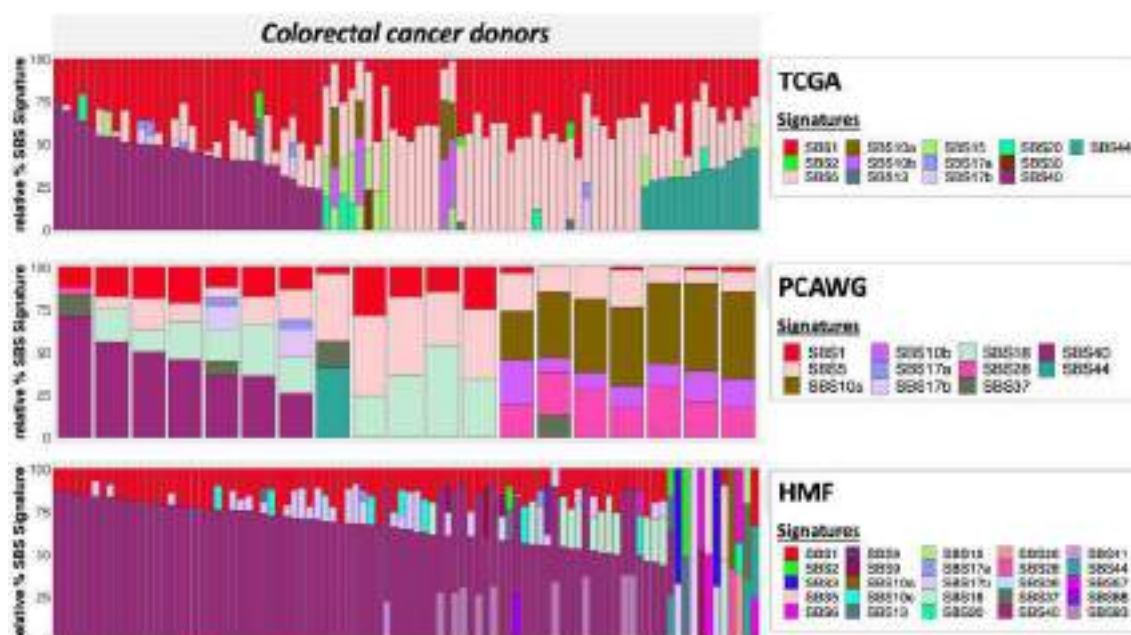


Figure 29. Percentage of SBS signatures found in colorectal tumours in the TCGA, PCAWG and HMF dataset. Each vertical bar in the plots refers to a tumour genome from a donor and the colours correspond to the SBS signatures found in their different proportions.

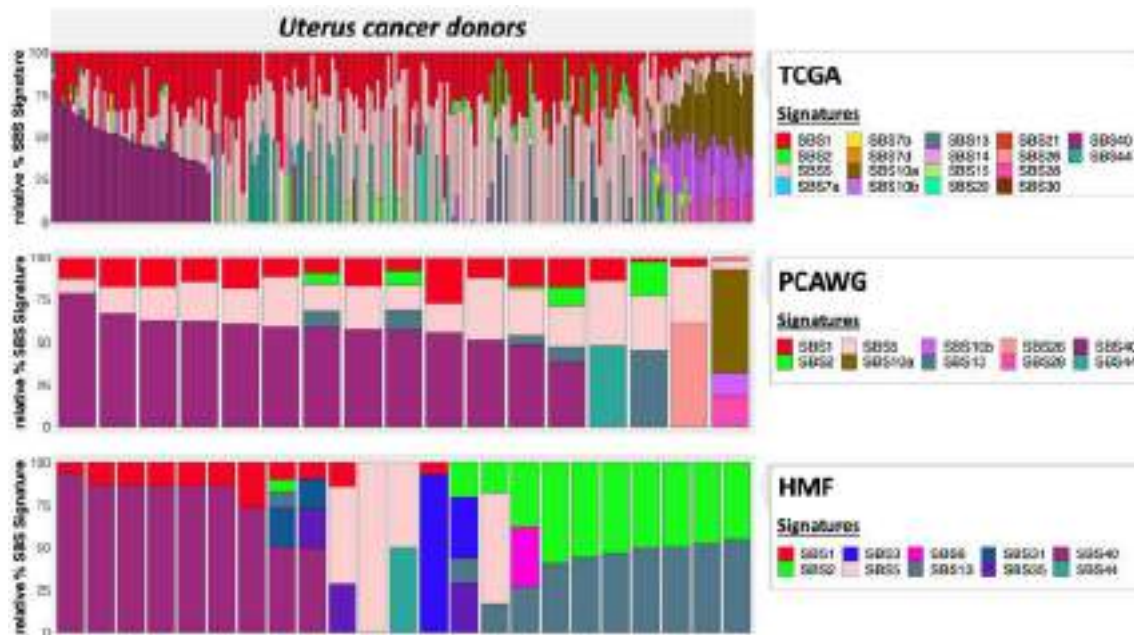


Figure 30. Percentage of SBS signatures found in uterus tumours in the TCGA, PCAWG and HMF dataset. Each vertical bar in the plots refers to a tumour genome from a donor and the colours correspond to the SBS signatures found in their different proportions.

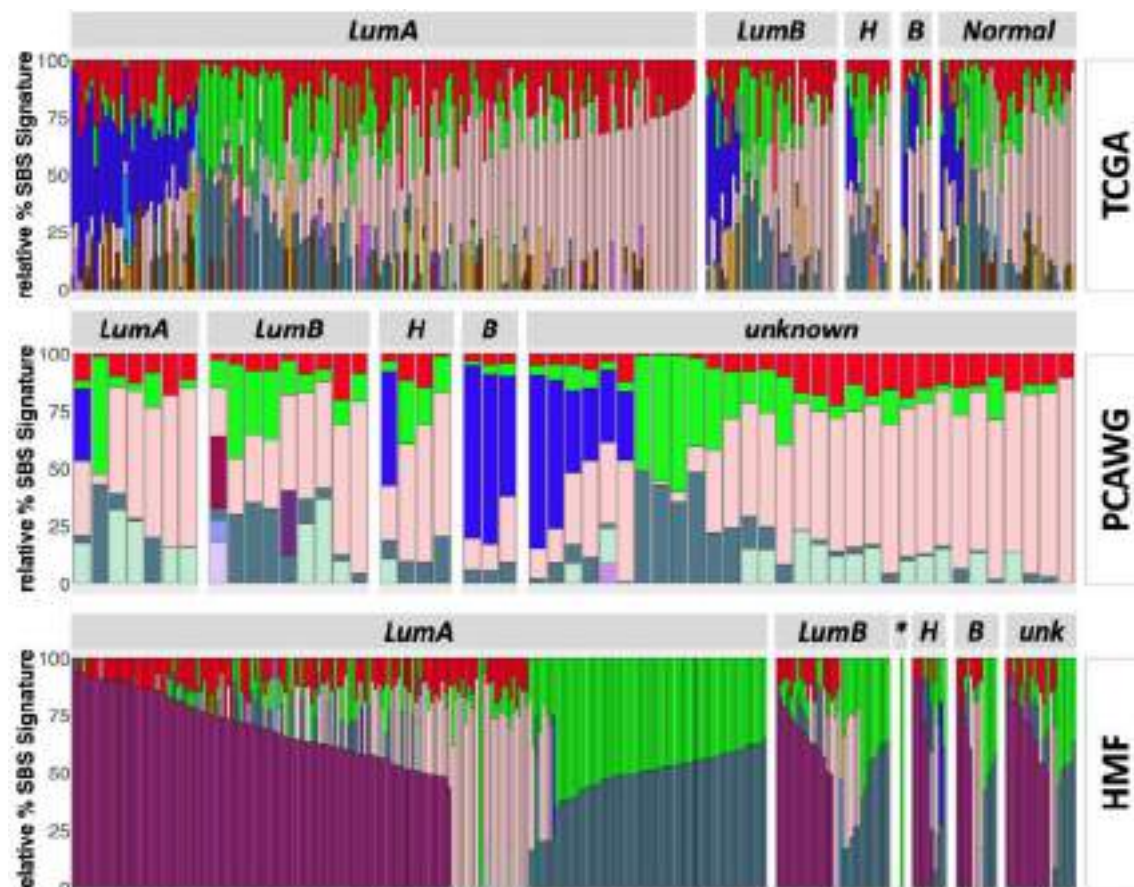


Figure 31. Percentage of SBS signatures across the different breast cancer molecular subtypes in the PCAWG, TCGA and HMF dataset. Each vertical bar in the plots refers to a tumour genome from a donor and the colours correspond to the SBS signatures found in their different proportions. LumA: Luminal A. LumB: Luminal B. H: Her2-enriched. B: Basal or Triple-Negative. Normal: Normal-like. *: Luminal A/Luminal B. Unk: unknown.

Take-home messages Chapter 1

- Skin cancer is the highest mutated cancer type in terms of SSMs. Considering SIMs, lung cancer is the highest mutated in primary tumours and oesophagus cancer in metastatic tumours. Among the different mutation subtypes, C>A mutations are the highest in lung cancer, C>G in bladder cancer and C>T in skin cancer.
- After curation of our joint dataset, the largest cohort is breast cancer with 11,159 breast tumour genomes.
- We could classify cancer genomes according to the dominant mutational signatures. We identify in primary breast, colorectal and uterus cancer that these cancer types shared as a dominant mutational signature SBS5, a clock-like signature. Uterus and colorectal primary tumours show three groups characterized by the same dominant mutational signatures (SBS40, Pol ϵ and defective DNA mismatch repair).
- In primary breast cancer a mutual exclusivity is observed between SBS3 and APOBEC signatures.
- In metastatic tumours from breast, uterus and colorectal cancer, we identify a group of donors that share SBS40 as the dominant mutational signature. SBS40 is the most dominant signature in almost all colorectal cancer samples. Breast and uterus cancer also share a group of donors characterized by SBS2/13 signatures (APOBEC activity).

5.2. CHAPTER 2. USE CASE IN PCAWG DATASET: RECURRENT SOMATIC MUTATIONS REVEAL NEW INSIGHTS INTO CONSEQUENCES OF MUTAGENIC PROCESSES IN CANCER

To illustrate the importance of studying the genomic landscape (Results - Chapter 1) to gain insights into the mutational processes in cancer, we studied the somatic mutations in the PCAWG dataset. We divided these mutations in 'recurrent', identical somatic mutations happening at exactly the same genomic location in two or more tumour genomes from different donors, and 'non-recurrent', the remaining mutations. We computed 13 features based on the recurrent somatic mutations found in 2,583 cancer genomes across 37 cancer types included in this dataset together with 29 other, general genomic features. Based on the total of 42 features we were able to group the samples into 16 clusters that capture clinically relevant cancer phenotypes.

This work has been published as: "Stobbe MD, Thun GA, Diéguez-Docampo A, Oliva M, Whalley JP, Raineri E and Gut IG (2019) Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer. *PLoS Comput Biol* 15(11): e1007496." The complete article and its Supplementary Material are available in Annex 1.

5.2.1. Recurrence is higher than expected by chance

There were 1,057,935 recurrent SSMs, which represent 2.44% of the total number of SSMs found in the PCAWG cohort. This were around five times higher (Fig A-I in Appendix 1 - S1 Text) than expected if only chance would drive recurrence (based on 5,000 simulations, Appendix 1 - S1 Text). For the six SSM subtypes (see Materials) the observed recurrence was around three (C>G and T>C SSMs) to twelve times (T>G SSMs) higher than expected by chance (Fig A-II in Appendix 1 - S1 Text). On tumour type level, we can either determine recurrence by only considering the samples from the same tumour type ('within tumour type') or across all samples ('pan-cancer'). In both cases, Kidney-RCC, Liver-HCC, Lung-AdenoCA and Lung-SCC stand out as the observed number of recurrent SSMs was only around three times (within tumour type) and around two times (pan-cancer) higher than expected by chance (Fig A-III+IV in Appendix 1 - S1 Text).

In contrast, the largest ratio was 86 times for recurrence 'within tumour type' (Prost-AdenoCA) and 7 times for recurrence 'pan-cancer' (Eso-AdenoCA).

5.2.2. Number of samples does not always correspond to the level of recurrence

To see the effect of the number of samples on recurrence, we looked at the overall recurrence within each tumour type (**Figure 32**). Although tumour types with more samples generally had a higher total number of recurrent mutations (**Figure 32A**), there were notable exceptions. For example, Liver-HCC has the most samples of all tumour types (314), but less recurrent SSMs and SIMs than six and five other tumour types, respectively. If we look at the percentage of recurrent mutations, even more tumour types overtake Liver-HCC as in this respect it ranks 9th and 10th in terms of SSMs and SIMs, respectively (**Figure 32B**). The opposite is true for Eso-AdenoCA (97 samples), which has a higher absolute number and percentage of recurrent SSMs than eight other tumour types that have more samples. Stomach-AdenoCA has the highest absolute number and percentage of recurrent SIMs of all tumour types, but less samples than 13 of them. One partial explanation for this is that a lower number of samples does not always translate to a lower total number of mutations (**Figure 32C**), even though the correlation is strong (Spearman's Rank correlation coefficient $r_s = 0.73$, $p = 2.8e-07$). However, even if the number of samples and the number of mutations are in line, the level of recurrence may still give a different picture. Liver-HCC, for instance, has also a higher total mutational load than Eso-AdenoCA ($1.2 \cdot 10^6$ and $7.9 \cdot 10^4$ more SSMs and SIMs, respectively), but still a lower level of recurrence.

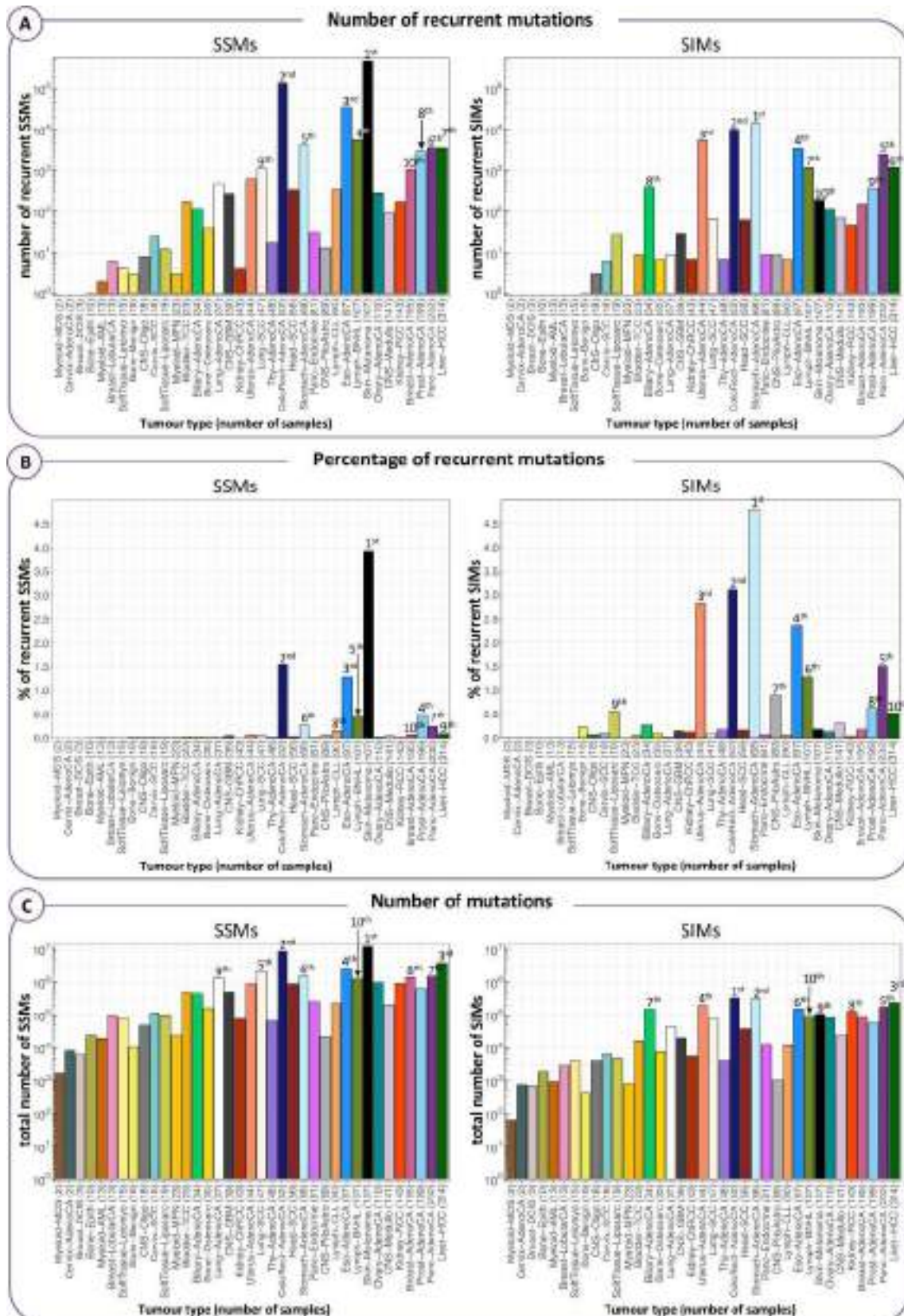


Figure 32. Recurrence within each tumour type in absolute numbers and percentages. The tumour types are ordered from the lowest to the highest number of samples. We labelled the top 10 ranking tumour types in terms of the following three values: (A) Absolute number of recurrent mutations, where recurrence is defined by considering each tumour type separately ('within tumour type' recurrence). (B) Percentage of recurrent mutations 'within tumour type'. (C) Total number of mutations, counting recurrent mutations only once.

5.2.3. General mutational characteristics versus recurrence

For each cancer genome, we computed 29 basic mutational characteristics that capture the effects of mutagenesis (*e.g.*, relative frequency of each SSM subtype) and 13 features capturing recurrence at different levels (Table A in S1 File (Appendix 1), see Methods). Recurrence for these features was determined based on the entire cohort. A detailed description of each of these 42 measures is available in Appendix 1 - S1 File. Based on the comparison of the recurrence-related features with the general ones (Appendix 1 - S2 Text), the key findings were that across the entire cohort: **1)** the correlation between mutational load and the absolute level of recurrence is stronger for SSMs ($r_s = 0.89$) than for SIMs ($r_s = 0.76$); **2)** the same correlation, but instead taking the percentage of recurrent mutations, is weak and negative for SSMs ($r_s = -0.21$) and non-significant for SIMs; **3)** relative recurrence for SIMs is higher than for SSMs; **4)** a particularly high percentage of C>T SSMs and 1 bp A/T deletions are recurrent (4.19% and 15.27%, respectively); **5)** there is a strong tendency for T>G SSMs to be recurrent despite their modest total number; **6)** there is a strong correlation between the level of recurrence for SIMs and the percentage of 1 bp SIMs in a long homopolymer context. Looking into the different tumour types, there were clear contrasts in terms of the associations between general and recurrence-related characteristics. For example, there is a statistically significant positive correlation between the number of mutations and the percentage recurrent for only two tumour types in the case of SSMs (Eso-AdenoCA: $r_s = 0.48$ and Skin-Melanoma: $r_s = 0.58$) and for seven types with respect to SIMs (most notably: Biliary-AdenoCA: $r_s = 0.71$ and Eso-AdenoCA: $r_s = 0.67$) (Fig D in Appendix 1 - S2 Text).

5.2.4. Recurrence characteristics divide the cohort

Next, we used the recurrence-based and general mutational features all together to see if we can uncover meaningful clusters of cancer genomes. As there were strong correlations between some of these features (**Figure 33**), we first performed a principal component analysis (PCA) to obtain independent features and reduce dimensionality (**Figure 34**). We took as many principal components (PCs) as needed to explain at least 80% of the variance in the data and considered the remaining PCs to capture noise. We

used this subset of PCs as input for hierarchical clustering [99]. The resulting hierarchical tree was cut at the desired height to obtain clusters. The centroids were computed for each cluster and used as input to the k-means consolidation step, which further improved the initial clustering (see Methods) [100]. To get a pan-cancer perspective we analysed all samples together.

The crude division into two clusters separated the cancer genomes with low relative levels of recurrent SIMs (*e.g.*, Liver-HCC, Kidney-RCC and Lung-SCC) from those with high levels (*e.g.*, ColoRect-AdenoCA, Eso-AdenoCA, Lymph-BNHL and Panc-AdenoCA) (Appendix 1 - S1 Fig). At three clusters, the relative level of recurrent SSMs split off a group of mainly Skin-Melanoma samples from the two other clusters. This cluster largely remained unchanged when increasing the number of clusters while the two other clusters continued to divide and became more specific to a tumour type or a particular mutational process. At the level of six clusters, for example, we saw a cluster split off that we can connect to microsatellite instability (MSI). We will discuss in further detail the division into 16 clusters (labelled from A to P), chosen as a trade-off between too many clusters, which would each be specific to just a handful of samples, and too few, which would result in loss of meaningful information (**Figure 35**). There are nine clusters (A, B, C, G, H, I, L, M and P) for which at least half of the samples are from the same tumour type. For another two clusters (O and N) samples from two tumour types constitute a majority. In the remaining five clusters (D, E, F, J and K) three or more tumour types are required for this. For each tumour type the percentage of samples in each of the 16 clusters is shown in Appendix 1 - S2 File. The association of each of the 42 features with the clusters is shown in **Figure 36**. The key characteristics of each cluster are shown in **Figure 35**. To facilitate a tight linkage of the clusters to mutational processes, we considered, in addition to the mutational features of a cancer genome, also tumour type assignment, microsatellite instability (MSI) status, immunoglobulin heavy-chain variable region gene (IGHV) mutation status (Lymph-CLL only) and tobacco smoking history of the donor (where available) (Appendix 1 - S3 Text). To provide further details on each cluster we integrated annotation based on GENCODE [130], Oncotator [131], driver predictions [132] [94], replication time [133] and mutational signatures [129]. A summary of this and further details are described in S3 Text in Appendix 1. In the following sections (5.2.5, 5.2.6, 5.2.7, 5.2.8, 5.2.9), we will show how the level of

recurrence can be indicative of the mutational processes, often in combination with the general features. Moreover, we show that our recurrence-based approach groups cancer genomes in a novel way that complements current classification approaches and captures clinically actionable cancer phenotypes.

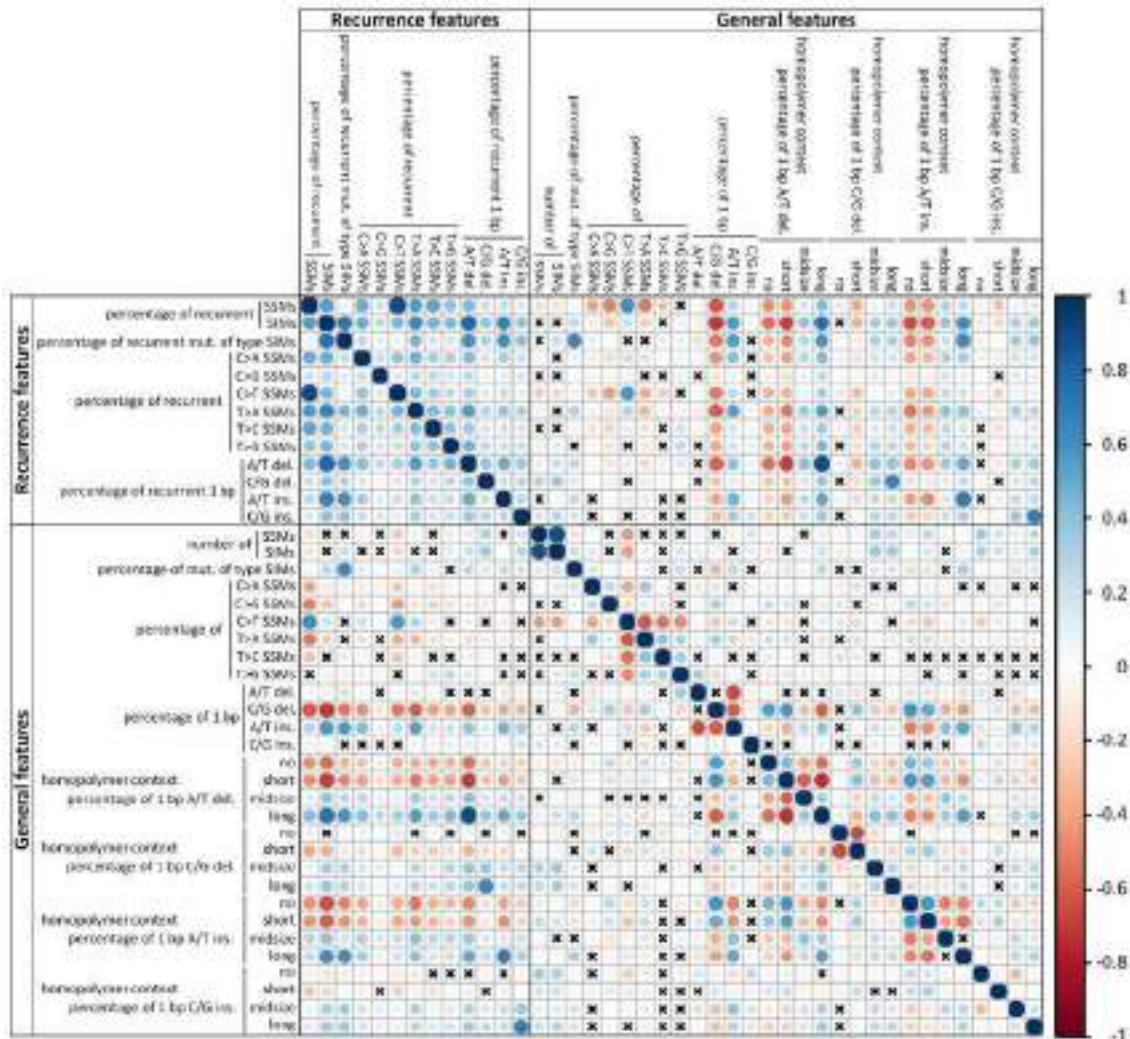


Figure 33. Spearman's rank correlation between the 42 mutational features. The colour of the circles indicate positive (blue) and negative (red) correlations, colour intensity represents correlation strength as measured by the Spearman's rank correlation coefficient. The size of the circle indicates the adjusted p-value with larger circles corresponding to lower p-values. The p-values were corrected for multiple testing using the Benjamini-Yekutieli method. Crosses indicate that the correlation is not significant (adjusted p-value > 0.05).

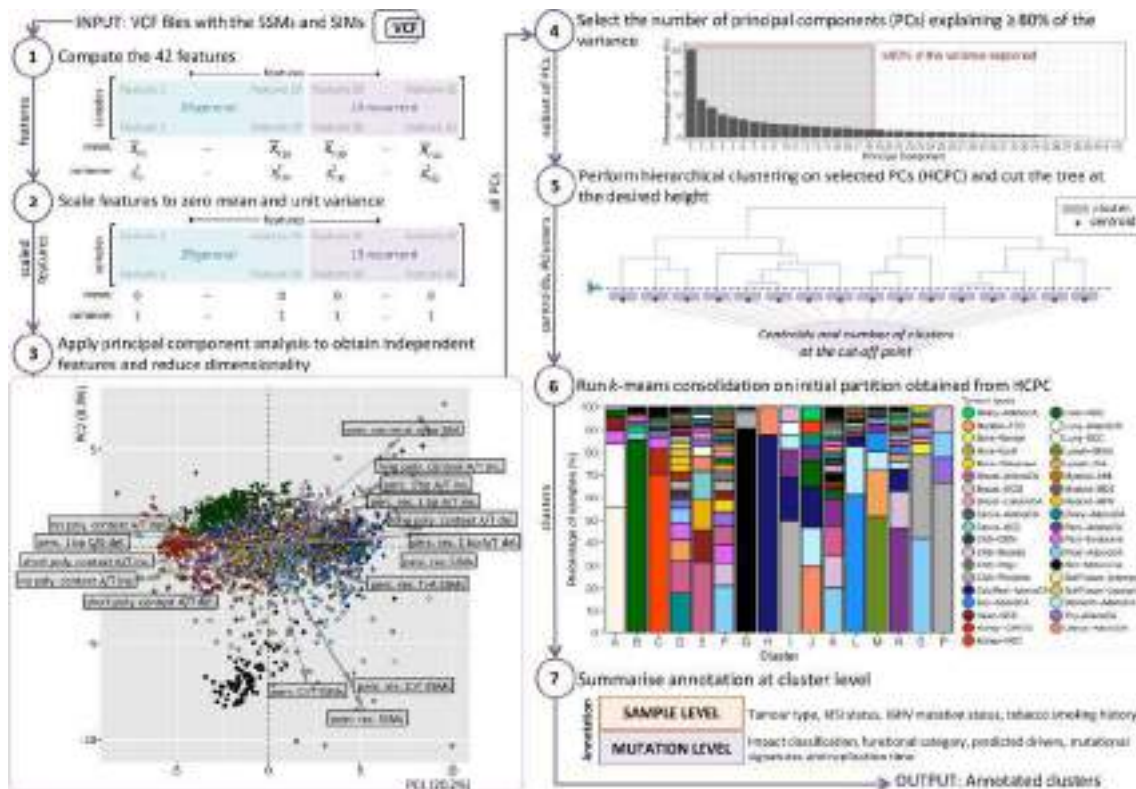


Figure 34. Workflow of the recurrence-based approach to group cancer genomes. The 42 features are described in detail in Appendix 1 - S1 File (Step 1). We scale all features to zero mean and unit variance to compensate for the differences between the ranges of the features (Step 2). The arrows in the PCA plot indicate the direction and level of contribution of the features that contribute above average to the first two PCs (Step 3). Seven of these features are related to recurrence. An interactive 3D version of the PCA plot is available here: <https://plot.ly/~biomedicalGenomicsCNAG/1.embed>. We take a subset of the PCs and consider the remaining PCs to capture noise (Step 4). For the hierarchical clustering we use the Euclidean distance as a dissimilarity measure and Ward’s method as the linkage criterion (Step 5). The results of the hierarchical clustering are used as a starting point for k-means clustering (Step 6). Some samples will in this step switch to a different cluster compared to the initial partition. This consolidation step is repeated a maximum of 10 times. Further details on the annotation of the clusters (Step 7) are described in Appendix 1 - S3 Text.

Cluster	Main tumour type(s)	Median number of		Max. of samples	Key characteristic(s)	Association with overall recurrence		Possible causative agent or mechanism
		SSMs	SIMs			SSMs	SIMs	
A	Lung-SCC	44,910	1,834	88	high % of = C>A SSMs = 1 bp C>G deletions	negative	negative	- telomerase - increased activity of cyclin dependent
B	Liver-HCC	11,846.5	578	324	high % of T>C SSMs	negative	negative	alcohol intake
C	Kidney-SCC	4,401	576	103	high % of 1 bp A/T insertions in or short homopolymer context	negative	negative	16 samples antitubercular
D	Ovary-AdenoCA, Breast-AdenoCA, Lymph-CLL, Pancre-Endocrine, Prostate-AdenoCA	3,494.5	149.5	142	* low % of 1 bp A/T insertions * high % of 1 bp C/G deletions	negative	negative	unknown
E	Breast-AdenoCA, Head-SCC, Bladder-TCG	18,828	382.5	98	high % of C>G SSMs	negative	no	increased activity of cyclin dependent
F	Prostate-AdenoCA, CNS-Medullo, Pancre- Endocrine, Breast-AdenoCA, Thy-AdenoCA	2,209	144	95	high % of = 1 bp C>G insertions in a long homopolymer context = recurrent 1 bp C/G insertions	no	no	unknown
G	Skin-Melanoma	88,002	647	87	high % of = C>T SSMs = recurrent SSMs = recurrent C>T SSMs	positive	negative	UV light
H	ColoRect-AdenoCA	822,214	8,568	8	high number of SSMs	positive	no	deregulated activity of polymersase
I	CNS-PiloAstro	125	8	16	high % of 1 bp C/G insertions	no	no	unknown!
J	Uterus-AdenoCA, Stomach-AdenoCA, GastroInt-AdenoCA	55,789	30,228	17	* high number of SSMs * mutations of type SIM = 1 bp C/G deletions in a midline homopolymer context	no	positive	microsatellite instability
K	Prostate-AdenoCA, CNS-Medullo, Breast-AdenoCA, Pancre-AdenoCA	3,205.5	291.5	222	high % of 1 bp A/T insertions	no	positive	unknown
L	End-AdenoCA	28,908	1,446.5	184	high % of = T>G SSMs = recurrent T>G SSMs	positive	positive	gastric acid (reflux)
M	Lymph-BNHL	7,065	416.5	184	high % of = recurrent 1 bp A/T deletions = 1 bp A/T deletions in a long homopolymer context = recurrent C>G SSMs	positive	positive	hypermutation of the immunoglobulin genes
N	Pancre-AdenoCA, CNS-Medullo	4,993	542	311	high % of = 1 bp A/T deletions in a long homopolymer context = recurrent 1 bp A/T deletions = recurrent mutations of type SSM	positive	positive	unknown
O	Prostate-AdenoCA, CNS-PiloAstro	182	11	43	high % of recurrent T>C, C>G & T>A SSMs	positive	positive	unknown!
P	CNS-PiloAstro	118	13	9	high % of = recurrent 1 bp C/G deletions = 1 bp A/T deletions in a long homopolymer context	positive	positive	unknown!

Figure 35. Key characteristics of the 16 clusters. Tumour types that form together $\geq 50\%$ of the cluster are listed. The legend for colours for the pie chart is provided in Figure 36. The key characteristics are those features with the strongest significantly negative or positive association with the cluster. Only if the association with overall recurrence is significant, the respective direction is indicated. 1 Cluster has a low median number of SSMs (<200) and SIMs (<20).

5.2.5. High levels of recurrent SSMs and low levels of recurrent SIMs characterize exposure to UV light

A positive association with overall recurrence of SSMs and more specifically with recurrence of C>T SSMs characterizes cluster G that mainly consists of Skin-Melanoma samples (**Figure 36**). The association is negative with the recurrence of SIMs. We link this cluster to mutagenesis induced by UV light (Appendix 1-S3 Text). The samples assigned to cluster G account by themselves for 60.7% of the total number of recurrent C>T SSMs. The combination of the high total number of SSMs per sample and the high percentage of C>T substitutions in this cluster is what contributes to the high level of recurrence. The mechanisms inherent to UV-light exposure further increase the probability of recurrence as it tends to result in C>T SSMs near energy sinks in the genome. The energy from UV-light-exposed DNA usually travels along the DNA strand until it arrives at the lowest energy point, a dT, particularly when it is next to a dC, which acts as energy barrier [134]. In agreement with this, for C>T mutations that are recurrent within this cluster there is a strong enrichment of a TTTCCT motif (the underlined C is mutated) (see

Methods). While the percentage of this motif in the genome is estimated to be only 0.4% of all 6-mers with a C at the central position, 4.5% and 19.5% of the non-recurrent and recurrent C>T SSMs, respectively, within this cluster are at this motif (**Figure 37**). An enrichment of a CTTCCG motif was found for frequently recurrent mutations in promoters in 38 melanoma samples [135]. In another set of 184 melanoma samples a CTTCCGG motif was found at the majority of ETS transcription factor binding sites (TFBSs) [136]. As the sequences are centred at the core consensus ETS binding motif TTCC, instead of at a mutation, the underlined nucleotide is the most frequently mutated base. In a subset of highly mutable ETS TFBSs the second C is the most mutated. These and our specific sequence motif help explain the observed high level of recurrence. Furthermore, a decreased activity level of the nucleotide excision repair pathway was detected in melanoma at active transcription factor binding sites and nucleosome embedded DNA compared to the flanking sites [18]. This increases local mutation rates and hence also increases the probability of recurrence.

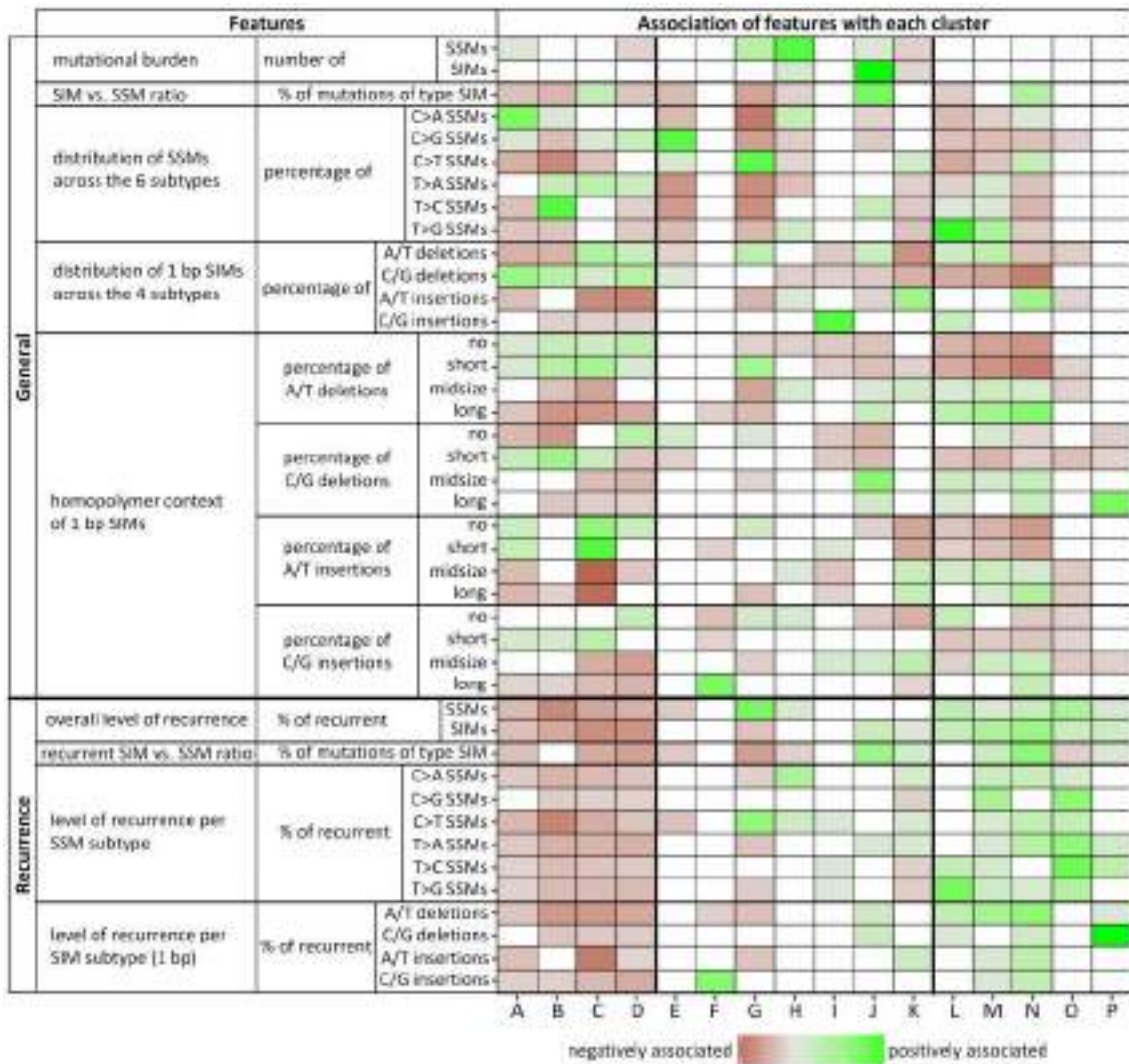


Figure 36. Overview of the 42 features and their association with each cluster. Red and green squares indicate statistically significant negative and positive associations, respectively, where the gradient indicates the strength of the association. White coloured squares indicate no significant association (adjusted p-value > 0.05). For deletions a ‘no homopolymer context’ means that the base next to the deleted one is not of the same type. For insertions this refers to a base inserted 5’ to either a base of a different type or a single base of the same type. Note that we do not have to consider preceding bases as all SIM calls were left aligned. A short homopolymer context is defined as a 2–4 bp mononucleotide repeat of the same type of base as the 1 bp SIM, midsize is 5–7 bp in length and long ≥ 8 bp.

5.2.6. High levels of recurrent SSMs characterize deregulated activity of Pol ϵ

A high level of recurrent SSMs also characterizes cluster H, specifically C>T and C>A SSMs. This cluster captures samples that can be considered ultra-hypermutators and their mutations are mainly caused by deregulated activity of polymerase epsilon (Pol ϵ) (Appendix 1 - S3 Text). These samples have a very high total number of C>A SSMs (median: 297,750) and the median percentage of recurrent C>A SSMs across the samples is 7.7%. Two thirds of all recurrent C>A SSMs in the entire cohort are also recurrent within only this cluster. The C>A mutations that are recurrent within this cluster are enriched for the motif TTCTTT, when considering only ungapped motifs (**Figure 37**, see Methods). Of the recurrent C>A SSMs 32.2% are at this motif, while for non-recurrent ones this is true for only 13.7% (χ^2 test: $p < 2.2e-16$). In the genome, the estimated percentage of this motif of all corresponding 6-mers (NNCNNN) is far smaller (0.6%), suggesting that effects of deregulated activity of Pol ϵ are most likely dependent on a sequence context exceeding a single neighbouring base on each side as also observed for whole-exome data by Martincorena *et al.* [137].

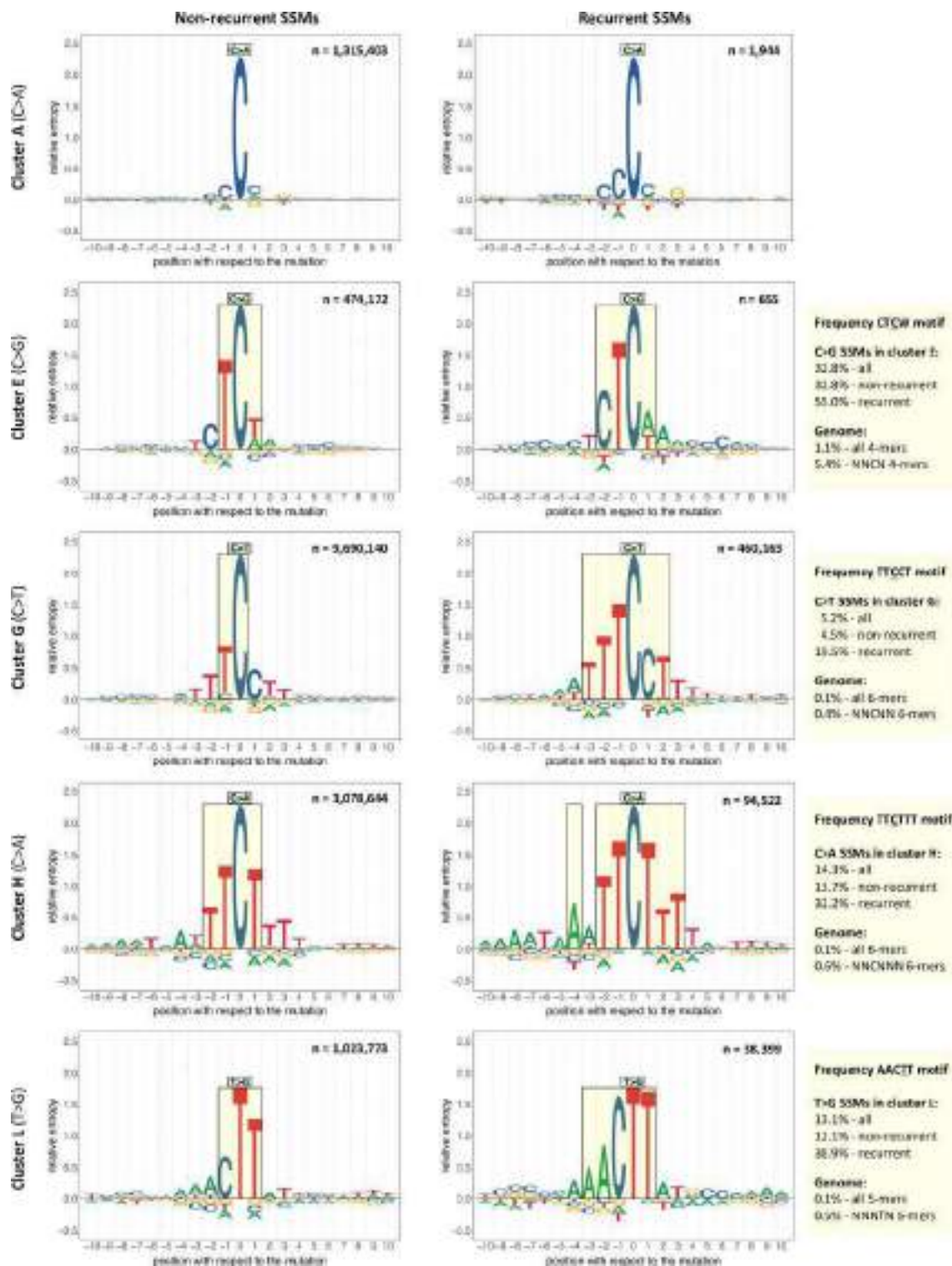


Figure 37. Enriched sequence motifs. The sequence logos represent the sequence context of ten bp 5' and 3' of the non-recurrent (left-side) or recurrent (right-side) mutations of the indicated cluster and SSM subtype. Here recurrence is defined as a mutation at the same genomic location in two or more samples from the same cluster. Each recurrent SSM is included only once to avoid giving extra weight to highly recurrent mutations. Relative entropy is used as a measure of information content (see Methods). Setting a threshold of 0.25 for the relative entropy results in the motifs highlighted in the rectangles. In the upper right corner of each sequence logo the number of mutations is indicated. To the right of the sequence logos are the percentages in which the enriched motif found for the recurrent SSMs is present in context of the mutations in the cluster and the corresponding k-mers in the genome (N = A, C, G or T). The enrichment for the motif for recurrent SSMs is in all four cases significantly higher than for the non-recurrent SSMs (χ^2 test: $p < 2.2e-16$).

5.2.7. High levels of recurrent SIMs characterize microsatellite instability

The highest level of recurrent SIMs across all clusters is observed for cluster J, which could be linked to a defective mismatch repair (MMR) pathway resulting in MSI (Appendix 1 - S3 Text). Of the 179,691 recurrent 1 bp SIMs in the entire cohort, almost half of them are recurrent when only considering this cluster. The very high median number of SIMs (30,228) in this cluster may play a role in the high level of recurrence. The key factor, however, is most likely the mutational process behind MSI, which is slipping of the DNA polymerase during replication of repetitive sequences and the lack of repair by the MMR pathway [138]. This not only explains the elevated number of SIMs [139], but also the association of this cluster with all SIM subtypes in the context of midsize-to-long homopolymers. As such homopolymers are scarce in the genome, the shift towards specifically altering them increases the probability of recurrence (Table F in Appendix 1 - S2 Text). Especially striking in this cluster is the proportion of 1 bp C/G deletions that are in the context of a midsize homopolymer (median: 73.2% vs. 8.4% for the other clusters combined, $p = 1.2e-12$). This translates to 6.0% recurrent 1 bp C/G deletions within this cluster versus <0.7% for any other cluster (Appendix 1 - S3 Text).

5.2.8. Positive association with recurrence of SSMs and SIMs: Gastric-acid exposure and hypermutation of immunoglobulin genes

Clusters L, M and N all positively associate with recurrence of both SSMs and SIMs. Cluster L, which for >80% consists of Eso-AdenoCA and Stomach-AdenoCA samples, can potentially be linked to gastric-acid exposure (Appendix 1 - S3 Text). The T>G and T>C SSMs that are recurrent within this cluster cover 45% and ~20%, respectively, of the total observed in the whole cohort. The median percentage of SSMs falling in late-replicating regions (Table C and Fig A in Appendix 1 - S3 Text) is significantly higher than in the rest of the clusters combined (75.2% vs. 61.0%, $p < 2.2e-16$). In general, the mutational load is expected to be higher in late-replicating regions as the MMR pathway is said to be less efficient there [140]. However, the question is why the effect is so strong in cluster L compared to the others (Fig B in Appendix 1 - S3 Text). It could be that transient single strand-DNA at stalled replication forks, whose formation has been suggested to be more prevalent in late-replicating regions [141], is particularly vulnerable to the mutagenicity

of acid-exposure. Alternatively, if the oxidative stress induced by gastric-acid exposure leads to the oxidation of dG in the dNTP pool [142], the use of error-prone DNA polymerases that incorporate the oxidized dG into the DNA [143] may be more frequent in late-replicating regions [144]. The strong shift towards late-replicating regions favours higher levels of recurrence. The same holds for the enrichment of the specific sequence context 'AACTT' that we observe for T>G mutations that are recurrent within this cluster (**Figure 37**, see Methods). Nearly 39% of the recurrent T>G SSMs are confined to this motif and ~12% of the non-recurrent ones (χ^2 test: $p < 2.2e-16$), which is still far higher than the estimated percentage of this motif in the genome (0.5% of all NNNTN 5-mers). For SIMs, the cluster has a positive association with recurrence for three out of the four SIM subtypes as well as with the same subtypes in a midsize and/or long homopolymer context. This suggests similar mechanisms as for cluster J. Finally, as observed for SSMs in this cluster, SIMs also show a tendency to fall into late-replicating regions (67.2%, Table C and Fig C in Appendix 1 - S3 Text). This may further add to the high level of recurrence for SIMs.

Cluster M, with mainly Lymph-BNHL and Lymph-CLL samples, is linked to the somatic hypermutation of the immunoglobulin genes (Appendix 1 - S3 Text). In the aforementioned tumour types, this process is indicative of memory B cells being the cell of origin as opposed to naïve B cells [145]. The cluster has positive associations with the level of recurrence for all six SSM subtypes. The association is particularly strong for C>G. Of all recurrent C>G SSMs, 10.7% can be found in this cluster alone. The high level of recurrence may partially be explained by the hypermutation observed in the limited area of the genome where the immunoglobulin genes are located. For SIMs, the cluster has positive associations with the level of recurrence for all four subtypes as well as with those subtypes in general when in a midsize and/or long homopolymer context. This cluster has the highest median percentage of SIMs in late-replicating regions (67.5% vs. 57.8% for the other cluster combined, $p < 2.2e-16$, Table C and Fig C in Appendix 1 - S3 Text), which may contribute to the high level of recurrence.

In cluster N, which consists of ~47% Panc-AdenoCA samples, the sources of mutagenesis are less clear, even after the inclusion of all annotation layers (Appendix 1 - S3 Text).

Except for C>G and T>C SSMs, the cluster has positive associations with the recurrence of all other subtypes of SSMs and every SIM subtype. This is especially noticeable as the median of the total number of mutations across samples is intermediate. A high percentage of the recurrent mutations are SIMs in this cluster with a median of 35.0%. This is far higher than for samples of the other clusters combined (median: 15.5%, $p < 2.2 \times 10^{-16}$). The positive associations with all SIM subtypes when in a midsize-to-long homopolymer context may point to a slippage-related mechanism (see also cluster J).

5.2.9. Negative association with recurrence: Tobacco-smoke exposure, alcohol use and increased activity of cytidine deaminases

There are also several mutagenic processes that are associated with low levels of recurrence (**Figure 36**) including those represented by clusters A, B, C and E. Cluster A, of which 84% are lung cancer samples, is linked to mutational processes induced by tobacco-smoke exposure (Appendix 1 - S3 Text). This cluster shows a positive association with the total number of SSMs and the percentage of C>A SSMs, the latter is a known consequence of tobacco-smoke exposure [146]. There are several factors that increase the probability of recurrence in this cluster, including the high total mutational load together with the high percentage of C>A SSMs and the enrichment of mutations in late-replicating regions (Appendix 1 - S3 Text). Also, tobacco-smoke induced mutations have been shown to be enriched in linker DNA (*i.e.*, DNA not wrapped around a nucleosome) [147], which constitute only between 10% and 25% of the genome in eukaryotes [148]. The key to explaining the lack of recurrence seems to be that there is little tendency to favour a specific sequence context for the C>A SSMs (**Figure 37**). This can also be observed in the 'tobacco smoking signature' [149], which is present in nearly 90% of the samples in this cluster (Appendix 1 - S3 Text). Unlike for several clusters mentioned above, there is a positive association with SIMs in short homopolymer contexts, which are more frequent in the genome than longer homopolymers, and the resulting distribution is therefore also more random. Note that cluster A also has a strong association with the percentage of total 1 bp C/G deletions, which has not been described previously as a possible consequence of tobacco-smoke exposure (Appendix 1 - S3 Text and S4 Text).

Cluster B, consisting of 85% Liver-HCC samples, is likely to be linked to mutational processes indirectly induced by excessive alcohol use (Appendix 1 - S3 Text). The level of recurrence is low despite the high number of samples of the same tumour type (277) and the consistent pattern of a high percentage of T>C SSMs (median: 31.7% vs. 14.6% in the other cluster combined, $p < 2.2e-16$). With regard to 1 bp SIMs, there is a positive association with a short homopolymer context, as for cluster A, with the exception of 1 bp A/T insertions.

In cluster C, in which ~82% are Kidney-RCC and Kidney-ChRCC samples, the mutational processes remain largely obscure except for a few samples that can be connected to aristolochic-acid exposure (Appendix 1 - S3 Text). Unlike for clusters A and B, the median number of SSMs across samples is relatively low. Furthermore, mutations are nearly equally spread between early- and late-replicating regions as only 53.9% of the SSMs and 47.5% of SIMs are in late (Table C, Figs B and C in Appendix 1 - S3 Text). SIMs are preferentially located in no or short homopolymer context, similar to clusters A and B.

In cluster E nearly one third are Breast-AdenoCA samples and key mutational characteristics point to the endogenous mutational process of increased activity of cytidine deaminases (Appendix 1 - S3 Text). There is a general paucity of associations with characteristics of recurrence. In line with this, the mutations in this cluster are nearly equally spread between early- and late-replicating regions of the genome (Table C, Figs B and C in Appendix 1 - S3 Text). The most outstanding feature of this cluster is the high percentage of C>G SSMs. This is the most rare substitution type, making the detection of recurrence unlikely, particularly if not confined to specific genomic regions. Interestingly though, the 655 C>G SSMs that are recurrent within this cluster are enriched for the motif CTCW (W = A or T) (**Figure 37**, see Methods). Very similar motifs have been described as being characteristic for deamination mediated by APOBEC3 [150]. The number of recurrent mutations is much lower than for the other motifs discussed. The CTCW motif is also shorter, more general and therefore relatively frequent in the genome (5.4% of all NNCN 4-mers), all possible causes for the lacking trend towards recurrence.

5.2.10. The added value of the recurrence-related features

The PCA shows that seven of the sixteen features that contribute above average to the first two PCs are related to recurrence (**Figure 34**). In addition, all 16 clusters have a statistically significant association with two or more recurrence-related features (**Figure 35**). The importance of the recurrence-related features is further demonstrated by the results of running the entire workflow (**Figure 34**) using only the general features. In this case we are no longer able to separate all ultra-hypermutator samples from the rest of the cohort (Appendix 1 - S2 Fig). Furthermore, the cluster linked to hypermutation of the immunoglobulin genes (cluster M) is dissolved, and the cluster possibly linked to gastric-acid exposure (cluster L) is less cancer-specific as it absorbs 90 samples of the dissolved cluster M and thereby nearly doubles in size. Another key difference is that only ~55% of the Lymph-CLL samples without hypermutation of the immunoglobulin genes are confined to a single cluster as opposed to ~86% when using all features.

Take-home messages Chapter 2 (Figure 38)

- Recurrence of somatic mutations was higher than expected by chance.
- The number of samples in the individual tumour types did not always correspond to the level of recurrence.
- The level of recurrence could not be fully explained by any of the following factors individually: mutational load, sequence context and genomic region.
- Level of recurrence can be indicative of the mutational processes in the sample:
 - UV light exposure is characterized by high levels of recurrent SSMs and low levels of recurrent SIMs.
 - Deregulated activity of Pol ϵ is characterized by high levels of recurrent SSMs.
 - Microsatellite instability is characterized by high levels of recurrent SIMs.

- Gastric-acid exposure and hypermutation of immunoglobulin genes have a positive association with recurrence of SSMs and SIMs.
- Tobacco-smoke exposure, alcohol use and increased activity of cytidine deaminases have a negative association with recurrence.
- The PCAWG dataset could be divided into 16 biologically relevant clusters using a combination of recurrence-based and general mutational characteristics.

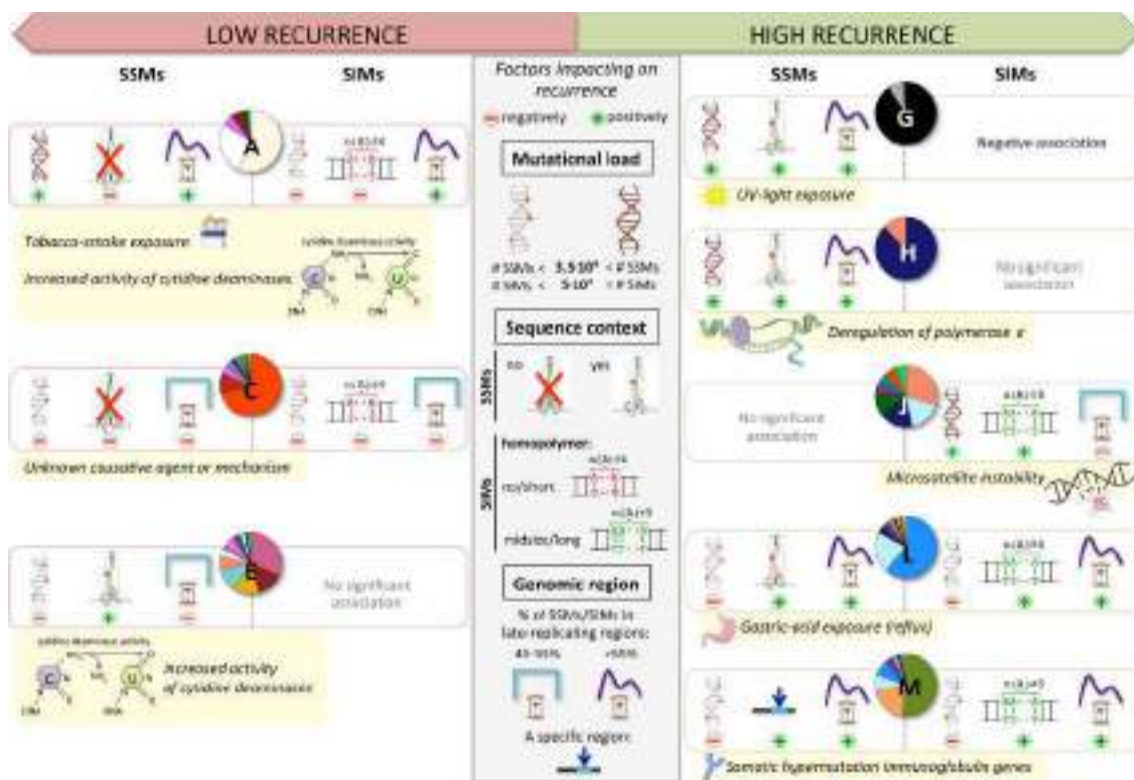


Figure 38. Summary of take-home messages: Factors impacting on recurrence in the context of the clusters. None of the three key factors (middle panel) that impact on recurrence individually explain the observed level of recurrence in the clusters. Whether a cluster has a relatively high or a comparatively lower mutational load is based on the median number of SSMs/SIMs across its samples (Figure 35). The actual specific sequence contexts for SSMs are shown in Figure 37. For cluster M there is enrichment for a specific sequence context as well, which is AGCT for C>G SSMs that are recurrent within this cluster (n = 949) (Appendix 1 - S3 Fig). For SIMs a homopolymer of A/T's is used to represent any type of homopolymer. Clusters A and C have a positive association to no and/or short homopolymer context for all types of 1 bp SIMs (red), while for clusters J, L and M this is the case for midsize and/or long homopolymer context (green) (Figure 36). For the replication time region, we compute the percentage of SSMs/SIMs that are in late-replicating regions (Appendix 1 - S3 Text). If this percentage is between 45–55%, then we consider the mutations to be nearly equally spread between early- and late-replicating regions of the genome. The specific region that is enriched in cluster M refers to the immunoglobulin genes. The recurrence in clusters A and G is also likely to be positively impacted by an increased mutation rate in a specific region as most of their samples are from a particular tumour type for which this has been reported. For lung cancer (cluster A) the mutation rate is increased in linker DNA [147] and for Skin-Melanoma (cluster G) at active transcription factor binding sites [18].

5.3. CHAPTER 3. CHARACTERIZATION OF AMINO ACID CHANGES RESULTING FROM SOMATIC MUTATIONS IN PROTEIN CODING GENES

For the amino acid changes that were translated from the missense mutations in our joint dataset we attributed eight features combining amino acid, evolutionary and structural properties. These protein features were selected because of being helpful to elucidate the effect that a mutation could have in the protein. The features were (1) the characteristic of the amino acid change (chemical and size change), (2) the solvent accessibility of the amino acid that was mutated, (3) the secondary structure in which the amino acid mutated was located, (4) the protein domain in which the amino acid mutated was located (if any), (5) specific site in which the amino acid mutated was involved in (if any), (6) the conservation of the amino acid that was mutated, (7) if the amino acid change belonged to a 3D-hotspot and (8) the change in the free energy of protein folding between the wild-type and mutated protein structure.

All features were collected taking advantage from UniProt annotation, The Protein Data Bank (PDB) [86] and Swiss-Model Repository [87], as well as from several tools such as FoldX [122], ConSurf [119] or mutation3D [30]. Using UniProt we indirectly collected information from other databases such as Pfam and Interpro regarding protein domains, or other proteomic databases in the case of functional sites. The pdb files obtained from the Protein Data Bank and Swiss-Model Repository, were not always identically formatted and manual curation was needed to extract the data correctly. The amino acid conservation was collected from ConSurfDB or computed with ConSurf (<https://consurfdb.tau.ac.il>) [119] when needed. The energy change on protein folding was computed using FoldX [122] and 3D hotspots were computed using mutation3D [30].

We focused on breast cancer as use case because it was the tumour type with the largest number of donors with missense mutations in the combined dataset, with exactly 9,306 donors with missense mutations. Across these donors there were 173,226 missense mutations or, considering unique mutations, a total of 159,430 unique missense mutations (**Figure 39**). These missense mutations were translated into 159,294 different amino acid changes that hit 18,523 different genes. We defined each amino acid change

with the eight protein features. The protein structure was available for less than 50% of the proteins affected by amino acid changes in this tumour type (48% of proteins in PCAWG dataset had a structure available, 47% in HMF, 48% in TCGA and 81% in B-CAST). However, despite the structures that were available, not all the amino acid changes could be mapped to a structure. Only 17% of amino acid changes in PCAWG and HMF, 19% in TCGA and 67% in B-CAST were the amino acids mapped to a structure and therefore, we only have the complete measurements of the eight features for them. For the amino acid changes that could not be mapped to a structure we have only the features that were not structural.

With all, first, we described the landscape of amino acid changes found in breast cancer and their characteristics based on eight protein features. Next, we performed a dimensionality reduction and clustering to look for groups of mutations that may share characteristics (**Figure 39**). Finally, we annotated whether mutations were drivers or not to look whether there was a difference in the features describing each group of mutations (considering 'driver mutations' one group and 'not a driver' the other group of mutations).

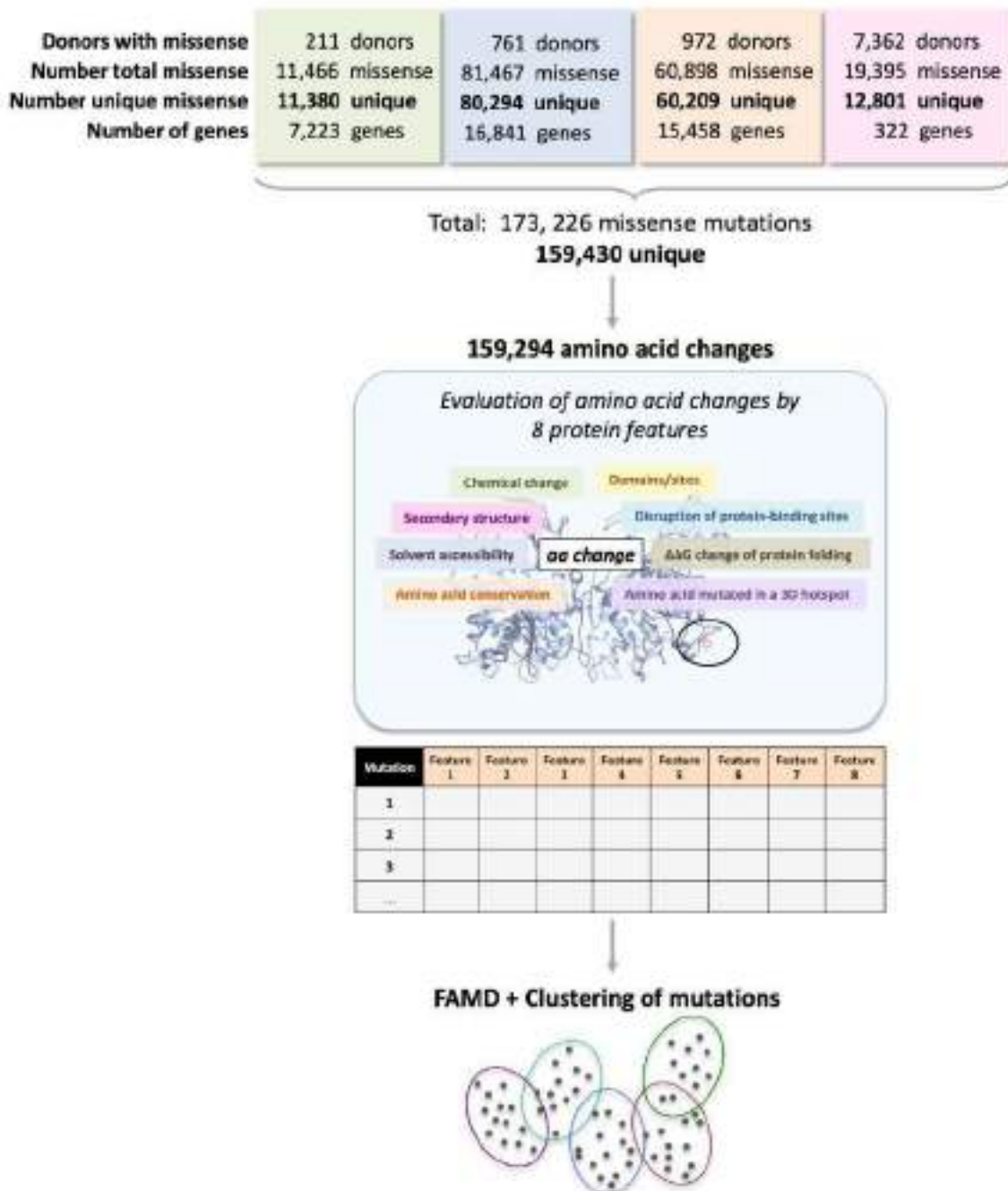


Figure 39. Workflow of the analysis of amino acid changes in breast cancer. Across 9,306 breast cancer donors (211 donors from PCAWG dataset, 761 from HMF, 972 from TCGA and 7,362 from B-CAST), we found 173,226 missense mutations in total, which were 159,430 considering unique mutations. These unique mutations were translated into 159,294 amino acid changes that were characterized by eight protein features (chemical change, secondary structure, solvent accessibility, amino acid conservation, domain, disruption of functional sites, change of the free energy of protein folding and belonging to a 3D hotspot). This data was used as input for a Factor Analysis of Mixed Data (FAMD) followed by hierarchical clustering to look for groups of mutations with the similar behaviour and a potential association to driver mutations.

5.3.1. Distribution of the amino acid changes found in breast tumours across the categories established for eight protein features

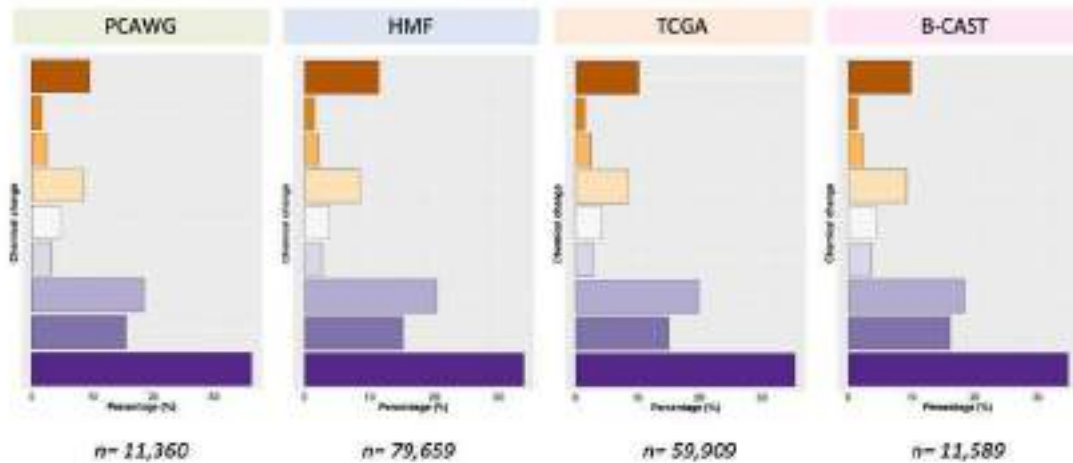
The twenty amino acids can be classified in four different groups depending on the charge of its side chain. As we described for the feature that we named as 'chemical change', depending on which is the original amino acid and which is the amino acid that arises from the mutation, we can determine different chemical changes (**Figure 40a**). The most frequent chemical change in breast cancer was a change between amino acids from the same category, followed by a loss of charge and a loss of polarity (**Figure 40b**). However, when we selected only the recurrent mutations (mutations that were found in more than one patient in the same dataset) and plotted the distribution of the chemical changes again, we could see that the 'change of charge' increased (**Figure 40c**). Under the assumption that recurrent mutations might be relevant for the cancer because they are coinciding across several patients, this may indicate that these 'change of charge' mutations might be more related to relevant mutation for the disease.

The most common amino acid mutation across all datasets was E>K (**Figure 41**). The second and third most common mutations (E>Q and D>N) in the overall dataset corresponded to what we observed for PCAWG, TCGA and HMF individually (**Figure 41**), but not for B-CAST. In the latter, the second most common amino acid mutation was H>R and the third, R>Q (**Figure 41d**). The fact that B-CAST is a panel of genes explains this difference in behaviour compared to the other datasets, since its composition is biased to the genes that are included in the panel. We could see that most of mutations in breast cancer happen in the secondary structure that is a loop (**Figure 42a**), next in an α -helix and the lowest number in a β -strand (**Figure 42a**). Regarding the solvent accessibility of the amino acid, mutations are more frequent in exposed amino acids (**Figure 42b**), except in B-CAST for which we saw a more equal distribution between exposed and buried amino acids. The percentage of mutations annotated as being part of a relevant site in the protein is consistent across PCAWG, HMF and TCGA, with around 33% of the mutations. From all amino acid changes in B-CAST, less than a 13% are affecting a functional site (**Figure 42c**). We had a total of 17,604 proteins that were mutated in breast cancer. Only 7,767 proteins had domains defined in its structure. A domain is a region that it is self-stabilizing and therefore can fold independently from the rest in the protein and be functional [151], for the proteins that there is no domain

annotation can be that it does not contain specific domains or that the structure is still unknown and the potential domains in the protein were not predicted nor elucidated yet. The protein domains more mutated in breast cancer were protein kinases followed by cadherin and Ig-Fibronectin Type III, in all datasets (**Figure 43**). This is expected since these are some of the more well-known domains and can be identified in different proteins and therefore more chances of being mutated. The conservation of the amino acids mutated in this tumour also tended to be high, considering the amino acids with a score of 7 or higher (in a range from 1 less conserved to 9 highly conserved).



b Subset of non recurrent missense mutations



c Subset of recurrent missense mutations

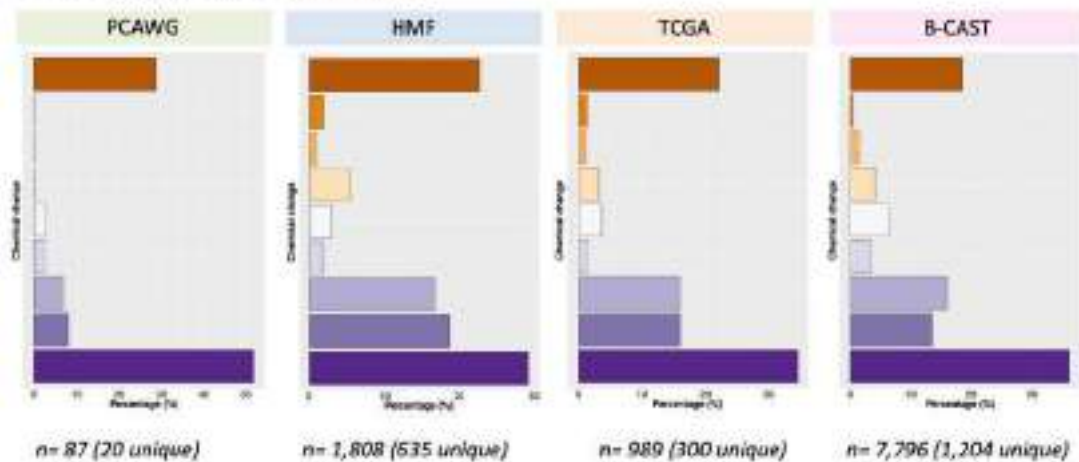


Figure 40. Distribution of amino acid changes found in breast tumours across the categories of 'Chemical change'. (a) Categories established depending on the amino acid change that is happening. (b) Distribution of the non-recurrent amino acid changes (amino acid mutations only found in one tumour genome in the cohort) across the different categories of 'chemical change' per dataset. (c) Distribution of the recurrent amino acid changes (amino acid mutations found in two or more tumour genomes in the cohort) across the different categories of 'chemical change' across datasets.

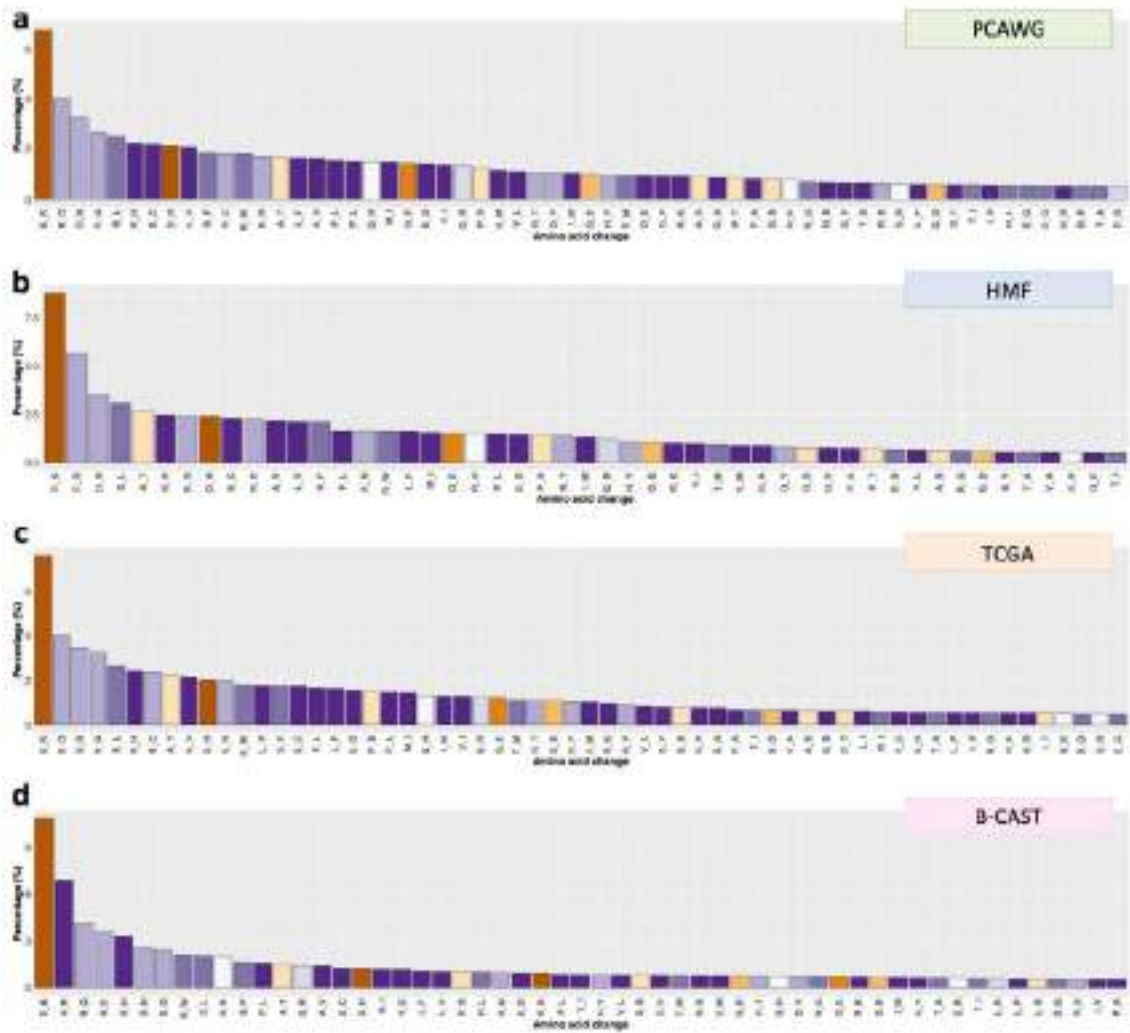



Figure 41. Percentage of the different amino acid changes found in breast cancer tumours. The 'x axis' indicates the top amino acid changes found in each dataset a) PCAWG, b) HMF, c) TCGA and d) B-CAST, e.g. glutamic acid to lysine is indicated as E_K. Each amino acid change is coloured by the chemical change, as defined in Figure 40a. The 'y axis' indicates the percentage of amino acid changes that involve that case of amino acids in each dataset.

a. Location in secondary structure

	PCAWG	HMF	TCGA	B-CAST
 Loop	6821 (53%)	48878 (51%)	38285 (55%)	6495 (51%)
 α -helix	3821 (33%)	28858 (36%)	21545 (39%)	4554 (36%)
 β -strand	1584	9785	7871	1662

b. Solvent accessibility

	PCAWG	HMF	TCGA	B-CAST
Exposed	6586 (57%)	48661 (61%)	35374 (59%)	6146 (48%)
Buried	3426 (30%)	23396 (29%)	18889 (31%)	5430 (42%)
NA	1423	7481	5575	1135

c. Mutation affecting a functional site

PCAWG	HMF	TCGA	B-CAST
3,602 / 11,466 (31.4%)	27,692 / 81,467 (34%)	20,041 / 60,898 (33.5%)	2,484 / 19,395 (12.8%)

Figure 42. Distribution of the amino acid changes found in breast cancer tumours across (a) location in secondary structure, (b) solvent accessibility and (c) location in a functional site.

The number of amino acid changes classified in each category is indicated together with the corresponding percentage per dataset (PCAWG, HMF, TCGA and B-CAST). **(a) Location in secondary structure.** The amino acid changes were classified depending on the secondary structure in which they were located. Three different types of secondary structure were considered: α -helix, β -strand and loop. **(b) Solvent accessibility.** The amino acid mutated was classified in buried, exposed or 'NA', the latter in the case of amino acids that could not be assigned to any category. **(c) Location in a functional site.** The number of mutations that were happening in an amino acid that is assigned to a functional site, such as protein binding site, DNA binding site or active site, is indicated from the total number of amino acid changes together with the percentage to which it corresponds.

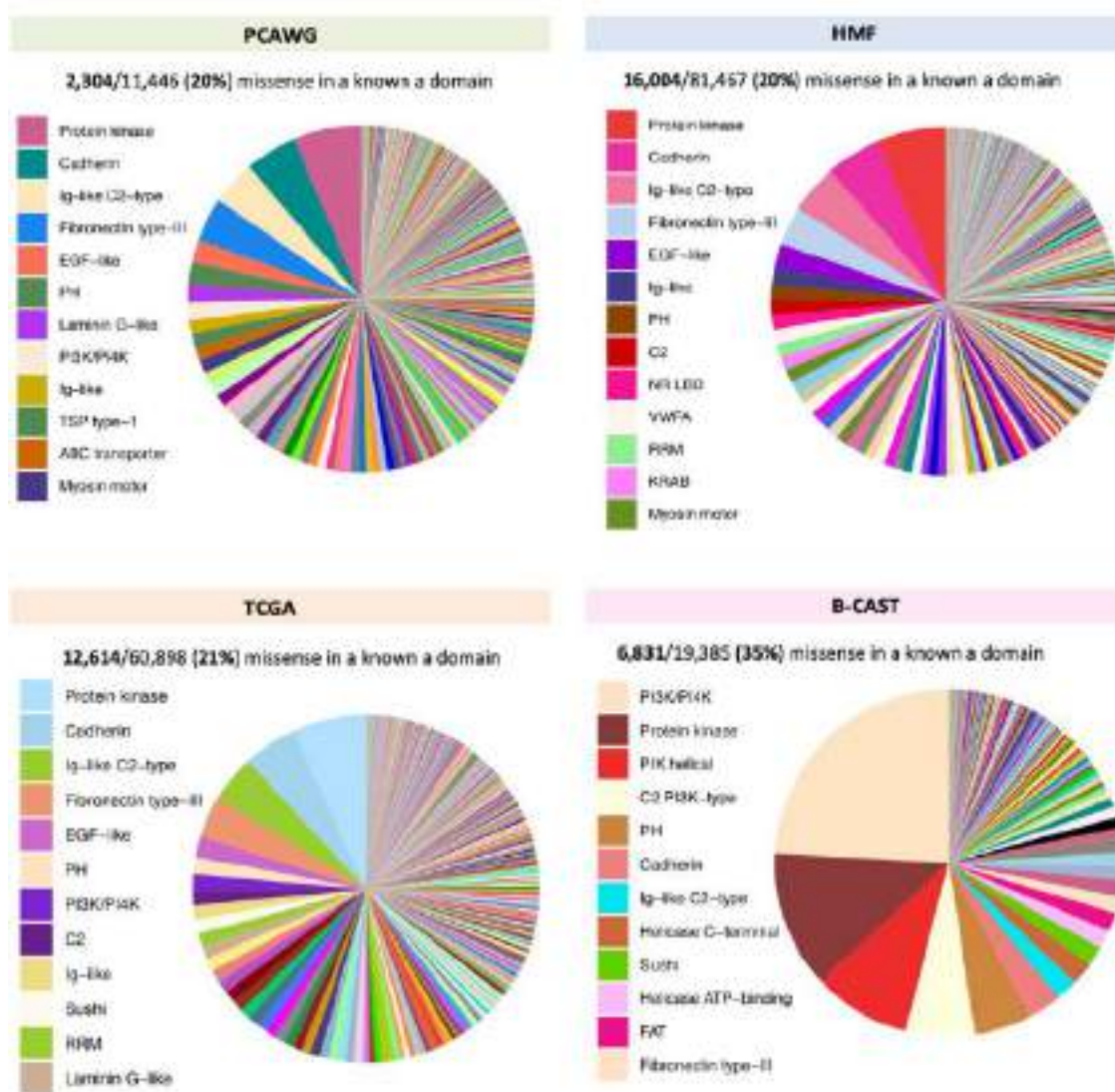


Figure 43. Distribution of the amino acid changes found in breast cancer tumours across protein domains. The number of amino acid changes that were assigned to a known domain is indicated together with the corresponding percentage among the amino acid changes found in each dataset (PCAWG, HMF, TCGA and B-CAST). The pie diagrams show the distribution of the amino acid changes across the different domains per dataset, indicating the legend for the top mutated domains in each dataset.

5.3.2. Proof of concept. Reduction method and clustering to find groups of relevant mutations

Using as input the amino acid changes with the eight different protein features by which we characterized these amino acid changes, we performed a Factor Analysis of Mixed Data (FAMD) followed by hierarchical clustering of the principal components obtained from the FAMD. We investigated whether we were obtaining different groups of mutations that were sharing the same characteristics across the different protein features or not. Furthermore, we annotated the mutations that were known or predicted as being a driver mutation. Groups of driver mutations could share characteristics and form clusters in which other less known mutations would be included and therefore, these less frequent mutations could be interesting targets for further investigation about their potential involvement in the disease.

Unfortunately, no coherent patterns were observed and no meaningful clusters could be identified. The difficulty of using categorical data in a dimensionality reduction method, since there are no intermediate values to go from one category to another, and the potential lack of more accurate features to characterize amino acid changes could be the reason of not being able to find different groups. We tried other dimensionality reduction methods such as Multiple Correspondence Analysis (MCA) or Multiple Factor Analysis (MFA), but all results were not conclusive. The variety of mutations and the variation in their categorization using these protein features might be responsible of the inability to find different clusters as we were expecting.

Take-home messages Chapter 3

- From the characterization of 159,294 amino acid changes across 9,306 breast cancer donors by eight protein features we can highlight that it was observed that most of the amino acid changes in breast cancer happen between amino acids of the same category. Most of the mutated amino acids are in the secondary structure that is a loop and the amino acid is exposed. Around 33% of the mutations are affecting a functional site in the datasets containing all

proteins, while only around 13% in the dataset that only contains a panel of 323 genes.

- The most frequent amino acid change in breast cancer is E>K (in each of the individual four datasets) followed by E>Q in PCAWG, TCGA and HMF and H>R in B-CAST. Different tendencies are also seen in B-CAST regarding other protein features. This might be explained by that its composition is biased to the genes that are included in the panel.
- Considering only the recurrent amino acid changes (mutations that were found in more than one patient in the same dataset) the proportion of 'change of charge' cases increase. These mutations might be more relevant for the disease.
- The protein features collected did not stratify mutations in relevant groups nor resulted in the characterization of driver mutations.

5.4. CHAPTER 4. LANDSCAPE OF PROTEIN CHANGES IN p110 α (PIK3CA) IN CANCER

Finding mutations in the same gene across and within cancer types could imply similarities in the origin and characteristics of a tumour and therefore, the possibility that the same treatment could be used for these patients [152]. However, not all mutations in the same gene have the same effect and therefore may imply different treatments are required [153]. For example, SF3B1 gene, which encodes a complex of the spliceosome, a macromolecular complex that splices the pre-mRNA, is differently mutated in different tumour types. Across the different tumour types the effect of the mutations on the patient's prognosis differs, while in uveal melanoma it is associated with a more favourable prognosis [154], in chronic lymphocytic leukaemia (CLL) it is linked to a more aggressive disease and shorter survival [155]. Also, studies have linked the genomic landscape of tumours with tumour immunity, identifying somatic mutations associated with immune infiltrates [156][157][158]. Therefore, combining the study of the mutations in the genome with other measurements such as the tumour microenvironment and other phenotypes associated to the tumours seems very informative. This could help to characterize the tumours, elucidate the right biological mechanism and improve the selection of treatment to fight the tumour successfully.

We investigated these aspects for PIK3CA. PIK3CA is a well-known gene involved in several cancers that encodes the p110 α protein. We looked at the landscape of protein changes found in this protein across all cancer types and focused further analyses on the tumours for which we had the highest number of donors with this gene mutated, which were breast, colorectal and uterus cancer. In addition, we focused in more detail on breast cancer and studied the association of PIK3CA mutations with the tumour immune microenvironment and clinical parameters of the tumour, such as grade, stage, hormone status or survival. We focused on breast cancer for this last part because it was the tumour type for which we had the highest number of donors and therefore we had more power to test our hypothesis.

5.4.1. Pan-cancer description of PIK3CA mutations

From the genomic mutations that were found in the PIK3CA gene in our dataset, we focused on the subset of coding mutations and excluded the silent mutations, since they do not result in a protein change.

5.4.1.1. Frequency of PIK3CA mutations across cancer types

PIK3CA was mutated in several cancer types and the proportion of mutated donors was different per dataset (**Figure 44**). In the PCAWG dataset (**Figure 44a**) the top 3 mutated cancer types were uterus, colorectal and breast, while in TCGA (**Figure 44b**) the ranking changed to uterus, breast and colorectal cancer. In HMF (**Figure 44c**), breast and uterus were the top mutated cancer types, while colorectal was in the 9th position. Considering all datasets, including B-CAST that had ~30% of breast cancer donors with PIK3CA mutation, and taking the mean per cancer type across the four datasets, the top mutated cancer types were uterus (~42%), followed by breast (~34%) and colorectal cancer (~29% of mutated donors).

The most frequent mutation type in PIK3CA gene was by far missense mutations (~95% in the joint dataset), followed by deletions (**Figure 45**). Only 0.17% of the mutations overall were nonsense mutations. Uterus cancer had with 1.41% the highest percentage of the three cancer types highlighted (**Figure 45**). Colorectal cancer had the highest percentage of deletions compared to breast and uterus cancer (**Figure 45**).

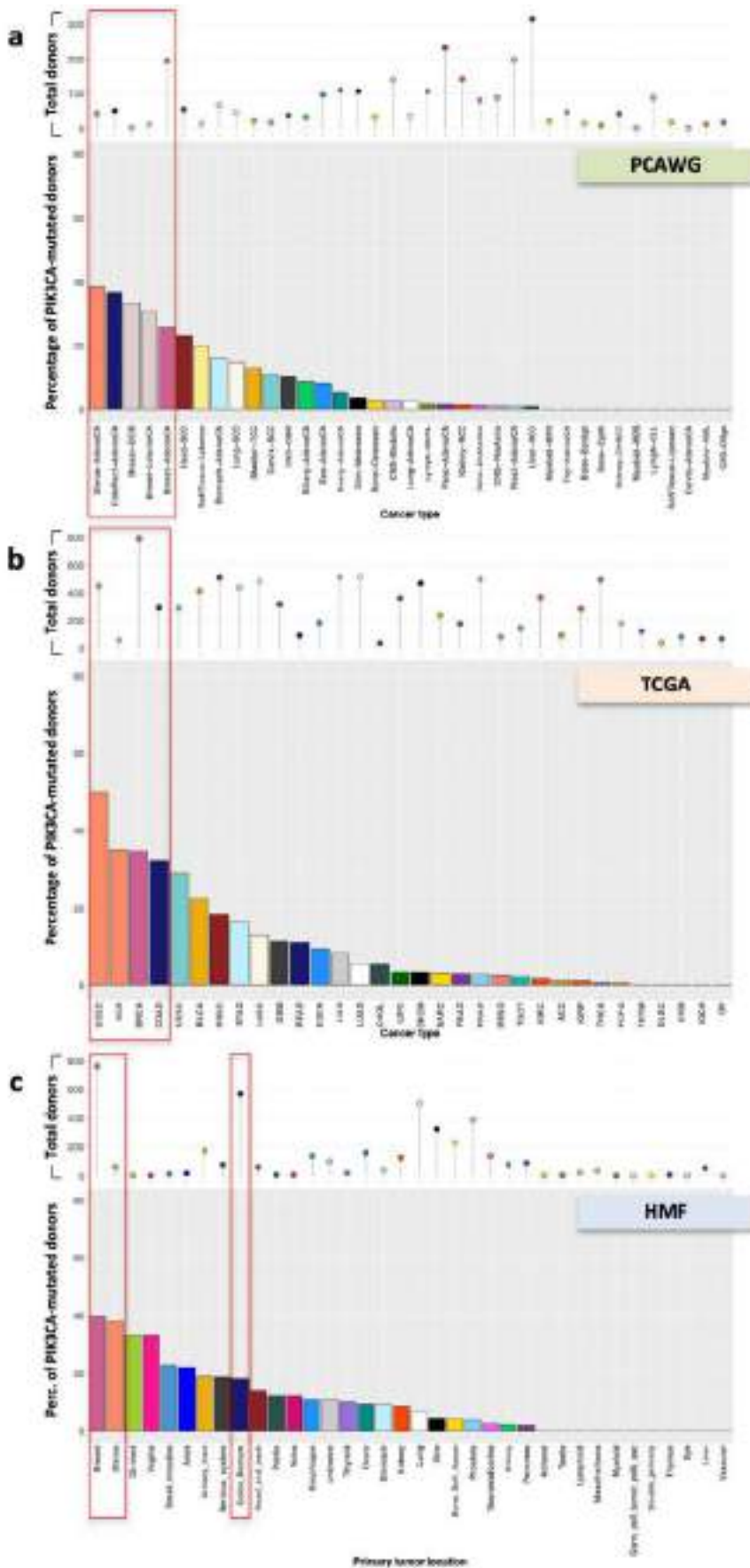


Figure 44. PIK3CA mutations in different datasets. Total number of donors (lollipop plot, above) and percentage of donors with PIK3CA mutation (barplot, below) per cancer type and per dataset: (a) PCAWG, (b) TCGA and (c) HMF dataset.

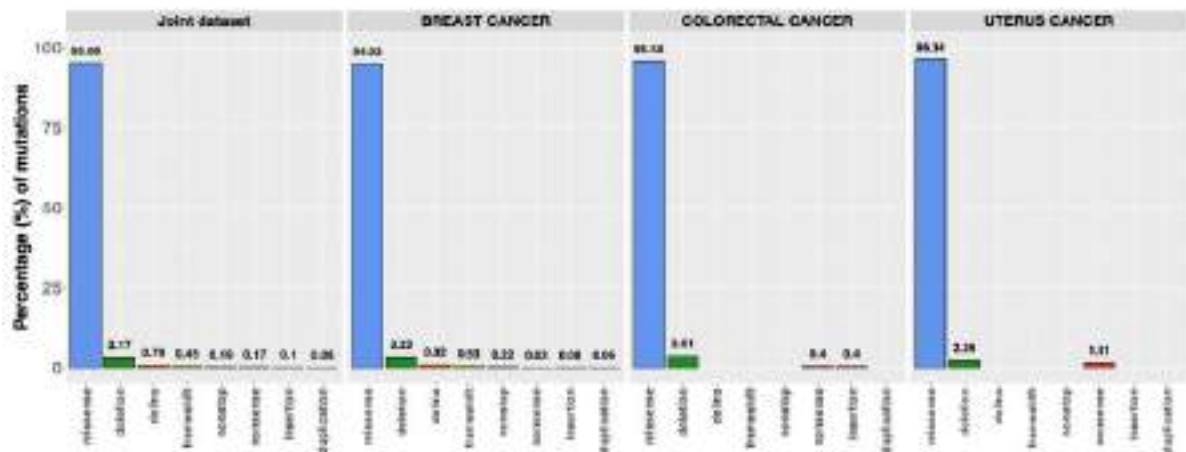


Figure 45. Mutation types in PIK3CA gene. Distribution of PIK3CA coding mutations (excluding silent mutations) across the different mutation types in the joint dataset (all cancer types included) and in the most frequently mutated cancer types inside the joint dataset (breast, colorectal and uterus cancer). ‘delins’: deletion followed by an insertion. ‘nonstop’: mutation that occur within the stop codon, changing the stop codon for a new amino acid, which leads to the continued and inappropriate translation of the mRNA making a protein longer than expected.

5.4.1.2. Description of protein changes in p110 α (PIK3CA)

From all coding mutations (excluding silent) in PIK3CA gene, 5,040 were missense mutations and were translated into 385 unique amino acid changes in the p110 α protein. Since the number of samples available for each cancer type was highly variable (**Figure 44**), we focused on the cancer types with the higher percentage of PIK3CA-mutated donors that we pointed out previously: breast, colorectal and uterus cancer.

Analysing the different protein features per amino acid change across breast, uterus and colorectal cancer, we saw that the highest proportion of amino acid changes were happening between amino acids classified in the same category (**Figure 46**), followed by the case of change of charge, in most of cases from an acidic to a basic amino acid in breast cancer. In colorectal and uterus cancer we also saw a proportion of mutations that involved a loss of charge, while this proportion was small in breast cancer. We saw that the mutations in the three cancer types were mainly happened between amino acids of the same or similar size (**Figure 47**). In breast cancer, a higher percentage of mutations affected a loop than an α -helix or β -strand, which was to a large degree due to the most frequent mutation in this gene in this cancer (H1047R). The histidine amino acid number 1047 is located in a loop. In colorectal and uterus cancer, a higher

percentage of mutations affected an α -helix (**Figure 48**). Across datasets, the highest percentage of mutations in PIK3CA affected residues that were considered conserved or with an intermediate level of conservation. Most of the amino acid changes affected exposed residues in all cancer types across the different datasets (**Figure 49**), what was expected since most amino acids in p110 α (PIK3CA) protein are classified as exposed. When we explored the conservation across cancer types, we saw more conserved residues affected in breast than in the other two cancer types. We explore this in other datasets, such as lung and bladder cancer and saw that the amino acids affected were also less conserved (**Figure 50**).

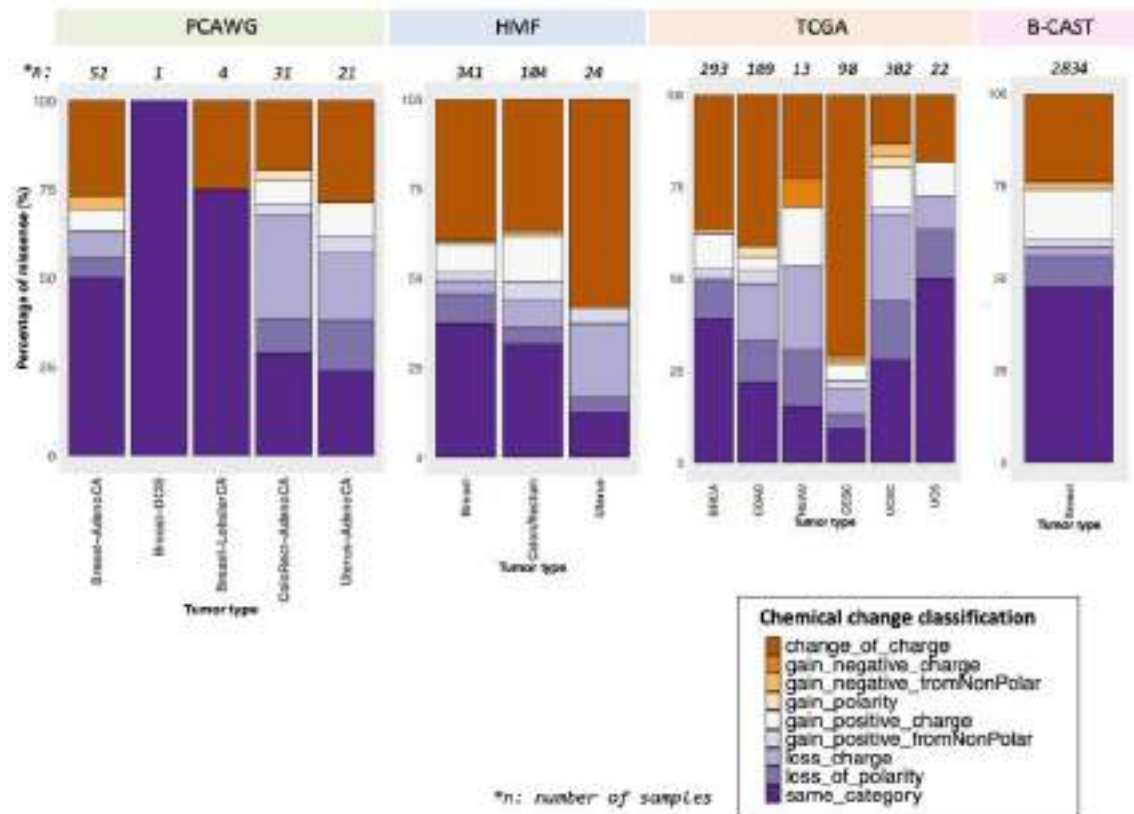


Figure 46. Proportion of the different categories of amino acid changes in p110 α (PIK3CA) protein in breast, colorectal and uterus cancer cohorts in the PCAWG, HMF, TCGA and B-CAST dataset. The number of samples per cohort is indicated above each bar.

Amino acid size change

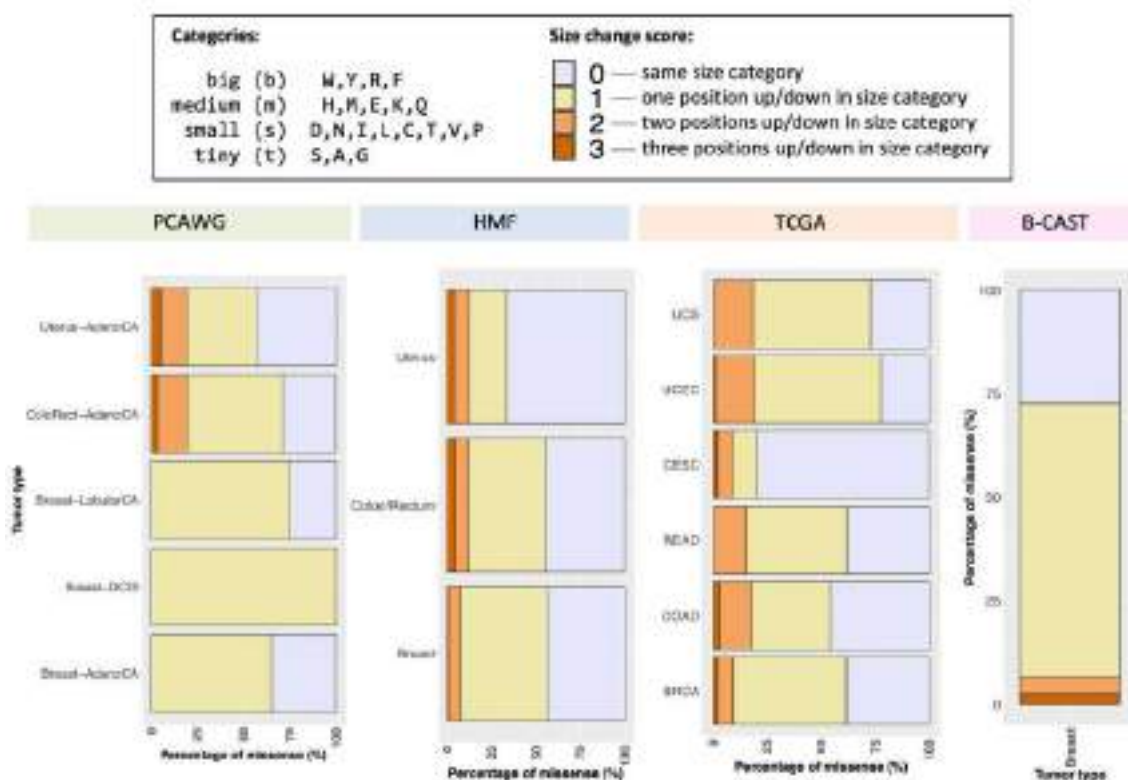


Figure 47. Proportion of the different categories of amino acid size change for the amino acid changes found in p110 α (PIK3CA) protein in breast, colorectal and uterus cancer cohorts in the PCAWG, HMF, TCGA and B-CAST dataset. The 20 amino acids are classified in big, medium, small or tiny and a score from 0 to 4 is defined depending on how big the change is.

Secondary structure

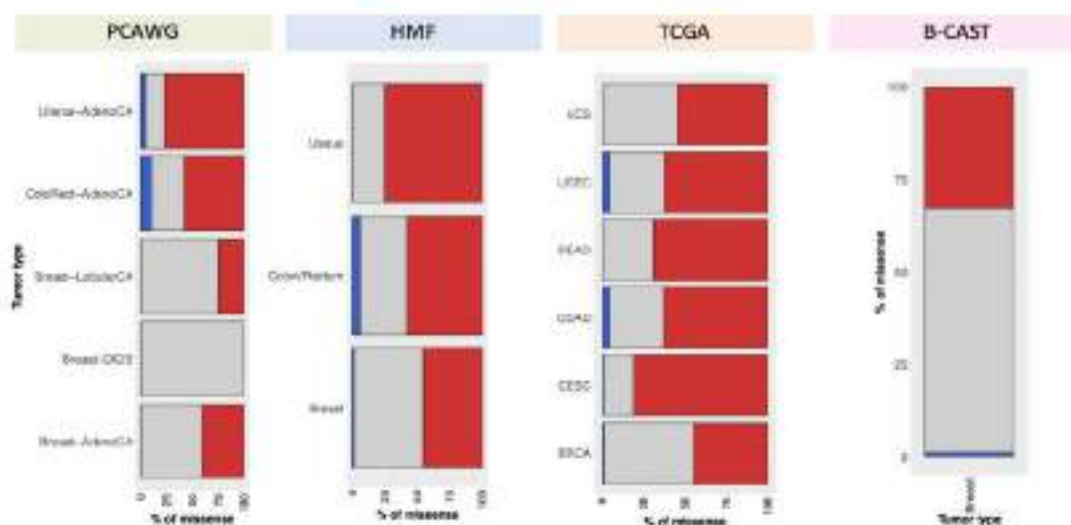


Figure 48. Proportion of type of secondary structure hit by amino acid changes found in p110 α (PIK3CA) protein in breast, colorectal and uterus cancer cohorts in the PCAWG, HMF, TCGA and B-CAST dataset.

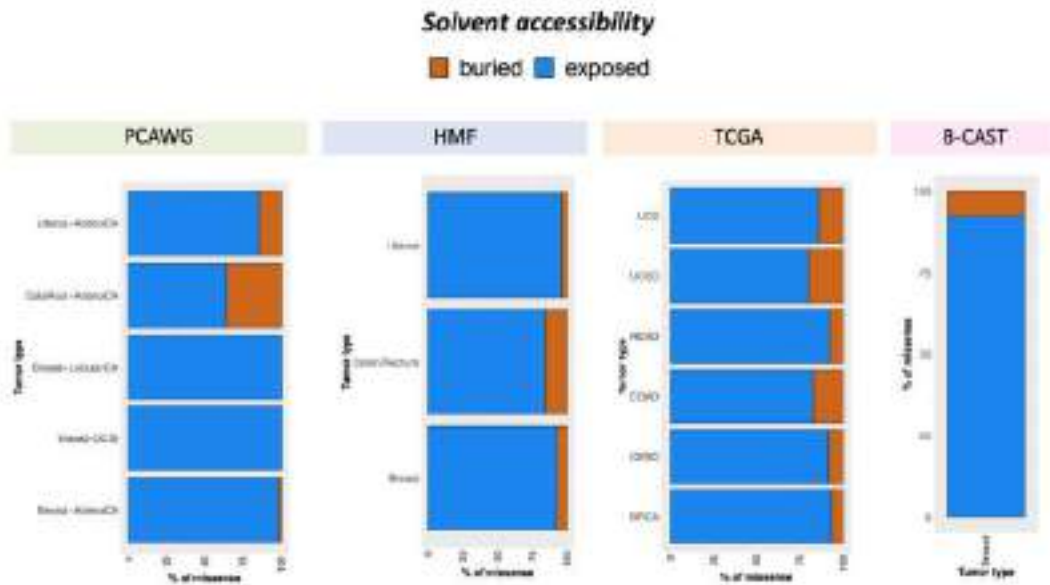


Figure 49. Proportion of amino acid changes found in p110 α (PIK3CA) protein happening in an exposed or buried amino acid in breast, colorectal and uterus cancer cohorts in the PCAWG, HMF, TCGA and B-CAST dataset.

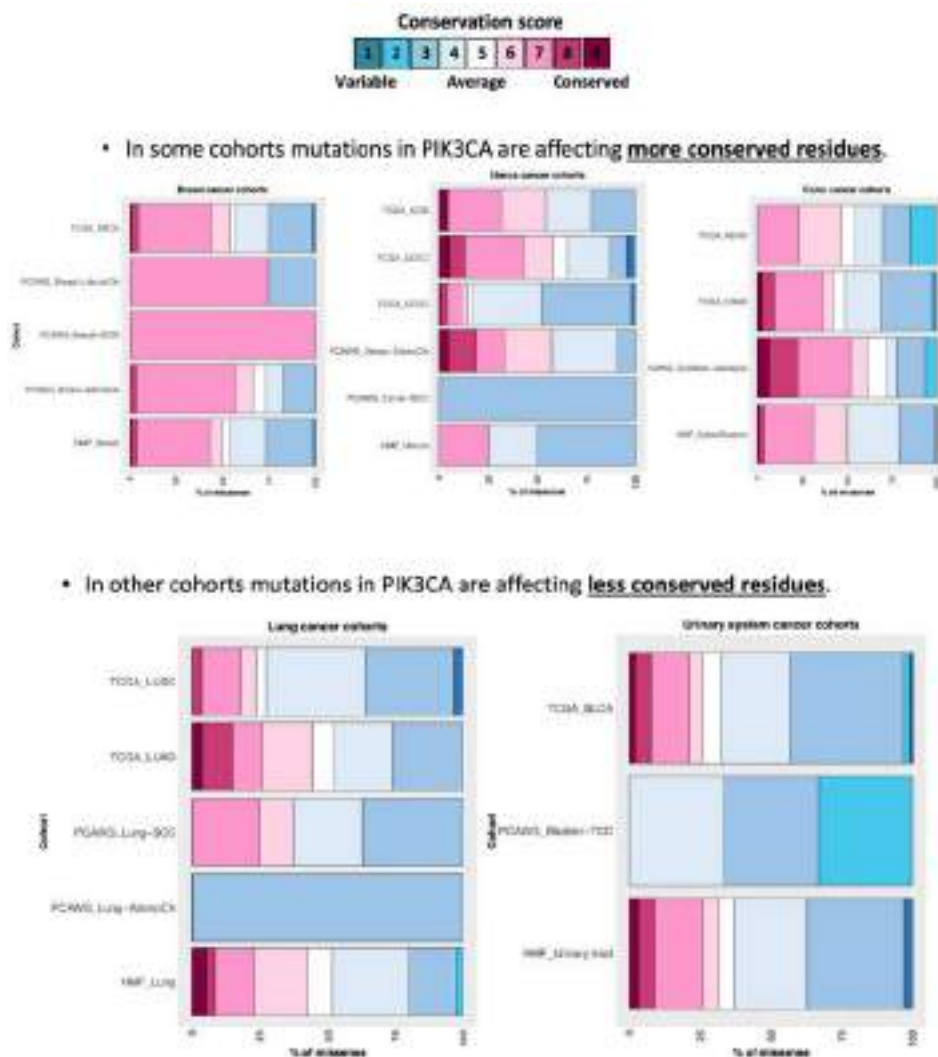


Figure 50. Proportion of amino acid conservation categories for the amino acid changes found in p110 α (PIK3CA) protein in different cohorts in the PCAWG, HMF, TCGA dataset.

In addition, mutation3D software was able to find several 3D-hotspots of mutations in the p110 α protein structure across the different datasets (Figure 51). Several of them had previously been reported in the literature [29][120]. The main 3D-hotspots identified were around the amino acids in position 1047, 545 and 345, respectively. These positions by their own have been identified as a hotspot mutation. A 3D-hotspot that included the residues 106, 107, 108 and 111 was also identified (Figure 52), which to our knowledge, has not yet been described in the literature. Amino acids G106 and K111 are already known to be frequently mutated in endometrial carcinomas [159]. The 3D-hotspots were located in regions of the p110 α protein structure that are known to be relevant to its protein function [120].

PCAWG			HMF			TCGA			B-CAST		
Cluster	Max. dist	P-value	Cluster	Max. dist	P-value	Cluster	Max. dist	P-value	Cluster	Max. dist	P-value
145, 363	8.84932	11.2176	465, 981, 363	9.17796	9.210	331, 296, 263, 304	6.10000	1.418	355, 354, 356	9.75453	0.779
102, 345, 348	9.49034	0.000	545, 546, 549	6.13896	5.418	545, 546, 547	5.31513	28.058	545, 545	1.58022	10
431, 538	8.64687	13	542, 539	7.98312	20	431, 436, 438	7.86643	21	431, 434	1.89058	18
451, 428	8.84394	13	432, 454, 438, 421	7.6643	0.812	725, 725, 722	8.59712	21	725, 730	6.24845	15
105, 544	9.79890	18.7084	726, 725	3.18094	19.36	979, 971, 968	9.67313	19	418, 420	6.75862	15
1063, 1045, 1047	5.54627	1.5836	417, 408	5.81537	20	18, 88, 98	8.99815	21	880, 885	4.88625	18
1875, 1071	6.51294	12	99, 98	8.42813	20	1004, 1091, 1097, 1093, 1090	5.47228	4.2127	39, 38, 88	7.6968	6.895
1067, 1069	7.46706	12	345, 471, 344, 343	9.32973	1.397				357, 344, 346, 378	6.51365	1.793
281, 245	5.81708	11.9066	1044, 1049, 1047	5.32586	4.538				1044, 1045, 1048, 1047	6.58939	2.49
48, 88, 18	7.9368	0.431	81, 81, 104, 69	8.38818	1.113				81, 76, 11, 81	9.101	2.36
191, 709	6.48809	11.950	1033, 1031	5.1491	20				1003, 1094	1.8829	18
95, 3	5.81925	11.5004	87, 89	3.78934	19.48				509, 297, 322, 188	9.25902	2.119
181, 184	3.81796	11.7004	186, 188	8.42781	20	2884, 3485, 3483	6.68418	21	438, 436, 1007	6.61339	6.748
111, 106, 107	6.17836	1.8900	1043, 1042, 1049	5.4931	4.821	491, 485, 357	6.15409	21	840, 890	9.25134	13
731, 732	2.86413	11.2780	725, 732	9.97878	20	528, 535, 532	6.18855	21	491, 489	8.28987	18
186, 578	9.41651	11.3008	208, 208	2.11872	20	121, 121, 138	8.13811	20.863	1025, 952	8.93275	18
			118, 134	4.15238	20	88, 87, 86, 85, 83	23		1035, 1023	1.22485	13
			174, 882	9.32587	20	674, 671, 684	8.812	21	213, 212, 115	1.89889	5.318
			91, 75	8.14877	20	92, 9, 91	8.75316	21	1025, 1021	6.54284	16
			52, 56	7.13180	20	181, 184, 60	5.99809	21	218, 124	4.95289	18
			113, 113, 115	6.10846	5.464	12, 11, 11	5.19444	20.897	919, 888, 887	6.82288	6.768
			786, 297	7.42233	18	881, 880, 888	8.89478	21	1007, 1006, 1005, 1010	5.63389	8.71
			814, 551	9.67233	20	379, 375, 388	5.45543	21	93, 75	6.14877	18
			456, 485	5.91596	20	681, 584, 558	7.71275	21	2600, 1027	1.89889	18
			1017, 1018	3.88273	19.48	165, 151, 658	6.87256	21	11, 13, 8	6.95121	6.818
			849, 490, 447	6.54261	5.924	84, 83	3.81513	30	681, 684	1.83788	18
			1015, 1011	6.14284	20	182, 188	8.79818	30	814, 811, 814	8.91213	4.888
						985, 981	5.86878	18	398, 330	8.73867	18
						551, 369	8.94580	30			
						1818, 1805, 1807	5.1477	20.893			
						745, 744	5.46078	30			
						38, 19	5.54068	30			

Figure 51. 3D-Hotspots of mutations on p110 α protein structure found by mutation3D per dataset (PCAWG, HMF, TCGA and B-CAST). 'Cluster': amino acids in p110 α protein included in the 3D-hotspot cluster. 'Max. distance': maximum distance between the amino acids that are belonging to the 3D-hotspot cluster. 'P-value': p-value computed empirically. The clusters with a p-value lower than 5% are highlighted because they would have a chance of 0.05 or lower of being wrong.

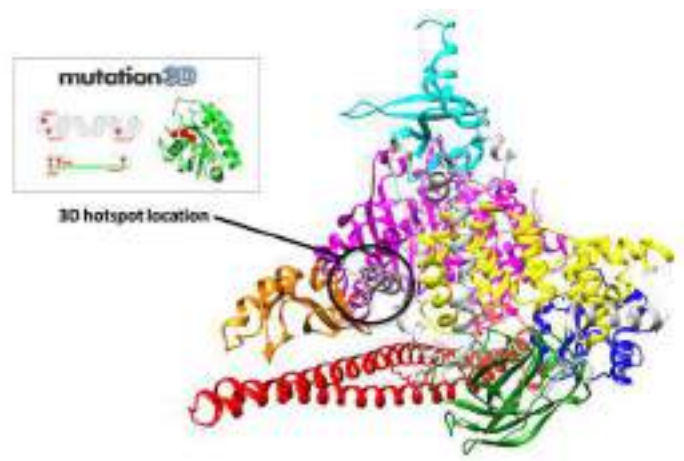


Figure 52. 3D-Cluster in the linker ABD-RBD including the amino acids 106, 107, 108 and 111.

Finally, a highlighted observation was at domain level. We observed that the distribution of amino acid changes across the protein domains across cancer types differed (**Figure 53**). For example, the kinase domain was the most frequent mutated domain in breast cancer in all datasets. We will focus on this different distribution across cancer types in depth in the next section (5.4.1.3)

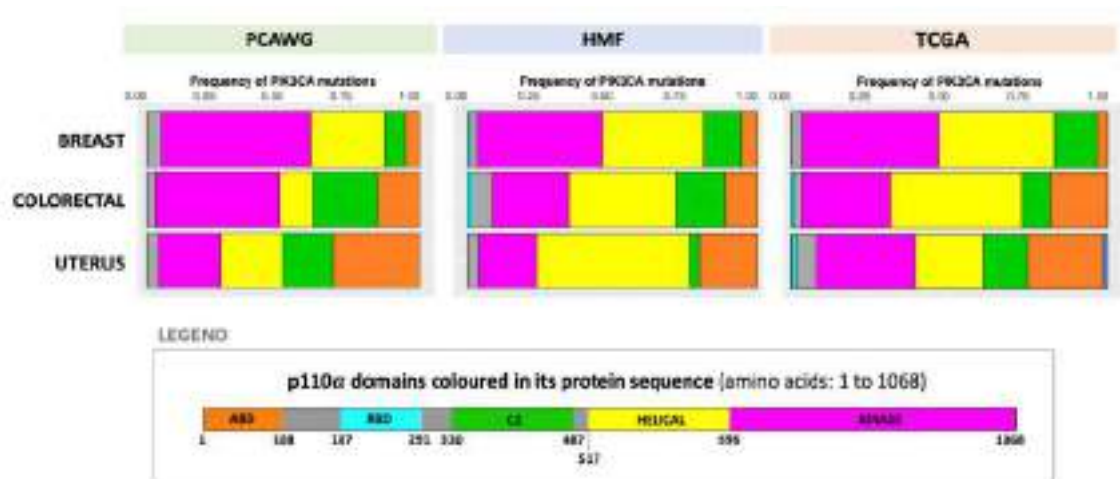


Figure 53. Proportion of mutations in each of the main domains of p110 α (PIK3CA) in breast, colorectal and uterus cancer in the PCAWG, HMF and TCGA dataset.

In addition, the evaluation of the change in the free energy of protein folding using FoldX [122] combined with the conservation of the amino acid that is mutated indicated that amino acid changes happening in residues more conserved or more variable were predicted to be destabilizing (**Figure 54**). There were no cases of stabilizing mutations (**Figure 54**). Some highly destabilizing mutations at the C2 PI3K-type domain, such as C378R/F/W/Y, were happening at residues very low conserved and still be potentially damaging, what indicates that conservation would have not been enough to elucidate the relevance of these mutations.

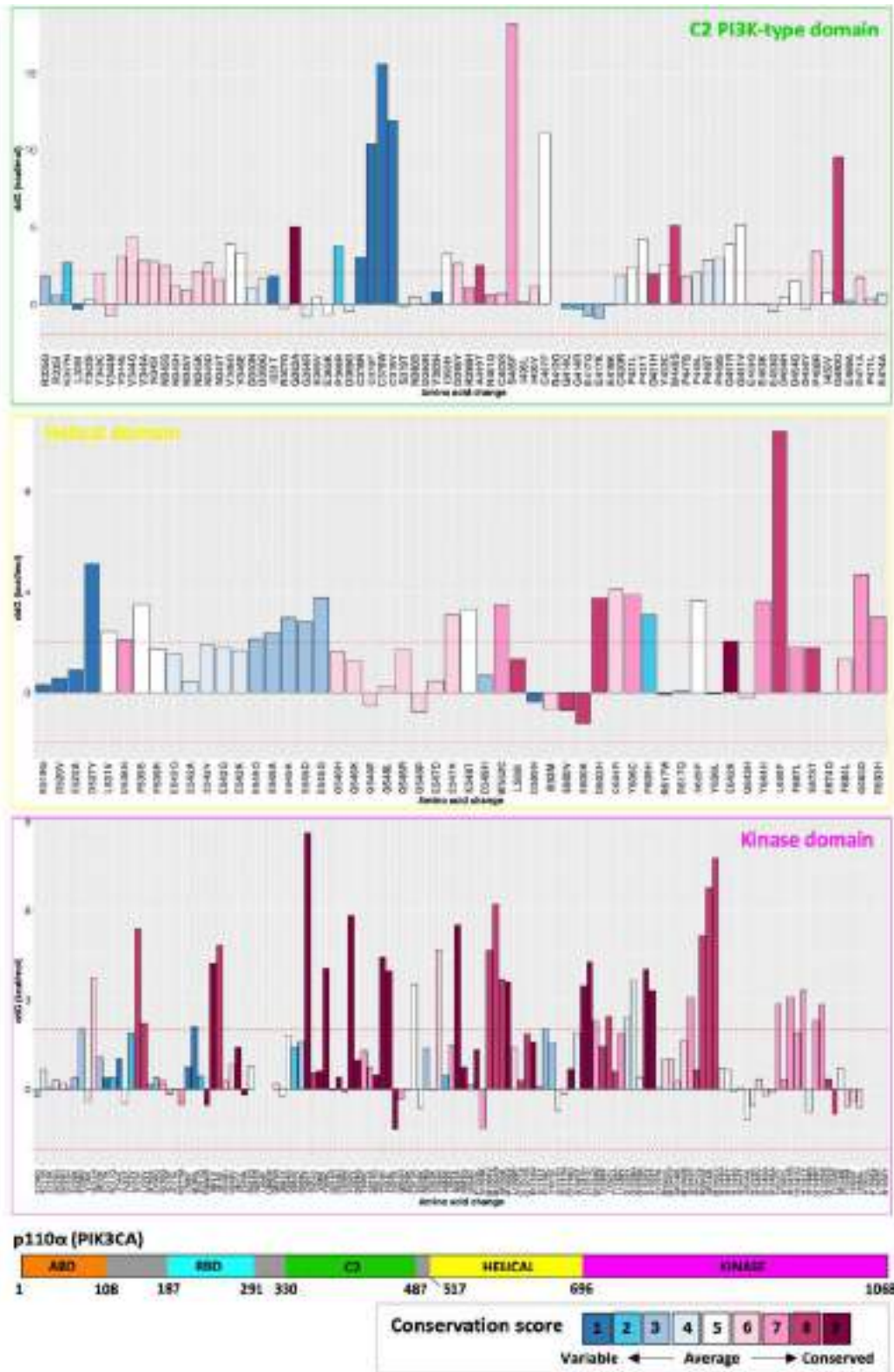


Figure 54. Change in the free energy of protein folding ($\Delta\Delta G$ in kcal/mol) upon amino acid changes in p110 α (PIK3CA) protein in C2 PI3K-type, helical and kinase domain. Red dotted line indicates the threshold by which the mutation is considered stabilizing or destabilizing (< -2 kcal/mol or > 2 kcal/mol, respectively). Values between -2 and 2 kcal/mol are considered not affecting protein stability. Bars are coloured by the score of conservation of the amino acid that is mutated.

5.4.1.3. The distribution of mutations along p110 α (PIK3CA) protein domains differs between breast, uterus and colorectal cancer

The distribution of mutations across the domains of p110 α was different among the most PIK3CA-mutated cancer types: breast, uterus and colorectal cancer. We can distinguish five main domains in the 1,068 amino acids long p110 α protein: ABD, RBD, C2 PI3K-type, helical and kinase domain (see *Introduction 1.5.1* and legend in **Figure 53**). **Figure 55** shows all PIK3CA coding in our joint dataset in the three cancer types: breast (**Figure 55a**), colorectal (**Figure 55b**) and uterus (**Figure 55c**) cancer. The distribution of mutations differs across these cancer types, except for the fact that we find the well-known hotspots mutated in all of them (residues 1047, 542 and 545). Breast cancer has the highest peak of mutations in the hotspot of the kinase domain (H1047), colorectal cancer has its highest peak at the hotspot at the helical domain (E545) and uterus cancer seems to be mutated with more equal frequency across the different hotspots. In addition, uterus had an increase in the proportion of mutations in the ABD compared to breast and colorectal cancer, with the amino acid 88 being the most mutated (**Figure 55a**). The stacked bars at the right of each 'lollipop' (**Figure 55a-c, right**) indicate the percentages of mutations in each region, where there were clear differences among the three cancer types. In breast cancer, there is a higher proportion of mutations in the kinase domain (pink). Mutations in the ABD domain were hardly present in breast cancer while they were higher in proportion in uterus cancer (**Figure 55c**) compared to breast and colorectal cancer (**Figure 55a** and **55b**). Uterus cancer had a higher proportion of mutations in the ABD domain and ABD-RBD linker. In colorectal cancer, helical mutations were the most abundant, due to the hotspot mutations for which they are enriched (E545K, E542K).

Aside from missense mutations, for which their effect on the protein is normally more uncertain, there were other mutation types that hit the p110 α protein (**Figure 56**). Deletions and insertions are clustered in the same positions across cancer types: end of ABD, linker ABD-RBD or C2 PI3K-type domain (**Figure 56**). Deletions clustered in the C2 PI3K-type domain have been suggested to be associated to sensitivity to PI3K inhibitors [160]. The proportion of mutations per domain with respect to the total mutations in each tumour type is represented in bars at the right of the lolliplos (**Figure 56**). Deletions and deletions followed by an insertion mainly affected the ABD, linker ABD-

RBD and the C2 PI3K-type domain in the three cancer types. These domains are not involved in the catalytic activity of the protein, but they are involved in the attachment to the regulator and membrane. Insertions and deletions can have a higher impact on protein compared to other mutations, and in this case it is suggested that these mutations have an effect on the interaction of p110 α protein with the regulator, leading to a constitutive activation because of the regulator not being able to bind and inhibit the activity [160]. In uterus cancer, a proportion of mutations, which were nonsense mutations, affected exclusively the kinase domain. These cases are expected to result in a non-active protein since the region where the catalytic activity occurs, the phosphorylation of PIP₂ to PIP₃ [41]. Frame-shift mutations after the kinase domain as well as non-stop mutations were only found in breast cancer (**Figure 56a**), which are expected to continue the translation of the protein sequence and result in a protein longer than the original.

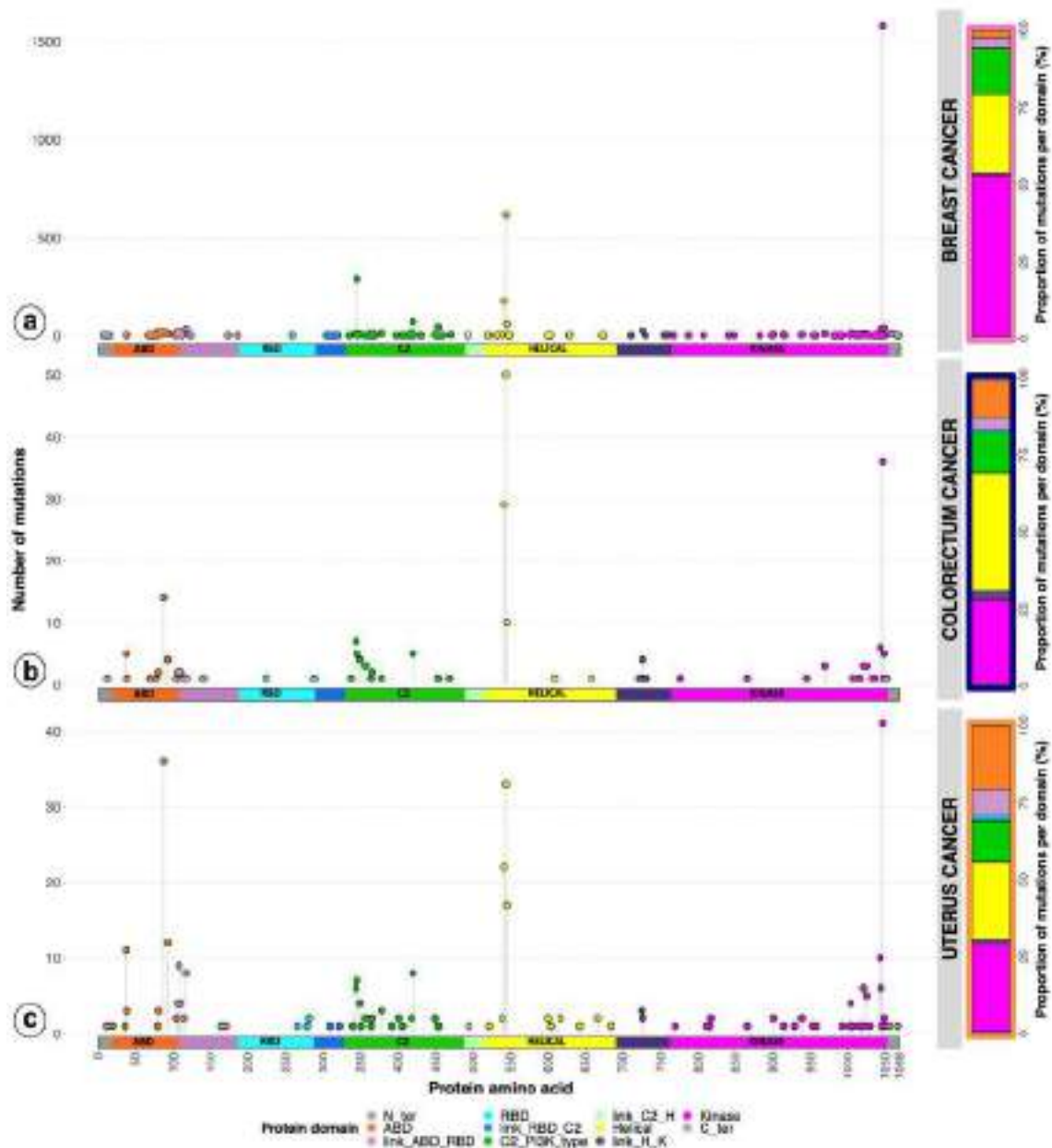


Figure 55. Number of missense mutations per position in the p110 α protein in three different cancer types: (a) breast, (b) colorectal and (c) uterus cancer. Colours indicate the region of the protein where the mutations are located. Note that the y axis scales are not the same, they are adapted to the values found in each cancer type. The vertical bar to the right of each cancer type name summarises the proportion of mutations that fall in each region. The number of samples included is 3,120, 201 and 244 for breast, colorectal and uterus cancer respectively.

N_ter: first residues of the protein from the N-terminal side; *ABD*: adaptor binding domain; *RBD*: Ras binding domain; *Helical*: helical domain, *Kinase*: kinase domain; *link_ABD_RBD*: linker residues between the ABD and RBD; *link_RBD_C2*: linker between the RBD and C2 PI3K type domain; *link_C2_H*: linker between the C2 PI3K type domain and the helical domain; *link_H_K*: link between the helical and kinase domain; *C_ter*: last residues of the protein until the C-terminal side.

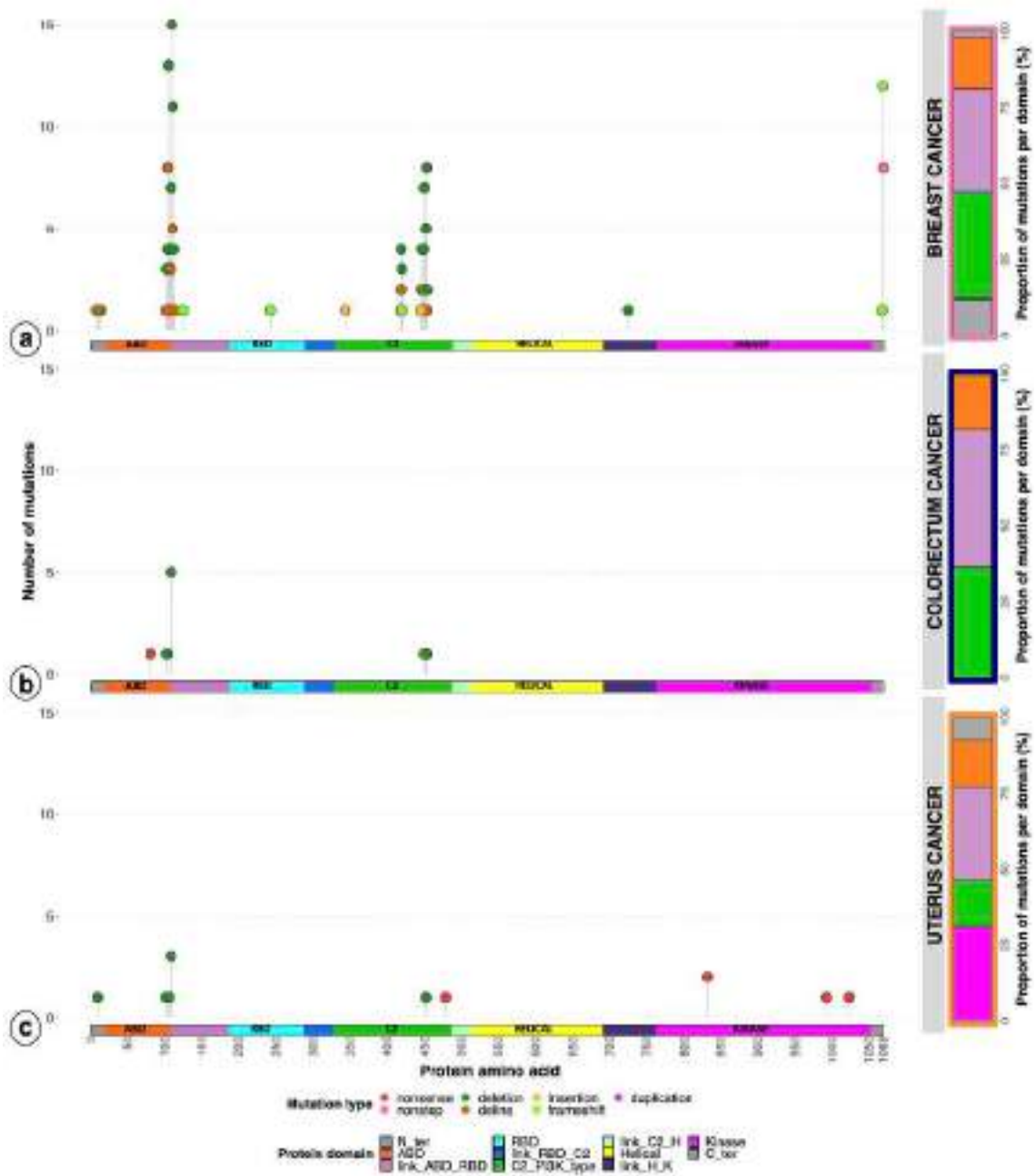


Figure 56. Number of non-missense coding mutations per position in the p110 α protein in three different cancer types: breast, colorectal and uterus cancer. Point colours indicate the type of mutation. Deletion, insertion and delins (deletion followed by an insertion) are in-frame; the same mutation types causing a frame-shift in the reading frame are annotated as “frameshift” and are changing the protein sequence afterwards. The vertical bar on the right of each cancer type name summarises the proportion of mutations falling in the different protein regions, indicated by different colours. The number of donors included is 180, 11 and 13 for breast, colorectal and uterus cancer respectively.

N_ter: first residues of the protein from the N-terminal side; *ABD*: adaptor binding domain; *RBD*: Ras binding domain; *Helical*: helical domain, *Kinase*: kinase domain; *link_ABD_RBD*: linker residues between the ABD and RBD; *link_RBD_C2*: linker between the RBD and C2 PI3K type domain; *link_C2_H*: linker between the C2 PI3K type domain and the helical domain; *link_H_K*: link between the helical and kinase domain; *C_ter*: last residues of the protein until the C-terminal side.

5.4.2. Underlying causes of PIK3CA mutations

We hypothesized that one possible explanation underlying the different distributions of mutations in PIK3CA across breast, uterus and colorectal cancer could be due to the different mutational processes that were active in each cancer type. We divide the samples into groups depending on the domain mutated (ABD, linker ABD-RBD, C2, helical and kinase domain), multiple domains mutated or no mutation in PIK3CA. We also look at other possible causes behind these observations such as differences in epigenomics, such as chromatin accessibility.

5.4.2.1. Mutational signatures can explain the different distribution of mutations across p110 α (PIK3CA) protein domains in uterus and colorectal cancer

We used mutational signatures as a proxy for the mutational processes that could be active in the tumours. **Table 7** shows the main mutational processes identified in the different cancer types. The activity of the APOBEC family of cytidine deaminases (SBS2 and SBS13) was the main mutational process in breast cancer together with the defective homologous recombination (HR) DNA damage repair pathway (SBS3, ID6). APOBEC signatures were also seen in some of the uterus cancer. Polymerase epsilon (Pol ϵ) exonuclease domain mutations (SBS10a/b, SBS28) that lead to a defective performance of this polymerase were present in uterus and colorectal cancer donors together with signature ID1, which is related to slippage during DNA replication of the replicated DNA strand. Defective Mismatch Repair (MMR) signatures (SBS26, SBS44, ID7) are mainly present in colorectal cancer and in some cases of uterus cancer. Signatures with unknown aetiology but that were suggested to relate to age in some studies [129], like SBS5 and SBS40, were present in all cancer types.

Principal Component Analysis (PCA) followed by hierarchical clustering on the principal components of the mutational signatures in the different tumour types showed in breast cancer an association between a group of donors with multiple mutations in PIK3CA and the presence of APOBEC mutational processes, while in uterus and colorectal cancer this association was to a mutational process related to Pol ϵ or defective DNA mismatch repair. The signature profile of the PIK3CA-mutated tumours split by single or multiple

mutations in p110 α (PIK3CA) protein is shown in **Figure 57**, **Figure 58** and **Figure 59**. The proportion of breast cancer donors that harbour multiple mutations in PIK3CA did not present a different mutational signature profile than those with a single mutation (**Figure 57**). We also observed clusters characterized by specific mutational processes enriched for a particular domain mutated (**Figure 60**). For the donors with PIK3CA mutations, the APOBEC signatures are related to the mutations in the helical domain (**Figure 60** – Cluster 2) while defective DNA mismatch repair and deregulated activity of Pol ϵ are more enriched for ABD mutations (**Figure 60** – Cluster 1 and 3). Mutational processes such as a defective DNA mismatch repair pathway and the ultrahypermutation due to the deregulated activity of Pol ϵ are uncommon in breast cancer, which might explain why there are so few ABD domain mutations in this type of cancer. Since both processes have a high number of mutations as consequence, the ABD mutations could be because of this. To confirm it, we tested for enrichment of PIK3CA mutations or specific PIK3CA domain mutations across groups of samples affected by different mutational processes that lead to a high number of somatic mutations. **Figure 61** shows these results with its corresponding odds ratio. No significant positive association was found between ABD mutations and other mutational processes that were involving a higher number of mutations, such as UV-light exposure, or an older age of the donor (**Figure 61**), so a high number of mutations seemed to not be always the explanation of the presence of ABD mutations.

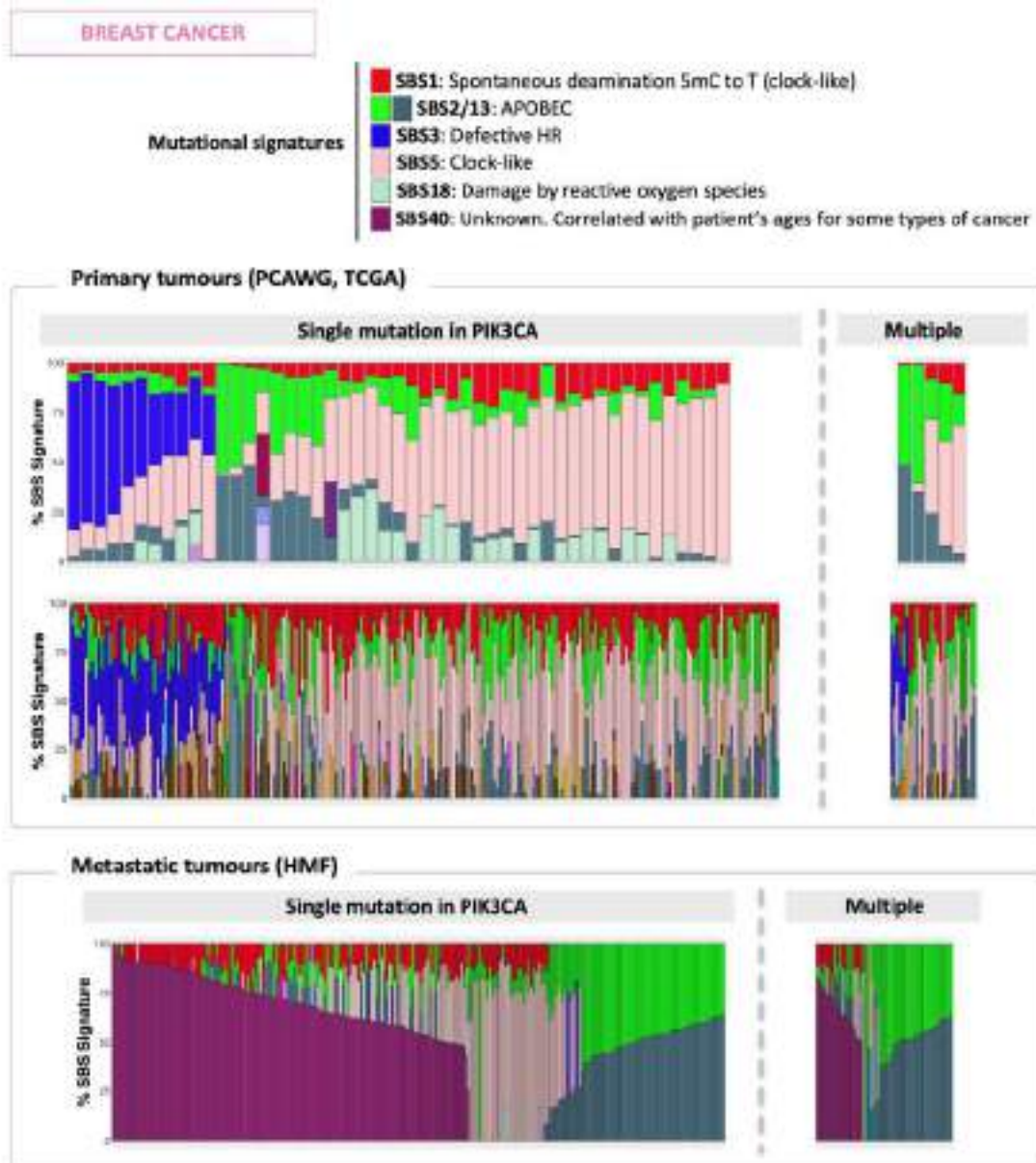


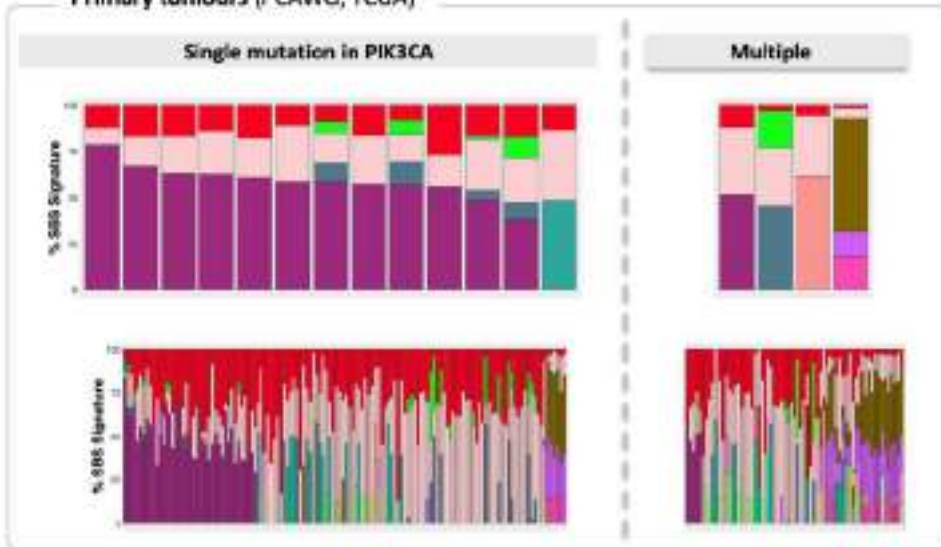
Figure 57. Mutational signatures PIK3CA-mutated breast tumours divided in single mutation vs. multiple mutations in p110 α (PIK3CA) protein. Each vertical bar in the plots refers to a tumour genome from a donor and the colours correspond to the SBS signatures found in their different proportions.

UTERUS CANCER

Mutational signatures

- SBS1: Spontaneous deamination 5mC to T (clock-like)
- SBS2, SBS13: APOBEC
- SBS3: Defective HR
- SBS5: Clock-like
- SBS10a/b: Pol α
- SBS28: Unknown but related to SBS10a/b
- SBS18: Damage by reactive oxygen species
- SBS40: Unknown. Correlated with patient's ages for some types of cancer
- SBS26, SBS44: Defective DNA MMR

Primary tumours (PCAWG, TCGA)



Metastatic tumours (HMF)



Figure 58. Mutational signatures PIK3CA-mutated uterus tumours divided in single mutation vs. multiple mutations in p110 α (PIK3CA) protein. Each vertical bar in the plots refers to a tumour genome from a donor and the colours correspond to the SBS signatures found in their different proportions.

COLORECTAL CANCER

Mutational signatures

- SBS1: Spontaneous deamination 5mC to T (clock-like)
- SBS2, SBS13: APOBEC
- SBS3: Defective HR
- SBS5: Clock-like
- SBS10a/b: Pol ε
- SBS28: Unknown but related to SBS10a/b
- SBS17b: Unknown
- SBS18: Damage by reactive oxygen species
- SBS40: Unknown. Correlated with patient's ages for some types of cancer
- SBS15, SBS26, SBS44: Defective DNA MMR

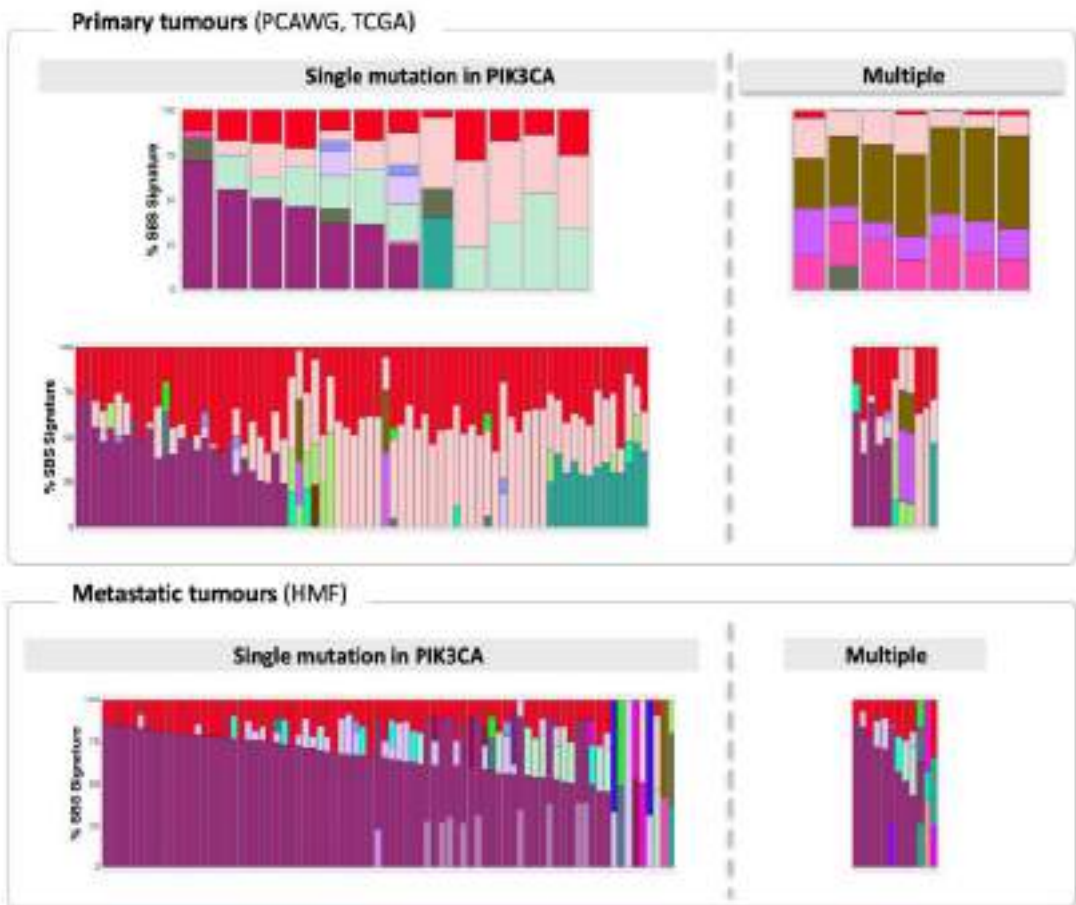


Figure 59. Mutational signatures PIK3CA-mutated colorectal tumours divided in single mutation vs. multiple mutations in p110 α (PIK3CA) protein. Each vertical bar in the plots refers to a tumour genome from a donor and the colours correspond to the SBS signatures found in their different proportions.

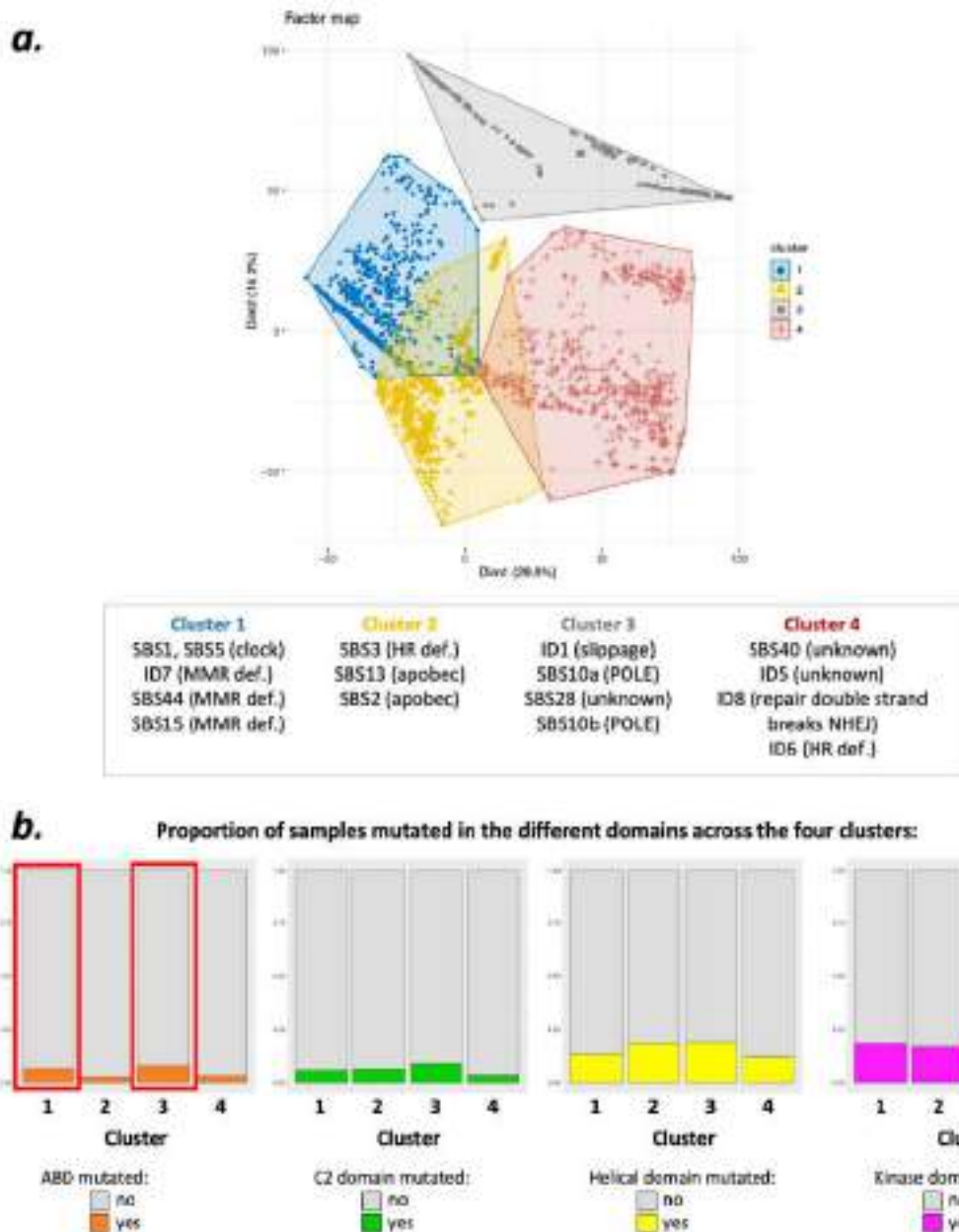


Figure 60. Principal Component Analysis (PCA) followed by hierarchical clustering of principal components of the mutational signatures describing breast and uterus cancer genomes from the PCAWG, TCGA and HMF dataset. a) PCA showing first vs. second dimension and main clusters found after hierarchical clustering of the principal components. Below, the mutational signatures that were associated to each of the four clusters are listed. ‘clock’: clock-like signature. ‘MMR def.’: DNA mismatch repair deficiency. ‘HR def.’: homologous recombination DNA damage repair deficiency. ‘slippage’: signature related to a potential slippage during DNA replication of the replicated DNA strand; substantial number of mutations of this signature are found in cancers with DNA mismatch repair deficiency [161]. ‘POLE’: Polymerase epsilon exonuclease domain mutations. ‘repair double strand breaks NHEJ’: signature that may involve repair of DNA double strand breaks by non-homologous DNA end-joining mechanisms [161]. **b)** For each of the four clusters, the proportion of cancer genomes harbouring a mutation in each of the different protein domains is indicated.

a.

		PIK3CA mutated	Not mutated	Significant (p-value < 0.05)	Odds Ratio
UV light exposure (SRS7a/b/c/d)	Yes	15	446	✓	0.2564311
	No	1,326	9,347		
POLE signature (SRS10a/b)	Yes	41	40	✓	8.326835
	No	1,200	9,751		
POLE exonuclease mutated	Yes	58	42	✓	12.61837
	No	1,177	9,751		
MSI	Yes	148	284	✓	6.363515
	No	1,093	9,589		
APOBEC (SRS1, SRS13)	Yes	256	578	✓	8.204659
	No	985	9,223		
Defective HR (E.g. BRCA1) SRS3	Yes	19	586	✗	-
	No	1,152	9,287		
Age >= 65 (n=10,843)	Yes	511	1,820	✓ (0.006711)	1.177717
	No	713	5,469		

b.

		AKD domain mutated	Not mutated	C2 type domain mutated	Not mutated	Helical domain mutated	Not mutated	Coiled domain mutated	Not mutated
UV light exposure (SRS7a/b/c/d)	Yes	✗		✗		5	456	7	456
	No	✗		✗		523	18,858	441	18,131
Fisher's Exact Test						✓ OR: 0.2337186		✓ (0.000) OR: 0.3534265	
POLE signature (SRS10a/b)	Yes	25	56	11	78	✗		24	57
	No	150	18,883	168	18,703	✗		425	18,528
Fisher's Exact Test		✓ OR: 52.08821		✓ OR: 10.5895		✗		✓ OR: 10.42395	
POLE exonuclease mutated	Yes	41	65	10	87	16	98	31	75
	No	134	18,794	152	18,778	512	18,416	418	18,518
Fisher's Exact Test		✓ OR: 56.86655		✓ OR: 25.46932		✓ OR: 5.61587		✓ OR: 10.38844	
MSI	Yes	52	388	28	324	43	389	61	291
	No	123	18,557	143	18,539	485	18,197	388	18,294
Fisher's Exact Test		✓ OR: 14.86634		✓ OR: 6.366534		✓ OR: 2.925334		✓ OR: 5.529625	
APOBEC (SRS1, SRS13)	Yes	✗		28	788	167	678	66	760
	No	✗		141	10,065	361	9,347	383	8,825
Fisher's Exact Test		✗		✓ OR: 2.469344		✓ OR: 6.810081		✓ OR: 2.227504	
Defective HR (E.g. BRCA1) SRS3	Yes	2	673	✗		✗		41	634
	No	173	10,180	✗		✗		408	8,951
Fisher's Exact Test		✓ (0.002) OR: 0.1745879		✗		✗		✓ (0.008) OR: 1.577162	
Age >= 65 (n=10,843)	Yes	✗		✗		225	3,515	✗	
	No	✗		✗		293	8,381	✗	
Fisher's Exact Test		✗		✗		✓ (0.000968) OR: 1.350195		✗	

Figure 61. Fisher's Exact test to test for associations of PIK3CA mutations with different mutational processes or conditions across cancer genomes from all cancer types in PCAWG and TCGA datasets. a) Associations of the different conditions with PIK3CA mutated or not mutated. **b)** Associations of the different conditions with the different PIK3CA domains mutated. The contingency tables used in the test are included together for significant results (✓ = significant). The non-significant comparisons are indicated with '✗' and the contingency table is not included). OR: Odds Ratio. OR equal or close to 1 indicates that there are no differences between the two conditions compared, the higher the OR the stronger is the association. The strongest associations are indicated with the OR in bold.

5.4.2.2. Potential cause of the different distribution of mutations

A possible factor that could drive differences, in terms of which domains are mutated in the different cancer types is epigenetics. We hypothesised that differences in epigenetics in the different tissues could result in differences in accessibility of the protein domains and, therefore, making mutations more likely in certain domains in one tissue than in the other.

First, we considered two epigenetic features, chromatin accessibility and methylation in healthy tissues. We searched for public data on chromatin accessibility such as ATACdb [162], CATlas [163] or EN-TEEx data portal (<https://www.encodeproject.org>) [164], but there was not sufficient data to have enough power to test our hypothesis in the tissues we were interested in. For methylation there was data available in the EN-TEEx project. We observed two positions differently methylated in breast (3 samples) and uterus (2 samples) (**Figure 62**). The positions that showed different values are 179,175,515 and 179,181,381 in chromosome 3 (GRCh38), which are located in the intron 1-2 (between the first and second exon) of the PIK3CA gene. However, based on this we cannot conclude if this could have an effect on the accessibility of any domain. Second, instead of normal tissue, we looked for differentially methylated probes in the samples of the PCAWG and TCGA datasets, for which we only had HM450K methylation arrays available (**Appendix 2**). The only probes related to PIK3CA in the HM450K methylation arrays were 5' upstream or in the intron between Exon1 and Exon2 of PIK3CA, so not in the gene body. Anyways, we did not find any of these probes differentially methylated in the different tissues depending on the protein domain mutated.

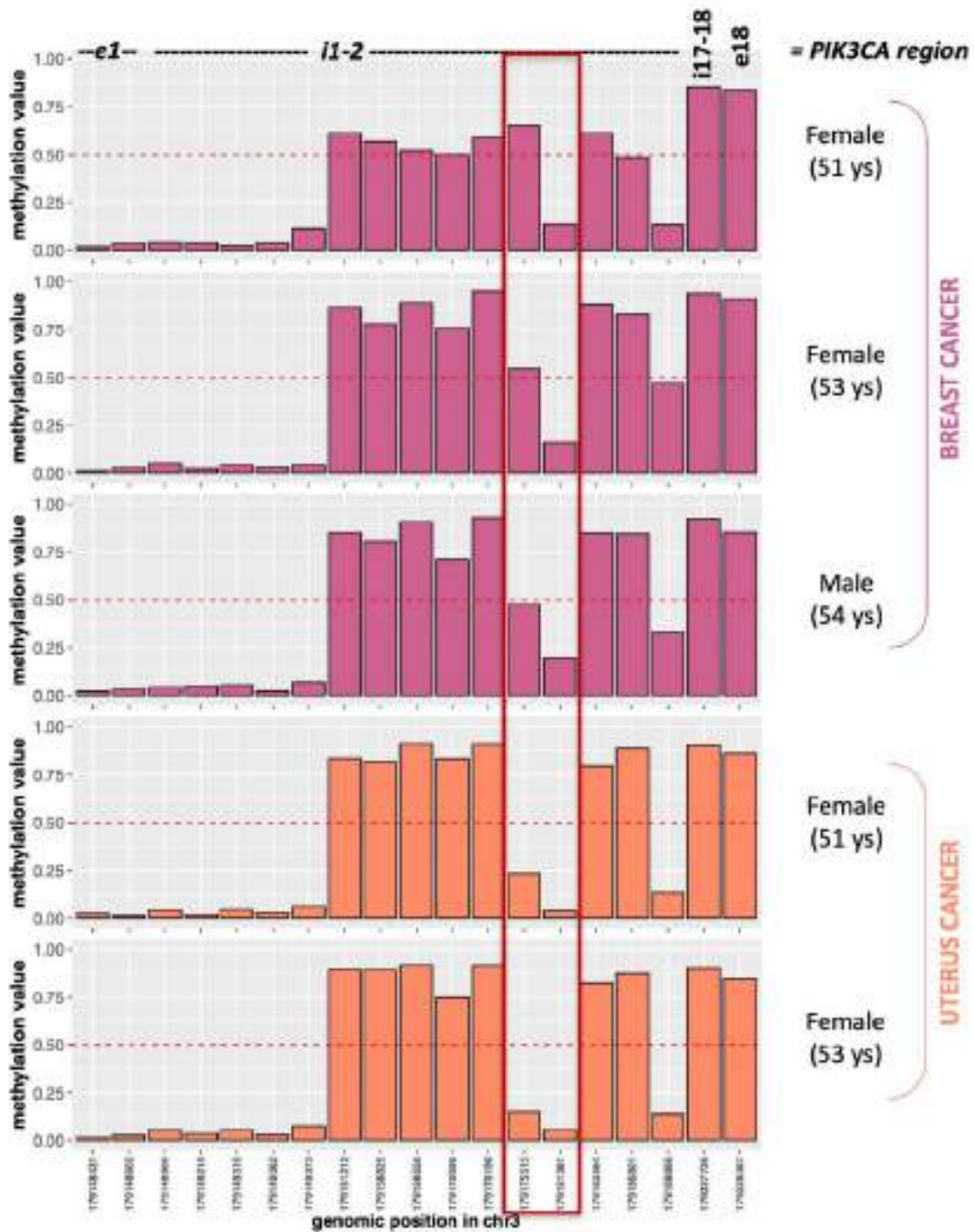


Figure 62. Methylation values of 19 positions in the PIK3CA gene in breast and uterus tissue. Annotation of the PIK3CA gene intron or exon in which these positions are located is indicated at the top of the plot. The red box indicates the two positions that seem to be less methylated in uterus compared to breast samples. Methylation data to do this plot was acquired from EN-TEX data portal.

5.4.3. Relationship between p110 α (PIK3CA) mutation and survival or other clinical parameters in cancer

The association of p110 α (PIK3CA) mutations with survival or other clinical features is controversial. Different associations can be found across different studies described in the literature. To our knowledge, most of studies evaluated differences between cancer genomes mutated vs. non-mutated in p110 α (PIK3CA). Only a few studies focused on differences depending on which p110 α (PIK3CA) domain is mutated, among them, most of the studies only consider the helical and kinase domain [165][166][167] and very few cases consider all domains [168]. We hypothesised that the disagreement in terms of associations might be due to the different mutations in the p110 α (PIK3CA) mutated tumour. First, we checked whether there were differences in survival depending on the p110 α (PIK3CA) mutation status, and also for mutations in specific domains in breast, uterus and colorectal cancer. Next, we tested for associations of clinical parameters, such as breast cancer subtype in breast cancer, tumour grade and stage of the tumour, with p110 α (PIK3CA) mutational status in general or with a specific domain mutated.

Survival analyses

Survival analysis was done in the TCGA, PCAWG and HMF dataset. In the TCGA dataset, we did see differences in uterus cancer (**Figure 63**). Censoring at both 5 and 15 years, survival in PIK3CA-mutated tumours in uterus cancer was significantly higher compared to non-mutated tumours (Cox proportional hazards (PH) regression at 5 years survival, HR<1, p-value=0.034 and Cox PH regression at 15 years survival, HR<1, p-value=0.04, respectively) (**Figure 63**). In breast and colorectal cancer, we did not find differences in survival between PIK3CA-mutated and non-mutated tumours. We also added age, ER and PR status (in the case of breast cancer) and sex (in the case of colon cancer) as variables to the different Cox regression models for testing survival in PIK3CA-mutated versus non-mutated tumours, to see if these variables could have an influence on the results, but still non-significant results were obtained in any of the cases. The same survival analyses in the PCAWG and HMF datasets did not show statistically significant results in any cancer type when comparing mutated and non-mutated tumours.

To evaluate survival depending on which p110 α (PIK3CA) domain was mutated, we performed a Cox PH regression including all mutated samples and as variable which domain was mutated and we also did independent regressions including two groups of tumours each time, *e.g.*, helical domain mutated versus kinase domain mutated tumours, kinase domain versus ABD mutated, etc. We did not see statistically significant differences in survival in any case in any dataset.

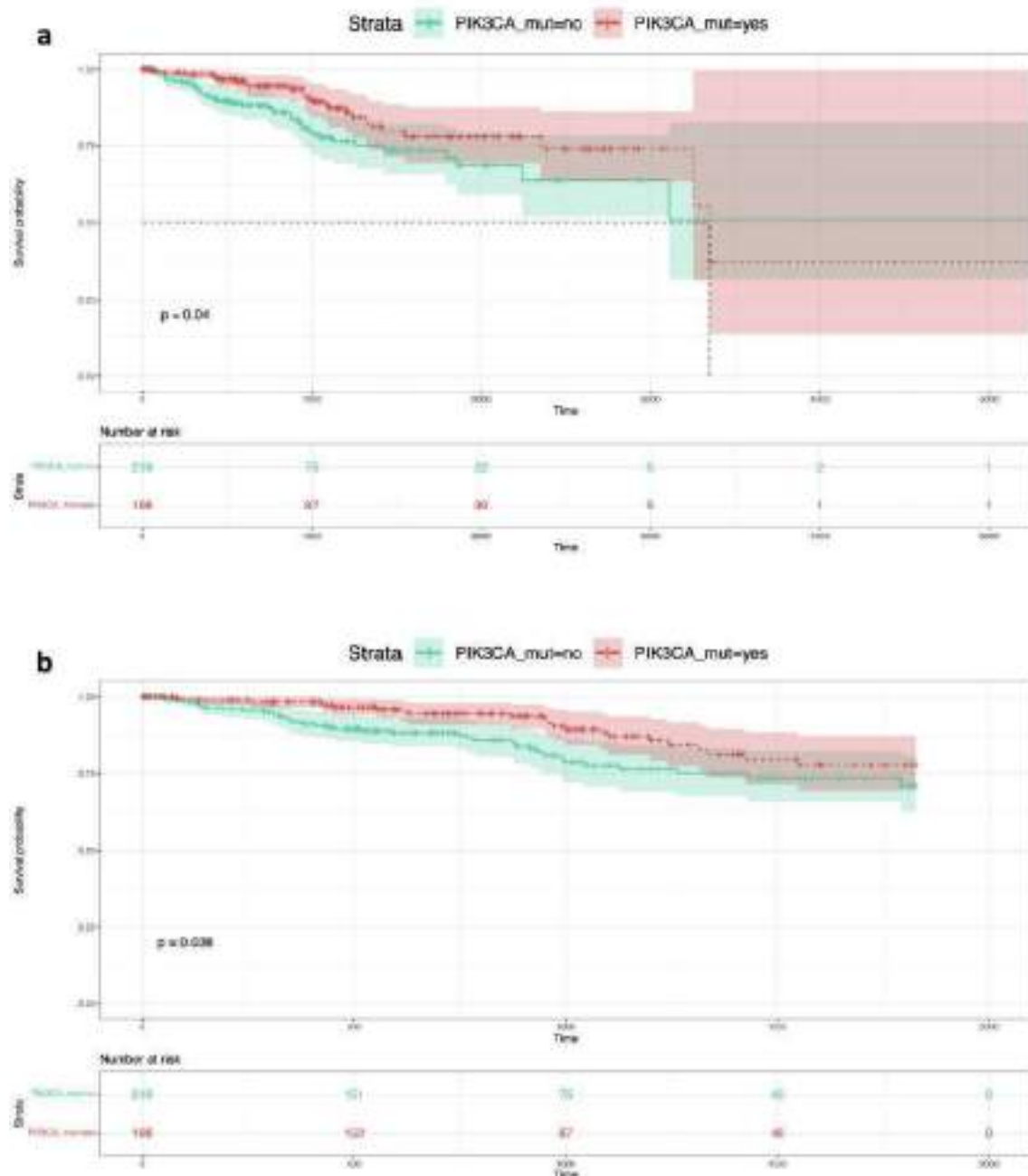


Figure 63. Kaplan-Meier curves for survival in PIK3CA-mutated versus non-mutated tumours in uterine cancer. (a) Survival analysis censored at 15 years and (b) censored at 5 years. The 'p' in each plot corresponds to the result of the long-rank test which coincides with the Cox regression results.

Associations of p110 α (PIK3CA) mutation with clinical parameters

The clinical parameters that were tested for relationship with p110 α (PIK3CA) mutation status were tumour grade, tumour stage and, in the case of breast cancer, ER, PR and HER2 status of the tumour. We also tested for relationships with the age of the patient or, in the case of colon cancer, the sex of the patient. In breast cancer, there were significantly more PIK3CA mutations in ER-positive tumours as well as HER2-negative, while no significant associations were found between PIK3CA mutations and tumour grade, tumour stage nor age of the patient. Association of p110 α (PIK3CA) mutation with the stage and grade of the tumour were also not significant in the case of uterus and colorectal cancer. No association with age was found neither in uterus nor in colorectal, and no association with sex was found in colorectal cancer.

We tested the same clinical parameters in the subset of PIK3CA-mutated tumours considering the p110 α (PIK3CA) domain mutated and we did not find significant associations.

5.4.4. Gene set enrichment analysis (GSEA) of p110 α mutated domains

We performed a differential expression analysis using DESeq2 (See *Methods*) between PIK3CA mutated and non-mutated tumours in breast and uterus cancer from the TCGA dataset. We identified 11,413 significantly differentially expressed (DE) genes (adjusted p-value<0.05) between breast tumours with PIK3CA mutated and non-mutated. From the total of DE genes, 3,049 genes (~27%) showed a higher expression in the PIK3CA mutated tumours with respect to the non-mutated. In the case of uterus cancer, 1,590 genes were significantly differentially expressed (adjusted p-value<0.05) between PIK3CA mutated and non-mutated tumours. From the total, 295 genes (~19%) showed higher expression in PIK3CA mutated tumours. For both breast and uterus tumours we also computed the differential expression between the different domains mutated as well as between tumours with a single mutation in PIK3CA and tumours with multiple mutations in this gene. We also used the same RNA-Seq data as input for GSEA tool with which the gene set enrichment analysis was performed. In breast cancer, samples with a kinase mutation showed significant enrichment of the 'PROTEIN SECRETION' gene set

compared to helical domain mutated samples. The protein secretion pathway is an essential molecular machinery for preparing and exporting proteins to the extracellular environment [169]. The kinase domain mutated breast cancer tumours could have an increase in this pathway and therefore be related to higher secretion of proteins. In uterus cancer, samples with C2 PI3K-type domain mutated, were enriched in 'DNA REPAIR' gene set compared to ABD, helical and kinase domain mutated samples. 'DNA REPAIR' gene set includes genes involved in DNA damage repair. The fact that it is enriched in C2 PI3K-type may mean there are more mistakes in these samples and the machinery is more expressed to get everything repaired. Deletions and insertions are frequent in this domain that might also increase the activation of the DNA repair pathways. Donors with multiple mutations in PIK3CA resulted in different gene sets enriched in breast and uterus cancer, except for three gene sets that were enriched in both cancer types: 'DNA REPAIR', 'MTORC1 SIGNALING' and 'UNFOLDED PROTEIN RESPONSE'. In the case of breast cancer seven gene sets in total were enriched in samples with multiple mutations in PIK3CA. 'FATTY ACID METABOLISM', 'CHOLESTEROL HOMEOSTASIS', 'PEROXISOME' and 'GLYCOLYSIS' were the gene sets only enriched in the case of breast cancer multiple mutated samples. In the case of uterus cancer, nine gene sets in total were enriched in samples with multiple mutations in PIK3CA. 'MYC TARGETS V1', 'OXIDATIVE PHOSPHORYLATION', 'PI3K AKT MTOR SIGNALING', 'MYC TARGETS V2', 'E2F TARGETS' and 'G2M CHECKPOINT' were the gene sets different to breast cancer.

5.4.5. Assessment of the tumour immune microenvironment in p110 α mutated and non-mutated tumours in breast cancer

Using the breast cancer cohort of the TCGA dataset, we aimed to determine the intra-tumoral immune landscape of primary breast tumours harbouring PIK3CA mutations and without mutations in this gene. Furthermore, in the subset of PIK3CA-mutated tumours, we analysed the differences of the tumour immune landscape depending on which of the domains of p110 α (PIK3CA) was mutated.

Exploiting a single-cell RNA-Seq data set of primary breast cancer [56], we performed a deconvolution of the TCGA bulk RNA-Seq data of 1,043 breast cancer donors using SPOTlight [85] (**Figure 64a**). We assessed the proportions of normal breast cells, tumour cells, stromal cells and immune cell populations inside the tumour samples. We excluded from this analysis samples that had multiple p110 α (PIK3CA) domains mutated, since they cannot be classified in just one group. Across all the bulk RNA-Seq samples, we were able to extract 16 different tumour, stromal and immune cell populations (**Figure 64b**) that were classified in three compartments: ‘cancer and normal breast cells’ (5 populations), ‘tumour stroma cells’ (4 populations) and ‘tumour immune cells’ (7 populations) (**Figure 64b-c**).

In the analysis of the stromal cell compartment, we observed that donors with mutated PIK3CA had a higher proportion of endothelial ($p=0.0013$) and endothelial lymphatic LYVE1 (Lymphatic vessel endothelial hyaluronan receptor 1) ($p=1.15\cdot 10^{-6}$) populations than the non-mutated ones. The two populations of Cancer Associated Fibroblasts (CAFs) assessed showed the same trend, being significantly higher in the PIK3CA mutated tumours (‘CAFs MSC iCAF-like’ $p=0.0025$; ‘CAF myCAF-like’ $p=3.49\cdot 10^{-7}$). Next, we assessed the main immune cell lineages divided into the innate immune cells (Dendritic Cells (DCs), macrophage and monocytes) and the adaptive immune cells (B cells, plasmablasts, T cells and NK, and cycling immune cells) in PIK3CA mutated and PIK3CA non-mutated tumours. We observed significant differences in proportions of three different immune cell populations: macrophages ($p=1.01\cdot 10^{-10}$), T cells & NK cells ($p=1.35\cdot 10^{-3}$) and cycling immune cells ($p=1.65\cdot 10^{-5}$) (**Figure 64c**).

Based on the significant differences observed in the proportion of immune cells, we investigated whether also immune gene signatures (**Table 8**) representing the molecular changes in different immune lineages could be altered depending on mutational status. We did the differential expression analyses of the genes contained in immune signatures related to several immune functions comparing PIK3CA-mutated and non-mutated tumours. The set of genes up-regulated in each immune signature in PIK3CA-mutated and non-mutated tumours were summarized in **Table 9** and the $\log_2(\text{foldchange})$ value and significance of all genes in all signatures are shown in **Figure 64d**. Global analyses of

genes sets up-regulated in PIK3CA-mutated revealed a distinctive immune gene profile related to innate immune responses, in particular, macrophages (SPP1) secreting cytokines (IL4, IL25, TGFB3, TGFB1, CXCL12) which are well known to induce immunosuppression in TME in breast cancer [57]. Also, genes linked with T helper (Th) 17 T cells (IKZF2, RORC, and CCR6) were significantly up-regulated. Higher infiltration of Th17 T cells was associated with breast cancer progression [170] (**Figure 64d**). In contrast, non-mutated tumours showed a gene profile highly related to intra-tumoral T cell cytotoxicity (GZMB, GNLY, IFNG and TNF), T cell infiltration (CXCL9, CXCL10, CXCL16 and CCL20) and T cell exhaustion (LAG3 and TOX).

Table 8. Gene signatures related to immune function: signature name, list of genes included in the signature and general description of the function in which they are involved in.

SIGNATURE NAME	List of genes	Description
Inhibitory receptors	C10orf54 (VISR), CD101, CD160, CD244, CTLA4, HAVCR2, LAG3, LAYN, PDCD1, TIGIT	Immune checkpoints that control T cell activation.
Transcription Factors (TFs)	BCL6, BTLA, CD200, EOMES, FOXP3, HIF1A, ID2, ID3, IKZF2, JAK1, JAK2, JAK3, NFKB1, PRDM1, RORC, SATB1, STAT3, TBX21, TCF7, TOX, TOX2, TYK2	TFs involved in differentiation processes of T cells and T cell exhaustion.
Effector/memory molecules	CD38, CD44, CD93, ENTPD1, FASLG, GNLY, GZMA, GZMB, GZMH, GZMK, ISG15, KLRB1, KLRG1, NKG7, PRF1, TNFSF10	Gene markers of effector and memory states in T cells.
Cytokines/Innate molecules	CEBPD, FCER1A, ICAM1, IFNG, IL10, IL1B, IL2, IL23A, IL25, IL4, IL6, IL6R, IL6ST, LIF, MAP3K8, MRC1, SEPP1, SIK1, SPP1, TGFB1, TGFB3, TLR2, TLR3, TNF	Cytokines and markers produced and expressed by innate cells (monocytes, macrophages and DCs).
Co-stimulatory molecules	CD27, CD28, ICOS, IL2RB, TNFRSF4, TNFRSF9, TNFSF14	Surface receptors that induce T cell activation.
Chemokines	CCL11, CCL19, CCL2, CCL20, CCL21, CCL22, CCL28, CCL5, CXCL10, CXCL12, CXCL13, CXCL16, CXCL2, CXCL9, IL8, XCL1, XCL2	Chemotactic cytokines (cell migration).
Chemokine receptors	CCR1, CCR2, CCR4, CCR5, CCR6, CCR7, CX3CR1, CXCR3, CXCR4, CXCR5, CXCR6	Chemotactic cytokines receptors.

Table 9. Significantly differentially expressed genes in each of the immune gene signatures between PIK3CA mutated and PIK3CA non-mutated tumours.

SIGNATURE NAME	Up-regulated in PIK3CA-mutated (MUT)	Up-regulated in PIK3CA non-mutated (WT)
Inhibitory receptors	LAYN (p<0.05)	LAG3 (p<0.05)
TFs	IKZF2, RORC, JAK2, NFKB1 (p<0.05)	TOX and SATB1 (p<0.05)
Effector/memory	TNFSF10, CD93 and ENTPD1 (p<0.05)	GZMB, CD38 and GNLY (p<0.05)
Cytokines/Innate	FCER1A, IL4, IL25, IL6ST, TLR3, SEPP1, CEBPD, TGFB3, TGFB1 (p<0.05)	IFNG, TNF, ICAM1 and MAP3K8 (p<0.05)
Co-stimulatory	No-up regulated genes	IL2RB and ICOS (p<0.05)
Chemokines	CXCL12, CCL11 and CCL22 (p<0.05)	CXCL9, CXCL10, CXCL16 and CCL20, (p<0.05).
Chemokine receptors	CX3CR1 and CCR6 (p<0.05)	CCR1 (p<0.05)

We did not find significant differences in the 11 different stromal and immune cell populations (Endothelial, Endothelial lymphatic LIVE1, CAFs MSC iCAF-like, CAFs myCAF-like, DC, Macrophage, Monocyte, B cell, Plasmablast, T & NK cells and Cycling) when we considered only the subset of PIK3CA mutated tumours and compared the groups of samples defined by which domain is mutated to each other (**Figure 65a**). One possible explanation is the heterogeneity inside of the immune populations, which may make it difficult to find differences. Therefore, in the two populations that we found significant differences between PIK3CA-mutated and non-mutated tumours (macrophages and, T cells and NK cells), we deconvoluted the PIK3CA-mutated tumours with the next level of annotation capturing the heterogeneity of macrophages, T cells and NK cells. We were able to identify five subpopulations of macrophages (**Table 6 Methods – Deconvolution 2** contains the name of the different subpopulations) (**Figure 65b**). In the case of T cells and NK cells, we split the population into CD8 and NK cytotoxic cells and CD4 T helper cells and did the deconvolution for each group. We were able to identify seven different subpopulations within the CD8 and NK cytotoxic cells and four subpopulations within the CD4 T helper cells (**Table 6 Methods – Deconvolutions 3 and 4** contains the names of the subpopulations) (**Figure 65c-d**). We observed differences in the proportion of different subpopulations of immune cells across the different p110 α (PIK3CA) domains mutated (**Figure 65e**). However, after multiple testing correction using Benjamini-Hochberg (BH), none of the differences were statistically significant. Due to this, to confirm the observed tendencies, we assessed the differentially expressed genes from the seven immune gene signatures previously described across the tumour with different domains mutated (**Figure 66, Figure 67, Figure 68**) to support the tendencies observed. Focusing on macrophages, we observed a tendency of a higher proportion of a subpopulation called macrophages APOE+ (Lipid associated macrophages 2 or LAM2 or LAM2:APOE+) in the linker ABD-RBD mutated tumours with respect to tumours mutated in the C2, helical and kinase domain. On the contrary, for the macrophages EGR1+ and macrophages FABP5+ (Lipid associated macrophages 1 or LAM1:FABP5+), we found a higher proportion in the C2, helical and kinase domain mutated tumours. The differential expression analysis showed an up-regulation of CEBPD ($p=0.09$) in linker ABD-RBD mutated tumours with respect to helical domain mutated ones. SPP1 ($p=0.003$) and CCL28 ($p=0.06$) were up-regulated in the helical domain mutated tumours

compared to the linker ABD-RBD mutated tumours. These genes defined the presence of a different profile of tumour associated macrophages in the linker ABD-RBD and the helical domain mutated tumours. Intriguingly, helical domain mutated tumours had a significantly higher expression of TNF compared to kinase domain mutated tumours ($p=0.02$). Also, the expression of IL6 was higher in helical compared to C2 domain mutated tumours ($p=0.075$). These results demonstrated a different profile in the subpopulations of macrophages depending on the domain that is mutated in the tumour. Within the CD8 and NK cytotoxic cell populations, we observed a tendency of a higher proportion of T cells CD8+ LAG3+ in the linker ABD-RBD mutated tumours with respect to the C2, helical and kinase domain mutated tumours. Other tendencies we observed are:

- Lower proportion of NK cell AREG+ population in the linker ABD-RBD mutated tumours with respect to helical domain mutated tumours.
- Higher proportion of T cell CD8+ GZMK+ population in the linker ABD-RBD mutated tumours with respect to the helical and kinase domain mutated ones, as well as a higher proportion in the helical mutated with respect to kinase mutated.
- Higher proportion of T cell CD8+ IFIT1 population in the C2 domain mutated with respect to linker ABD-RBD, helical and kinase domain mutated tumours.

We also assessed the differentially expressed immune gene signatures in the different protein domains. We identified LAG3 as a top marker with higher expression in the linker ABD-RBD mutated tumours compared to helical and kinase domain mutated ones ($p=0.09$) (**Figure 66**). These results showed that linker ABD-RBD mutated tumours have a higher degree of CD8+ T cell infiltration of exhausted cells, specifically, a population of T cells CD8+ expressing LAG3.

Finally, for T helper cell populations we saw a tendency of a higher proportion of T regulatory cells (T-regs_FOXP3 or Tregs) in the linker ABD-RBD domain mutated tumours with respect to tumours with a mutation in the C2, helical or kinase domain. We assessed the differential expression of the genes in our immune gene signatures. We observed a significantly higher expression of FOXP3 in the Transcription Factors (TFs)

signature in the linker ABD-RBD when compared to helical domain mutated tumours ($p=0.08$). Another gene supporting this higher proportion of Tregs in the linker ABD-RBD mutated tumours is a higher expression of IL23A ($p=0.0005$) compared to kinase mutated tumours. Although not significant, we also see a high log₂-fold change for the expression of TIGIT and CTLA4 in the linker ABD-RBD against both helical and kinase domain mutated tumours.

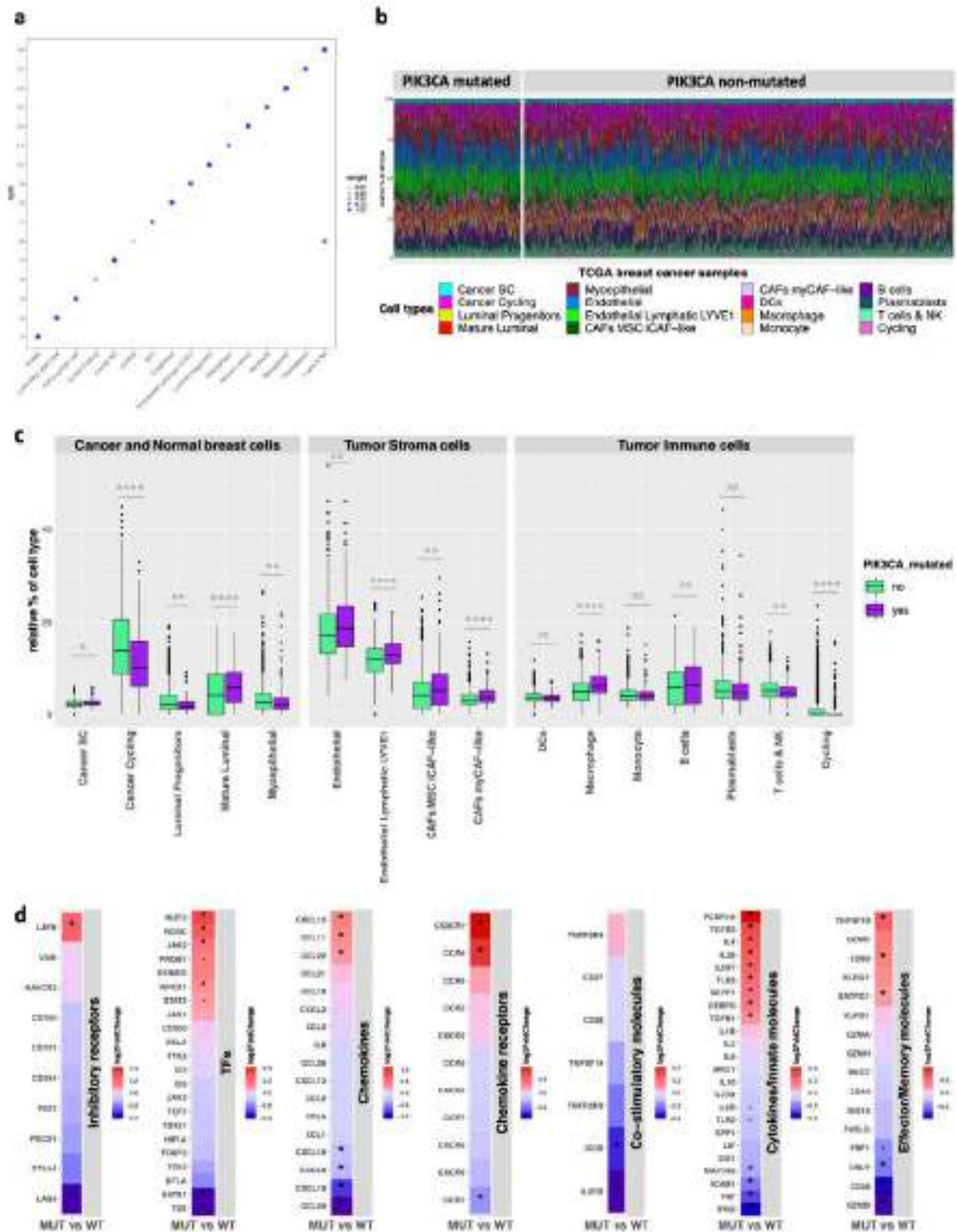


Figure 64. Deconvolution of TME in PIK3CA mutated and non-mutated breast cancer. **a.** Topic profiles from SPOTlight. **b.** Deconvolution of primary breast cancer tumours split by the mutational status of PIK3CA. Each line indicates the relative proportion of the different cell types found in each tumour. **c.** Comparison of the relative proportion of each cell type in mutated versus non-mutated tumours. Significance of the comparisons are indicated: *ns* (p-value>0.05), * (p-value<=0.05), ** (p-value<=0.01), *** (p-value<=0.001) and **** (p-value<=0.0001). **d.** Differential expression of genes included in the immune gene signatures. PIK3CA mutated tumours versus non-mutated.

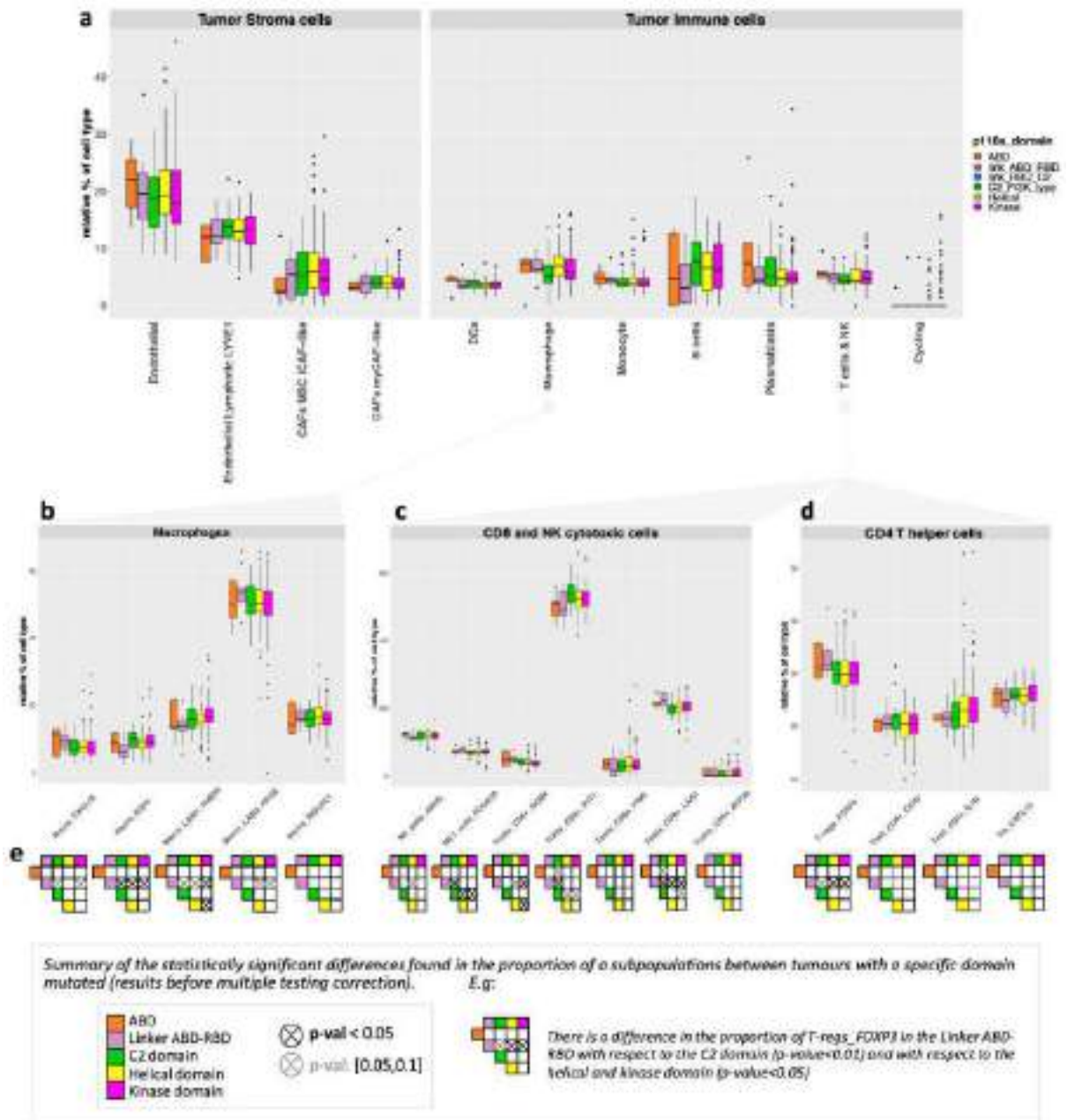
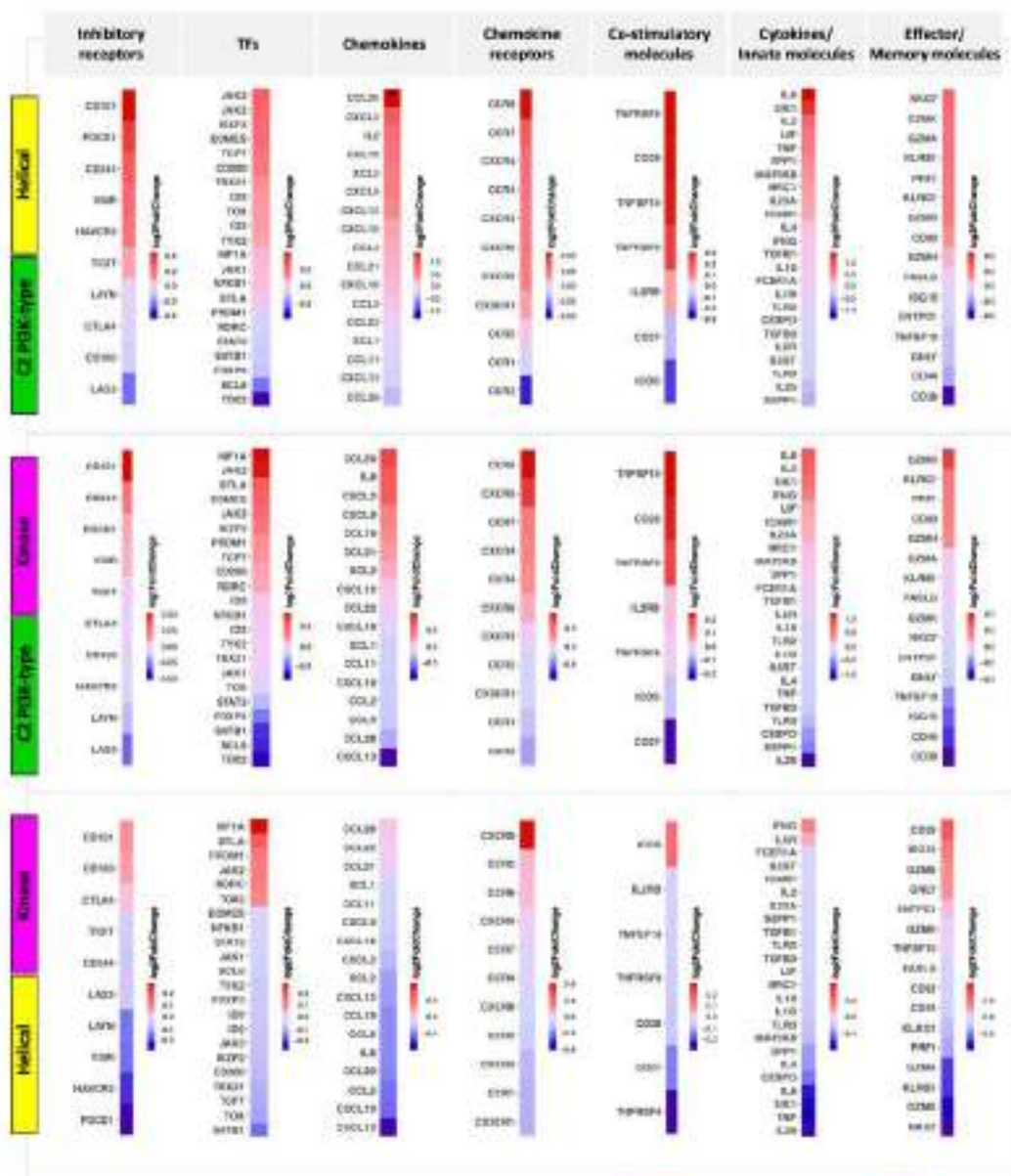


Figure 65. Deconvolution of TME in PIK3CA mutated tumours depending on the domain mutated.

- Proportions of tumour stroma and immune cells obtained after the deconvolution.
- Proportions of the different subtypes of Macrophages obtained after the deconvolution.
- Proportions of the different subtypes of CD8 and NK cytotoxic cells obtained after the deconvolution.
- Proportions of the different subtypes of CD4 T helper cells obtained after the deconvolution.
- Statistical results for the pairwise comparisons across the different domains within cell subtype. P-values lower than 0.1 before multiple testing correction are indicated with crosses (**black cross** indicates a p-value<0.05, **grey cross** indicates 0.05 < p-value < 0.1).



[+] value = Higher expression in group 1 (above)
 [-] value = Higher expression in group 2 (below)

Figure 67. Differential expression of immune gene signatures: all possible pairwise comparisons of C2, helical and kinase domain mutated tumours. The immune gene signatures are indicated at the top of the figure and are seven: ‘inhibitory receptors’, ‘transcription factors’ (TFs), ‘chemokines’, ‘chemokine receptors’, ‘co-stimulatory molecules’, ‘cytokines/innate molecules’ and ‘effector/memory molecules’. To the left of each box are the groups included in the differential expression analysis, group 1 at the top and group 2 below. Colour from red (higher expressed in group1) to blue (higher expressed in group 2) shows the log2-foldchange resulting from the differential expression analysis of each comparison. Asterisk (*) indicates a p-value <0.05. Dot (.) indicates a p-value<0.1.

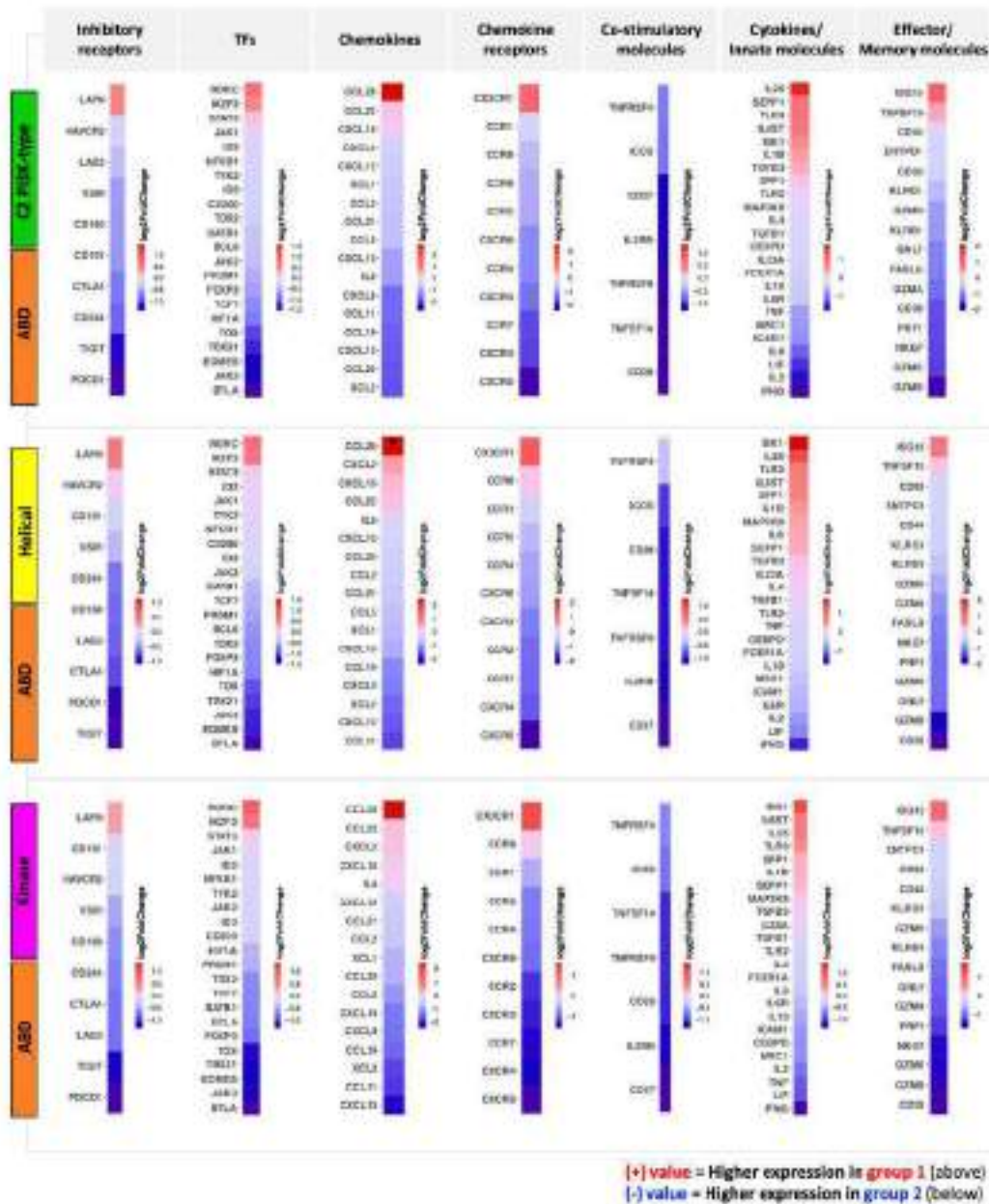


Figure 68. Differential expression of immune gene signatures: ABD domain mutated tumours against all other domains. The immune gene signatures are indicated at the top of the figure and are seven: ‘inhibitory receptors’, ‘transcription factors’ (TFs), ‘chemokines’, ‘chemokine receptors’, ‘co-stimulatory molecules’, ‘cytokines/innate molecules’ and ‘effector/memory molecules’. To the left of each box are the groups included in the differential expression analysis, group 1 at the top and group 2 below. Colour from red (higher expressed in group1) to blue (higher expressed in group 2) shows the log2-foldchange resulting from the differential expression analysis of each comparison. Asterisk (*) indicates a p-value <0.05. Dot (.) indicates a p-value <0.1.

Take-home messages Chapter 4

- The most frequent mutation type in PIK3CA gene was missense mutation (>95%). There were 385 unique amino acid changes resulting from these missense mutations.
- Uterus, breast and colorectal cancer were the most frequently PIK3CA-mutated cancer types.
- There was a different distribution of mutations along the p110 α (PIK3CA) protein domains in these three mutated cancer types. Breast cancer had a higher proportion of mutations in the kinase domain, colorectal cancer in the helical domain and uterus in the ABD domain and linker ABD-RBD.
- PIK3CA mutations in the ABD domain in uterus could be related to mutational processes such as defective DNA damage repair or hypermutation activity of Pol ϵ . The hotspot mutation in the helical domain could be associated with the activity of APOBEC family of cytidine deaminases.
- No significant positive association was found between ABD mutations and other mutational processes that were involving a higher number of mutations, such as UV-light exposure, or an older age of the donor.
- In uterus cancer from the TCGA dataset, there was a significantly higher survival in the PIK3CA-mutated compared with non-mutated tumours. There were no differences in survival in breast and colorectal cancer. In the PCAWG and HMF datasets no differences were observed in any of the cancer types analysed. At protein domain level in all datasets, no differences in survival were observed in any cancer type.
- In breast cancer there were significantly more PIK3CA mutations were in ER-positive tumours in breast cancer. No associations were found between the

PIK3CA mutational status and tumour grade, tumour stage nor age of the patient.

- The tumour microenvironment of breast cancer between PIK3CA mutated and non-mutated tumours showed different stromal composition as well as differences in the immune populations analysed.
- The tumour microenvironment of breast PIK3CA-mutated tumours showed a significantly higher proportion of stromal cells and macrophages, and lower proportion of T and NK cells compared with breast tumours without PIK3CA mutation.
- The analysis of the proportion of subpopulations of macrophages, T and NK cells in breast PIK3CA-mutated tumours showed different tendencies depending on which protein domain was mutated.
- In the breast tumours with the linker ABD-RBD mutated, we identified an exhausted profile in T cells, characterized by a significantly higher expression of LAG3.

6. DISCUSSION

A joint dataset of whole genome, whole exome and panel sequencing data from primary and metastatic tumours that summed up a total of 25,499 cancer genomes across over 40 cancer types was studied. This dataset consists of four cohorts: the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset, the Hartwig Medical Foundation (HMF) dataset, The Cancer Genome Atlas (TCGA) dataset and the Breast-Cancer STratification study (B-CAST) dataset. The massive increase of sequencing data and its public availability enables the extensive study of the variation in cancer to expand our understanding [171][172]. One caveat, however, is that there is currently no gold standard for calling mutations and different mutation calling pipelines are being used. This makes joining different datasets a challenge [173]. However, it goes beyond the scope of our project to recall the mutations for 25,499 cancer genomes, which would include downloading massive amounts of data. As our focus has largely been on substitutions, the differences between pipelines are much smaller compared to indels [174][171]. The vocabulary used by the different datasets also poses a challenge. Especially when within a cohort there is no standard used either. This can lead to typographical errors, different words used to refer to the same concept, or the same word used for different concepts. Adhering to standards like the one proposed by Musen *et al.* [175] will therefore be essential to reduce the overhead needed to homogenize the metadata. Another challenge is that the more datasets are combined and the more donors that participate the probability increases that the same donor is part of multiple cohorts. As it goes against privacy standards to explicitly look for this overlap, a mechanism will need to be put in place to be able to identify these cases.

The pan-cancer landscape of somatic mutations at the level of substitutions, insertions and deletions was described with a view to identifying pertinent features. This shows that in PCAWG, TCGA and HMF the most mutated cancer type considering SSMs is skin cancer. Considering SIMs, lung cancer is the highest mutated in primary tumours and oesophagus cancer in metastatic tumours. The distribution of the six SSM subtypes characterizes individual cancer types and is consistent between primary and metastatic tumours. For example, C>A mutations are the highest in lung, C>G in bladder and C>T in

skin cancer, which is consistent across the individual datasets. Based on the somatic mutations, an assessment of the landscape of mutational signatures in primary and metastatic tumours was also done. Within a specific tumour type these signatures can be used to stratify donors into distinct groups, which could be highly informative for selecting the best treatment. Focused on breast, colorectal and uterus cancer, that summed up to 3,601 donors with WGS or WES data, interesting results arose. Across these three cancer types, there are groups of cancer genomes showing mutational signature profiles in common, as well as different ones between primary and metastatic tumours. For example, in breast cancer, using the dominant signature, three groups are identified across 1,903 cancer genomes: SBS3 or defective homologous recombination DNA damage repair; SBS2/13 or APOBEC activity and SBS5 clock-like signature. This division of breast cancer tumours according to the different mutational signatures has already been reported by Denkert *et al.* [176] in their study of 405 patients. The same finding in our analysis across four times this number of patients can confirm this observation. In metastatic breast tumours, the group of SBS3 or defective homologous recombination DNA damage repair was not seen. Primary tumours from uterus and colorectal cancer also showed a group of donors with the SBS5 clock-like signature as dominating, like it was found in breast cancer, while they had other groups characterized by SBS10 or Pol ϵ hypermutation activity, SBS44 or defective DNA mismatch repair and, a last group characterized by SBS40 clock-like signature. The hypermutation activity of Pol ϵ has been reported in uterus and colorectal cancer and has been related to a good prognosis and favourable responses to immunotherapies [177].

To show the relevance of studying the genomic landscape of tumours, the PCAWG dataset was used as use case. Mutations found in the DNA of a tumour are expected to be largely unique to each tumour as there are three billion places in the DNA that can be mutated. However, despite these odds, across the cancer genomes of 2,583 participants available in PCAWG covering 37 tumour types, a total of over a million non-unique mutations were observed. The analysis of the genomic landscape of the PCAWG dataset, based on 42 features either based on all or only the recurrent mutations, shows how this can be used to stratify cancer genomes into clinically relevant groups. The division into 16 clusters and their characteristics could be valuable for complementing

current classification schemes, which are mainly based on histology and organ of origin. We can assign a new sample to one of our 16 clusters by first projecting it onto the PCA space based on the PCAWG cohort. Next, we use the first 18 principal components to compute the Euclidean distance to the centroid of each of the 16 clusters and assign the sample to the nearest one. If there are multiple clusters with a minimum difference in distance to the new sample, then to select one cluster we use the sequence motifs (**Figure 37**) and various layers of annotation (Appendix 1 - S3 Text) like replication time. Ultimately, whole-genome sequencing should be able to replace multiple diagnostic tests currently in use and make diagnoses more accurate. One example illustrating the value of the clusters found towards this goal is the MSI phenotype linked to one of the clusters. For these patients, immunotherapy may be beneficial [178] while adjuvant chemotherapy may not be needed [179]. A second example of an actionable phenotype that we capture with one of our clusters is ultra-hypermutation (cluster H), which has also been related to beneficial results from immunotherapy [180][181]. A third example is the somatic hypermutation of the immunoglobulin genes, which identifies memory B-cells as the cell of origin in the case of lymphomas. This has been linked to a less aggressive form of Lymph-CLL and more favourable prognosis [144], which may in turn influence treatment selection. Without explicitly analysing the immunoglobulin genes [182], we were largely able to separate the Lymph-CLL samples with somatic hypermutation (cluster M) from those without (cluster D). The characteristics of the former group include a high percentage of recurrent C>G SSMs and 1 bp A/T deletions. A final example relates to those Eso-AdenoCA samples that are assigned to cluster L, which have a high percentage of T>C as well as T>G SSMs and a higher total mutational load than Eso-AdenoCA samples not assigned to this cluster. Eso-AdenoCA samples with the characteristics of cluster L have also been suggested to benefit from immunotherapy [183]. The same treatment option may therefore be prioritized for the 22 Stomach-AdenoCA samples that are also in cluster L. Similarly, a refined investigation of tumour samples that do not cluster with the vast majority of its own kind may ideally point to differences in disease prognosis or treatment response and even has the potential to define novel subtypes or reveal misclassification. Such an analysis would be especially worthwhile for the ~20% or less samples from Kidney-RCC, Liver-HCC, Lung-SCC or Lymph-BNHL that are not assigned to the main cluster. Another possible application of

this classification scheme is to assign a metastatic sample with unknown primary site to a cluster to shed light on the possible tissue of origin or pan-cancer characteristics like MSI.

To go beyond the genomic landscape, the amino acid changes resulting from the somatic mutations were assessed by computing eight protein features, obtained from a combination of amino acid, evolutionary and structural properties. This was used to have an overview of the profiles observed within breast cancer. The most frequent amino acid change in this tumour type in each of the four datasets is glutamic acid to lysine (E>K), followed by glutamic acid to glutamine (E>Q) in PCAWG, HMF and TCGA dataset, but not in B-CAST. In B-CAST the second most frequent amino acid change is histidine to arginine (H>R). The difference is due to the fact that a subset of 323 genes was sequenced in contrast to all genes in the other three datasets. The high frequency of the H>R amino acid change is largely explained by a frequent hotspot mutation in the luminal A subtype, which constitutes nearly 60% of the B-CAST dataset. B-CAST behaves also different to the other datasets in that around 50% percent of the amino acids mutated are exposed and the other 50% buried in the structure, while in PCAWG, HMF and TCGA almost 60% of the amino acids mutated are exposed. Coinciding across the four datasets, over 50% of the amino acid changes are happening to an amino acid located in a loop in the protein structure and the protein domains more mutated in were protein kinases followed by cadherin and Ig-Fibronectin Type III. After dimensionality reduction followed by clustering based on the eight protein features no well-defined clusters were found. Also, after the annotation of the mutations that are known as drivers in the original data, there was not any clear pattern or association of the different drivers with specific features. Therefore, these protein features did not help to uncover groups of mutations sharing characteristics that could be associated to known drivers. As a limitation, some amino acid changes could not be analysed because the limited availability of protein structures. Even if a protein structure is available for a protein, it is often not complete and therefore the exact amino acid of interest is missing in the structure. Moreover, depending on the feature that is measured, there is the need of a protein structure of high quality, which further limits the number of amino acid changes that can be analysed. This is the case for the input of FoldX, which requires a resolution

of less than 2Å to ensure that the software models the amino acid change and does the computation of the change of the free energy of folding correctly. The recent availability of AlphaFold [36] predicted structures could increase the number of amino acid changes that could be evaluated at structure level. Therefore, the incorporation of these predicted structures might increase substantially the data to get insights from this type of analysis.

One of the most mutated genes with missense mutations across 11,159 breast cancer patients is PIK3CA. Investigating if this gene is also frequently mutated in other tumour types than breast cancer across the 25,499 cancer genomes, uncovered its presence mutated in several tumour types, in particular in colorectal and uterus cancer, as it has been previously observed [184][185]. Focus on the assessment of the eight protein features for the protein changes in the p110 α protein, encoded by PIK3CA, elucidated differences in the proportion of mutations across the different protein domains in breast, colorectal and uterus cancer. Deciphering the underlying causes of the different distribution of mutations across protein domains could provide information on the different mechanisms affected in different cancer types. We investigate potential underlying causes of the different mutations and relate mutational processes such as hypermutation activity of Pol ϵ or defective DNA damage repair in uterus cancer to mutations in the ABD domain. The lack of available data to investigate other potential causes, such as epigenetics, did not allow us to establish a mechanism leading to differentially mutated domains.

The survival analysis in uterus cancer in the TCGA dataset shows a higher survival rate in patients with PIK3CA mutated tumour compared with patients with tumours without a PIK3CA mutation, in line with previous reports [186]. This result could help to predict the prognosis of this group of patients and the likely course of the disease [186]. When this analysis is extended to the different mutated domains, there is no significant differences in survival in any of the cancer types. The same survival analyses on PCAWG and HMF datasets did not show statistically significant results in any cancer type when comparing mutated and non-mutated tumours nor the comparisons between the different domains mutated. Our data was not sufficient powered to demonstrate other

associations. In the case of PCAWG dataset, this is likely due to the small sample size, which has an even bigger impact when splitting according to the p110 α domain mutated. In the case of HMF dataset, there were few samples with survival data in the case of uterus cancer. Partly this might be because the HMF is a relative new dataset as the foundation behind it started in 2015 (<https://www.hartwigmedicalfoundation.nl>). The data they provide regarding survival comes from trials or studies that are still running, such as CPCT (ClinicalTrials.gov Identifier: NCT01855477) with an estimated end date next year (2023) or DRUP (ClinicalTrials.gov Identifier: NCT02925234) with an estimated end date in 2027. This means that a follow up of 5 or more years is not available yet for all donors. Moreover, not all studies within the HMF dataset gave permission for survival data to be shared.

Breast cancer is the most diagnosed cancer type in the world [1], and it is characterized by high heterogeneity [60], which makes stratifying patients even more essential. A better stratification of patients is important to customize the treatment strategy and to improve the management of this disease [187]. In the attempt to associate PIK3CA mutations to different clinical features, such as hormone receptor status, tumour grade, tumour stage and age, a clear association was found between the ER-positive status and having a PIK3CA mutation, as has been previously reported [188]. No other significant results were found between PIK3CA mutation and the rest of clinical parameter analysed in the different cancer types. Knowledge of the intra-tumoral heterogeneity in breast cancer is also important since it facilitates immune evasion, clonal survival and therapy resistance [55]. Immunotherapy is an emerging therapy with promising results, lower toxicity than other strategies and high accuracy [189][190] that can be applied in some cases depending on the cellular composition of the tumour. For the TNBC subtype it is known that there is immune cell infiltration and the use of immunotherapy has already been incorporated in the clinic in metastatic cases of this breast cancer subtype [191]. For other breast cancer subtypes such as HER2-positive or HR-positive the knowledge of the immune component of the tumour microenvironment is limited [192]. The presence of tumour-infiltrating lymphocytes (TILs) in HER2-positive has been suggested to be linked to a favourable prognosis, while its significance in breast ER-positive tumours remains uncertain [84]. Tumour-infiltrating lymphocyte composition, organization and

PD-1/ PD-L1 expression are linked in breast cancer [193]. PD-1/PD-L1 axis is one of the mechanisms by which tumour cells evade the cytotoxic immune response [194]. PIK3CA mutations has been already associated with PD-L1 expression in other cancer types such as cervical cancer, suggesting the potential use of PD-L1 inhibitors to fight these tumours [195]. With a focus on finding the potential association of PIK3CA mutations and the tumour microenvironment (TME), the intra-tumoral heterogeneity across breast cancer donors from the TCGA dataset was analysed. The cellular composition of the TME in breast cancer is significantly different in breast tumours with a PIK3CA mutation compared to those without. PIK3CA mutated tumours with a significantly higher proportion of stromal cells and macrophages, and lower proportion of T and NK cells compared to non-mutated breast tumours suggests that different immunotherapy strategies could be applied [58][196][197].

All stroma cell populations, endothelial cells and CAFs, are in a significantly higher proportion in PIK3CA mutated tumours compared to not mutated. It has been suggested that signals from the microenvironment control CAF differentiation or migration [198]. CAFs are critical for cancer occurrence and progression because of their versatile roles in extracellular matrix remodelling, blood vessel formation, immune response, and, in turn, promotion of cancer cell proliferation, migration and invasion [198]. Indeed, it has been reported that CAFs lead to reprogramming of blood monocytes towards immune suppressive lipid associated macrophages (LAMs), which inhibit T-cell activation and proliferation [199].

PIK3CA mutated tumours showed a significantly higher proportion of macrophages and lower proportion of T and NK cell. These populations were investigated at a more detailed level of annotation to be more precise in the changes observed in subpopulations, also considering the stratification of tumours according to the p110 α (PIK3CA) domain that was mutated. For the analysis of the tumour immune microenvironment (TIME) in PIK3CA-mutated breast tumours at the level of which p110 α (PIK3CA) domain was mutated, some tendencies are observed but no significant results are seen after multiple testing correction using Benjamini-Hochberg. This could be due to the small sample size, so the expression of different immune signature was

used to investigate the tendencies observed. Tumours mutated in the linker ABD-RBD seemed to have a different profile compared with tumours with mutations in the C2 PI3K-type, helical and kinase domain, which show a more similar profile among them.

In the TIME of tumours mutated in C2, helical and kinase domain, the macrophages EGR1+ and macrophages FABP5+ (Lipid associated macrophages 1 or LAM1:FABP5+) are the ones in higher proportion. FABP is reported as a functional marker of pro-tumour macrophages [200]. Tumour associated macrophages (TAMs) are increasingly recognized as major contributors to the metastatic progression of breast cancer and enriched levels of TAMs often correlate with poor prognosis [57]. Survival analysis using the METABRIC40 cohort showed that the LAM1:FABP5 signature correlates with worse survival [56].

The TIME of tumours mutated in the link ABD-RBD is characterized by three main aspects. First, a higher proportion of the subtype of macrophages APOE+ (Lipid associated macrophages 2 or LAM2 or LAM2:APOE+), which is characterized by the expression of APOE. This subpopulation has been associated with immunosuppression in breast cancer as well as in other cancer types [201][199]. Second, a higher proportion of a subpopulation of exhausted T cells characterized by a significantly higher expression of lymphocyte-activation gene 3 (LAG3). This has been described as sign of exhaustion together with the expression of T-cell immunoglobulin and mucin-domain containing 3 (TIM3) and cytotoxic T lymphocyte-associated antigen 4 (CTLA4) [82]. Exhausted CD8+ T cells express these inhibitory receptors contributing to resistance in anti-PD1 treatment [82]. The identification of this kind of profile is interesting because there are therapies under development to avoid the exhaustion of T cells expressing this marker [202][203]. A third aspect that characterizes these link ABD-RBD mutated tumours is a higher proportion of T regulatory cells (T-regs_FOXP3 or Tregs), which enhances the suppression of the anti-tumour immunity. In summary, this shows an environment where tumour cells block successfully the immune system that has been related to a poor prognosis [204][82].

The integration of 25,499 cancer genomes from four datasets enabled us to create a pan-cancer landscape of somatic mutations that gave insights into the mutational burden, *i.e.* substitutions and insertions/deletions, as well as the distribution of mutation types across cancer genomes from different cancer types. This joint dataset allowed us to have a large sample size to study mutational processes, using mutational signatures as a proxy, and find groups of patients defined by different mutational signatures. Using PCAWG as a use case we showed the relevance of studying the genomic landscape of tumours. The study of 42 genomic features computed based on all somatic mutations and only the recurrent ones, divided 2,583 patients covering 37 cancer types into 16 clusters that can be linked to several actionable clinical phenotypes. New samples could be assigned to one of the defined clusters and the accuracy of the diagnosis could be increased in some cases such as with the identification of MSI or ultra-hypermutation, in which case patients might benefit from immunotherapy. This could also help to the development of a generic and personalized cancer diagnostic test that only uses the mutations found in the tumour. At protein level, when studying eight protein features for each of the 159,294 amino acid changes resulting from the somatic mutations, we could not uncover well-defined groups within breast cancer nor did these features characterize drivers. However, when we focused on the p110 α protein encoded by PIK3CA, we could show that breast cancer had a higher proportion of mutations in the kinase domain of this protein, colorectal cancer in the helical domain and uterus in the ABD domain and linker ABD-RBD. The enrichment of ABD domain and linker ABD-RBD mutations in uterus and colorectal cancer could be related to defective DNA damage repair or hypermutation activity of Pol ϵ . Focused on breast cancer, our results showed different tumour immune microenvironments in tumours with different PIK3CA mutated domains. Particularly, we uncovered that tumours mutated in the linker ABD-RBD have an exhausted T cell population characterized by the expression of LAG3. Tumours with mutations in the C2 PI3K-type, helical and kinase domain we found to be enriched in myeloid populations with a gene profile similar to immunosuppressive macrophages [56]. It is known that PIK3CA, as a oncogene, can promote tumourigenesis by providing tumour cells the advantage to avoid the antitumoral response by the immune system [205]. Our analysis suggests that different p110 α (PIK3CA) domains mutated might be related to promote tumourigenesis by different immune escape

mechanisms. Identification of these mechanisms can improve the selection of the optimal combination strategy to increase the efficacy of immunotherapy. In conclusion, our analysis shows that knowledge at genomic level, such as number of mutations, recurrence and mutational processes, as well as at protein level, such as differences in amino acid mutations, together with the study of the tumour microenvironment, provide new insights into cancer mechanisms. Our results contribute to stratifying patients in biologically relevant groups and thereby help personalise treatment strategies.

7. CONCLUSIONS

- The integration of 25,499 cancer genomes from four datasets enable us to create a pan-cancer landscape of somatic mutations that gave insights into the total number of mutations, *i.e.* substitutions and insertions/deletions, as well as mutation types across cancer genomes.
- Using PCAWG as a use case of our joint dataset, the study of different features computed based on all mutations and only the recurrent ones, enable to delineate various mutational processes, uncover new mutational manifestations and characterize several actionable clinical phenotypes in a novel way.
- From our joint dataset, we translate somatic mutations into their corresponding amino acid changes and characterize them by eight protein features. Focused on breast cancer we identify that most of the amino acid changes happen between amino acids of the same category. However, considering only the recurrent amino acid changes (mutations that were found in more than one patient in the same dataset) the proportion of 'change of charge' cases increase. Most of the mutated amino acids are in the secondary structure that is a loop and the amino acid is exposed. Around 33% of the mutations are affecting a functional site in the PCAWG, TCGA and HMF datasets, while only around 13% in B-CAST.

- The dimensionality reduction followed by clustering of amino acid mutations in breast cancer characterized by eight protein features did not group mutations that could be associated with being a driver.
- The exploration of amino acid changes in PIK3CA shows a different distribution of disease relevant mutated domains across cancer types.
- Underlying causes of the different distribution of mutations across domains can be different mutational processes. The case of a higher proportion of mutations in ABD domain in uterus cancer seems to be linked to the deregulated activation of Pol ϵ or deficiency of DNA mismatch repair pathways.
- Tumours with different PIK3CA mutated domains show differences in the tumour immune microenvironments in breast cancer.
- Tumours mutated in the linker ABD-RBD have an exhausted T cell population characterized by the expression of LAG3.
- Tumours with mutations in the C2 PI3K-type, helical and kinase domain have an enrichment of myeloid populations with a gene profile similar to immunosuppressive macrophages.

8. REFERENCES

- [1] H. Sung *et al.*, “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.
- [2] D. L. Narayanan, R. N. Saladi, and J. L. Fox, “Ultraviolet radiation and skin cancer,” *Int. J. Dermatol.*, vol. 49, no. 9, pp. 978–986, 2010, doi: 10.1111/j.1365-4632.2010.04474.x.
- [3] T. Walser *et al.*, “Smoking and lung cancer: The role of inflammation,” *Proc. Am. Thorac. Soc.*, vol. 5, no. 8, pp. 811–815, 2008, doi: 10.1513/pats.200809-100TH.
- [4] N. V. Volkova *et al.*, “Mutational signatures are jointly shaped by DNA damage and repair,” *Nat. Commun.*, vol. 11, no. 1, 2020, doi: 10.1038/s41467-020-15912-7.
- [5] Y. H. Woo and W. H. Li, “DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes,” *Nat. Commun.*, vol. 3, 2012, doi: 10.1038/ncomms1982.
- [6] P. Polak *et al.*, “Cell-of-origin chromatin organization shapes the mutational landscape of cancer,” *Nature*, vol. 518, no. 7539, pp. 360–364, 2015, doi: 10.1038/nature14221.
- [7] O. Pich, F. Muiños, R. Sabarinathan, I. Reyes-Salazar, A. Gonzalez-Perez, and N. Lopez-Bigas, “Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes,” *Cell*, vol. 175, no. 4, pp. 1074–1087.e18, 2018, doi: 10.1016/j.cell.2018.10.004.
- [8] L. B. Alexandrov *et al.*, “Signatures of mutational processes in human cancer,” *Nature*, vol. 500, no. 7463, pp. 415–421, 2013, doi: 10.1038/nature12477.
- [9] M. S. Lawrence *et al.*, “Mutational heterogeneity in cancer and the search for new cancer-associated genes,” *Nature*, vol. 499, no. 7457, pp. 214–218, 2013, doi: 10.1038/nature12213.
- [10] S. Turajlic *et al.*, “Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis,” *Lancet Oncol.*, vol. 18, no. 8, pp. 1009–1021, 2017, doi: 10.1016/S1470-2045(17)30516-8.
- [11] P. Iengar, “An analysis of substitution, deletion and insertion mutations in cancer genes,” *Nucleic Acids Res.*, vol. 40, no. 14, pp. 6401–6413, 2012, doi: 10.1093/nar/gks290.
- [12] G. P. Pfeifer, M. F. Denissenko, M. Olivier, N. Tretyakova, S. S. Hecht, and P. Hainaut, “Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers,” *Oncogene*, vol. 21–48, no. 6, pp. 7435–7451, 2002, doi: 10.1038/sj.onc.1205803.
- [13] M. R. Stratton, P. J. Campbell, and P. A. Futreal, “The cancer genome,” *Nature*, vol. 458, no. 7239, pp. 719–724, 2009, [Online]. Available: <http://dx.doi.org/10.1038/nature07943>
- [14] S. Horn *et al.*, “and Sporadic Melanoma,” no. February, pp. 959–961, 2013.
- [15] H. FW, H. E, X. MJ, K. GV, C. L, and G. LA., “Highly recurrent TERT promoter mutations in human melanoma,” *Science (80-.)*, vol. 339, no. 6122, pp. 957–959, 2013, doi: 10.1126/science.1229259.Highly.
- [16] N. Weinhold, A. Jacobsen, N. Schultz, C. Sander, and W. Lee, “Genome-wide analysis of noncoding regulatory mutations in cancer,” *Nat. Genet.*, vol. 46, no.

- 11, pp. 1160–1165, 2014, doi: 10.1038/ng.3101.
- [17] D. Perera, R. C. Poulos, A. Shah, D. Beck, J. E. Pimanda, and J. W. H. Wong, “Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes,” *Nature*, vol. 532, no. 7598, pp. 259–263, 2016, doi: 10.1038/nature17437.
- [18] R. Sabarinathan, L. Mularoni, J. Deu-Pons, A. Gonzalez-Perez, and N. López-Bigas, “Nucleotide excision repair is impaired by binding of transcription factors to DNA,” *Nature*, vol. 532, no. 7598, pp. 264–267, 2016, doi: 10.1038/nature17661.
- [19] E. Rheinbay *et al.*, “Analyses of non-coding somatic drivers in 2,658 cancer whole genomes,” *Nature*, vol. 578, no. 7793, pp. 102–111, 2020, doi: 10.1038/s41586-020-1965-x.
- [20] M. Deraitus and K. Freeman, “Essentials of cell biology,” p. 475, 2001, doi: 10.1145/634295.634339.
- [21] C. J. Needham, J. R. Bradford, A. J. Bulpitt, M. A. Care, and D. R. Westhead, “Predicting the effect of missense mutations on protein function: Analysis with Bayesian networks,” *BMC Bioinformatics*, vol. 7, pp. 1–14, 2006, doi: 10.1186/1471-2105-7-405.
- [22] T. W. Fitzgerald *et al.*, “Large-scale discovery of novel genetic causes of developmental disorders,” *Nature*, vol. 519, no. 7542, pp. 223–228, 2015, doi: 10.1038/nature14135.
- [23] R. M. Carr *et al.*, “乳鼠心肌提取 HHS Public Access,” *Physiol. Behav.*, vol. 176, no. 1, pp. 139–148, 2016, doi: 10.1002/anie.201208344.Nonproteinogenic.
- [24] G. Karki, “Amino Acids: Characteristics and Classification of amino acids.”
- [25] H. R. B. and F. E. C. Oliver Lichtarge, “Laurence Sterne, Tristram Shandy,” *J. Mol. Biol.*, vol. 257, pp. 342–358, 1996, doi: 10.1002/9781405165327.ch39.
- [26] C. Debès, M. Wang, G. Caetano-Anollé, and F. Gräter, “Evolutionary Optimization of Protein Folding,” *PLoS Comput. Biol.*, vol. 9, no. 1, 2013, doi: 10.1371/Citation.
- [27] B. Niu *et al.*, “Protein-structure-guided discovery of functional mutations across 19 cancer types,” *Nat. Genet.*, vol. 48, no. 8, pp. 827–837, 2016, doi: 10.1038/ng.3586.
- [28] C. Tokheim *et al.*, “Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure,” *Cancer Res.*, vol. 76, no. 13, pp. 3719–3731, 2016, doi: 10.1158/0008-5472.CAN-15-3190.
- [29] A. Kamburov *et al.*, “Comprehensive assessment of cancer missense mutation clustering in protein structures,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 40, pp. E5486–E5495, 2015, doi: 10.1073/pnas.1516373112.
- [30] M. J. Meyer *et al.*, “mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome,” *Hum. Mutat.*, vol. 37, no. 5, pp. 447–456, 2016, doi: 10.1002/humu.22963.
- [31] A. Hijikata, T. Tsuji, M. Shionyu, and T. Shirai, “Decoding disease-causing mechanisms of missense mutations from supramolecular structures,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–8, 2017, doi: 10.1038/s41598-017-08902-1.
- [32] Z. Zhang, M. A. Miteva, L. Wang, and E. Alexov, “Analyzing effects of naturally occurring missense mutations,” *Comput. Math. Methods Med.*, vol. 2012, 2012, doi: 10.1155/2012/805827.
- [33] H. Fujiwara, K. I. Tatsumi, S. Tanaka, M. Kimura, O. Nose, and N. Amino, “A novel V59E missense mutation in the sodium iodide symporter gene in a family with

- iodide transport defect," *Thyroid*, vol. 10, no. 6, pp. 471–474, 2000, doi: 10.1089/thy.2000.10.471.
- [34] A. De la Vieja, C. S. Ginter, and N. Carrasco, "The Q267E mutation in the sodium/iodide symporter (NIS) causes congenital iodide transport defect (ITD) by decreasing the NIS turnover number," *J. Cell Sci.*, vol. 117, no. 5, pp. 677–687, 2004, doi: 10.1242/jcs.00898.
- [35] E. Krissinel, "On the relationship between sequence and structure similarities in proteomics," *Bioinformatics*, vol. 23, no. 6, pp. 717–723, 2007, doi: 10.1093/bioinformatics/btm006.
- [36] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, doi: 10.1038/s41586-021-03819-2.
- [37] S. Jean and A. A. Kiger, "Classes of phosphoinositide 3-kinases at a glance," *J. Cell Sci.*, vol. 127, no. 5, pp. 923–928, 2014, doi: 10.1242/jcs.093773.
- [38] G. Hoxhaj and B. D. Manning, "The PI3K–AKT network at the interface of oncogenic signalling and cancer metabolism," *Nat. Rev. Cancer*, vol. 20, no. 2, pp. 74–88, 2020, doi: 10.1038/s41568-019-0216-7.
- [39] C. Massacesi *et al.*, "PI3K inhibitors as new cancer therapeutics: Implications for clinical trial design," *Onco. Targets. Ther.*, vol. 9, pp. 203–210, 2016, doi: 10.2147/OTT.S89967.
- [40] B. Bilanges, Y. Posor, and B. Vanhaesebroeck, "PI3K isoforms in cell signalling and vesicle trafficking," *Nat. Rev. Mol. Cell Biol.*, vol. 20, no. 9, pp. 515–534, 2019, doi: 10.1038/s41580-019-0129-z.
- [41] J. E. Burke and R. L. Williams, "Synergy in activating class I PI3Ks," *Trends Biochem. Sci.*, vol. 40, no. 2, pp. 88–100, 2015, doi: 10.1016/j.tibs.2014.12.003.
- [42] D. Zardavas, W. A. Phillips, and S. Loi, "PIK3CA mutations in breast cancer: Reconciling findings from preclinical and clinical data," *Breast Cancer Res.*, vol. 16, no. 1, pp. 1–10, 2014, doi: 10.1186/bcr3605.
- [43] M. Zhang, H. Jang, and R. Nussinov, "PI3K Driver Mutations: A Biophysical Membrane-Centric Perspective," *Cancer Res.*, vol. 81, no. 2, pp. 237–247, 2021, doi: 10.1158/0008-5472.CAN-20-0911.
- [44] J. E. Burke, O. Perisic, G. R. Masson, O. Vadas, and R. L. Williams, "Oncogenic mutations mimic and enhance dynamic events in the natural activation of phosphoinositide 3-kinase p110 α (PIK3CA)," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 38, pp. 15259–15264, 2012, doi: 10.1073/pnas.1205508109.
- [45] M. Gymnopoulos, M. A. Elsliger, and P. K. Vogt, "Rare cancer-specific mutations in PIK3CA show gain of function," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 13, pp. 5569–5574, 2007, doi: 10.1073/pnas.0701005104.
- [46] A. Alqahtani, H. S. K. Ayeshe, and H. Halawani, "PIK3CA gene mutations in solid malignancies: Association with clinicopathological parameters and prognosis," *Cancers (Basel)*, vol. 12, no. 1, pp. 1–18, 2020, doi: 10.3390/cancers12010093.
- [47] H. Wu *et al.*, "The distinct clinicopathological and prognostic implications of PIK3CA mutations in breast cancer patients from central China," *Cancer Manag. Res.*, vol. 11, pp. 1473–1492, 2019, doi: 10.2147/CMAR.S195351.
- [48] M. Barbareschi *et al.*, "Different prognostic roles of mutations in the helical and kinase domains of the PIK3CA gene in breast carcinomas," *Clin. Cancer Res.*, vol. 13, no. 20, pp. 6064–6069, 2007, doi: 10.1158/1078-0432.CCR-07-0266.

- [49] A. Akinleye, P. Avvaru, M. Furqan, Y. Song, and D. Liu, "As Cancer Therapeutics," pp. 1–17, 2013.
- [50] R. Huang and P. K. Zhou, *DNA damage repair: historical perspectives, mechanistic pathways and clinical translation for targeted cancer therapy*, vol. 6, no. 1. Springer US, 2021. doi: 10.1038/s41392-021-00648-7.
- [51] R. Mishra, H. Patel, S. Alanazi, M. K. Kilroy, and J. T. Garrett, "PI3K inhibitors in cancer: Clinical implications and adverse effects," *Int. J. Mol. Sci.*, vol. 22, no. 7, 2021, doi: 10.3390/ijms22073464.
- [52] Q. Yang, P. Modi, T. Newcomb, C. Quéva, and V. Gandhi, "Idelalisib: First-in-class PI3K delta inhibitor for the treatment of chronic lymphocytic leukemia, small lymphocytic leukemia, and follicular lymphoma," *Clin. Cancer Res.*, vol. 21, no. 7, pp. 1537–1542, 2015, doi: 10.1158/1078-0432.CCR-14-2034.
- [53] F. Leenhardt, M. Alexandre, and W. Jacot, "Alpelisib for the treatment of PIK3CA-mutated, hormone receptor-positive, HER2-negative metastatic breast cancer," *Expert Opin. Pharmacother.*, vol. 22, no. 6, pp. 667–675, 2021, doi: 10.1080/14656566.2021.1873952.
- [54] S. Wang, K. Xie, and T. Liu, "Cancer Immunotherapies: From Efficacy to Resistance Mechanisms – Not Only Checkpoint Matters," *Front. Immunol.*, vol. 12, no. July, pp. 1–17, 2021, doi: 10.3389/fimmu.2021.690112.
- [55] A. K. Palucka and L. M. Coussens, "The Basis of Oncoimmunology," *Cell*, vol. 164, no. 6, pp. 1233–1247, 2016, doi: 10.1016/j.cell.2016.01.049.
- [56] S. Z. Wu *et al.*, "A single-cell and spatially resolved atlas of human breast cancers," *Nat. Genet.*, vol. 53, no. 9, pp. 1334–1347, 2021, doi: 10.1038/s41588-021-00911-1.
- [57] A. C. Little *et al.*, "IL-4/IL-13 stimulated macrophages enhance breast cancer invasion via rho-GTPase regulation of synergistic VEGF/CCL-18 signaling," *Front. Oncol.*, vol. 9, no. MAY, pp. 1–13, 2019, doi: 10.3389/fonc.2019.00456.
- [58] Y. Wang, K. C. C. Johnson, M. E. Gatti-Mays, and Z. Li, *Emerging strategies in targeting tumor-resident myeloid cells for cancer immunotherapy*, vol. 15, no. 1. BioMed Central, 2022. doi: 10.1186/s13045-022-01335-y.
- [59] S. S. Chandran *et al.*, "Immunogenicity and therapeutic targeting of a public neoantigen derived from mutated PIK3CA," *Nat. Med.*, vol. 28, no. 5, pp. 946–957, 2022, doi: 10.1038/s41591-022-01786-3.
- [60] F. Lüönd, S. Tiede, and G. Christofori, "Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression," *Br. J. Cancer*, vol. 125, no. 2, pp. 164–175, 2021, doi: 10.1038/s41416-021-01328-7.
- [61] M. A. Aleskandarany, M. E. Vandenberghe, C. Marchiò, I. O. Ellis, A. Sapino, and E. A. Rakha, "Tumour Heterogeneity of Breast Cancer: From Morphology to Personalised Medicine," *Pathobiology*, vol. 85, no. 1–2, pp. 23–34, 2018, doi: 10.1159/000477851.
- [62] K. Hemminki and C. Granström, "Morphological types of breast cancer in family members and multiple primary tumours: Is morphology genetically determined?," *Breast Cancer Res.*, vol. 4, no. 4, pp. 1–6, 2002, doi: 10.1186/bcr444.
- [63] N. Harbeck *et al.*, *Breast cancer*, vol. 5, no. 1. 2019. doi: 10.1038/s41572-019-0111-2.
- [64] A. Burguin *et al.*, "Breast Cancer Treatments: Updates and New Challenges",

- Journal of Personalized Medicine*, vol. 11(8), 808, 2021 doi: <https://doi.org/10.3390/jpm11080808>.
- [65] S. Kalli, A. Semine, S. Cohen, S. P. Naber, S. S. Makim, and M. Bahl, "American joint committee on cancer's staging system for breast cancer, eighth edition: What the radiologist needs to know," *Radiographics*, vol. 38, no. 7, pp. 1921–1933, 2018, doi: 10.1148/rg.2018180056.
- [66] K. H. Kensler *et al.*, "PAM50 molecular intrinsic subtypes in the nurses' health Study cohorts," *Cancer Epidemiol. Biomarkers Prev.*, vol. 28, no. 4, pp. 798–806, 2019, doi: 10.1158/1055-9965.EPI-18-0863.
- [67] A. Spitale, P. Mazzola, D. Soldini, L. Mazzucchelli, and A. Bordoni, "Breast cancer classification according to immunohistochemical markers: Clinicopathologic features and short-term survival analysis in a population-based study from the South of Switzerland," *Ann. Oncol.*, vol. 20, no. 4, pp. 628–635, 2009, doi: 10.1093/annonc/mdn675.
- [68] A. Prat *et al.*, "Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer," *Breast Cancer Res.*, vol. 12, no. 5, 2010, doi: 10.1186/bcr2635.
- [69] M. Hergueta-Redondo, J. Palacios, A. Cano, and G. Moreno-Bueno, "'New' molecular taxonomy in breast cancer," *Clin. Transl. Oncol.*, vol. 10, no. 12, pp. 777–785, 2008, doi: 10.1007/s12094-008-0290-x.
- [70] M. J. Engstrøm *et al.*, "Molecular subtypes, histopathological grade and survival in a historic cohort of breast cancer patients.," *Breast Cancer Res. Treat.*, vol. 140, no. 3, pp. 463–473, 2013, doi: 10.1007/s10549-013-2647-2.
- [71] A. Hennigs *et al.*, "Prognosis of breast cancer molecular subtypes in routine clinical care: A large prospective cohort study," *BMC Cancer*, vol. 16, no. 1, pp. 1–9, 2016, doi: 10.1186/s12885-016-2766-3.
- [72] S. M. Fragomeni, A. Sciallis, J. S. Jeruss, B. C. Unit, A. Arbor, and A. Arbor, "HHS Public Access," vol. 27, no. 1, pp. 95–120, 2019, doi: 10.1016/j.soc.2017.08.005.Molecular.
- [73] National Cancer Institute, "Cancer Stat Facts: Female Breast Cancer Subtypes", [Online]. Available: <https://seer.cancer.gov/statfacts/html/breast-subtypes.html>
- [74] S. Masood, "Breast cancer subtypes: Morphologic and biologic characterization," *Women's Heal.*, vol. 12, no. 1, pp. 103–119, 2016, doi: 10.2217/whe.15.99.
- [75] S. Loibl, P. Poortmans, M. Morrow, C. Denkert, and G. Curigiano, "Breast cancer," *Lancet*, vol. 397, no. 10286, pp. 1750–1769, 2021, doi: 10.1016/S0140-6736(20)32381-3.
- [76] Y. M. Sohn, K. Han, and M. Seo, "Immunohistochemical Subtypes of Breast Cancer: Correlation with Clinicopathological and Radiological Factors," *Iran. J. Radiol.*, vol. 13, no. 4, 2016, doi: 10.5812/iranjradiol.31386.
- [77] B. Golia, H. R. Singh, and G. Timinszky, "Poly-ADP-ribosylation signaling during DNA damage repair," *Front. Biosci. - Landmark*, vol. 20, no. 3, pp. 440–457, 2015, doi: 10.2741/4318.
- [78] H. Farmer *et al.*, "Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy," *Nature*, vol. 434, no. 7035, pp. 917–921, 2005, doi: 10.1038/nature03445.
- [79] C. D. Murin, "Considerations of Antibody Geometric Constraints on NK Cell Antibody Dependent Cellular Cytotoxicity," *Front. Immunol.*, vol. 11, no. July, pp.

- 1–27, 2020, doi: 10.3389/fimmu.2020.01635.
- [80] V. Salemme, G. Centonze, F. Cavallo, P. Defilippi, and L. Conti, “The Crosstalk Between Tumor Cells and the Immune Microenvironment in Breast Cancer: Implications for Immunotherapy,” *Front. Oncol.*, vol. 11, no. March, pp. 1–20, 2021, doi: 10.3389/fonc.2021.610303.
- [81] G. P. Dunn, L. J. Old, and R. D. Schreiber, “The three Es of cancer immunoediting,” *Annu. Rev. Immunol.*, vol. 22, no. 4, pp. 329–360, 2004, doi: 10.1146/annurev.immunol.22.012703.104803.
- [82] Q. Lei, D. Wang, K. Sun, L. Wang, and Y. Zhang, “Resistance Mechanisms of Anti-PD1/PDL1 Therapy in Solid Tumors,” *Front. Cell Dev. Biol.*, vol. 8, no. July, 2020, doi: 10.3389/fcell.2020.00672.
- [83] O. Demaria, S. Cornen, M. Daëron, Y. Morel, R. Medzhitov, and E. Vivier, “Harnessing innate immunity in cancer therapy,” *Nature*, vol. 574, no. 7776, pp. 45–56, 2019, doi: 10.1038/s41586-019-1593-5.
- [84] K. El Bairi *et al.*, “The tale of TILs in breast cancer: A report from The International Immuno-Oncology Biomarker Working Group,” *npj Breast Cancer*, vol. 7, no. 1, 2021, doi: 10.1038/s41523-021-00346-1.
- [85] M. Elosua-Bayes, P. Nieto, E. Mereu, I. Gut, and H. Heyn, “SPOTlight: Seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes,” *Nucleic Acids Res.*, vol. 49, no. 9, p. E50, 2021, doi: 10.1093/nar/gkab043.
- [86] S. K. Burley *et al.*, “RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences,” *Nucleic Acids Res.*, vol. 49, no. 1, pp. D437–D451, 2021, doi: 10.1093/nar/gkaa1038.
- [87] S. Bienert *et al.*, “The SWISS-MODEL Repository-new features and functionality,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D313–D319, 2017, doi: 10.1093/nar/gkw1132.
- [88] W. H., “ggplot2: Elegant Graphics for Data Analysis,” *Springer-Verlag New York*, vol. ISBN 978-3, no. [https://ggplot2.tidyverse.org.](https://ggplot2.tidyverse.org/), 2016.
- [89] S. M. Ashiqul, “SigProfilerSingleSample,” 2018. <https://github.com/AlexandrovLab/SigProfilerSingleSample>
- [90] L. B. Alexandrov *et al.*, “The repertoire of mutational signatures in human cancer,” *Nature*, vol. 578, no. 7793, pp. 94–101, 2020, doi: 10.1038/s41586-020-1943-3.
- [91] S. M. A. Islam *et al.*, “Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor,” *Cell Genomics*, p. 100179, 2022, doi: 10.1016/j.xgen.2022.100179.
- [92] P. J. Campbell *et al.*, “Pan-cancer analysis of whole genomes,” *Nature*, vol. 578, no. 7793, pp. 82–93, 2020, doi: 10.1038/s41586-020-1969-6.
- [93] J. P. Whalley *et al.*, “Framework for quality assessment of whole genome cancer sequences,” *Nat. Commun.*, vol. 11, no. 1, p. 5040, 2020, doi: 10.1038/s41467-020-18688-y.
- [94] E. Rheinbay *et al.*, “Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes,” *bioRxiv*, 2017, doi: 10.1101/237313.
- [95] M. Costello *et al.*, “Discovery and characterization of artifactual mutations in deep

- coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation," *Nucleic Acids Res.*, vol. 41, no. 6, pp. e67–e67, 2013, doi: 10.1093/nar/gks1443.
- [96] Y. Fan *et al.*, "MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data," *Genome Biol.*, vol. 17, no. 1, p. 178, 2016, doi: 10.1186/s13059-016-1029-6.
- [97] V. Moncunill *et al.*, "Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads," *Nat. Biotechnol.*, vol. 32, p. 1106, 2014, doi: 10.1038/nbt.3027.
- [98] B. J. Knaus and N. J. Grünwald, "vcfr: a package to manipulate and visualize variant call format data in R," *Mol. Ecol. Resour.*, vol. 17, no. 1, pp. 44–53, 2017, doi: 10.1111/1755-0998.12549.
- [99] S. Lê, J. Josse, and F. Husson, "FactoMineR: An R Package for Multivariate Analysis," *J. Stat. Softw.*, vol. 25, no. 1, p. 18, 2008, doi: 10.18637/jss.v025.i01.
- [100] F. Husson, J. Josse, and J. Pages, "Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data," *Appl. Math. Dep.*, pp. 1–17, 2010.
- [101] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, "Information content of binding sites on nucleotide sequences," *J. Mol. Biol.*, vol. 188, no. 3, pp. 415–431, 1986, [Online]. Available: <http://www.sciencedirect.com/science/article/pii/0022283686901658>
- [102] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951, [Online]. Available: <http://www.jstor.org/stable/2236703>
- [103] O. Wagih, "ggseqlogo: a versatile R package for drawing sequence logos," *Bioinformatics*, vol. 33, no. 22, pp. 3645–3647, 2017, doi: 10.1093/bioinformatics/btx469.
- [104] L. Zappia and A. Oshlack, "Clustering trees: a visualization for evaluating clusterings at multiple resolutions," *Gigascience*, vol. 7, no. 7, 2018, doi: 10.1093/gigascience/giy083.
- [105] J. A. Marsh and S. A. Teichmann, "Relative solvent accessible surface area predicts protein conformational changes upon binding," *Structure*, vol. 19, no. 6, pp. 859–867, 2011, doi: 10.1016/j.str.2011.03.010.
- [106] B.-S. Inc., "Amino Acid Masses Table." <https://www.biosyn.com/tew/amino-acid-masses-tables.aspx>
- [107] S. Ahmad, M. M. Gromiha, H. Fawareh, and A. Sarai, "ASAView: Database and tool for solvent accessibility representation in proteins," *BMC Bioinformatics*, vol. 5, pp. 1–5, 2004, doi: 10.1186/1471-2105-5-51.
- [108] C. Savojardo, M. Manfredi, P. L. Martelli, and R. Casadio, "Solvent Accessibility of Residues Undergoing Pathogenic Variations in Humans: From Protein Structures to Protein Sequences," *Front. Mol. Biosci.*, vol. 7, no. January, pp. 1–9, 2021, doi: 10.3389/fmolb.2020.626363.
- [109] Z. Lyu, Z. Wang, F. Luo, J. Shuai, and Y. Huang, "Protein Secondary Structure Prediction With a Reductive Deep Learning Method," *Front. Bioeng. Biotechnol.*, vol. 9, no. June, pp. 1–8, 2021, doi: 10.3389/fbioe.2021.687426.
- [110] L. PAULING, R. B. COREY, and H. R. BRANSON, "The structure of proteins; two

- hydrogen-bonded helical configurations of the polypeptide chain.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 37, no. 4, pp. 205–211, 1951, doi: 10.1073/pnas.37.4.205.
- [111] B. Zhao *et al.*, "DescribePROT: Database of amino acid-level protein structure and function predictions," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D298–D308, 2021, doi: 10.1093/nar/gkaa931.
- [112] M. K. Basu, E. Poliakov, and I. B. Rogozin, "Domain mobility in proteins: Functional and evolutionary implications," *Brief. Bioinform.*, vol. 10, no. 3, pp. 205–216, 2009, doi: 10.1093/bib/bbn057.
- [113] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, "Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt," *Nat. Protoc.*, vol. 4, no. 8, pp. 1184–1191, 2009, doi: 10.1038/nprot.2009.97.
- [114] M. Blum *et al.*, "The InterPro protein families and domains database: 20 years on," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D344–D354, 2021, doi: 10.1093/nar/gkaa977.
- [115] C. J. A. Sigrist *et al.*, "New and continuing developments at PROSITE," *Nucleic Acids Res.*, vol. 41, no. D1, pp. 344–347, 2013, doi: 10.1093/nar/gks1067.
- [116] J. Mistry *et al.*, "Pfam: The protein families database in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D412–D419, 2021, doi: 10.1093/nar/gkaa913.
- [117] I. Letunic, S. Khedkar, and P. Bork, "SMART: Recent updates, new developments and status in 2020," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D458–D460, 2021, doi: 10.1093/nar/gkaa937.
- [118] E. F. Pettersen *et al.*, "UCSF Chimera - A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004, doi: 10.1002/jcc.20084.
- [119] H. Ashkenazy *et al.*, "ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W344–W350, 2016, doi: 10.1093/NAR/GKW408.
- [120] J. Gao *et al.*, "3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets," *Genome Med.*, vol. 9, no. 1, pp. 1–13, 2017, doi: 10.1186/s13073-016-0393-x.
- [121] S. Khan and M. Vihinen, "Performance of protein stability predictors," *Hum. Mutat.*, vol. 31, no. 6, pp. 675–684, 2010, doi: 10.1002/humu.21242.
- [122] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, "The FoldX web server: An online force field," *Nucleic Acids Res.*, vol. 33, no. SUPPL. 2, 2005, doi: 10.1093/nar/gki387.
- [123] J. Pagès, "Analyse Factorielle de Donnees Mixtes," *Rev. Stat. Appl.*, vol. 4, pp. 93–111, 2004.
- [124] F. Husson and J. Josse, "Principal Component Methods - Hierarchical Clustering - Partitional Clustering: Why Would We Need to Choose for Visualizing Data," 2010.
- [125] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, pp. 1–21, 2014, doi: 10.1186/s13059-014-0550-8.
- [126] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, pp. 15545–15550, 2005, doi: 10.1073/pnas.0506580102.
- [127] M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, and B. Spiegelman, "PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately

- downregulated in human diabetes," *Nat. Genet.*, vol. 34, no. 3, pp. 267–273, 2003.
- [128] F. G. Berger and C. R. Boland, *Novel Approaches to Colorectal Cancer*, Volume 151. 2021.
- [129] L. Alexandrov *et al.*, "The Repertoire of Mutational Signatures in Human Cancer," *bioRxiv*, p. 322859, 2018, doi: 10.1101/322859.
- [130] J. Harrow *et al.*, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome Res*, vol. 22, no. 9, pp. 1760–1774, 2012, doi: 10.1101/gr.135350.111.
- [131] A. H. Ramos *et al.*, "Oncotator: Cancer Variant Annotation Tool," *Hum Mut*, vol. 36, no. 4, pp. E2423–E2429, 2015, doi: 10.1002/humu.22771.
- [132] R. Sabarinathan *et al.*, "The whole-genome panorama of cancer drivers," *bioRxiv*, 2017, doi: 10.1101/190330.
- [133] R. S. Hansen *et al.*, "Sequencing newly replicated DNA reveals widespread plasticity in human replication timing," *Proc. Natl. Acad. Sci.*, vol. 107, no. 1, pp. 139–144, 2010, doi: 10.1073/pnas.0912402107.
- [134] I. G. Gut, P. D. Wood, and R. W. Redmond, "Interaction of Triplet Photosensitizers with Nucleotides and DNA in Aqueous Solution at Room Temperature," *J. Am. Chem. Soc.*, vol. 118, no. 10, pp. 2366–2373, 1996, doi: 10.1021/ja9519344.
- [135] N. J. Fredriksson, K. Elliott, S. Filges, J. Van den Eynden, A. Ståhlberg, and E. Larsson, "Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature," *PLOS Genet.*, vol. 13, no. 5, p. e1006773, 2017, doi: 10.1371/journal.pgen.1006773.
- [136] P. Mao *et al.*, "ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma," *Nat. Commun.*, vol. 9, no. 1, p. 2626, 2018, doi: 10.1038/s41467-018-05064-0.
- [137] I. Martincorena *et al.*, "Universal Patterns of Selection in Cancer and Somatic Tissues," *Cell*, vol. 171, no. 5, pp. 1029–1041.e21, 2017, doi: 10.1016/j.cell.2017.09.042.
- [138] H. Ellegren, "Microsatellites: simple sequences with complex evolution," *Nat. Rev. Genet.*, vol. 5, p. 435, 2004, doi: 10.1038/nrg1348.
- [139] J. X. Sun *et al.*, "A direct characterization of human mutation based on microsatellites," *Nat Genet*, vol. 44, no. 10, pp. 1161–1165, 2012, [Online]. Available: <http://dx.doi.org/10.1038/ng.2398>
- [140] F. Supek and B. Lehner, "Differential DNA mismatch repair underlies mutation rate variation across the human genome," *Nature*, vol. 521, no. 7550, pp. 81–84, 2015, doi: 10.1038/nature14173.
- [141] Y. Yang, J. Sterling, F. Storici, M. A. Resnick, and D. A. Gordenin, "Hypermutability of Damaged Single-Strand DNA Formed at Double-Strand Breaks and Uncapped Telomeres in Yeast *Saccharomyces cerevisiae*," *PLOS Genet.*, vol. 4, no. 11, p. e1000264, 2008, doi: 10.1371/journal.pgen.1000264.
- [142] M. Tomkova, J. Tomek, S. Kriaucionis, and B. Schuster-Böckler, "Mutational signature distribution varies with DNA replication timing and strand asymmetry," *Genome Biol.*, vol. 19, no. 1, p. 129, 2018, doi: 10.1186/s13059-018-1509-y.
- [143] H. Kamiya, "Mutations Induced by Oxidized DNA Precursors and Their Prevention by Nucleotide Pool Sanitization Enzymes," *Genes Environ.*, vol. 29, no. 4, pp. 133–140, 2007, doi: 10.3123/jemsge.29.133.

- [144] V. B. Seplyarskiy, G. A. Bazykin, and R. A. Soldatov, "Polymerase ζ Activity Is Linked to Replication Timing in Humans: Evidence from Mutational Signatures," *Mol. Biol. Evol.*, vol. 32, no. 12, pp. 3158–3172, 2015, doi: 10.1093/molbev/msv184.
- [145] T. J. Hamblin, Z. Davis, A. Gardiner, D. G. Oscier, and F. K. Stevenson, "Unmutated Ig V_H Genes Are Associated With a More Aggressive Form of Chronic Lymphocytic Leukemia," *Blood*, vol. 94, no. 6, pp. 1848–1854, 1999, [Online]. Available: <http://www.bloodjournal.org/content/bloodjournal/94/6/1848.full.pdf>
- [146] G. P. Pfeifer, M. F. Denissenko, M. Olivier, N. Tretyakova, S. S. Hecht, and P. Hainaut, "Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers," *Oncogene*, vol. 21, p. 7435, 2002, doi: 10.1038/sj.onc.1205803.
- [147] O. Pich, F. Muiños, R. Sabarinathan, I. Reyes-Salazar, A. Gonzalez-Perez, and N. Lopez-Bigas, "Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes," *Cell*, vol. 175, no. 4, pp. 1074–1087.e18, 2018, doi: 10.1016/j.cell.2018.10.004.
- [148] E. Segal *et al.*, "A genomic code for nucleosome positioning," *Nature*, vol. 442, no. 7104, pp. 772–778, 2006, doi: 10.1038/nature04979.
- [149] L. B. Alexandrov *et al.*, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, no. 7463, pp. 415–421, 2013, doi: 10.1038/nature12477.
- [150] K. Chan *et al.*, "An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers," *Nat. Genet.*, vol. 47, p. 1067, 2015, doi: 10.1038/ng.3378.
- [151] S. Uchida, "Databases and software to make your research life easier," *Annot. New Genes*, pp. 7–47, 2012, doi: 10.1533/9781908818126.7.
- [152] Z. Zhang *et al.*, "Neoantigen: A New Breakthrough in Tumor Immunotherapy," *Front. Immunol.*, vol. 12, no. April, pp. 1–9, 2021, doi: 10.3389/fimmu.2021.672356.
- [153] J. Cortés, E. Calvo, A. Vivancos, J. Perez-Garcia, J. A. Recio, and J. Seoane, "New approach to cancer therapy based on a molecularly defined cancer classification," *CA. Cancer J. Clin.*, vol. 64, no. 1, pp. 70–74, 2014, doi: 10.3322/caac.21211.
- [154] S. et al. Yavuzigitoglu, "Uveal Melanomas with SF3B1 Mutations: A Distinct Subclass Associated with Late-Onset Metastases," *Ophthalmology*, vol. 123, no. 5, pp. 1118–1128, doi: <https://doi.org/10.1016/j.ophtha.2016.01.023>.
- [155] Y. Wan and C. J. Wu, "SF3B1 mutations in chronic lymphocytic leukemia," *Blood*, vol. 121, no. 23, pp. 4627–4634, 2013, doi: 10.1182/blood-2013-02-427641.
- [156] W. Caleb Rutledge *et al.*, "Tumor-infiltrating lymphocytes in glioblastoma are associated with specific genomic alterations and related to transcriptional class," *Clin. Cancer Res.*, vol. 19, no. 18, pp. 4951–4960, 2013, doi: 10.1158/1078-0432.CCR-13-0551.
- [157] P. Charoentong *et al.*, "Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade," *Cell Rep.*, vol. 18, no. 1, pp. 248–262, 2017, doi: 10.1016/j.celrep.2016.12.019.
- [158] M. D. Wellenstein and K. E. de Visser, "Cancer-Cell-Intrinsic Mechanisms Shaping the Tumor Immune Landscape," *Immunity*, vol. 48, no. 3, pp. 399–416, 2018, doi: 10.1016/j.immuni.2018.03.004.
- [159] M. L. Rudd *et al.*, "A unique spectrum of somatic PIK3CA (p110 α) mutations within

- primary endometrial carcinomas,” *Clin. Cancer Res.*, vol. 17, no. 6, pp. 1331–1340, 2011, doi: 10.1158/1078-0432.CCR-10-0540.
- [160] S. Croessmann *et al.*, “PIK3CA C2 domain deletions hyperactivate phosphoinositide 3-kinase (PI3K), generate oncogene dependence, and are exquisitely sensitive to PI3Ka inhibitors,” *Clin. Cancer Res.*, vol. 24, no. 6, pp. 1426–1435, 2018, doi: 10.1158/1078-0432.CCR-17-2141.
- [161] S. A. Forbes *et al.*, “COSMIC: Exploring the world’s knowledge of somatic mutations in human cancer,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D805–D811, 2015, doi: 10.1093/nar/gku1075.
- [162] F. Wang *et al.*, “ATACdb: A comprehensive human chromatin accessibility database,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D55–D64, 2021, doi: 10.1093/nar/gkaa943.
- [163] K. Zhang *et al.*, “A single-cell atlas of chromatin accessibility in the human genome,” *Cell*, vol. 184, no. 24, pp. 5985–6001.e19, 2021, doi: 10.1016/j.cell.2021.10.024.
- [164] C. A. Davis *et al.*, “The Encyclopedia of DNA elements (ENCODE): Data portal update,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D794–D801, 2018, doi: 10.1093/nar/gkx1081.
- [165] S. Mjos *et al.*, “PIK3CA exon9 mutations associate with reduced survival, and are highly concordant between matching primary tumors and metastases in endometrial cancer,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, 2017, doi: 10.1038/s41598-017-10717-z.
- [166] A. Voutsina, A. Kalikaki, V. Mitsi, E. Stathopoulos, V. Georgoulas, and D. Mavroudis, “Association of PIK3CA kinase domain mutations with poor prognosis in operable breast cancer,” *J. Clin. Oncol.*, vol. 29:15_suppl, 2011.
- [167] Y. Tang *et al.*, “PIK3CA gene mutations in the helical domain correlate with high tumor mutation burden and poor prognosis in metastatic breast carcinomas with late-line therapies,” *Aging (Albany. NY)*, vol. 12, no. 2, pp. 1577–1590, 2020, doi: 10.18632/aging.102701.
- [168] A. J. Li *et al.*, “PIK3CA and TP53 mutations predict overall survival of stage II/III colorectal cancer patients,” *World J. Gastroenterol.*, vol. 24, no. 5, pp. 631–640, 2018, doi: 10.3748/wjg.v24.i5.631.
- [169] A. Feizi, F. Gatto, M. Uhlen, and J. Nielsen, “Human protein secretory pathway genes are expressed in a tissue-specific pattern to match processing demands of the secretome,” *npj Syst. Biol. Appl.*, vol. 3, no. 1, pp. 1–9, 2017, doi: 10.1038/s41540-017-0021-4.
- [170] V. Karpisheh *et al.*, “The role of Th17 cells in the pathogenesis and treatment of breast cancer,” *Cancer Cell Int.*, vol. 22, no. 1, pp. 1–13, 2022, doi: 10.1186/s12935-022-02528-8.
- [171] ICGC/TCGA-Consortium, “Pan-cancer analysis of whole genomes,” *Nature*, vol. 578, no. 7793, pp. 82–93, 2020, doi: 10.1038/s41586-020-1969-6.
- [172] P. Priestley *et al.*, “Pan-cancer whole-genome analyses of metastatic solid tumours,” *Nature*, vol. 575, no. 7781, pp. 210–216, 2019, doi: 10.1038/s41586-019-1689-y.
- [173] K. Ellrott *et al.*, “Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines,” *Cell Syst.*, vol. 6, no. 3, pp. 271–281.e7, 2018, doi: 10.1016/j.cels.2018.03.002.

- [174] T. Schneider, G. H. Smith, M. R. Rossi, C. E. Hill, and L. Zhang, "Validation of a Customized Bioinformatics Pipeline for a Clinical Next-Generation Sequencing Test Targeting Solid Tumor-Associated Variants," *J. Mol. Diagnostics*, vol. 20, no. 3, pp. 355–365, 2018, doi: 10.1016/j.jmoldx.2018.01.007.
- [175] M. A. Musen, M. J. O'Connor, E. Schultes, M. M. Romero, J. Hardi, and J. Graybeal, "Modeling community standards for metadata as templates makes data FAIR," *Sci. Data*, vol. 9, no. 696, pp. 1–15, 2022, doi: 10.48550/arXiv.2208.02836.
- [176] C. Denkert *et al.*, "Reconstructing tumor history in breast cancer: signatures of mutational processes and response to neoadjuvant chemotherapy," *Ann. Oncol.*, vol. 32, no. 4, pp. 500–511, 2021, doi: 10.1016/j.annonc.2020.12.016.
- [177] X. Xing, N. Jin, and J. Wang, "Polymerase Epsilon-Associated Ultramutagenesis in Cancer," *Cancers (Basel)*, vol. 14, no. 6, pp. 1–13, 2022, doi: 10.3390/cancers14061467.
- [178] Y. Xiao and G. J. Freeman, "The Microsatellite Instable (MSI) Subset of Colorectal Cancer is a particularly good candidate for checkpoint blockade immunotherapy," *Cancer Discov.*, vol. 5, no. 1, pp. 16–18, 2015, doi: 10.1158/2159-8290.CD-14-1397.
- [179] Z. Saridaki, J. Souglakos, and V. Georgoulas, "Prognostic and predictive significance of MSI in stages II/III colon cancer," *World J. Gastroenterol.*, vol. 20, no. 22, pp. 6809–6814, 2014, doi: 10.3748/wjg.v20.i22.6809.
- [180] M. Schlesner and R. Eils, "Hypermutation takes the driver's seat," *Genome Med.*, vol. 7, no. 1, p. 31, 2015, doi: 10.1186/s13073-015-0159-x.
- [181] V. Heong, N. Ngoi, and D. S. P. Tan, "Update on immune checkpoint inhibitors in gynecological cancers," *J. Gynecol. Oncol.*, vol. 28, no. 2, p. e20, 2017, doi: 10.3802/jgo.2017.28.e20.
- [182] X. S. Puente *et al.*, "Non-coding recurrent mutations in chronic lymphocytic leukaemia," *Nature*, vol. 526, no. 7574, pp. 519–524, 2015, doi: 10.1038/nature14666.
- [183] M. Secrier *et al.*, "Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance," *Nat. Genet.*, vol. 48, p. 1131, 2016, doi: 10.1038/ng.3659.
- [184] G. Cathomas, "PIK3CA in colorectal cancer," *Front. Oncol.*, vol. 4 MAR, no. March, pp. 16–19, 2014, doi: 10.3389/fonc.2014.00035.
- [185] B. Bianco, C. P. Barbosa, C. M. Trevisan, A. S. Laganà, and E. Montagna, "Endometrial cancer: A genetic point of view," *Transl. Cancer Res.*, vol. 9, no. 12, pp. 7706–7715, 2020, doi: 10.21037/tcr-20-2334.
- [186] D. I. Lin, "Improved survival associated with somatic PIK3CA mutations in copy-number low endometrioid endometrial adenocarcinoma," *Oncol. Lett.*, vol. 10, no. 5, pp. 2743–2748, 2015, doi: 10.3892/ol.2015.3702.
- [187] S. Han, W. H. Shuen, W. W. Wang, E. Nazim, and H. C. Toh, "Tailoring precision immunotherapy: Coming to a clinic soon," *ESMO Open*, vol. 5, p. e000631, 2020, doi: 10.1136/esmoopen-2019-000631.
- [188] B. Pang *et al.*, "Prognostic role of PIK3CA mutations and their association with hormone receptor expression in breast cancer: A meta-analysis," *Sci. Rep.*, vol. 4, pp. 1–9, 2014, doi: 10.1038/srep06255.
- [189] S. Tan, D. Li, and X. Zhu, "Cancer immunotherapy: Pros, cons and beyond," *Biomed. Pharmacother.*, vol. 124, no. December 2019, 2020, doi:

- 10.1016/j.biopha.2020.109821.
- [190] K. Kiyotani, Y. Toyoshima, and Y. Nakamura, "Personalized immunotherapy in cancer precision medicine," *Cancer Biol. Med.*, vol. 18, no. 4, pp. 955–965, 2021, doi: 10.20892/j.issn.2095-3941.2021.0032.
- [191] Y. Abdou, A. Goudarzi, J. X. Yu, S. Upadhaya, B. Vincent, and L. A. Carey, "Immunotherapy in triple negative breast cancer: beyond checkpoint inhibitors," *npj Breast Cancer*, vol. 8, no. 1, 2022, doi: 10.1038/s41523-022-00486-y.
- [192] I. Schlam *et al.*, "The tumor immune microenvironment of primary and metastatic HER2– positive breast cancers utilizing gene expression and spatial proteomic profiling," *J. Transl. Med.*, vol. 19, no. 1, pp. 1–14, 2021, doi: 10.1186/s12967-021-03113-9.
- [193] L. Buisseret *et al.*, "Tumor-infiltrating lymphocyte composition, organization and PD-1/PD-L1 expression are linked in breast cancer," *Oncoimmunology*, vol. 6, no. 1, 2017, doi: 10.1080/2162402X.2016.1257452.
- [194] Z. Liu, X. Yu, L. Xu, Y. Li, and C. Zeng, "Current insight into the regulation of PD-L1 in cancer," *Exp. Hematol. Oncol.*, vol. 11, no. 1, pp. 1–16, 2022, doi: 10.1186/s40164-022-00297-8.
- [195] M. He, Y. Wang, G. Zhang, K. Cao, M. Yang, and H. Liu, "The prognostic significance of tumor-infiltrating lymphocytes in cervical cancer," *J. Gynecol. Oncol.*, vol. 32, no. 3, pp. 1–16, 2021, doi: 10.3802/jgo.2021.32.e32.
- [196] Y. Wu, M. Yi, M. Niu, Q. Mei, and K. Wu, "Myeloid-derived suppressor cells: an emerging target for anticancer immunotherapy," *Mol. Cancer*, vol. 21, no. 1, p. 184, 2022, doi: 10.1186/s12943-022-01657-y.
- [197] Y. Zhang *et al.*, "Single-cell analyses reveal key immune cell subsets associated with response to PD-L1 blockade in triple-negative breast cancer," *Cancer Cell*, vol. 39, no. 12, pp. 1578-1593.e8, 2021, doi: 10.1016/j.ccell.2021.09.010.
- [198] F. Wu *et al.*, "Signaling pathways in cancer-associated fibroblasts and targeted therapy for cancer," *Signal Transduct. Target. Ther.*, vol. 6, no. 1, pp. 1–35, 2021, doi: 10.1038/s41392-021-00641-0.
- [199] E. Timperi *et al.*, "Lipid-Associated Macrophages Are Induced by Cancer-Associated Fibroblasts and Mediate Immune Suppression in Breast Cancer," *Cancer Res*, vol. 82, no. 18, pp. 3291–3306, 2022, doi: <https://doi.org/10.1158/0008-5472.CAN-22-1427>.
- [200] J. Hao *et al.*, "Expression of adipocyte/macrophage fatty acid-binding protein in tumor-associated macrophages promotes breast cancer progression," *Cancer Res.*, vol. 78, no. 9, pp. 2343–2355, 2018, doi: 10.1158/0008-5472.CAN-17-2465.
- [201] D. Khantakova, S. Brioschi, and M. Molgora, "Exploring the Impact of TREM2 in Tumor-Associated Macrophages," *Vaccines*, vol. 10, no. 6, 2022, doi: 10.3390/vaccines10060943.
- [202] T. Maruhashi, D. Sugiura, I. M. Okazaki, and T. Okazaki, "LAG-3: from molecular functions to clinical applications," *J. Immunother. cancer*, vol. 8, no. 2, 2020, doi: 10.1136/jitc-2020-001014.
- [203] S. S.-A. Natalija Budimir, Graham D. Thomas, Joseph S. Dolina, "Reversing T-cell Exhaustion in Cancer: Lessons Learned from PD-1/PD-L1 Immune Checkpoint Blockade," *Cancer Immunol Res*, vol. 10, no. 2, pp. 146–153, 2022, [Online]. Available: <https://doi.org/10.1158/2326-6066.CIR-21-0515>
- [204] G. J. Bates *et al.*, "Quantification of regulatory T cells enables the identification of

- high-risk breast cancer patients and those at risk of late relapse," *J. Clin. Oncol.*, vol. 24, no. 34, pp. 5373–5380, 2006, doi: 10.1200/JCO.2006.05.9584.
- [205] N. B. Collins *et al.*, "PI3K activation allows immune evasion by promoting an inhibitory myeloid tumor microenvironment," *J. Immunother. Cancer*, vol. 10, no. 3, pp. 1–12, 2022, doi: 10.1136/jitc-2021-003402.

9. SUPPLEMENTARY INFORMATION

APPENDIX 1. Article Stobbe *et al.* (2021) and Supplementary material

Stobbe MD, Thun GA, Diéguez-Docampo A, Oliva M, Whalley JP, Raineri E, et al. (2019) **Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer.** *PLoS Comput Biol* 15(11): e1007496.

RESEARCH ARTICLE

Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer

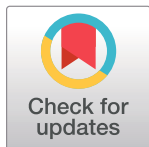
Miranda D. Stobbe¹, Gian A. Thun¹, Andrea Diéguez-Docampo¹, Meritxell Oliva¹^{‡a}, Justin P. Whalley¹^{‡b}, Emanuele Raineri¹, Ivo G. Gut^{1,2*}

1 CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain, **2** Universitat Pompeu Fabra (UPF), Barcelona, Spain

^{‡a} Current address: Institute for Genomics and Systems Biology & Department of Genetic Medicine, University of Chicago, Chicago, IL, United States of America

^{‡b} Current address: Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

* ivo.gut@cnag.crg.eu



OPEN ACCESS

Citation: Stobbe MD, Thun GA, Diéguez-Docampo A, Oliva M, Whalley JP, Raineri E, et al. (2019) Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer. *PLoS Comput Biol* 15(11): e1007496. <https://doi.org/10.1371/journal.pcbi.1007496>

Editor: Jian Ma, Carnegie Mellon University, UNITED STATES

Received: April 30, 2019

Accepted: October 22, 2019

Published: November 25, 2019

Copyright: © 2019 Stobbe et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data underlying our findings and the code for the workflow and sequence logos are provided here: <https://github.com/biomedicalGenomicsCNAG/RecurrentMutations>. The primary data and part of the metadata (tumour type, tobacco smoking history, MSI classification, impact classification, predicted drivers, mutational signatures) can be obtained from the PCAWG consortium through the procedure described here: <http://docs.icgc.org/pcawg/data/>. The IGHV status is available here: <https://www.nature.com/articles/>

Abstract

The sheer size of the human genome makes it improbable that identical somatic mutations at the exact same position are observed in multiple tumours solely by chance. The scarcity of cancer driver mutations also precludes positive selection as the sole explanation. Therefore, recurrent mutations may be highly informative of characteristics of mutational processes. To explore the potential, we use recurrence as a starting point to cluster >2,500 whole genomes of a pan-cancer cohort. We describe each genome with 13 recurrence-based and 29 general mutational features. Using principal component analysis we reduce the dimensionality and create independent features. We apply hierarchical clustering to the first 18 principal components followed by k-means clustering. We show that the resulting 16 clusters capture clinically relevant cancer phenotypes. High levels of recurrent substitutions separate the clusters that we link to UV-light exposure and deregulated activity of POLE from the one representing defective mismatch repair, which shows high levels of recurrent insertions/deletions. Recurrence of both mutation types characterizes cancer genomes with somatic hypermutation of immunoglobulin genes and the cluster of genomes exposed to gastric acid. Low levels of recurrence are observed for the cluster where tobacco-smoke exposure induces mutagenesis and the one linked to increased activity of cytidine deaminases. Notably, the majority of substitutions are recurrent in a single tumour type, while recurrent insertions/deletions point to shared processes between tumour types. Recurrence also reveals susceptible sequence motifs, including TT[C>A]TTT and AAC[T>G]T for the POLE and ‘gastric-acid exposure’ clusters, respectively. Moreover, we refine knowledge of mutagenesis, including increased C/G deletion levels in general for lung tumours and specifically in midsize homopolymer sequence contexts for microsatellite instable tumours. Our findings are an important step towards the development of a generic cancer diagnostic test for clinical practice based on whole-genome sequencing that could replace multiple diagnostics currently in use.

[nature14666#supplementary-information](https://doi.org/10.1371/journal.pcbi.1007496.g001). The tobacco smoking history for a subset of the donors was retrieved from the TCGA webportal (<https://tcga-data.nci.nih.gov/>). The GENCODE annotation v19 used for the functional category was downloaded from the Release history webpage from GENCODE: <https://www.encodegenes.org/human/releases.html>. The replication time data was obtained from the website of University of California, Santa Cruz (UCSC): <http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeUwRepliSeq>.

Funding: We acknowledge the support of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) through the Instituto de Salud Carlos III and the 2014-2020 Smart Growth Operating Program, to the EMBL partnership and co-financing with the European Regional Development Fund (MINECO/FEDER, BIO2015-71792-P - awarded to IGG). We also acknowledge the support of the Centro de Excelencia Severo Ochoa, and the Generalitat de Catalunya through the Departament de Salut, Departament d'Empresa i Coneixement and the CERCA Programme. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare that a European Patent Application relating the described methodology has been filed, and is pending, on behalf of the Center for Genomic Regulation. The authors have declared that no further competing interests exist.

Author summary

Mutations found in the DNA of a tumour are expected to be largely unique to each tumour as there are three billion places in the DNA that can be mutated. However, despite these odds, in a cancer study with 2,583 participants covering 37 tumour types we observe in total over a million non-unique mutations. Based on this observation, we hypothesize that these mutations can be highly informative of the biological processes that caused them. Using characteristics of these non-unique mutations and general statistics like the total number of mutations, we classify the tumours into 16 groups. These groups not only delineate various mutational processes, but also characterize them in more detail. Moreover, we can link the groups to several clinically actionable phenotypes. Our work is a crucial step towards the development of a generic and personalized cancer diagnostic test that only uses the mutations found in the tumour.

Introduction

Mutational processes induced by exogenous sources and/or endogenous mechanisms determine the mutational burden of a cell. They each leave their own genomic fingerprint that differs in terms of the number, types and distribution of mutations. Cancer cells usually show higher mutation rates than normal cells due to elevated cell proliferation and lack of proper DNA repair. The mutations accumulated before, during and after the oncogenic transformation may result in a mutational load exceeding several thousand per cancer genome [1]. Even with such a high burden, the sheer size of the human genome with over three billion bp still makes it improbable that by chance alone identical somatic mutations are found at exactly the same genomic location in two or more cancer patients. Such mutations we will henceforth refer to as being 'recurrent'. Positive selection is one possible explanation for the recurrence of mutations. Recurrent mutations or often more general, recurrently mutated genes and regulatory elements, are used in the prediction of cancer drivers that provide a growth advantage to the cell [2]. However, the number of mutations per cancer genome that so far has been identified as being under positive selection is very small [3, 4] and the discussion on what are sufficient conditions for driver mutations to cause cancer is on-going [5, 6]. Instead of focusing on driver mutations, we hypothesize that recurrent mutations may be highly informative of the non-randomness of mutagenesis and provide a different way to group cancer genomes. In support of this, at both megabase as well as local scale cancer-specific patterns of the non-random distribution of mutations have been well described [7]. For instance, mutation rate is influenced by replication time [8], is linked to epigenomic features [9], shows a periodic pattern around nucleosomes [10], and can depend strongly on the 5' and 3' flanking base as shown in mutational signatures for several mutational processes [11]. This enrichment of mutations in specific genomic regions or sequence contexts increases the probability of recurrence as does the number of mutations per sample, which also varies across mutagenic processes.

We use recurrence as a starting point for a systematic analysis of cancer genomes from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium [12]. This cohort study, brought together by an initiative of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), covers 37 tumour types from 2,583 donors (S1 Table) and is the largest publicly available dataset of its kind. It allows a comprehensive pan-cancer analysis of recurrence in particular since the somatic mutation calling pipeline was identical across all genomes. Moreover, the whole-genome sequencing data that is available for all

donors provides a more complete view than whole-exome sequencing data that so far has been used for large-scale pan-cancer analyses [13]. To make full use of the whole-genome sequencing data and analyse recurrence in an unbiased way, we take here a purely data driven approach that is independent of the completeness and correctness of current genome annotations. Hereby we will focus on Somatic Single-base Mutations (SSMs) and Somatic Insertion/deletion Mutations (SIMs). We first confirm that the number of recurrent mutations is far higher than expected by chance alone and shed light on the relationship between recurrence and the number of samples. Next, we analyse recurrence in the context of general mutational characteristics that capture the effect of mutational processes on the genome. Finally, these general features together with recurrence-related features form the base for clustering cancer genomes in a novel way and determine what recurrence can tell us about mutagenesis. To help interpret the recurrence observed in the 16 identified clusters, link clusters to potential mutational processes and provide further details of each cluster, we use various types of metadata, including tumour type information, driver predictions, and replication time. As a result, we are not only able to refine the mutational consequences of many exposure-specific processes, but also capture clinically relevant phenotypes by using hitherto unused, but easily obtainable mutational features from whole-genome sequences.

Results

Recurrence is higher than expected by chance

There are 1,057,935 recurrent SSMs, which represent 2.44% of the total number of SSMs found in the PCAWG cohort. This is around five times higher (Fig A-I in [S1 Text](#)) than expected if only chance would drive recurrence (based on 5,000 simulations, [S1 Text](#)). For the six SSM subtypes (see [Methods](#)) the observed recurrence is around three (C>G and T>C SSMs) to twelve times (T>G SSMs) higher than expected by chance (Fig A-II in [S1 Text](#)). On tumour type level, we can either determine recurrence by only considering the samples from the same tumour type ('within tumour type') or across all samples ('pan-cancer'). In both cases, Kidney-RCC, Liver-HCC, Lung-AdenoCA and Lung-SCC stand out as the observed number of recurrent SSMs is only around three times (within tumour type) and around two times (pan-cancer) higher than expected by chance (Fig A-III+IV in [S1 Text](#)). In contrast, the largest ratio is 86 times for recurrence 'within tumour type' (Prost-AdenoCA) and 7 times for recurrence 'pan-cancer' (Eso-AdenoCA).

Number of samples does not always correspond to the level of recurrence

To see the effect of the number of samples on recurrence, we look at the overall recurrence within each tumour type ([Fig 1](#)). Although tumour types with more samples generally have a higher total number of recurrent mutations ([Fig 1A](#)), there are notable exceptions. For example, Liver-HCC has the most samples of all tumour types (314), but less recurrent SSMs and SIMs than six and five other tumour types, respectively. If we look at the percentage of recurrent mutations, even more tumour types overtake Liver-HCC as in this respect it ranks 9th and 10th in terms of SSMs and SIMs, respectively ([Fig 1B](#)). The opposite is true for Eso-AdenoCA (97 samples), which has a higher absolute number and percentage of recurrent SSMs than eight other tumour types that have more samples. Stomach-AdenoCA has the highest absolute number and percentage of recurrent SIMs of all tumour types, but less samples than 13 of them. One partial explanation for this is that a lower number of samples does not always translate to a lower total number of mutations ([Fig 1C](#)), even though the correlation is strong (Spearman's Rank correlation coefficient $r_s = 0.73$, $p = 2.8e-07$). However, even if the number of samples and the number of mutations are in line, the level of recurrence may still give a

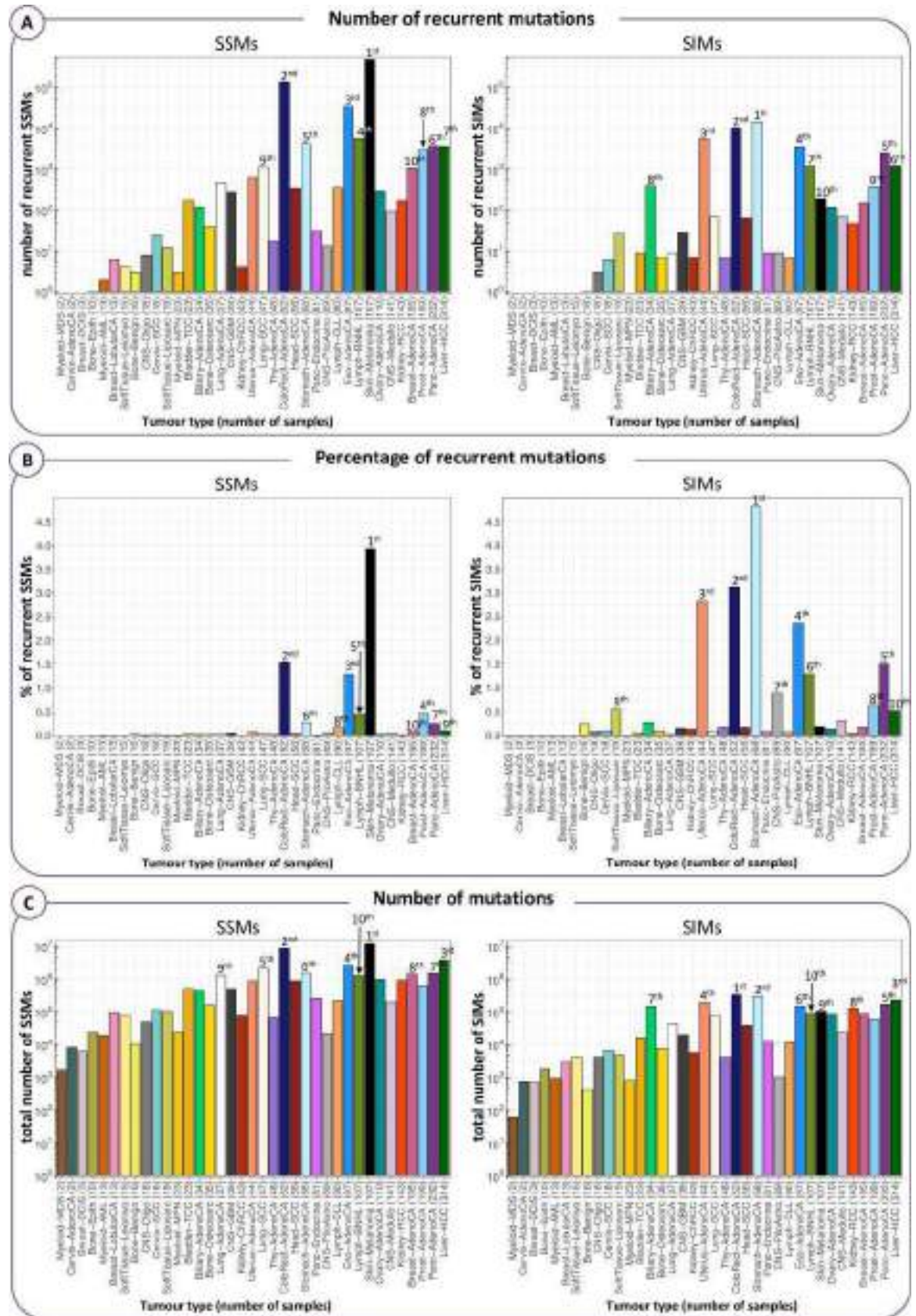


Fig 1. Recurrence within each tumour type in absolute numbers and percentages. The tumour types are ordered from the lowest to the highest number of samples. We labelled the top 10 ranking tumour types in terms of the following three values: (A) Absolute number of recurrent mutations, where recurrence is defined by considering each tumour type separately ('within tumour type' recurrence). (B) Percentage of recurrent mutations 'within tumour type'. (C) Total number of mutations, counting recurrent mutations only once.

<https://doi.org/10.1371/journal.pcbi.1007496.g001>

different picture. Liver-HCC, for instance, has also a higher total mutational load than Eso-AdenoCA ($1.2 \cdot 10^6$ and $7.9 \cdot 10^4$ more SSMs and SIMs, respectively), but still a lower level of recurrence.

General mutational characteristics versus recurrence

For each cancer genome, we compute 29 basic mutational characteristics that capture the effects of mutagenesis (e.g. relative frequency of each SSM subtype) and 13 features capturing recurrence at different levels (Table A in [S1 File](#), see [Methods](#)). Recurrence for these features is determined based on the entire cohort. A detailed description of each of these 42 measures is available in [S1 File](#). Based on the comparison of the recurrence-related features with the general ones ([S2 Text](#)), the key findings are that across the entire cohort: 1) the correlation between mutational load and the absolute level of recurrence is stronger for SSMs ($r_S = 0.89$) than for SIMs ($r_S = 0.76$); 2) the same correlation, but instead taking the percentage of recurrent mutations, is weak and negative for SSMs ($r_S = -0.21$) and non-significant for SIMs; 3) relative recurrence for SIMs is higher than for SSMs; 4) a particularly high percentage of C>T SSMs and 1 bp A/T deletions are recurrent (4.19% and 15.27%, respectively); 5) there is a strong tendency for T>G SSMs to be recurrent despite its modest total number; 6) there is a strong correlation between the level of recurrence for SIMs and the percentage of 1 bp SIMs in a long homopolymer context. Looking into the different tumour types, there are clear contrasts in terms of the associations between general and recurrence-related characteristics. For example, there is a statistically significant positive correlation between the number of mutations and the percentage recurrent for only two tumour types in the case of SSMs (Eso-AdenoCA: $r_S = 0.48$ and Skin-Melanoma: $r_S = 0.58$) and for seven types with respect to SIMs (most notably: Biliary-AdenoCA: $r_S = 0.71$ and Eso-AdenoCA: $r_S = 0.67$) (Fig D in [S2 Text](#)).

Recurrence characteristics divide the cohort

Next, we use the recurrence-based and general mutational features all together to see if we can uncover meaningful clusters of cancer genomes. As there are strong correlations between some of these features ([Fig 2](#)), we first perform a principal component analysis (PCA) to obtain independent features and reduce dimensionality ([Fig 3](#)). We take as many principal components (PCs) as needed to explain at least 80% of the variance in the data and consider the remaining PCs to capture noise. We use this subset of PCs as input for hierarchical clustering [[14](#)]. The resulting hierarchical tree is cut at the desired height to obtain clusters. The centroids are computed for each cluster and used as input to the k-means consolidation step, which further improves the initial clustering (see [Methods](#)) [[15](#)]. To get a pan-cancer perspective we analyse all samples together.

The crude division into two clusters separates the cancer genomes with low relative levels of recurrent SIMs (e.g. Liver-HCC, Kidney-RCC and Lung-SCC) from those with high levels (e.g. ColoRect-AdenoCA, Eso-AdenoCA, Lymph-BNHL and Panc-AdenoCA) ([S1 Fig](#)). At three clusters, the relative level of recurrent SSMs splits off a group of mainly Skin-Melanoma samples from the two other clusters. This cluster largely remains unchanged when increasing the number of clusters while the two other clusters continue to divide and become more specific to a tumour type or a particular mutational process. At the level of six clusters, for example, we

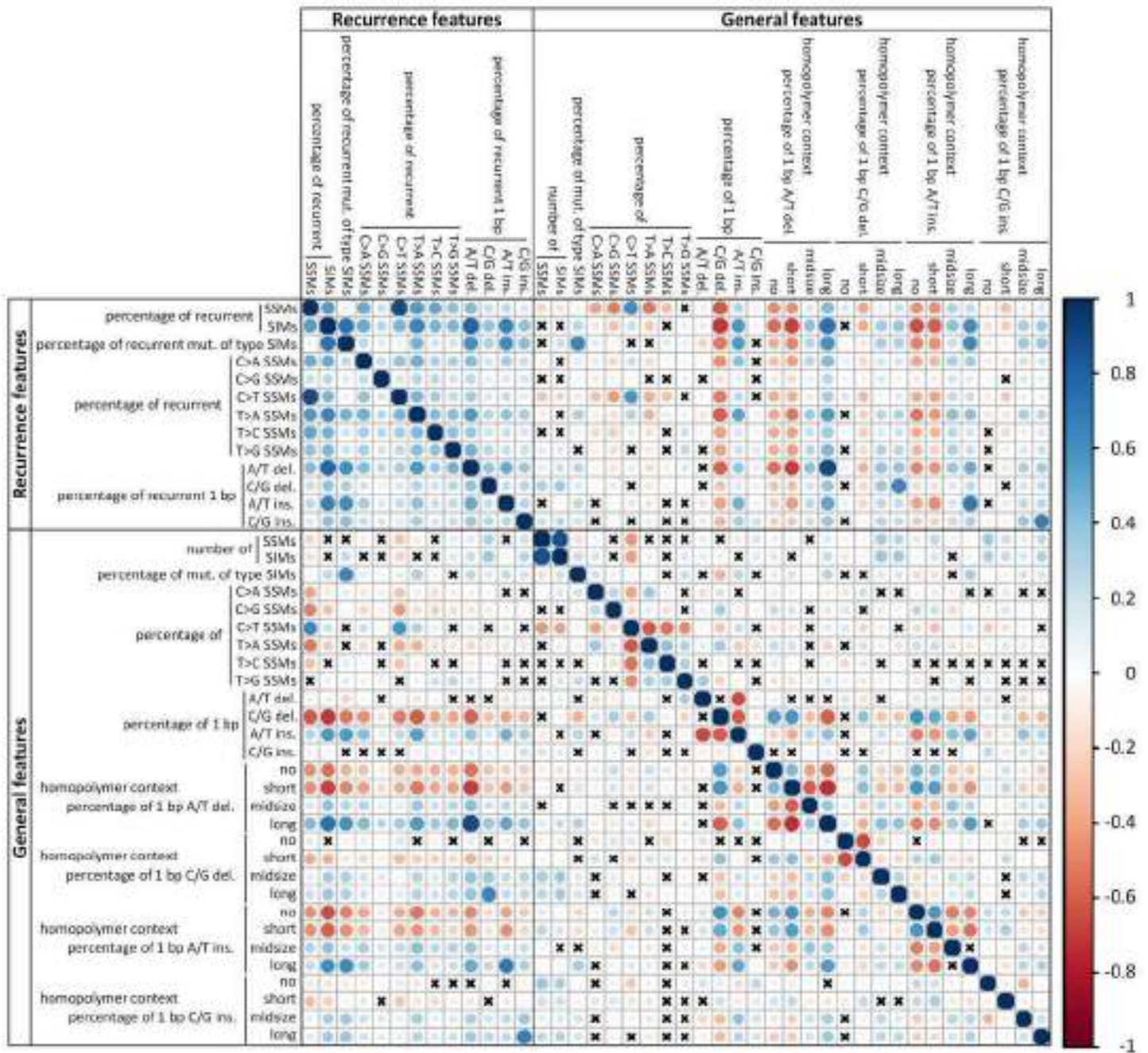


Fig 2. Spearman's rank correlation between the 42 mutational features. The colour of the circles indicate positive (blue) and negative (red) correlations, colour intensity represents correlation strength as measured by the Spearman's rank correlation coefficient. The size of the circle indicates the adjusted p-value with larger circles corresponding to lower p-values. The p-values were corrected for multiple testing using the Benjamini-Yekutieli method. Crosses indicate that the correlation is not significant (adjusted p-value > 0.05).

<https://doi.org/10.1371/journal.pcbi.1007496.g002>

see a cluster split off that we can connect to microsatellite instability (MSI). We will discuss in further detail the division into 16 clusters, chosen as a trade-off between too many clusters, which would each be specific to just a handful of samples, and too few, which would result in loss of meaningful information (Fig 4). There are nine clusters (A, B, C, G, H, I, L, M and P)

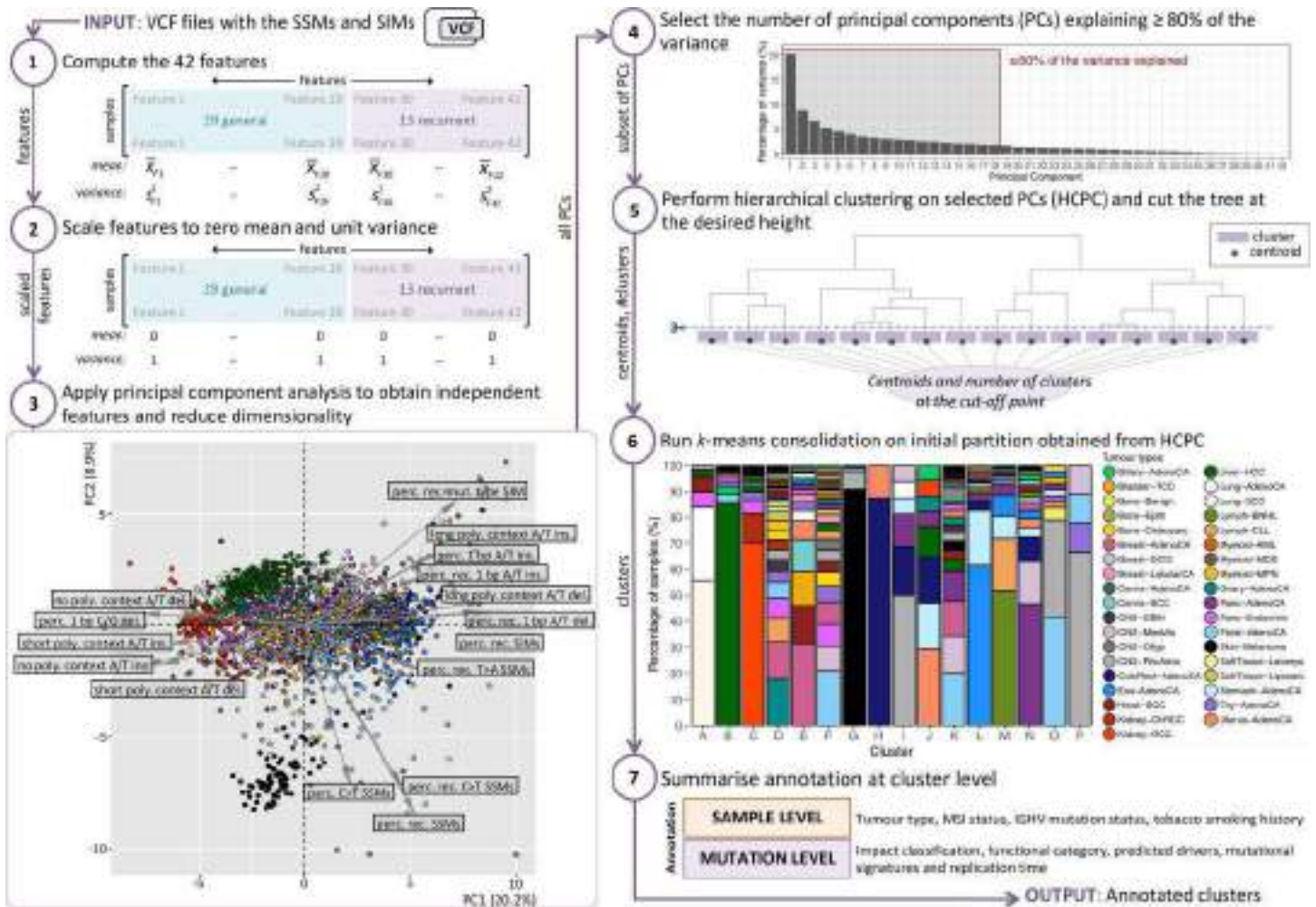


Fig 3. Workflow of the recurrence-based approach to group cancer genomes. The 42 features are described in detail in S1 File (Step 1). We scale all features to zero mean and unit variance to compensate for the differences between the ranges of the features (Step 2). The arrows in the PCA plot indicate the direction and level of contribution of the features that contribute above average to the first two PCs (Step 3). Seven of these features are related to recurrence. An interactive 3D version of the PCA plot is available here: <https://plot.ly/~biomedicalGenomicsCNAG/1.embed>. We take a subset of the PCs and consider the remaining PCs to capture noise (Step 4). For the hierarchical clustering we use the Euclidean distance as a dissimilarity measure and Ward’s method as the linkage criterion (Step 5). The results of the hierarchical clustering are used as a starting point for k-means clustering (Step 6). Some samples will in this step switch to a different cluster compared to the initial partition. This consolidation step is repeated a maximum of 10 times. Further details on the annotation of the clusters (Step 7) are described in S3 Text.

<https://doi.org/10.1371/journal.pcbi.1007496.g003>

for which at least half of the samples are from the same tumour type. For another two clusters (O and N) samples from two tumour types constitute a majority. In the remaining five clusters (D, E, F, J and K) three or more tumour types are required for this. For each tumour type the percentage of samples in each of the 16 clusters is shown in S2 File. The association of each of the 42 features with the clusters is shown in Fig 5. The key characteristics of each cluster are shown in Fig 4. To facilitate a tight linkage of the clusters to mutational processes, we consider, in addition to the mutational features of a cancer genome, also tumour type assignment, microsatellite instability (MSI) status, immunoglobulin heavy-chain variable region gene (IGHV) mutation status (Lymph-CLL only) and tobacco smoking history of the donor (where available) (S3 Text). To provide further details on each cluster we integrate annotation based on GENCODE [16], Oncotator [17], driver predictions [3, 18], replication time [19] and mutational signatures [20]. A summary of this and further details are described in S3 Text. In the

Cluster	Main tumour type(s)	Median number of		Num. of samples	Key characteristic(s)	Association with overall recurrence		Possible causative agent or mechanism
		SSMs	SIMs			SSMs	SIMs	
A	Lung-SQC	44,910	1,634	60	high % of • C>A SSMs • 1 bp C/G deletions	negative	negative	- tobacco smoke - increased activity of cytidine deaminases
B	Liver-HCC	11,046.5	575	324	high % of T>C SSMs	negative	negative	alcohol intake
C	Kidney-REC	4,491	576	196	high % of 1 bp A/T insertions in no or short homopolymer context	negative	negative	10 samples; aristolochic acid
D	Ovary-AdenoCA, Breast-AdenoCA, Lymph-CLL, Panc-Endocrine, Prost-AdenoCA	3,484.5	189.5	502	• low % of 1 bp A/T insertions • high % of 1 bp C/G deletions	negative	negative	unknown
E	Breast-AdenoCA, Head-SQC, Bladder-TCC	10,426	362.5	98	high % of C>G SSMs	negative	no	increased activity of cytidine deaminases
F	Prost-AdenoCA, CNS-Medullo, Panc-Endocrine, Breast-AdenoCA, Thy-AdenoCA	2,389	144	95	high % of • 1 bp C/G insertions in a long homopolymer context • recurrent 1 bp C/G insertions	no	no	unknown
G	Skin-Melanoma	89,002	647	87	high % of • C>T SSMs • recurrent SSMs • recurrent C>T SSMs	positive	negative	UV-light
H	ColoRect-AdenoCA	822,314	9,160	8	high number of SSMs	positive	no	deregulated activity of polymerase ϵ
I	CNS-PiloAstro	125	5	16	high % of 1 bp C/G insertions • high number of SIMs	no	no	unknown ¹
J	Uterus-AdenoCA, Stomach-AdenoCA, ColoRect-AdenoCA	55,789	30,228	17	high % of • mutations of type SIM • 1 bp C/G deletions in a midsize homopolymer context	no	positive	microsatellite instability
K	Prost-AdenoCA, CNS-Medullo, Breast-AdenoCA, Panc-AdenoCA	3,155.5	281.5	522	high % of 1 bp A/T insertions	no	positive	unknown
L	Eso-AdenoCA	24,906	1,446.5	104	high % of • T>G SSMs • recurrent T>G SSMs	positive	positive	gastric acid (reflux)
M	Lymph-BNHL	7,065	416.5	184	high % of • recurrent 1 bp A/T deletions • 1 bp A/T deletions in a long homopolymer context • recurrent C>G SSMs	positive	positive	hypermutation of the immunoglobulin genes
N	Panc-AdenoCA, CNS-Medullo	4,993	542	311	high % of • 1 bp A/T deletions in a long homopolymer context • recurrent 1 bp A/T deletions • recurrent mutations of type SIM	positive	positive	unknown
O	Prost-AdenoCA, CNS-PiloAstro	182	11	43	high % of recurrent T>C, C>G & T>A SSMs	positive	positive	unknown ¹
P	CNS-PiloAstro	118	13	9	high % of • recurrent 1 bp C/G deletions • 1 bp A/T deletions in a long homopolymer context	positive	positive	unknown ¹

Fig 4. Key characteristics of the 16 clusters. Tumour types that form together $\geq 50\%$ of the cluster are listed. The legend for colours for the pie chart is provided in Fig 3. The key characteristics are those features with the strongest significantly negative or positive association with the cluster. Only if the association with overall recurrence is significant, the respective direction is indicated. ¹Cluster has a low median number of SSMs (<200) and SIMs (<20).

<https://doi.org/10.1371/journal.pcbi.1007496.g004>

following sections we will show how the level of recurrence can be indicative of the mutational processes, often in combination with the general features. Moreover, we show that our recurrence-based approach groups cancer genomes in a novel way that complements current classification approaches and captures clinically actionable cancer phenotypes.

High levels of recurrent SSMs and low levels of recurrent SIMs characterize exposure to UV light

A positive association with overall recurrence of SSMs and more specifically with recurrence of C>T SSMs characterizes cluster G that mainly consists of Skin-Melanoma samples (Fig 5). The association is negative with the recurrence of SIMs. We link this cluster to mutagenesis induced by UV light (S3 Text). The samples assigned to cluster G account by themselves for 60.7% of the total number of recurrent C>T SSMs. The combination of the high total number of SSMs per sample and the high percentage of C>T substitutions in this cluster is what contributes to the high level of recurrence. The mechanisms inherent to UV-light exposure further increase the probability of recurrence as it tends to result in C>T SSMs near energy sinks in the genome. The energy from UV-light-exposed DNA usually travels along the DNA strand

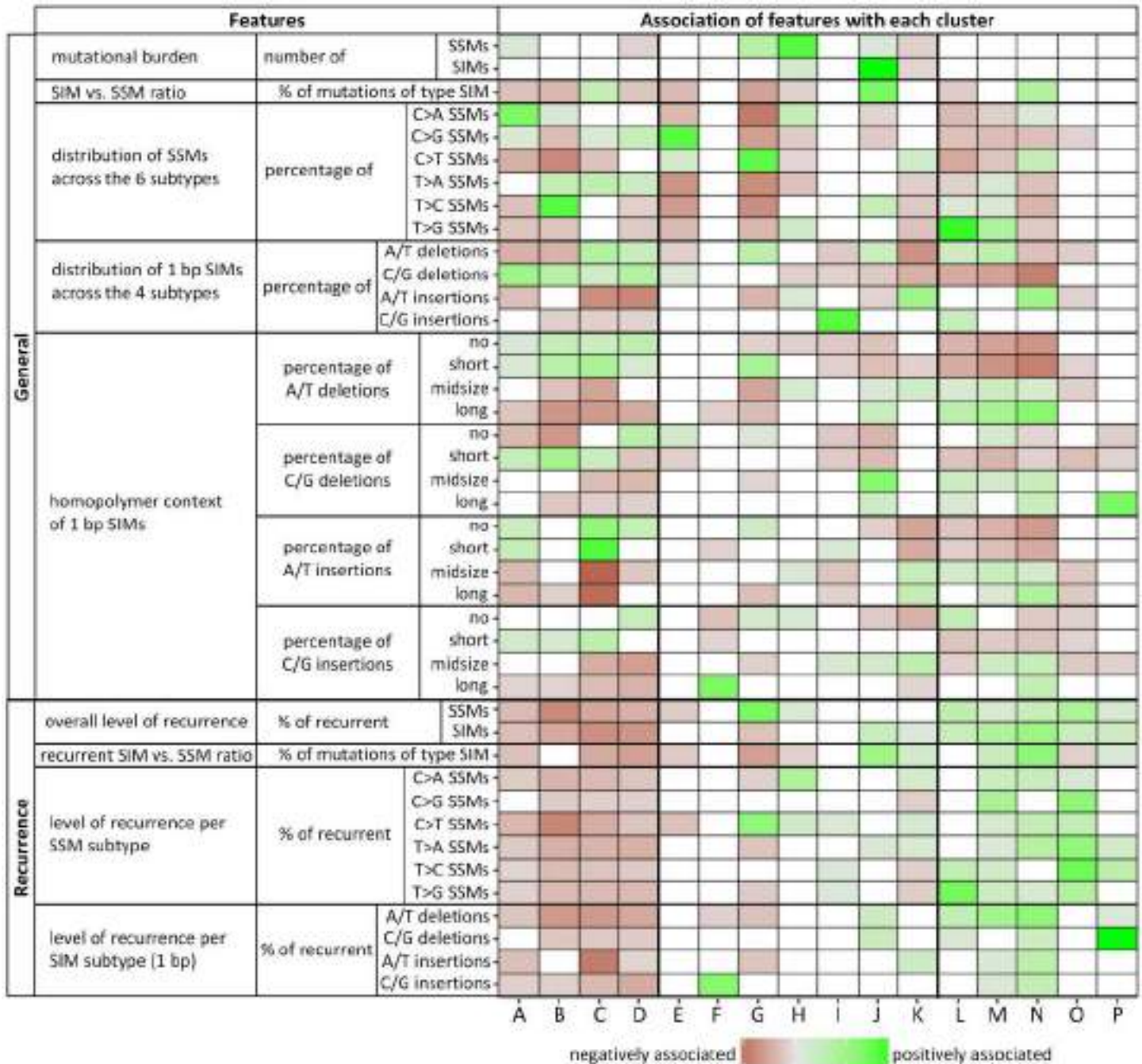


Fig 5. Overview of the 42 features and their association with each cluster. Red and green squares indicate statistically significant negative and positive associations, respectively, where the gradient indicates the strength of the association. White coloured squares indicate no significant association (adjusted p-value > 0.05). For deletions a 'no homopolymer context' means that the base next to the deleted one is not of the same type. For insertions this refers to a base inserted 5' to either a base of a different type or a single base of the same type. Note that we do not have to consider preceding bases as all SIM calls were left aligned. A short homopolymer context is defined as a 2–4 bp mononucleotide repeat of the same type of base as the 1 bp SIM, midsize is 5–7 bp in length and long ≥ 8 bp.

<https://doi.org/10.1371/journal.pcbi.1007496.g005>

until it arrives at the lowest energy point, a dT, particularly when it is next to a dC, which acts as energy barrier [21]. In agreement with this, for C>T mutations that are recurrent within this cluster there is a strong enrichment of a TTTCCCT motif (the underlined C is mutated)

(see [Methods](#)). While the percentage of this motif in the genome is estimated to be only 0.4% of all 6-mers with a C at the central position, 4.5% and 19.5% of the non-recurrent and recurrent C>T SSMs, respectively, within this cluster are at this motif ([Fig 6](#)). An enrichment of a CTTCCG motif was found for frequently recurrent mutations in promoters in 38 melanoma samples [22]. In another set of 184 melanoma samples a CTTCCGG motif was found at the majority of ETS transcription factor binding sites (TFBSs) [23]. As the sequences are centred at the core consensus ETS binding motif TTCC, instead of at a mutation, the underlined nucleotide is the most frequently mutated base. In a subset of highly mutable ETS TFBSs the second C is the most mutated. These and our specific sequence motif help explain the observed high level of recurrence. Furthermore, a decreased activity level of the nucleotide excision repair pathway was detected in melanoma at active transcription factor binding sites and nucleosome embedded DNA compared to the flanking sites [24]. This increases local mutation rates and hence also increases the probability of recurrence.

High levels of recurrent SSMs characterize deregulated activity of POLE

A high level of recurrent SSMs also characterizes cluster H, specifically C>T and C>A SSMs. This cluster captures samples that can be considered ultra-hypermutators and their mutations are mainly caused by deregulated activity of POLE ([S3 Text](#)). These samples have a very high total number of C>A SSMs (median: 297,750) and the median percentage of recurrent C>A SSMs across the samples is 7.7%. Two thirds of all recurrent C>A SSMs in the entire cohort are also recurrent within only this cluster. The C>A mutations that are recurrent within this cluster are enriched for the motif TTCTTT, when considering only ungapped motifs ([Fig 6](#), see [Methods](#)). Of the recurrent C>A SSMs 32.2% are at this motif, while for non-recurrent ones this is true for only 13.7% (χ^2 test: $p < 2.2e-16$). In the genome, the estimated percentage of this motif of all corresponding 6-mers (NNCNNN) is far smaller (0.6%), suggesting that effects of deregulated activity of POLE are most likely dependent on a sequence context exceeding a single neighbouring base on each side as also observed for whole-exome data by Martincorena *et al.* [25].

High levels of recurrent SIMs characterize microsatellite instability

The highest level of recurrent SIMs across all clusters is observed for cluster J, which could be linked to a defective mismatch repair (MMR) pathway resulting in MSI ([S3 Text](#)). Of the 179,691 recurrent 1 bp SIMs in the entire cohort, almost half of them are recurrent when only considering this cluster. The very high median number of SIMs (30,228) in this cluster may play a role in the high level of recurrence. The key factor, however, is most likely the mutational process behind MSI, which is slipping of the DNA polymerase during replication of repetitive sequences and the lack of repair by the MMR pathway [26]. This not only explains the elevated number of SIMs [27], but also the association of this cluster with all SIM subtypes in the context of midsize-to-long homopolymers. As such homopolymers are scarce in the genome, the shift towards specifically altering them increases the probability of recurrence (Table F in [S2 Text](#)). Especially striking in this cluster is the proportion of 1 bp C/G deletions that are in the context of a midsize homopolymer (median: 73.2% vs. 8.4% for the other clusters combined, $p = 1.2e-12$). This translates to 6.0% recurrent 1 bp C/G deletions within this cluster versus <0.7% for any other cluster ([S3 Text](#)).

Positive association with recurrence of SSMs and SIMs: Gastric-acid exposure and hypermutation of immunoglobulin genes

Clusters L, M and N all positively associate with recurrence of both SSMs and SIMs. Cluster L, which for >80% consists of Eso-AdenoCA and Stomach-AdenoCA samples, can potentially be

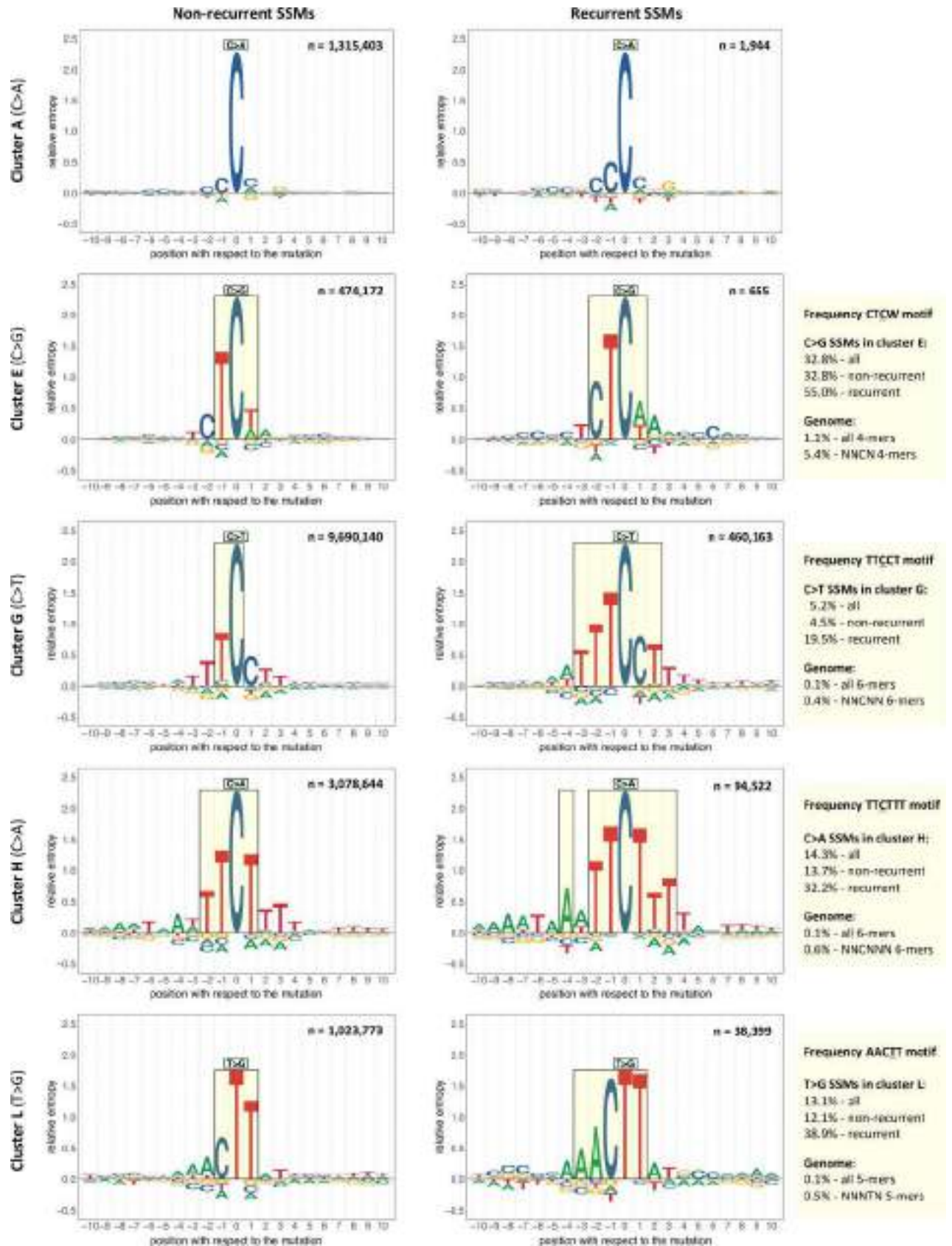


Fig 6. Enriched sequence motifs. The sequence logos represent the sequence context of ten bp 5' and 3' of the non-recurrent (left-side) or recurrent (right-side) mutations of the indicated cluster and SSM subtype. Here recurrence is defined as a mutation at the same genomic location in two or more samples from the same cluster. Each recurrent SSM is included only once to avoid giving extra weight to highly recurrent mutations. Relative entropy is used as a measure of information content (see [Methods](#)). Setting a threshold of 0.25 for the relative entropy results in the motifs highlighted in the rectangles. In the upper right corner of each sequence logo the number of mutations is indicated. To the right of the sequence logos are the percentages in which the enriched motif found for the recurrent SSMs is present in context of the mutations in the cluster and the corresponding k-mers in the genome (N = A, C, G or T). The enrichment for the motif for recurrent SSMs is in all four cases significantly higher than for the non-recurrent SSMs (χ^2 test: $p < 2.2e-16$).

<https://doi.org/10.1371/journal.pcbi.1007496.g006>

linked to gastric-acid exposure ([S3 Text](#)). The T>G and T>C SSMs that are recurrent within this cluster cover 45% and ~20%, respectively, of the total observed in the whole cohort. The median percentage of SSMs falling in late-replicating regions (Table C and Fig A in [S3 Text](#)) is significantly higher than in the rest of the clusters combined (75.2% vs. 61.0%, $p < 2.2e-16$). In general, the mutational load is expected to be higher in late-replicating regions as the MMR pathway is said to be less efficient there [28]. However, the question is why the effect is so strong in cluster L compared to the others (Fig B in [S3 Text](#)). It could be that transient single strand-DNA at stalled replication forks, whose formation has been suggested to be more prevalent in late-replicating regions [29], is particularly vulnerable to the mutagenicity of acid-exposure. Alternatively, if the oxidative stress induced by gastric-acid exposure leads to the oxidation of dG in the dNTP pool [30], the use of error-prone DNA polymerases that incorporate the oxidized dG into the DNA [31] may be more frequent in late-replicating regions [32]. The strong shift towards late-replicating regions favours higher levels of recurrence. The same holds for the enrichment of the specific sequence context 'AACTT' that we observe for T>G mutations that are recurrent within this cluster (Fig 6, see [Methods](#)). Nearly 39% of the recurrent T>G SSMs are confined to this motif and ~12% of the non-recurrent ones (χ^2 test: $p < 2.2e-16$), which is still far higher than the estimated percentage of this motif in the genome (0.5% of all NNNTN 5-mers). For SIMs, the cluster has a positive association with recurrence for three out of the four SIM subtypes as well as with the same subtypes in a midsize and/or long homopolymer context. This suggests similar mechanisms as for cluster J. Finally, as observed for SSMs in this cluster, SIMs also show a tendency to fall into late-replicating regions (67.2%, Table C and Fig C in [S3 Text](#)). This may further add to the high level of recurrence for SIMs.

Cluster M, with mainly Lymph-BNHL and Lymph-CLL samples, is linked to the somatic hypermutation of the immunoglobulin genes ([S3 Text](#)). In the aforementioned tumour types, this process is indicative of memory B cells being the cell of origin as opposed to naïve B cells [33]. The cluster has positive associations with the level of recurrence for all six SSM subtypes. The association is particularly strong for C>G. Of all recurrent C>G SSMs, 10.7% can be found in this cluster alone. The high level of recurrence may partially be explained by the hypermutation observed in the limited area of the genome where the immunoglobulin genes are located. For SIMs, the cluster has positive associations with the level of recurrence for all four subtypes as well as with those subtypes in general when in a midsize and/or long homopolymer context. This cluster has the highest median percentage of SIMs in late-replicating regions (67.5% vs. 57.8% for the other cluster combined, $p < 2.2e-16$, Table C and Fig C in [S3 Text](#)), which may contribute to the high level of recurrence.

In cluster N, which consists of ~47% Panc-AdenoCA samples, the sources of mutagenesis are less clear, even after the inclusion of all annotation layers ([S3 Text](#)). Except for C>G and T>C SSMs, the cluster has positive associations with the recurrence of all other subtypes of SSMs and every SIM subtype. This is especially noticeable as the median of the total number of mutations across samples is intermediate. A high percentage of the recurrent mutations are SIMs in this cluster with a median of 35.0%. This is far higher than for samples of the other

clusters combined (median: 15.5%, $p < 2.2e-16$). The positive associations with all SIM subtypes when in a midsize-to-long homopolymer context may point to a slippage-related mechanism (see also cluster J).

Negative association with recurrence: Tobacco-smoke exposure, alcohol use and increased activity of cytidine deaminases

There are also several mutagenic processes that are associated with low levels of recurrence (Fig 5) including those represented by clusters A, B, C and E. Cluster A, of which 84% are lung cancer samples, is linked to mutational processes induced by tobacco-smoke exposure (S3 Text). This cluster shows a positive association with the total number of SSMs and the percentage of C>A SSMs, the latter is a known consequence of tobacco-smoke exposure [34]. There are several factors that increase the probability of recurrence in this cluster, including the high total mutational load together with the high percentage of C>A SSMs and the enrichment of mutations in late-replicating regions (S3 Text). Also, tobacco-smoke induced mutations have been shown to be enriched in linker DNA (*i.e.* DNA not wrapped around a nucleosome) [10], which constitute only between 10% and 25% of the genome in eukaryotes [35]. The key to explaining the lack of recurrence seems to be that there is little tendency to favour a specific sequence context for the C>A SSMs (Fig 6). This can also be observed in the ‘tobacco smoking signature’ [11], which is present in nearly 90% of the samples in this cluster (S3 Text). Unlike for several clusters mentioned above, there is a positive association with SIMs in short homopolymer contexts, which are more frequent in the genome than longer homopolymers, and the resulting distribution is therefore also more random. Note that cluster A also has a strong association with the percentage of total 1 bp C/G deletions, which has not been described previously as a possible consequence of tobacco-smoke exposure (S3 Text and S4 Text).

Cluster B, consisting of 85% Liver-HCC samples, is likely to be linked to mutational processes indirectly induced by excessive alcohol use (S3 Text). The level of recurrence is low despite the high number of samples of the same tumour type (277) and the consistent pattern of a high percentage of T>C SSMs (median: 31.7% vs. 14.6% in the other cluster combined, $p < 2.2e-16$). With regard to 1 bp SIMs, there is a positive association with a short homopolymer context, as for cluster A, with the exception of 1 bp A/T insertions.

In cluster C, in which ~82% are Kidney-RCC and Kidney-ChRCC samples, the mutational processes remain largely obscure except for a few samples that can be connected to aristolochic-acid exposure (S3 Text). Unlike for clusters A and B, the median number of SSMs across samples is relatively low. Furthermore, mutations are nearly equally spread between early- and late-replicating regions as only 53.9% of the SSMs and 47.5% of SIMs are in late (Table C, Figs B and C in S3 Text). SIMs are preferentially located in no or short homopolymer context, similar to clusters A and B.

In cluster E nearly one third are Breast-AdenoCA samples and key mutational characteristics point to the endogenous mutational process of increased activity of cytidine deaminases (S3 Text). There is a general paucity of associations with characteristics of recurrence. In line with this, the mutations in this cluster are nearly equally spread between early- and late-replicating regions of the genome (Table C, Figs B and C in S3 Text). The most outstanding feature of this cluster is the high percentage of C>G SSMs. This is the rarest substitution type, making the detection of recurrence unlikely, particularly if not confined to specific genomic regions. Interestingly though, the 655 C>G SSMs that are recurrent within this cluster are enriched for the motif CTCW (W = A or T) (Fig 6, see Methods). Very similar motifs have been described as being characteristic for deamination mediated by APOBEC3 [36]. The number of recurrent mutations is much lower than for the other motifs discussed. The CTCW motif is also shorter,

more general and therefore relatively frequent in the genome (5.4% of all NNCN 4-mers), all possible causes for the lacking trend towards recurrence.

The added value of the recurrence-related features

The PCA shows that seven of the sixteen features that contribute above average to the first two PCs are related to recurrence (Fig 3). In addition, all 16 clusters have a statistically significant association with two or more recurrence-related features (Fig 5). The importance of the recurrence-related features is further demonstrated by the results of running the entire workflow (Fig 3) using only the general features. In this case we are no longer able to separate all ultra-hypermutable samples from the rest of the cohort (S2 Fig). Furthermore, the cluster linked to hypermutation of the immunoglobulin genes (cluster M) is dissolved, and the cluster possibly linked to gastric-acid exposure (cluster L) is less cancer-specific as it absorbs 90 samples of the dissolved cluster M and thereby nearly doubles in size. Another key difference is that only ~55% of the Lymph-CLL samples without hypermutation of the immunoglobulin genes are confined to a single cluster as opposed to ~86% when using all features.

Discussion

Only a very small percentage of the 1,057,935 recurrent SSMs and 186,576 recurrent SIMs in the PCAWG cohort are expected to be purely by chance. We estimate based on simulations that only around 0.47% of the SSMs would be recurrent if no biological factors would play a role, which is less than one fifth of the observed 2.44%. Technical artefacts could contribute to the level of recurrence, but although they can never be fully excluded, the PCAWG consortium has made a great effort to minimise false positive calls. A consensus was taken of the individual results from multiple somatic mutation callers, followed by the application of various filters to remove, *e.g.*, germline variants [12] (see Methods). This resulted in a conservative, but reliable dataset of somatic mutations. Increasing the size of the cohort may change the percentage of recurrent mutations, but in which direction depends on the tumour type of the additional samples, their mutational burden and importantly the mutational processes underlying the observed mutations.

Recurrence is considered an important indication that a mutation might be under selective pressure in protein-coding regions [37, 38]. Hence, by focusing on recurrence we are inherently not only looking at the mutational consequences of mutational and repair processes, but also at positively selected mutations. One way that has been used to reduce the influence of the latter is to count all recurrent mutations only once [39]. However, in our approach, as we describe each individual cancer genome with the 42 features, this is not an option as we would not know to which samples to add this single count for each recurrent mutation. Instead, we would need to leave out all recurrent mutations, but this would even be more rigorous. In either case, it also implies that over a million mutations are assumed to be under positive selection. Besides the fact that recurrence is not a sufficient condition for positive selection [37], it may not even be a necessary one in a dataset of the size of our cohort [3, 38]. Another option is to remove all predicted driver mutations. In total there are only 4,223 predicted driver mutations that are either SSMs or SIMs, which constitutes just 0.009% of the total amount of mutations. It is, therefore, unlikely that leaving them out will affect the general features. Their effect on the percentage of overall recurrence is also negligible (-0.001% for SSMs and +0.002% for SIMs), partly because only ~12% of the predicted driver mutations are recurrent within the PCAWG cohort. Based on the overall statistics, removing the predicted driver mutations will also hardly affect the recurrence-related features of individual cancer genomes and consequently not result in any noticeable change in the uncovered clusters. As identifying the driver

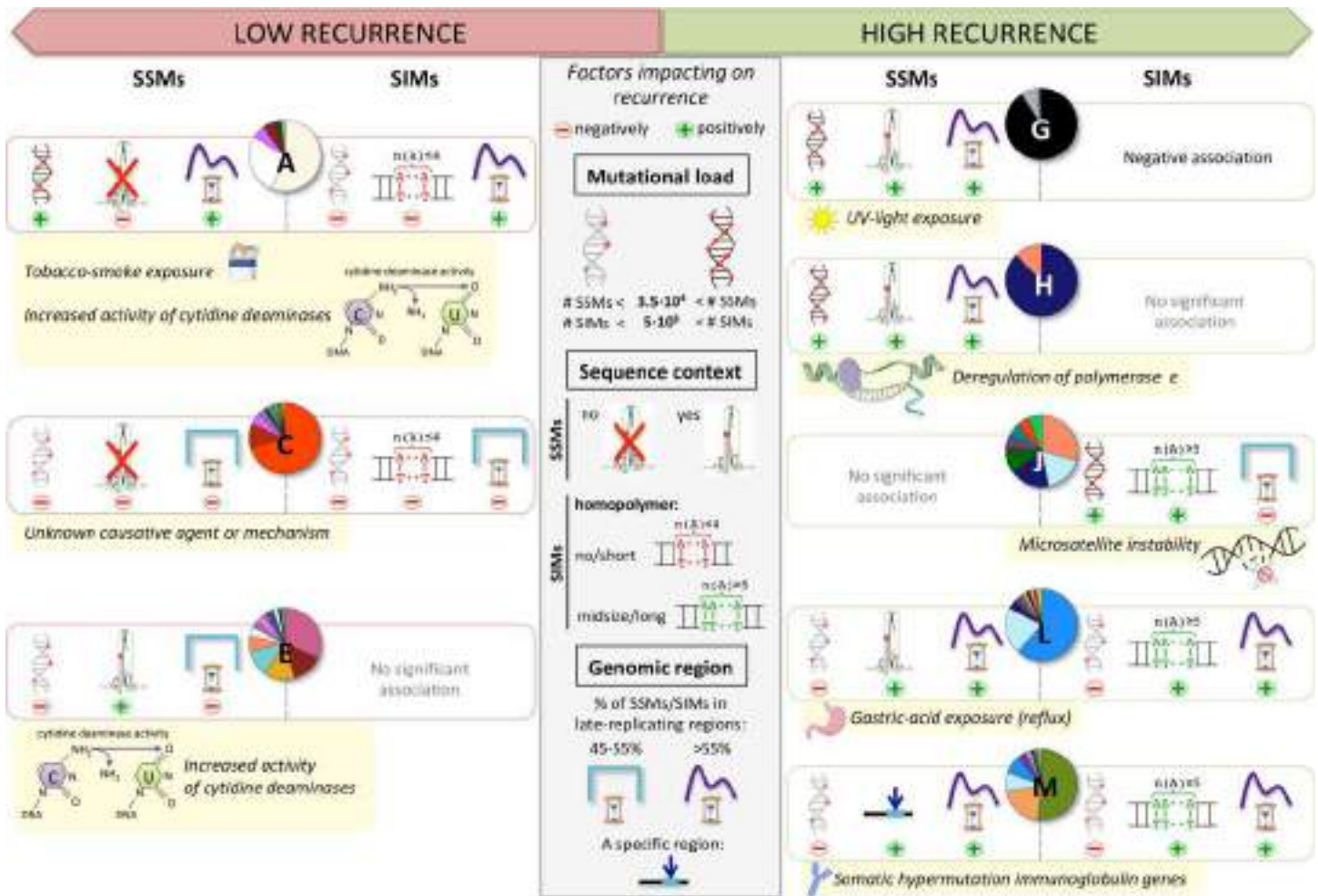


Fig 7. Factors impacting on recurrence in the context of the clusters. None of the three key factors (middle panel) that impact on recurrence individually explain the observed level of recurrence in the clusters. Whether a cluster has a relatively high or a comparatively lower mutational load is based on the median number of SSMs/SIMs across its samples (Fig 4). The actual specific sequence contexts for SSMs are shown in Fig 6. For cluster M there is enrichment for a specific sequence context as well, which is AGCT for C>G SSMs that are recurrent within this cluster (n = 949) (S3 Fig). For SIMs a homopolymer of A/T's is used to represent any type of homopolymer. Clusters A and C have a positive association to no and/or short homopolymer context for all types of 1 bp SIMs (red), while for clusters J, L and M this is the case for midsize and/or long homopolymer context (green) (Fig 5). For the replication time region we compute the percentage of SSMs/SIMs that are in late-replicating regions (S3 Text). If this percentage is between 45–55%, then we consider the mutations to be nearly equally spread between early- and late-replicating regions of the genome. The specific region that is enriched in cluster M refers to the immunoglobulin genes. The recurrence in clusters A and G is also likely to be positively impacted by an increased mutation rate in a specific region as the majority of their samples are from a particular tumour type for which this has been reported. For lung cancer (cluster A) the mutation rate is increased in linker DNA [10] and for Skin-Melanoma (cluster G) at active transcription factor binding sites [24].

<https://doi.org/10.1371/journal.pcbi.1007496.g007>

mutations is, in addition, far from unambiguous and a dynamic area of research [3, 18], it is of limited practicality to our workflow to remove them. Of note, the impact of positive selection might be greater when analyzing only the exome [39] as there are less mutations in total and the large majority of drivers is found in protein-coding loci [3, 18].

Mutational load, enrichment of mutations in a specific sequence context or in specific parts of the genome all impact on recurrence. However, none of these factors provide individually a universal explanation for the observed levels of recurrence per cluster (Fig 7). For example, the cluster linked to tobacco-smoke exposure has a very low percentage of recurrence, despite the high mutational load, the enrichment of mutations in late-replicating regions and increased mutation rate in linker DNA. The absence of a preferred sequence context likely plays a role in

this. The short and non-specific motif found in samples with increased activity of cytidine deaminases (CTCW) is also not sufficient by itself to result in high levels of recurrence. For causative agents like UV light and deregulated activity of POLE, however, the high total number of mutations combined with the observed 6 bp specific sequence context does lead to high levels of recurrent SSMs. For the cluster linked to gastric-acid exposure, the number of SSMs is much lower than for the clusters linked to those two agents or tobacco-smoke exposure. Nevertheless, it has a high level of recurrence, likely because of the 5 bp sequence motif for T>G SSMs and the three times higher occurrence of SSMs in late-replicating regions than in early. One possible caveat here is that replication timing is a process with very high plasticity across cell types [19], and taking the median timing across the available five cancer cell lines (S3 Text) may individually lead to non-adequate interpretations. A typical example for the potential impact of an elevated local mutation rate on the proportion of recurrence is the hypermutation of the immunoglobulin genes in memory B cells. As mutations detected in several lymphoma samples are largely confined to those genes, their modest total number of mutations still results in a high relative level of recurrence. Finally, in the case of the MSI samples, the slippage of the DNA polymerase during replication of repetitive sequences, combined with a lack of repair capacity results in a high percentage of SIMs in a midsize-to-long homopolymer context. This coincides with a high level of recurrence for SIMs, despite the relatively equal distribution of SIMs between early- and late-replicating regions that we observe and that has been reported before [28]. Associations with the much more frequent short homopolymers do not translate into high level of SIM recurrence, not even in the case of a high number of total SIMs (e.g. as observed in the ‘tobacco-smoke exposure’ cluster). The effect of the sequence context may be stronger for SIMs than for SSMs. This would explain the ~3.6 fold higher percentage of recurrent SIMs (8.69%) versus SSMs (2.44%), despite the fact that there are 20 times more SSMs. Unlike for SSMs, the actual position of an insertion/deletion in a homopolymer cannot be determined, contributing to loss in resolution and a higher likelihood of recurrence. In summary, we infer that the non-randomness in the distribution of mutations strongly depends on the causative agent. Consequently, recurrence is generally able to cluster the genomes in a way that shows clear associations with tumour type assignments and mutational processes. For SSMs 60.0% is only recurrent in one particular tumour type, while for SIMs this percentage is 10.7% (S2 Table). This suggests a higher resemblance of mutational patterns within tumour types for SSMs than for SIMs. In contrast, 79.8% of the recurrent SIMs (versus 37.1% for SSMs) can only be detected in a pan-cancer approach, pointing to shared mutational processes which allow us to group samples in a more tumour type independent way. The recurrence-related features based on these recurrent SSMs and SIMs are key to our ability to cluster the cancer genomes into biologically relevant clusters. If we only use the general features we lose important information about mutational processes (S2 Fig).

The simple general mutational features, the different types of annotation and the uncovered sequence motifs do provide a deeper understanding of several mutational processes (S3 Text). For instance, MSI samples (cluster J) have a particularly high percentage of 1 bp C/G deletions in the context of midsize homopolymers. We also see a strong shift towards the presence of SIMs compared to SSMs resulting in a high absolute and relative number of SIMs. Ultra-hypermutators (cluster H) form a mirror image in this respect as we observe a shift in the opposite direction, resulting instead in a high absolute and relative number of SSMs. Another difference is that in cluster H there is a significantly higher percentage of mutations in late-replicating regions than for cluster J (SSMs: 60.2% vs. 52.8%, $p = 0.0011$, SIMs: 66.7% vs. 51.3%, $p = 1.8e-06$). The mutational processes induced by tobacco-smoke exposure (cluster A), whose link to an increased percentage of C>A SSMs is well-known, are also associated with a high percentage of 1 bp C/G deletions (S4 Text). A third example is the high percentage of 1 bp A/T

insertions in context of a short homopolymer observed for cluster C that mainly consists of Kidney-RCC and Kidney-ChRCC samples. For this cluster there is also a nearly equal distribution of mutations between early and late-replicating regions, which is in contrast to what is generally observed for cancer genomes [8] with the exception of MSI samples [28]. However, unlike for MSI genomes, for cluster C a deficient MMR pathway can most likely not explain it. Deficient translesion synthesis has been shown in yeast to also lead to a more equal distribution [40]. In the opposite direction, the cluster possibly linked to gastric-acid exposure (cluster L) has an unexpectedly strong tendency of both SSMs and SIMs to be in late-replicating regions compared to all other clusters, which could point to the extensive usage of error-prone polymerases. The sequence motif (AACTT) found for the T>G SSMs recurrent within this cluster ($n = 38,399$, 38.9% with the motif) provides another interesting characteristic (Fig 6). Only 8.9% of the T>G SSMs recurrent in the 2,479 samples not in cluster L ($n = 25,318$) are confined to this motif. An important contributor to the recurrent T>G SSMs not in cluster L is the cluster linked to the deregulated activity of POLE (cluster H). The T>G SSMs that are recurrent within cluster H ($n = 11,553$) are instead enriched for the sequence motif AAATTTAT (S4 Fig). There are some interesting parallels between cluster H and L. First, for both holds that the Eso-AdenoCA and ColoRect-AdenoCA samples that form the majority of cluster L and H, respectively, have a higher median number of SSMs than samples from the same tumour types not assigned to the respective clusters (Eso-AdenoCA: 29,302.5 vs. 11,404, $p = 1.3e-09$, ColoRect-AdenoCA: 850,298 vs. 15,045, $p = 1.5e-08$). Second, changes to the dNTP pool are in both cases likely linked to the observed mutations together with the more frequent usage of alternative (error-prone) polymerases (cluster L) or a polymerase with a deregulated activity (cluster H). Third, the sequence motifs found for both clusters exceed the single neighbouring base. The latter is the case for all sequence motifs that we found (Fig 6) and also none of them have the same number of bases on both sides of the mutated position. These two observations and the motifs themselves are also important to take into account when estimating the background mutation rate used in *e.g.* driver prediction [25, 37]. The motifs point to an increased mutational probability of individual bases [22] that is context-specific and characteristic for certain mutational processes. This has primarily been shown and taken into account for a sequence context of a single neighbouring base [37] or, less frequently, for an equal number of several bases at both sides of the mutation [25]. As we extract these motifs based on recurrent mutations there is a possibility that positive selection plays a role. However, this is likely negligible as the number of recurrent, predicted driver mutations is only 427 when considering all six SSM subtypes together.

Several of our clusters are linked to cancer phenotypes that are relevant for treatment and/or have prognostic value. Our division into 16 clusters and their characteristics could, therefore, be valuable for complementing current classification schemes, which are mainly based on histology and organ of origin. We can assign a new sample to one of our 16 clusters by first projecting it onto the PCA space based on the PCAWG cohort. Next, we use the first 18 principal components to compute the Euclidean distance to the centroid of each of the 16 clusters and assign the sample to the nearest one. If there are multiple clusters with a minimum difference in distance to the new sample, then to select one cluster we use the sequence motifs (Fig 6) and various layers of annotation (S3 Text) like replication time. Ideally, we would use only the samples in the 'reference set', which currently is the PCAWG cohort, to compute the recurrence-related features for a new sample. However, ~90% and ~72% of the recurrent SSMs and SIMs, respectively, in this set are only recurrent in two samples (Fig F in S2 Text). Therefore, the recurrence-based features of the new sample might be underestimated in which case the sample is also less likely to be assigned to clusters that have a positive association with recurrence. Instead we would need to include the new sample for computing recurrence, which

could also affect the recurrence-related features for some samples in the reference set. This might result in changes in the clustering, but the impact of a single sample is most likely minimal. Of note, the interdependence of samples in terms of the recurrence-related features also makes cross-validation difficult. The level of recurrence is not high enough to compute recurrence for the training and test set separately, and even a leave-one-out strategy would create dependence between the two sets. We hypothesize that, by increasing the size of the reference set, we will reach at a certain point a plateau in terms of recurrence. This would enable us to compute the recurrence-based features for a new sample using only the reference set. A larger dataset would also allow further insights into the non-randomness of mutational processes, especially of those that are not active across a large set of samples or that are only observed in specific tumour types for which the number of samples is currently limited. Efforts are, in fact, already on their way to expand the PCAWG dataset with more whole-genome sequences from ICGC and other consortia.

Given that incorporating whole-genome sequencing in a clinical setting is gaining traction as evidenced by projects like Genomics England (www.genomicsengland.co.uk) and the Hartwig Medical Foundation (www.hartwigmedicalfoundation.nl), analyses making full use of this kind of data are urgently needed. Ultimately, whole-genome sequencing can then replace multiple diagnostic tests currently in use and make diagnoses more accurate. One example illustrating the value of our clusters towards this goal is the MSI phenotype linked to cluster J. For these patients, immunotherapy may be beneficial [41] while adjuvant chemotherapy may not be needed [42]. To classify a cancer genome as MSI, we can use our 42 features to determine whether or not a sample belongs to cluster J, as detailed above. A high percentage of 1 bp C/G deletions in a midsize homopolymer is, however, even by itself already a strong indication for MSI. The MSI phenotype cluster J captures, forms a possible alternative to either explicitly identifying all microsatellite alterations between tumour and normal tissue [43] or using specific markers to detect alterations in five or seven of them like the Bethesda markers [44]. There are also 10 mutational signatures linked to a deficient MMR pathway of which seven are based on single base substitutions, two on doublet base substitutions and one on small indels [20]. Two more indel-based signatures (ID1 and ID2) that are found in nearly all cancer genomes, are linked to a deficient MMR pathway if they contribute >10,000 indels. Signatures look at mutational processes at mutation level rather than sample level. A non-zero contribution of an individual MSI-linked signature or a high contribution (>10,000) of ID1 and ID2 is not sufficient to classify a sample as MSI given that this naïve approach would result in 368 possible candidates. Instead it requires a combination of signatures and/or thresholds on the amount of mutations contributed to the sample to be able to use the signatures for MSI classification. A second example of an actionable phenotype that we capture with one of our clusters is ultra-hypermutation (cluster H), which has also been related to beneficial results from immunotherapy [45, 46]. A third example is the somatic hypermutation of the immunoglobulin genes, which identifies memory B-cells as the cell of origin in the case of lymphomas. This has been linked to a less aggressive form of Lymph-CLL and more favourable prognosis [33], which may in turn influence treatment selection. Without explicitly analysing the immunoglobulin genes [47], we were largely able to separate the Lymph-CLL samples with somatic hypermutation (cluster M) from those without (cluster D). The characteristics of the former group include a high percentage of recurrent C>G SSMs and 1 bp A/T deletions. A final example relates to those Eso-AdenoCA samples that are assigned to cluster L, which have a high percentage of T>C as well as T>G SSMs and a higher total mutational load than Eso-AdenoCA samples not assigned to this cluster. Eso-AdenoCA samples with the characteristics of cluster L have also been suggested to benefit from immunotherapy [48]. The same treatment option may therefore be prioritized for the 22 Stomach-AdenoCA samples that are also in cluster L.

Similarly, a refined investigation of tumour samples that do not cluster with the vast majority of its own kind may ideally point to differences in disease prognosis or treatment response and even has the potential to define novel subtypes or reveal misclassification. Such an analysis would be especially worthwhile for the ~20% or less samples from Kidney-RCC, Liver-HCC, Lung-SCC or Lymph-BNHL that are not assigned to the main cluster. Another possible application of our classification scheme is to assign a metastatic sample with unknown primary site to a cluster to shed light on the possible tissue of origin or pan-cancer characteristics like MSI.

In conclusion, we provide here a comprehensive analysis of somatic mutations in cancer genomes irrespective of tumour type using 42 features with a truly pan-cancer focus. This allows us to include tumour types with very few samples for which individual analysis is little informative. Moreover, information can be borrowed across the entire data set enabling the detection of processes present in multiple tumour types. We let the genome prioritize what is important by using position-specific recurrence and by considering features that do not depend on the completeness and correctness of current genome annotations. This has enabled us to delineate various mutational processes, uncover new mutational manifestations and characterize several actionable clinical phenotypes in a novel way. Findings from this and similar analyses in the future will be of utmost importance for the goal to tailor treatment to the individual patient.

Methods

PCAWG cohort – quality control

We used the cohort of cancer genomes assembled by the PCAWG project [12] of the ICGC and TCGA. For every donor, whole-genome sequencing data was available for a normal-tumour pair and all samples were analysed uniformly. A detailed description of the quality control is provided in the PCAWG marker paper [12]. In short, 176 samples were excluded for various reasons as part of the quality control, most commonly because of contamination with RNA. Samples of another 75 donors were of borderline quality for various reasons, including a high percentage of paired reads mapping to different chromosomes [12, 49]. We decided not to include the samples of those donors, which left us with genomic data of 2,583 donors covering 37 tumour types (S1 Table). The distribution of the samples across the tumour types is also indicated in S1 Table. In case there were multiple tumour samples for the same donor, we selected a single sample following the decision made within the consortium. To make the decision five criteria were used as described by the PCAWG Drivers and Functional Interpretation Group [18]. In order of importance, they prioritized the sample: 1) of a primary tumour over metastatic and recurrent ones; 2) with a OxoG score over 40, which indicates low levels of oxidative damage artefacts [50]; 3) with the highest quality according to the star rating system [49]; 4) with RNA-Seq data available; 5) with the lowest level of contamination with foreign DNA. If none of these criteria led to the selection of a single sample, a random selection was made.

PCAWG cohort – mutation calls

The description of the procedure for the mutation calls is provided in the marker paper of the PCAWG consortium [12]. In brief, the sequenced reads of the respective normal and tumour sample pairs were aligned with BWA-MEM to the GRCh37/h19 genome. Four mutation calling pipelines were run on the resulting BAM-files for each normal/tumour sample pair. The pipelines used for calling SSMs were MuSE [51] and three in-house pipelines developed at the Deutsches Krebsforschungszentrum (DKFZ) in collaboration with the European Molecular Biology Laboratory (EMBL), Wellcome Sanger Institute and Broad Institute, respectively. A

consensus set was built by keeping those calls on which two or more callers agreed. SIMs were called by SMuFIN [52] and three pipelines developed by the same institutes as mentioned for SSMs. The consensus was determined by stacked logistic regression instead, as the level of agreement between the callers was lower than for SSMs. Furthermore, the SIM calls were left aligned to make them comparable across samples. Several filters were applied to both the SSM and SIM calls to remove, among other things, calls due to oxidative damage artefacts [50] and germline variants. Great care was taken by the consortium to reduce the number of false positive mutation calls, resulting in a reliable dataset that is believed to be a conservative representation of the true set of mutations.

Definition of mutations

For SSMs there are 16 possible subtypes. However, we can neither detect substitutions with a base of the same type (*e.g.* A>A) nor do we usually know on which strand the (pre-)mutagenic event happened first (*e.g.* A>C is equivalent to T>G on the other strand). Therefore, we combined the substitutions that are each other's reverse complement and refer to them by the pyrimidine of the mutated base pair: C>A, C>G, C>T, T>A, T>C and T>G. We regarded substitutions directly next to each other (median number across samples: 25) as separate single base events since, aside from the very limited numbers, in several cases the individual callers only supported one single base event, and only the consensus resulted in a multiple base substitution call. For 1 bp SIMs, these are the four subtypes A/T deletions, C/G deletions, A/T insertions and C/G insertions, as analogously to SSMs, we cannot determine on which strand the (pre-) mutagenic event happened first.

Features describing each cancer genome

We computed 29 general features and 13 related to recurrence (Table A in [S1 File](#)) to characterize different aspects of the somatic mutations in a cancer genome. We used the *vcfR* package in R to read in the VCF files [53]. The general features comprised the number of SSMs and SIMs (two features), the percentage of SIMs with respect to the total number of mutations (one feature), the distribution of SSMs and SIMs across the different subtypes (six and four features, respectively), and the homopolymer context of 1 bp SIMs for each of the four subtypes (four times four features). We used the *BCFtools* (version 1.5) to compute recurrence using the VCF files as input. Recurrence was captured by the overall percentage of recurrent SSMs and SIMs (two features), percentage of recurrent mutations of type SIM (one feature) and recurrence per SSM and SIM subtype (six and four features, respectively). The homopolymer context is not included in the recurrence features, as the number of recurrent SIMs is too low to stratify into 16 additional features. Except for the number of SSMs and SIMs, all other 40 features were in percentages.

Principal Component Analysis and hierarchical clustering on Principal Components

The R package *FactoMineR* (v1.41) was used for the PCA [14]. All input features for the PCA were scaled to zero mean and unit variance to account for the differences between the ranges of the features, especially with respect to the two features in absolute terms versus the ones in terms of percentages. The first 18 PCs explained together over 80% of the variance of the data. The remaining components were assumed to mostly represent noise in the data. The PCs were used as input to the 'hierarchical clustering on principal components' (HCPC) function from the *FactoMineR* package. The Euclidean distance was used as a measure of dissimilarity and the Ward criterion for linkage. We cut the hierarchical clustering tree at various heights to see

a more global down to a more specific division of the samples. The HCPC function includes a consolidation step in the form of k-means clustering [15], which uses the centroids of the hierarchical clustering as a starting point. This consolidation step was repeated a maximum of 10 times. The k-means clustering increased the variance between clusters from 17.5 to 18.9. Other advantages of this hybrid approach are that it reduces the sensitivity of k-means clustering to outliers and the initial centroids are selected in an informed way instead of at random. As a consequence of this step, some samples were finally assigned to a different cluster than after the hierarchical clustering. A ‘v test’, included in the FactoMineR package, was used to determine which features were significantly associated with each cluster. This test compares the mean of a particular feature in a cluster to the overall mean in the dataset. We corrected the p-values of all ‘v tests’ for multiple testing using the Benjamini-Yekutieli method. A feature is considered to be significantly associated to a cluster if the adjusted p-value < 0.05.

Detection and enrichment of motifs

We collected for clusters A, E, G, H, L and M all SSMs of the subtype that is the most characteristic. This is C>A for clusters A and H, C>G for cluster E and M, C>T for cluster G and T>G for cluster L. In addition, we looked at T>G SSMs in cluster H to compare them to cluster L. Next, we extracted from the reference genome (GRCh37/h19) the ten adjacent bases in 5’ and 3’ direction of the mutation using the Rsamtools package in R. We used the extracted sequence context as input to construct two sequence logos per cluster: one for the mutations that are recurrent within the cluster and one for those that are not. We include each recurrent mutation only once to avoid giving extra weight to highly recurrent mutations. As a measure of information content we used the relative entropy [54, 55], which is defined for position *i* by:

$$RE_i = \sum_{b \in \{A,C,G,T\}} f(b_i) \log_2 \frac{f(b_i)}{P(b)}$$

Here, $f(b_i)$ stands for the frequency of base *b* (A, C, G or T) in position *i* and $P(b)$ stands for the prior probability of base *b* as determined by the frequency in the human genome (GRCh37/h19). The height of each base in the sequence plot is proportional to $f(b_i) \log_2 \frac{f(b_i)}{P(b)}$. A positive value corresponds to an enrichment of the base with respect to the prior probability and a negative value to a depletion. The relative entropy (RE_i) is zero, if all four bases are observed with the same frequency as the prior in position *i*. We set 0.25 as a threshold for RE_i to define the enriched motif. Furthermore, we computed per cluster the percentages of all, non-recurrent and recurrent SSMs that were in the sequence context that was found to be enriched in the recurrent SSMs. To estimate the percentage of the respective motifs in the human genome, we first slid a window of the same size (*k*) as the motif across the genome with a shift equal to the length of the motif and counted all possible k-mers. Next, we added to this the counts retrieved in the same way for the reverse complement of the reference sequence (corresponding to the opposing strand), since we also combined the reverse complements for each of the SSM subtypes. From this we computed the percentage of the enriched motif with respect to all k-mers and to the k-mer with the base that is mutated in the enriched motif at the same position.

Statistical tests

The correlation between every possible pair of the 42 features was measured by the Spearman’s rank correlation coefficient using the R package Hmisc (v4.1–1). Multiple testing correction of the p-values of all correlation tests (including those in S2 Text) was done by the Benjamini-Yekutieli method. For the other correlations mentioned we also used the Spearman’s rank correlation coefficient.

We used the Wilcoxon rank-sum test with continuity correction as the test of significance for differences in features observed between clusters.

The different proportions of sequence motifs between recurrent and non-recurrent SSMs were assessed by using χ^2 tests.

Plots

Figs 1, 3, 5 and 6, the pie charts in Fig 4 and the plots in Supporting Information, except for S1, were made using the R package ggplot2 (v3.0.0). Fig 6, S3 Fig and S4 Fig additionally required ggseqlogo (v0.1) [56] and Fig 2 was made with the use of the R package corrplot (v0.84). Fig 7 was made using Microsoft PowerPoint and we also included images from the Servier Medical Art website (<http://smart.servier.com/>). The ‘clustering tree’ in S1 Fig was made using the clustree R package [57]. We have manually replaced the nodes in the tree with the pie diagram showing the distribution of tumour types in each cluster. For the colours of the different tumour types we have made use of the script provided by the PCAWG consortium, available at: <https://github.com/ICGC-TCGA-PanCancer/pcawg-colour-palette>.

Supporting information

S1 Fig. Clustering tree showing tumour type distribution for 2 to 20 clusters. The clustering tree shows how clusters evolve across different clustering resolutions ranging from 2 to 20 clusters. For example, cluster G splits off from the rest of the cohort at a resolution of three clusters and remains largely unchanged in higher resolutions. We have marked for each of our 16 clusters the clustering resolutions across which they remain largely stable, *i.e.* the Jaccard similarity index between a cluster at resolution 16 and one at a higher or lower resolution is at least 0.85. The number under each cluster indicates the number of samples in that particular cluster. The colour of an arrow indicates the number of samples the two connected clusters have in common. The transparency of the arrow indicates the proportion of samples the two connected clusters have in common with respect to the cluster at the higher resolution. Only arrows representing a proportion of more than 0.1 are shown. Consequently, the number of samples in a cluster at a certain clustering resolution may not match with the connected cluster(s) at a higher resolution. Note that the clustering shown is the result after the k-means clustering step.

(PDF)

S2 Fig. PCA and clustering with and without the recurrence-related features. When using only the 29 general features for the PCA (A), the first two PCs explain less variance than when using all 42 features for the PCA (B) (27.5% vs. 29.1%). The features indicated in the two PCA plots are those that contribute above average to the first two PCs. The subsequent clustering also differs as shown in (C) and (D). Without using the recurrence-related features, only five of the eight samples linked to ultra-hypermutation (D – cluster H) are in a separate cluster (C – cluster VIII). Also the cluster linked to hypermutation of the immunoglobulin genes (D – cluster M) is dissolved as evidenced by the fact that the samples are spread across eight clusters (C – clusters III, IV, VI, XI, XII, XIII, XIV and XV). One consequence of this is that only 19 of the 40 the Lymph-CLL samples with hypermutation are in the same cluster as opposed to 36 when using all features (E). In addition, the largest fraction of cluster M ends up in a cluster with Eso-AdenoCA and Stomach-AdenoCA samples (C – cluster XII), making that cluster less cancer-specific than when using all features (D – cluster L). The Lymph-CLL samples without hypermutation of the immunoglobulin genes are also no longer largely confined to a single cluster (E). Moreover, the samples with and without hypermutation end up more

often in the same cluster than when recurrence-related features are also used.
(PDF)

S3 Fig. Enriched sequence motifs for C>G SSMs in cluster M. The sequence logos represent the sequence context of ten bp 5' and 3' of the non-recurrent (left-side) or recurrent (right-side) C>G mutations of cluster M. Here recurrence is defined as a mutation at the same genomic location in two or more samples from cluster M. Relative entropy is used as a measure of information content (see [Methods](#)). Setting a threshold of 0.25 for the relative entropy results in the motifs highlighted in the rectangles. In the upper right corner of both sequence logos the number of mutations is indicated. To the right of the sequence logos are the percentages in which the enriched motif found for the recurrent C>G SSMs is present in context of the mutations in the cluster and the corresponding k-mers in the genome (N = A, C, G or T). The enrichment for the motif for recurrent C>G SSMs is significantly higher than for the non-recurrent C>G SSMs (χ^2 test: $p < 2.2e-16$).

(TIF)

S4 Fig. Enriched sequence motifs for T>G SSMs in cluster H. The sequence logos represent the sequence context of ten bp 5' and 3' of the non-recurrent (left-side) or recurrent (right-side) T>G mutations of cluster H. Here recurrence is defined as a mutation at the same genomic location in two or more samples from cluster H. Relative entropy is used as a measure of information content (see [Methods](#)). Setting a threshold of 0.25 for the relative entropy results in the motifs highlighted in the rectangles. In the upper right corner of both sequence logos the number of mutations is indicated. To the right of the sequence logos are the percentages in which the enriched motif found for the recurrent T>G SSMs is present in context of the mutations in the cluster and the corresponding k-mers in the genome (N = A, C, G or T). The enrichment for the motif for recurrent T>G SSMs is significantly higher than for the non-recurrent T>G SSMs (χ^2 test: $p < 2.2e-16$).

(TIF)

S1 Table. Tumour type abbreviation, full name and number of samples.

(PDF)

S2 Table. Recurrence in pan-cancer context and within tumour type(s).

(PDF)

S1 Text. Estimation of the levels of recurrence when purely driven by chance.

(PDF)

S2 Text. Recurrence versus general mutational characteristics.

(PDF)

S3 Text. Detailed cluster-specific descriptions.

(PDF)

S4 Text. Smoking history and related mutational subtypes.

(PDF)

S1 File. Characteristic plots summarising each of the 42 features.

(PDF)

S2 File. Sample distribution per tumour type across the 16 clusters.

(PDF)

Acknowledgments

We would like to thank the PCAWG consortium for providing the somatic mutation calls, driver predictions, mutational signatures, MSI status, impact classification and clinical data as well as their support throughout the project.

Author Contributions

Conceptualization: Ivo G. Gut.

Formal analysis: Miranda D. Stobbe, Emanuele Raineri.

Funding acquisition: Ivo G. Gut.

Investigation: Miranda D. Stobbe, Gian A. Thun, Andrea Diéguez-Docampo, Meritxell Oliva.

Methodology: Miranda D. Stobbe, Justin P. Whalley.

Resources: Ivo G. Gut.

Software: Miranda D. Stobbe.

Supervision: Ivo G. Gut.

Visualization: Miranda D. Stobbe, Andrea Diéguez-Docampo.

Writing – original draft: Miranda D. Stobbe.

Writing – review & editing: Miranda D. Stobbe, Gian A. Thun, Andrea Diéguez-Docampo, Ivo G. Gut.

References

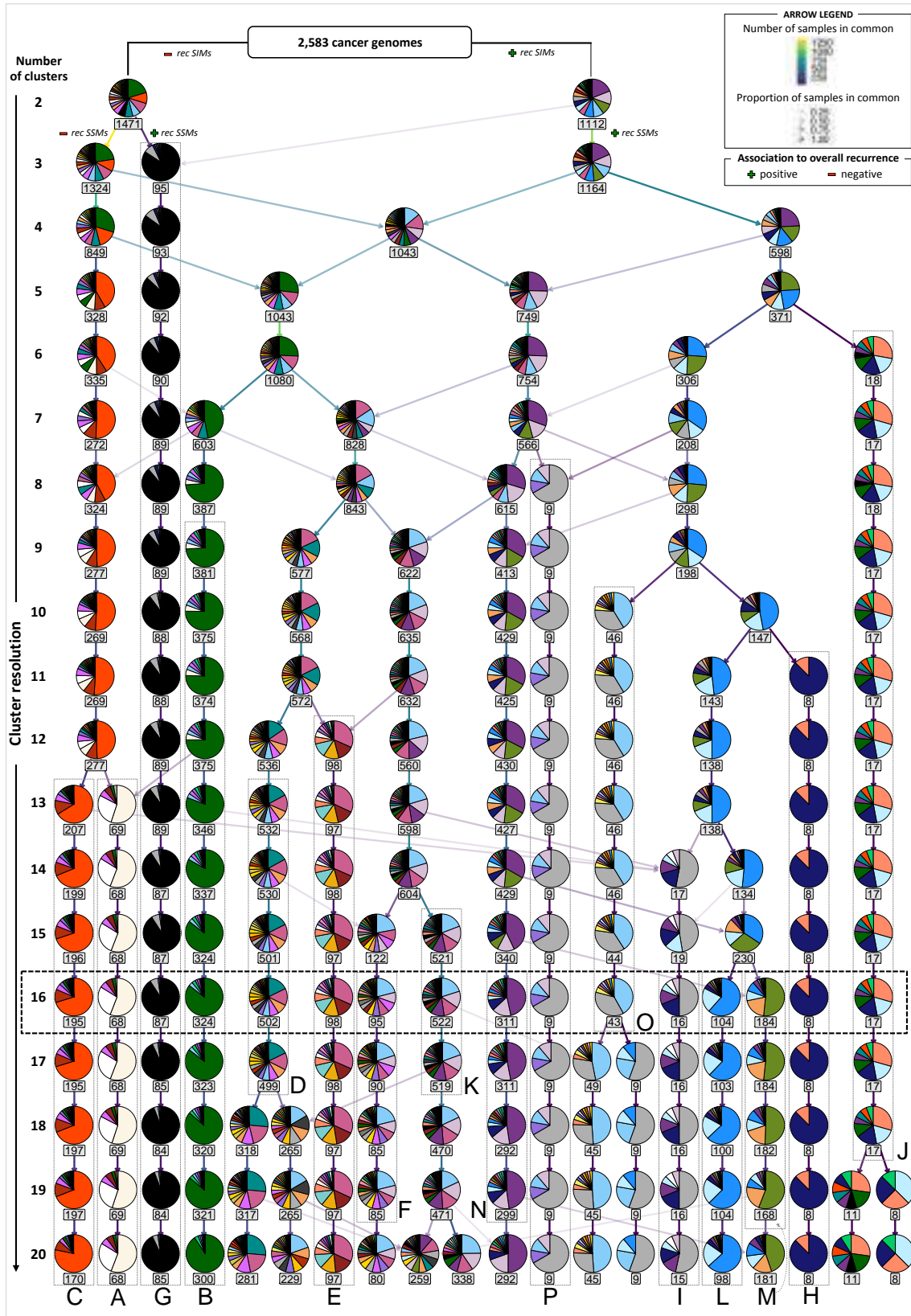
1. Fredriksson NJ, Ny L, Nilsson JA, Larsson E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet.* 2014; 46:1258. <https://doi.org/10.1038/ng.3141> PMID: 25383969
2. Ostrow SL, Barshir R, DeGregori J, Yeger-Lotem E, Hershberg R. Cancer Evolution Is Associated with Pervasive Positive Selection on Globally Expressed Genes. *PLOS Genet.* 2014; 10(3):e1004239. <https://doi.org/10.1371/journal.pgen.1004239> PMID: 24603726
3. Sabarinathan R, Pich O, Martincorena I, Rubio-Perez C, Juul M, Wala J, et al. The whole-genome panorama of cancer drivers. *bioRxiv.* 2017. <https://doi.org/10.1101/190330>
4. Tomasetti C, Marchionni L, Nowak MA, Parmigiani G, Vogelstein B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci USA.* 2015; 112(1):118–23. <https://doi.org/10.1073/pnas.1421839112> PMID: 25535351
5. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. *Science.* 2018; 362(6417):911–7. <https://doi.org/10.1126/science.aau3879> PMID: 30337457
6. Ciccarelli FD. Mutations differ in normal and cancer cells of the oesophagus. *Nature.* 2019; 565(7739):301–3. <https://doi.org/10.1038/d41586-018-07737-8> PMID: 30643303
7. Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell.* 2019; 177(1):101–14. <https://doi.org/10.1016/j.cell.2019.02.051> PMID: 30901533
8. Woo YH, Li W-H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat Commun.* 2012; 3:1004. <https://doi.org/10.1038/ncomms1982> PMID: 22893128
9. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature.* 2015; 518(7539):360–4. <https://doi.org/10.1038/nature14221> PMID: 25693567
10. Pich O, Muñios F, Sabarinathan R, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N. Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell.* 2018; 175(4):1074–87.e18. <https://doi.org/10.1016/j.cell.2018.10.004> PMID: 30388444

11. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. *Nature*. 2013; 500(7463):415–21. <https://doi.org/10.1038/nature12477> PMID: 23945592
12. Campbell PJ, Getz G, Stuart JM, Korbel JO, Stein LD. Pan-cancer analysis of whole genomes. *bioRxiv*. 2017. <https://doi.org/10.1101/162784>
13. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell*. 2018; 173(2):305–20.e10. <https://doi.org/10.1016/j.cell.2018.03.033> PMID: 29625049
14. Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*. 2008; 25(1):18. Epub 2008-03-18. <https://doi.org/10.18637/jss.v025.i01>
15. Husson F, Josse J, Pages J. Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data. *Applied Mathematics Department*. 2010:1–17.
16. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*. 2012; 22(9):1760–74. Epub 2012/09/08. <https://doi.org/10.1101/gr.135350.111> PMID: 22955987
17. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, et al. Oncotator: Cancer Variant Annotation Tool. *Hum Mutat*. 2015; 36(4):E2423–E9. <https://doi.org/10.1002/humu.22771> PMID: 25703262
18. Rheinbay E, Nielsen MM, Abascal F, Tiao G, Hornshøj H, Hess JM, et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. *bioRxiv*. 2017. <https://doi.org/10.1101/237313>
19. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci USA*. 2010; 107(1):139–44. <https://doi.org/10.1073/pnas.0912402107> PMID: 19966280
20. Alexandrov L, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Boot A, et al. The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*. 2018:322859. <https://doi.org/10.1101/322859>
21. Gut IG, Wood PD, Redmond RW. Interaction of Triplet Photosensitizers with Nucleotides and DNA in Aqueous Solution at Room Temperature. *J Am Chem Soc*. 1996; 118(10):2366–73. <https://doi.org/10.1021/ja9519344>
22. Fredriksson NJ, Elliott K, Filges S, Van den Eynden J, Ståhlberg A, Larsson E. Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. *PLOS Genetics*. 2017; 13(5):e1006773. <https://doi.org/10.1371/journal.pgen.1006773> PMID: 28489852
23. Mao P, Brown AJ, Esaki S, Lockwood S, Poon GMK, Smerdon MJ, et al. ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nature Communications*. 2018; 9(1):2626. <https://doi.org/10.1038/s41467-018-05064-0> PMID: 29980679
24. Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*. 2016; 532(7598):264–7. <https://doi.org/10.1038/nature17661> PMID: 27075101
25. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. 2017; 171(5):1029–41.e21. <https://doi.org/10.1016/j.cell.2017.09.042> PMID: 29056346
26. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet*. 2004; 5:435. <https://doi.org/10.1038/nrg1348> PMID: 15153996
27. Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, et al. A direct characterization of human mutation based on microsatellites. *Nat Genet*. 2012; 44:1161. <https://doi.org/10.1038/ng.2398> PMID: 22922873
28. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*. 2015; 521(7550):81–4. <https://doi.org/10.1038/nature14173> PMID: 25707793
29. Yang Y, Sterling J, Storici F, Resnick MA, Gordenin DA. Hypermutability of Damaged Single-Strand DNA Formed at Double-Strand Breaks and Uncapped Telomeres in Yeast *Saccharomyces cerevisiae*. *PLOS Genet*. 2008; 4(11):e1000264. <https://doi.org/10.1371/journal.pgen.1000264> PMID: 19023402
30. Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biology*. 2018; 19(1):129. <https://doi.org/10.1186/s13059-018-1509-y> PMID: 30201020
31. Kamiya H. Mutations Induced by Oxidized DNA Precursors and Their Prevention by Nucleotide Pool Sanitization Enzymes. *Genes and Environment*. 2007; 29(4):133–40. <https://doi.org/10.3123/jemsge.29.133>

32. Seplyarskiy VB, Bazykin GA, Soldatov RA. Polymerase ζ Activity Is Linked to Replication Timing in Humans: Evidence from Mutational Signatures. *Molecular Biology and Evolution*. 2015; 32(12):3158–72. <https://doi.org/10.1093/molbev/msv184> PMID: 26376651
33. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated Ig V μ Genes Are Associated With a More Aggressive Form of Chronic Lymphocytic Leukemia. *Blood*. 1999; 94(6):1848–54. PMID: 10477713
34. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene*. 2002; 21:7435. <https://doi.org/10.1038/sj.onc.1205803> PMID: 12379884
35. Segal E, Fondudfe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, et al. A genomic code for nucleosome positioning. *Nature*. 2006; 442(7104):772–8. <https://doi.org/10.1038/nature04979> PMID: 16862119
36. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet*. 2015; 47:1067. <https://doi.org/10.1038/ng.3378> PMID: 26258849
37. Brown A-L, Li M, Goncarenco A, Panchenko AR. Finding driver mutations in cancer: Elucidating the role of background mutational processes. *PLOS Computational Biology*. 2019; 15(4):e1006981. <https://doi.org/10.1371/journal.pcbi.1006981> PMID: 31034466
38. Yang W, Bang H, Jang K, Sung MK, Choi JK. Predicting the recurrence of noncoding regulatory mutations in cancer. *BMC Bioinformatics*. 2016; 17(1):492. <https://doi.org/10.1186/s12859-016-1385-y> PMID: 27912731
39. Goncarenco A, Rager SL, Li M, Sang Q-X, Rogozin IB, Panchenko AR. Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic Acids Research*. 2017; 45(W1):W514–W22. <https://doi.org/10.1093/nar/gkx367> PMID: 28472504
40. Lang GI, Murray AW. Mutation Rates across Budding Yeast Chromosome VI Are Correlated with Replication Timing. *Genome Biology and Evolution*. 2011; 3:799–811. <https://doi.org/10.1093/gbe/evr054> PMID: 21666225
41. Xiao Y, Freeman GJ. The Microsatellite Instable (MSI) Subset of Colorectal Cancer is a particularly good candidate for checkpoint blockade immunotherapy. *Cancer Discov*. 2015; 5(1):16–8. <https://doi.org/10.1158/2159-8290.CD-14-1397> PMID: 25583798
42. Saridaki Z, Souglakos J, Georgoulas V. Prognostic and predictive significance of MSI in stages II/III colon cancer. *World J Gastroenterol*. 2014; 20(22):6809–14. <https://doi.org/10.3748/wjg.v20.i22.6809> PMID: 24944470
43. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics*. 2014; 30(7):1015–6. <https://doi.org/10.1093/bioinformatics/btt755> PMID: 24371154
44. Umar A, Boland CR, Terdiman JP, Syngal S, Chapelle Adl, Rüschoff J, et al. Revised Bethesda Guidelines for Hereditary Nonpolyposis Colorectal Cancer (Lynch Syndrome) and Microsatellite Instability. *J Natl Cancer Inst*. 2004; 96(4):261–8. <https://doi.org/10.1093/jnci/djh034> PMID: 14970275
45. Schlesner M, Eils R. Hypermutation takes the driver's seat. *Genome Med*. 2015; 7(1):31. <https://doi.org/10.1186/s13073-015-0159-x> PMID: 25821521
46. Heong V, Ngoi N, Tan DSP. Update on immune checkpoint inhibitors in gynecological cancers. *J Gynecol Oncol*. 2017; 28(2):e20. <https://doi.org/10.3802/jgo.2017.28.e20> PMID: 28028993
47. Puente XS, Bea S, Valdes-Mas R, Villamor N, Gutierrez-Abril J, Martin-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015; 526(7574):519–24. <https://doi.org/10.1038/nature14666> PMID: 26200345
48. Secrier M, Li X, de Silva N, Eldridge MD, Contino G, Bornschein J, et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat Genet*. 2016; 48:1131. <https://doi.org/10.1038/ng.3659> PMID: 27595477
49. Whalley JP, Buchhalter I, Rheinbay E, Raine KM, Kleinheinz K, Stobbe MD, et al. Framework For Quality Assessment Of Whole Genome, Cancer Sequences. *bioRxiv*. 2017. <https://doi.org/10.1101/140921>
50. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*. 2013; 41(6):e67–e. <https://doi.org/10.1093/nar/gks1443> PMID: 23303777
51. Fan Y, Xi L, Hughes DST, Zhang J, Zhang J, Futreal PA, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*. 2016; 17(1):178. <https://doi.org/10.1186/s13059-016-1029-6> PMID: 27557938

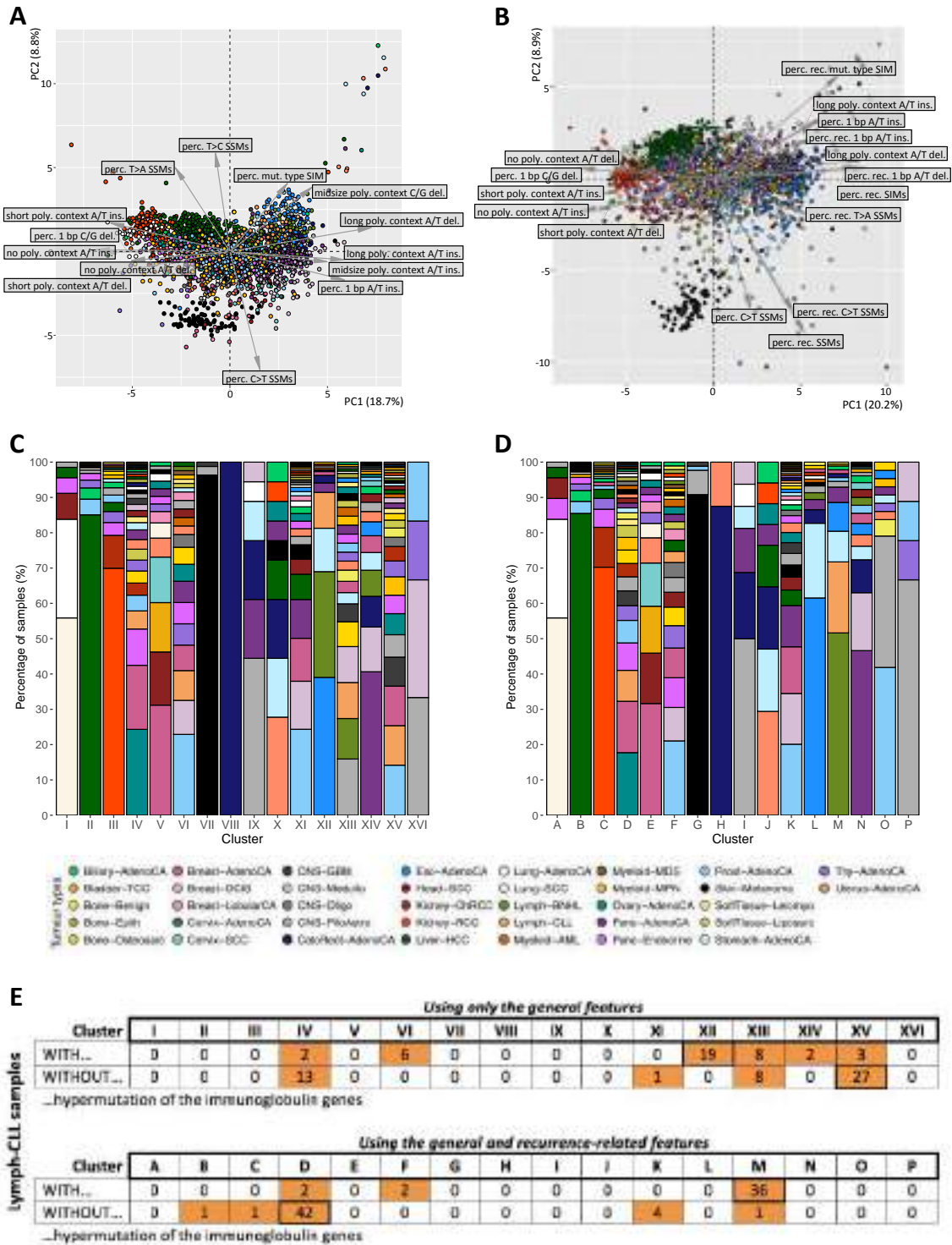
52. Moncunill V, Gonzalez S, Beà S, Andrieux LO, Salaverria I, Royo C, et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat Biotechnol.* 2014; 32:1106. <https://doi.org/10.1038/nbt.3027> PMID: 25344728
53. Knaus BJ, Grünwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources.* 2017; 17(1):44–53. <https://doi.org/10.1111/1755-0998.12549> PMID: 27401132
54. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. *J Mol Biol.* 1986; 188(3):415–31. [https://doi.org/10.1016/0022-2836\(86\)90165-8](https://doi.org/10.1016/0022-2836(86)90165-8) PMID: 3525846
55. Kullback S, Leibler RA. On Information and Sufficiency. *Ann Math Statist.* 1951; 22(1):79–86.
56. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics.* 2017; 33(22):3645–7. <https://doi.org/10.1093/bioinformatics/btx469> PMID: 29036507
57. Zappia L, Oshlack A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience.* 2018; 7(7). <https://doi.org/10.1093/gigascience/giy083>

SUPPLEMENTARY FIGURES: S1, S2, S3, S4



S1 Fig. Clustering tree showing tumour type distribution for 2 to 20 clusters. (Continue in next page)

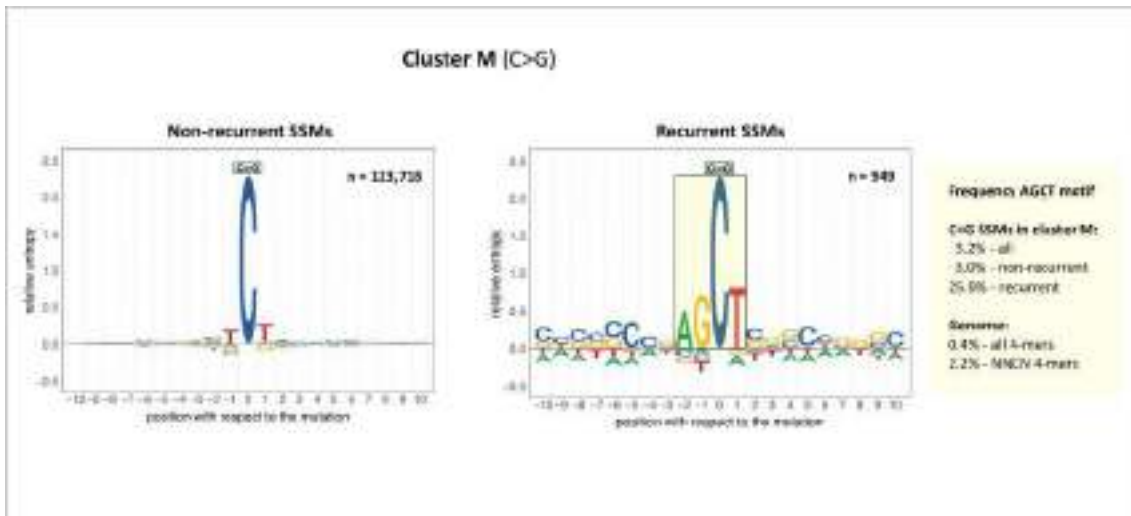
The clustering tree shows how clusters evolve across different clustering resolutions ranging from 2 to 20 clusters. For example, cluster G splits off from the rest of the cohort at a resolution of three clusters and remains largely unchanged in higher resolutions. We have marked for each of our 16 clusters the clustering resolutions across which they remain largely stable, *i.e.* the Jaccard similarity index between a cluster at resolution 16 and one at a higher or lower resolution is at least 0.85. The number under each cluster indicates the number of samples in that particular cluster. The colour of an arrow indicates the number of samples the two connected clusters have in common. The transparency of the arrow indicates the proportion of samples the two connected clusters have in common with respect to the cluster at the higher resolution. Only arrows representing a proportion of more than 0.1 are shown. Consequently, the number of samples in a cluster at a certain clustering resolution may not match with the connected cluster(s) at a higher resolution. Note that the clustering shown is the result after the k-means clustering step.



S2 Fig. PCA and clustering with and without the recurrence-related features.

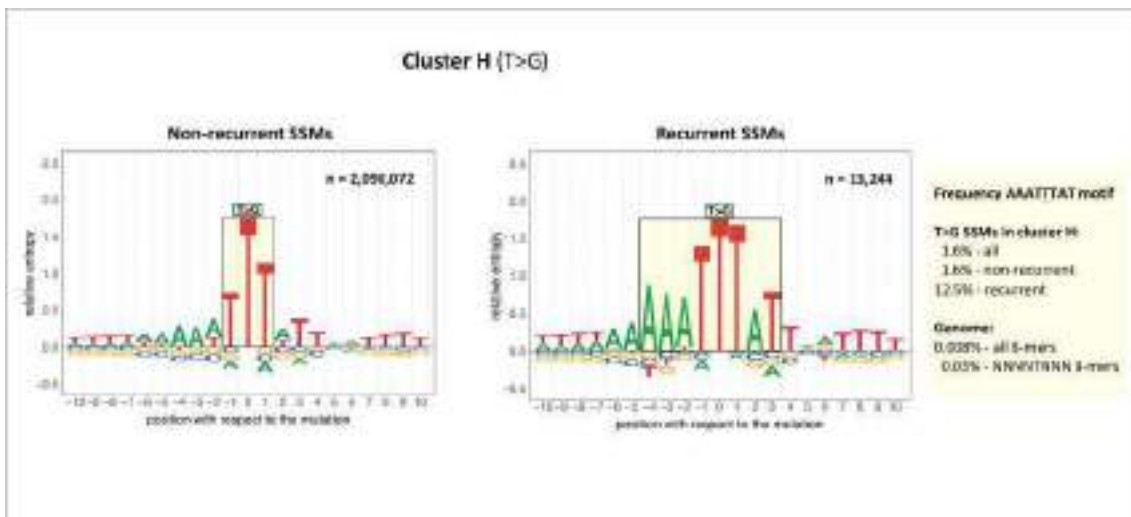
When using only the 29 general features for the PCA (A), the first two PCs explain less variance than when using all 42 features for the PCA (B) (27.5% vs. 29.1%). The features indicated in the two PCA plots are those that contribute above average to the first two PCs. The subsequent clustering also differs as shown in (C) and (D). Without using the recurrence-related features, only five of the eight samples linked to ultra-hypermutation (D – cluster H) are in a separate cluster (C – cluster VIII). Also the cluster linked to hypermutation of the immunoglobulin genes (D – cluster M) is dissolved as evidenced by the fact that the samples are spread across eight clusters (C – clusters III, IV, VI, XI, XII, XIII, XIV and XV). One consequence of this is that only 19 of the 40 the Lymph-CLL samples with hypermutation are in the same cluster as opposed to 36 when using all features (E). In addition, the largest fraction of cluster M ends up in a cluster with Eso-AdenoCA and Stomach-AdenoCA samples (C – cluster XII), making that cluster less cancer-

specific than when using all features (D – cluster L). The Lymph-CLL samples without hypermutation of the immunoglobulin genes are also no longer largely confined to a single cluster (E). Moreover, the samples with and without hypermutation end up more often in the same cluster than when recurrence-related features are also used.



S3 Fig. Enriched sequence motifs for C>G SSMs in cluster M.

The sequence logos represent the sequence context of ten bp 5' and 3' of the non-recurrent (left-side) or recurrent (right-side) C>G mutations of cluster M. Here recurrence is defined as a mutation at the same genomic location in two or more samples from cluster M. Relative entropy is used as a measure of information content (see Methods). Setting a threshold of 0.25 for the relative entropy results in the motifs highlighted in the rectangles. In the upper right corner of both sequence logos the number of mutations is indicated. To the right of the sequence logos are the percentages in which the enriched motif found for the recurrent C>G SSMs is present in context of the mutations in the cluster and the corresponding k-mers in the genome (N = A, C, G or T). The enrichment for the motif for recurrent C>G SSMs is significantly higher than for the non-recurrent C>G SSMs (χ^2 test: $p < 2.2e-16$).



S4 Fig. Enriched sequence motifs for T>G SSMs in cluster H.

The sequence logos represent the sequence context of ten bp 5' and 3' of the non-recurrent (left-side) or recurrent (right-side) T>G mutations of cluster H. Here recurrence is defined as a mutation at the same genomic location in two or more samples from cluster H. Relative entropy is used as a measure of information content (see Methods). Setting a threshold of 0.25 for the relative entropy results in the motifs highlighted in the rectangles. In the upper right corner of both sequence logos the number of mutations is indicated. To the right of the sequence logos are the percentages in which the enriched motif found for the recurrent T>G SSMs is present in context of the mutations in the cluster and the corresponding k-mers in the genome (N = A, C, G or T). The enrichment for the motif for recurrent T>G SSMs is significantly higher than for the non-recurrent T>G SSMs (χ^2 test: $p < 2.2e-16$).

SUPPLEMENTARY TABLE: S1, S2

S1 Table. Tumour type abbreviation, full name and number of samples.

Abbreviation	Full name	Number of samples
Biliary-AdenoCA	biliary adenocarcinoma	34
Bladder-TCC	bladder transitional cell carcinoma	23
Bone-Benign	benign neoplasm of the bone	16
Bone-Epith	epithelial neoplasm of bone	10
Bone-Osteosarc	bone osteosarcoma	35
Breast-AdenoCA	breast adenocarcinoma	195
Breast-DCIS	breast ductal carcinoma in situ	3
Breast-LobularCA	breast lobular carcinoma	13
Cervix-AdenoCA	cervical adenocarcinoma	2
Cervix-SCC	cervical squamous cell carcinoma	18
CNS-GBM	central nervous system - glioblastoma multiforme	39
CNS-Medullo	central nervous system - medulloblastoma	141
CNS-Oligo	central nervous system - oligodendroglioma	18
CNS-PiloAstro	central nervous system - pilocytic astrocytoma	89
ColoRect-AdenoCA	colorectal adenocarcinoma	52
Eso-AdenoCA	oesophageal adenocarcinoma	97
Head-SCC	head/neck squamous cell carcinoma	56
Kidney-ChRCC	chromophobe renal cell carcinoma	43
Kidney-RCC	renal cell carcinoma	143
Liver-HCC	hepatocellular carcinoma	314
Lung-AdenoCA	lung adenocarcinoma	37
Lung-SCC	lung squamous cell carcinoma	47
Lymph-BNHL	B-cell non-Hodgkin lymphoma	107
Lymph-CLL	chronic lymphocytic leukaemia	90
Myeloid-AML	acute myeloid leukaemia	13
Myeloid-MDS	myelodysplastic syndromes	2
Myeloid-MPN	myeloproliferative neoplasm	23
Ovary-AdenoCA	ovarian adenocarcinoma	110
Panc-AdenoCA	pancreatic adenocarcinoma	232
Panc-Endocrine	pancreatic endocrine neoplasm	81
Prost-AdenoCA	prostate adenocarcinoma	199
Skin-Melanoma	skin melanoma	107
SoftTissue-Leiomyo	soft tissue leiomyosarcoma	15
SoftTissue-Liposarc	soft tissue liposarcoma	19
Stomach-AdenoCA	stomach adenocarcinoma	68
Thy-AdenoCA	thyroid adenocarcinoma	48
Uterus-AdenoCA	uterus adenocarcinoma	44

S2 Table. Recurrence in pan-cancer context and within tumour type(s).

Recurrent in	Unique to tumour type(s) in which it is recurrent	Percentage of recurrent	
		SSMs	SIMs
pan-cancer context only		37.1%	79.8%
	Yes	60.0%	10.7%
single tumour type	No	2.8%	8.2%
	Yes	0.1%	0.3%
multiple tumour types	No	0.05%	1.0%

Overview of the percentages of SSMs and SIMs that are recurrent in a pan-cancer setting only, within a single tumour type and in multiple tumour types.

SUPPLEMENTARY TEXT: S1, S2, S3, S4

S1 Text. Estimation of the levels of recurrence when purely driven by chance. (Next section in this document)

S2 Text. Recurrence versus general mutational characteristics.

Available at: <https://doi.org/10.1371/journal.pcbi.1007496.s008>

S3 Text. Detailed cluster-specific descriptions.

Available at: <https://doi.org/10.1371/journal.pcbi.1007496.s009>

S4 Text. Smoking history and related mutational subtypes.

Available at: <https://doi.org/10.1371/journal.pcbi.1007496.s010>

S1 Text. Estimation of the levels of recurrence when purely driven by chance.

Estimation of the levels of recurrence when purely driven by chance

For the estimation of the levels of recurrence if only chance was the driving force we performed the following simulation in which we only take C+G content into account. All other factors that may influence the probability of recurrence (*e.g.* replication time) did not match our definition of chance. For each cancer genome we randomly sampled the same number of SSMs as had been observed in the sample and also kept the counts for each of the six SSM subtypes the same. To take into account the C+G content of the human genome, random numbers were sampled for the C>A/G/T SSMs within the range of 1 to 1,144,530,852, which corresponds to the number of C/G bases in the GRCh37/h19 genome. Once a number had been selected it could not be selected again for the same cancer genome. The same was done for the T>A/C/G mutations, where we sampled numbers within the range of 1 to 1,716,796,279. Simulations were repeated 5,000 times and for each simulation we computed the recurrence overall, recurrence per SSM subtype and for each tumour type the recurrence 'within tumour type' and 'pan-cancer' (Fig A). Only for the recurrence within tumour type there were cases for which there were simulations with an equal or higher number of recurrent SSMs than observed. For three tumour types (Breast-DCIS, Cervix-AdenoCA and Myeloid-MDS) the observed number of recurrent SSMs was zero and nearly all simulated values were also zero (<0.5% were higher). For another five tumour types (Bone-Epith, Breast- LobularCA, Kidney-ChRCC, Myeloid-AML and SoftTissue-Leiomyo) between 2 and 186 of the 5,000 simulated values were equal or higher.

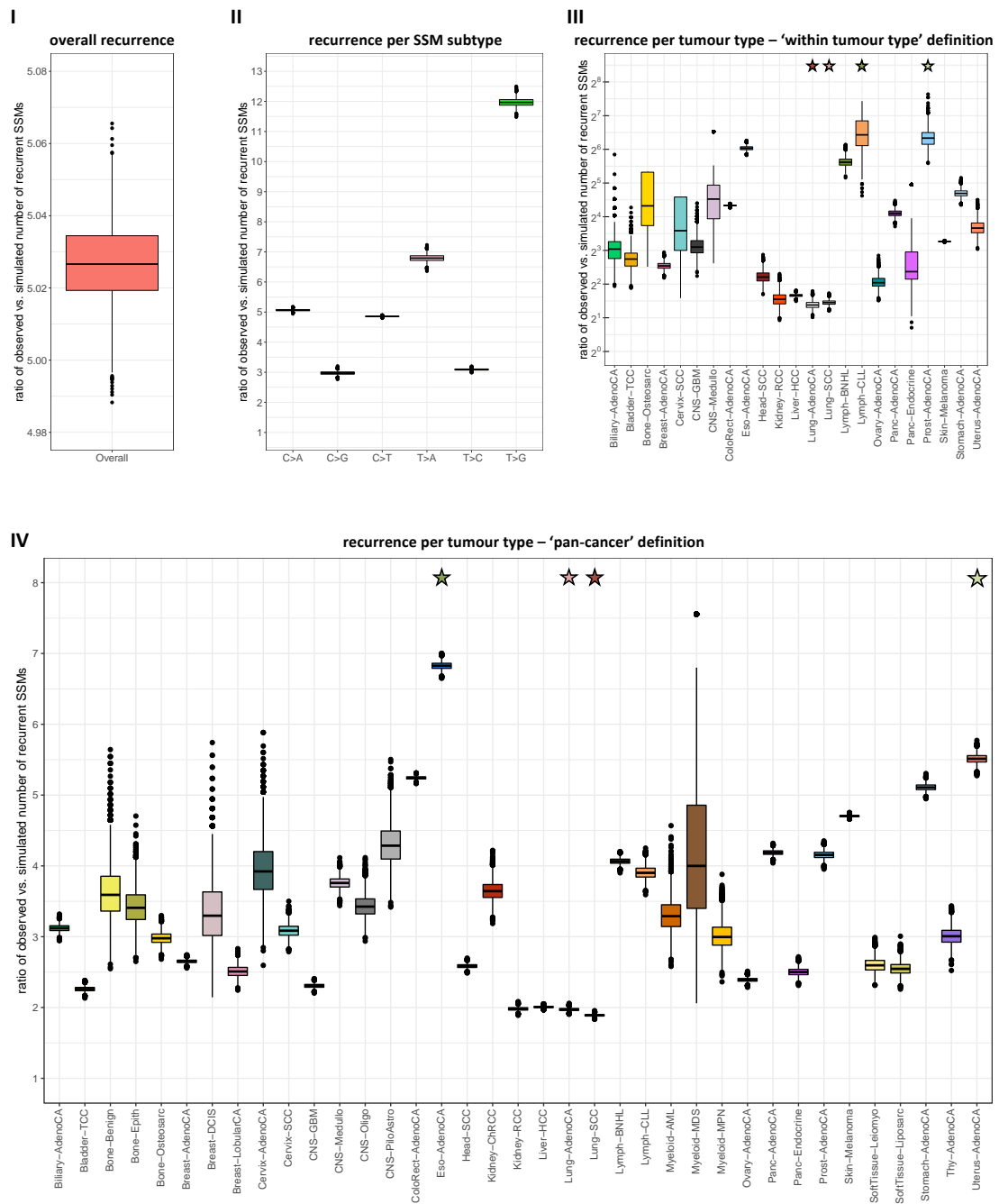


Fig A. Observed recurrence of SSMs versus what is expected by chance.

Each boxplot shows the ratio of the observed number of recurrent SSMs and the number of recurrent SSMs calculated in the simulation (N=5,000) in the following settings: (I) overall recurrence; (II) recurrence for each of the six SSM subtypes; (III) recurrence per tumour type using the ‘within tumour type’ definition; (IV) recurrence per tumour type using the ‘pan-cancer’ definition. The dark green stars at the top of plots III and IV indicate the tumour types with the highest median and the light green stars the second highest. The dark red ones indicate the lowest median and the light red ones the second lowest. For visualization purposes we left out in plot III the results of 14 tumour types for which >40% of the simulations resulted in zero recurrent SSMs, which led to a ratio that is infinite. For the boxplots of Bone-Osteosarc, Cervix-SCC, CNS-Medullo, Lymph-CLL and Panc-Endocrine we left out between 0.4% and 21.7% of the simulations in which no recurrent SSMs were found. In plot IV we left out for visualization purposes the results of 77 simulations for Myeloid-MDS that were all between 8.5 and 17.

SUPPLEMENTARY FILE: S1, S2

S1 File. Characteristic plots summarising each of the 42 features.

S2 File. Sample distribution per tumour type across the 16 clusters.

S1 File. Characteristic plots summarising each of the 42 features.

Characteristic plots summarising each of the 42 features

Each cancer genome is described by 42 features (Table A). We display graphical representations for each feature (Fig A to I) and show absolute numbers in most cases on the y-axis (where applicable). We refer to a value as being an outlier if it is above the third quartile plus 1.5 times the interquartile range ($Q3+1.5 \times IQR$). We describe the main observations below the individual plots.

Table A. Overview of the 42 mutational features describing each cancer genome.

General features	mutational burden	number of	SSMs
			SIMs
	SIM vs. SSM ratio	% of mutations of type SIM	
	distribution of SSMs across the 6 subtypes	percentage of	C>A SSMs
			C>G SSMs
			C>T SSMs
			T>A SSMs
			T>C SSMs
			T>G SSMs
	distribution of 1 bp SIMs across the 4 subtypes	percentage of	A/T deletions
			C/G deletions
			A/T insertions
			C/G insertions
	homopolymer context of 1 bp SIMs	% of A/T deletions	no
			short
			midsize
			long
		% of C/G deletions	no
			short
midsize			
long			
% of A/T insertions		no	
		short	
		midsize	
		long	
% of C/G insertions		no	
		short	
		midsize	
		long	
Recurrence features	overall level of recurrence	% of recurrent	SSMs
			SIMs
	recurrent SIM vs. SSM ratio	% of recurrent mutations of type SIM	
	level of recurrence per SSM subtype	% of recurrent	C>A SSMs
			C>G SSMs
			C>T SSMs
			T>A SSMs
			T>C SSMs
			T>G SSMs
	level of recurrence per SIM subtype (1 bp)	% of recurrent	A/T deletions
C/G deletions			
A/T insertions			
C/G insertions			

Overview of the 29 general features and the 13 features related to recurrence that are used as input for the PCA. For deletions a 'no homopolymer context' means that the base next to the one that is deleted is not of the same type. For insertions a 'no homopolymer context' refers to a base that is inserted 5' to a base of a different type or a single base of the same type. Note that we do not have to consider the preceding bases as all SIM calls were left aligned. A short homopolymer context is defined as a 2-4 bp mononucleotide repeat of the same base as the 1 bp SIM, midsize is 5-7 bp in length and long ≥ 8 bp.

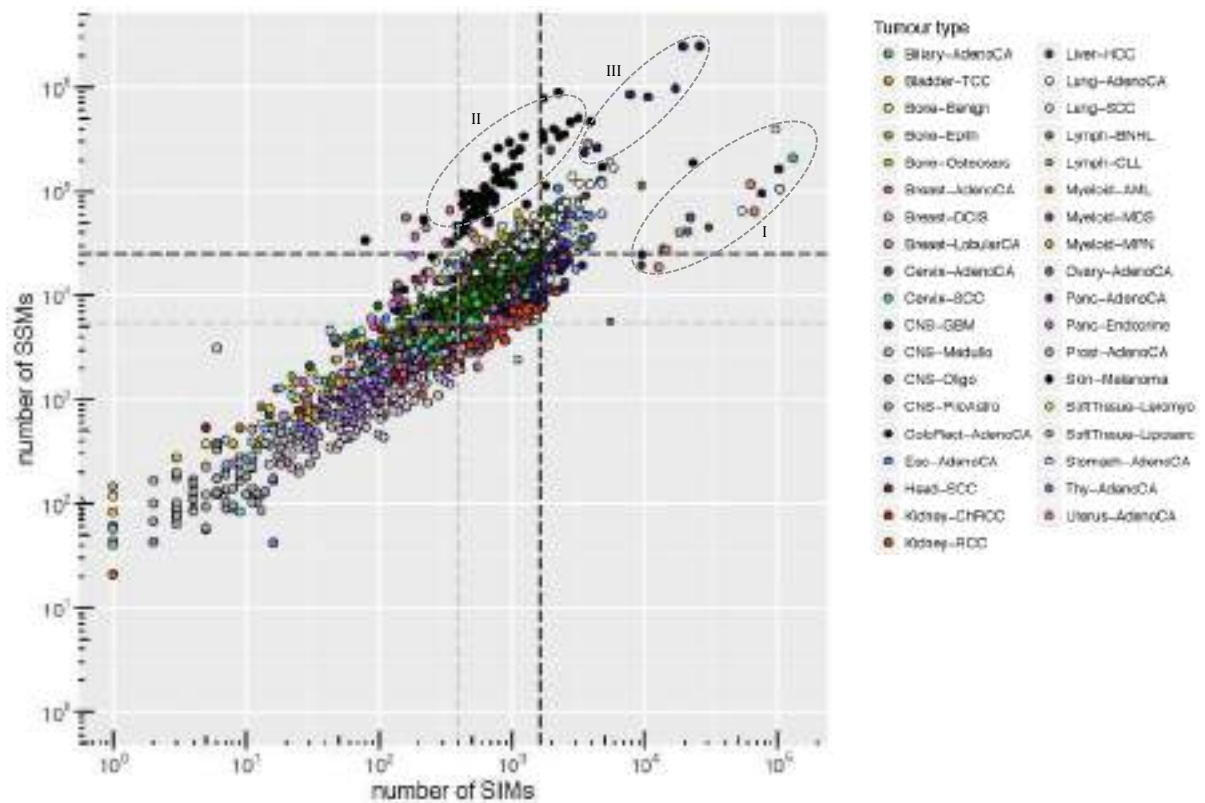


Fig A. Overall mutational burden in terms of SIMs and SSMs per sample.

The two grey lines indicate the median number of SIMs and SSMs, respectively, across the entire cohort. The black lines indicate the $Q3+1.5 \times IQR$. For SIMs there are 184 outliers, the highest number of samples are from Eso-AdenoCA (22.3%), followed by ColoRect-AdenoCA (13.6%) and Lung-SCC (13.0%). For Eso-AdenoCA this corresponds to 42.3% of the samples, 48.1% for ColoRect-AdenoCA and 51.1% for the Lung-SCC. Highlighted in the plot (I) are samples with a high mutational load, which have a particularly high proportion of SIMs. For SSMs there are 255 outliers of which the highest number of samples are from Skin-Melanoma (29.8%), followed by Eso-AdenoCA (16.1%) and Lung-SCC (14.9%). This corresponds for Skin-Melanoma to 71.0% of the samples, 42.3% for Eso-AdenoCA and 80.9% for Lung-SCC. The outliers of Skin-Melanoma (II) are above the bulk of the samples by having a higher proportion of SSMs. There are 122 samples that are outliers in terms of SIMs and SSMs of which the highest number of samples are from Eso-AdenoCA (23.0%), followed by Lung-SCC (19.7%) and Skin-Melanoma (11.5%). The eight samples highlighted in the plot (III) have a very high number of SSMs, but a lower proportion of SIMs compared to the samples highlighted in I.

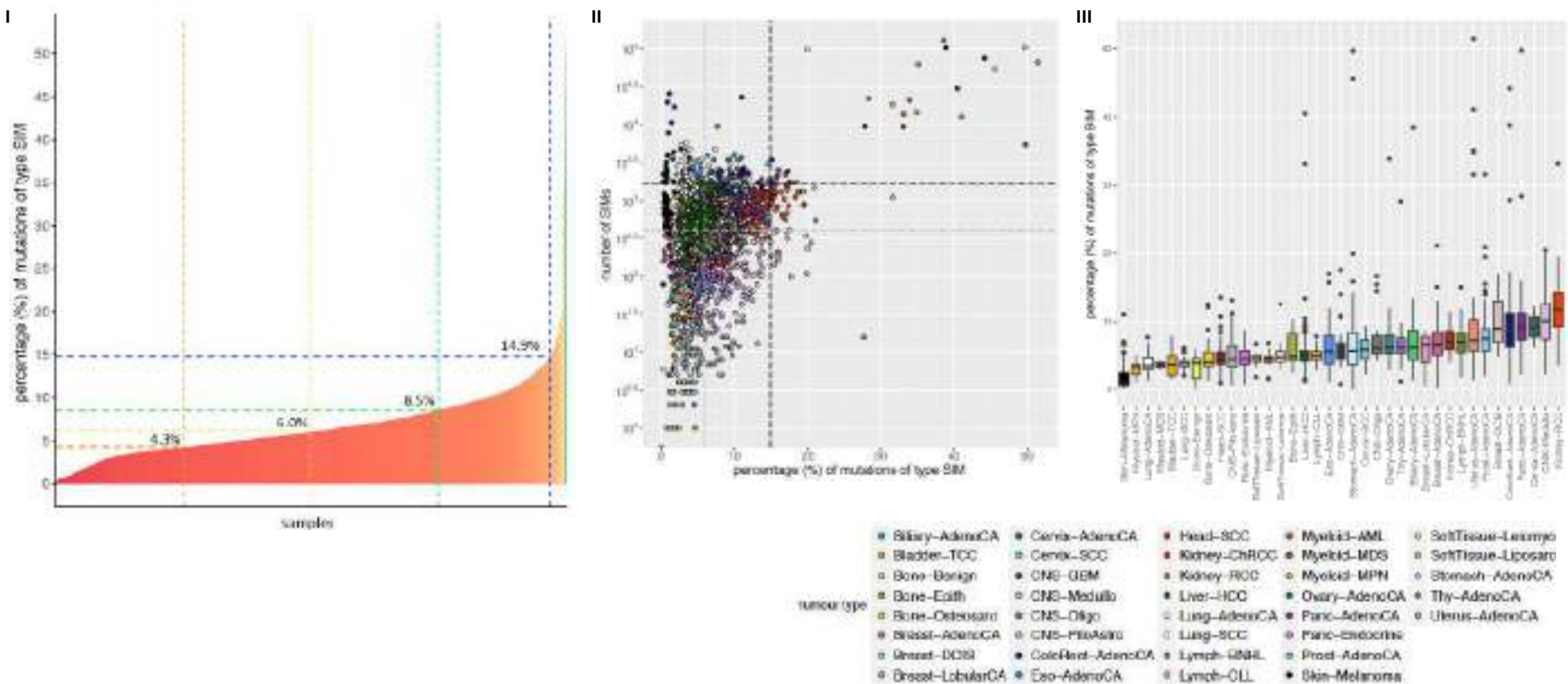


Fig B. The percentage of mutations of type SIM per sample.

(I) The percentage of mutations of type SIM is, with the exception of one Uterus-AdenoCA sample, below 50%. The yellow line indicates the median percentage of mutations of type SIM across the dataset (6.0%). To the right of the vertical yellow line the samples have a percentage above the median. The orange (4.3%) and green (8.5%) lines indicate the first and third quartile, respectively. The Q1-1.5xIQR is equal to 0% and is not shown. The blue line indicates the Q3+1.5xIQR (14.9%) to the right of which samples are outliers. (II) The percentage of mutations of type SIM versus the number of SIMs per sample. The grey lines indicate the medians and the black lines indicate the Q3+1.5xIQR. There are 32 samples from 11 different tumour types that are outliers in terms of percentage and absolute number. This includes 6 samples of ColoRect-AdenoCA and 5 samples each of Uterus-AdenoCA and Kidney-RCC. (III) Boxplots representing the percentage of mutations of type SIM show considerable variability among tumour types. They are ordered according to the median.

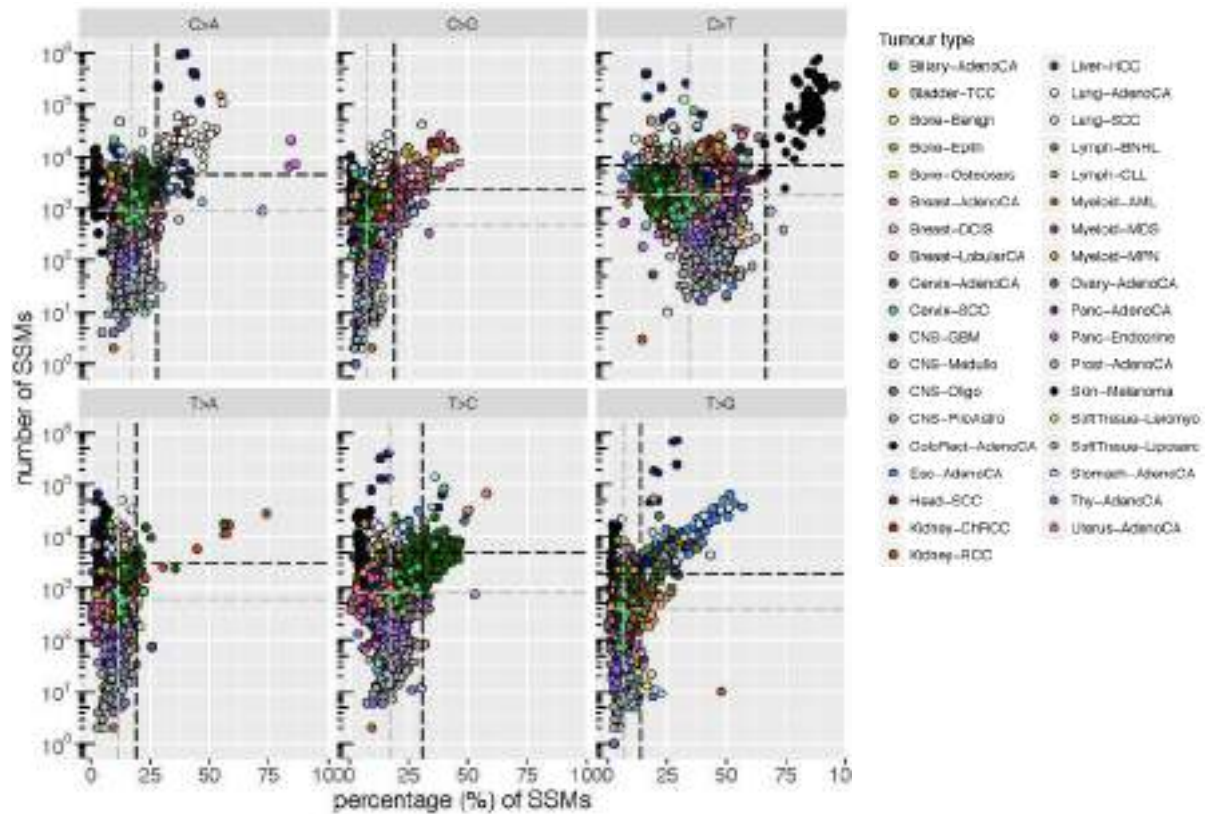


Fig C. Absolute and relative number of SSMs across the six subtypes.

Shown for each sample are the percentage of SSMs of the indicated subtype and the corresponding absolute number. Per sample the six percentages sum up to 100%. The grey lines indicate the medians and the black lines the $Q3+1.5 \times IQR$ based on, for the vertical lines, the percentage of SSMs of the particular subtype, and for the horizontal lines, the absolute numbers. The median percentage across the entire dataset is highest for C>T (34.2%), followed by C>A and T>C (both 17.0%), T>A (11.5%), C>G (7.7%) and T>G (6.6%). For each of the six subtypes there are a number of samples that are outliers in terms of percentage and absolute number. For the C>A SSMs there are 78 outliers from eight different tumour types of which the highest number of samples are from Lung-SCC (46.2%), followed by Lung-AdenoCA (24.4%) and ColoRect-AdenoCA (16.7%). This corresponds for Lung-SCC to 76.6% of the samples, 51.4% for Lung-AdenoCA and 25% for ColoRect-AdenoCA. There are 84 outliers for C>G SSMs from 11 different tumour types of which the highest number of samples are from Breast-AdenoCA (32.1%), followed by Bladder-SCC and Head-SCC (17.9% for both). This corresponds for Breast-AdenoCA to 13.8% of the samples, 65.2% for Bladder-SCC and 26.8% for Head-SCC. For the C>T SSMs there are 80 outliers of which 79 are from Skin-Melanoma and 1 from CNS-GBM. For Skin-Melanoma this corresponds to 73.8% of the samples. For T>A SSMs there are only 11 outliers of which 6 are from Liver-HCC and 5 from Kidney-RCC. For the T>C SSMs there are 85 outliers from 7 different tumour types of which 87.1% are from Liver-HCC. This corresponds to 23.6% of the total number of Liver-HCC samples. Finally, for T>G SSMs there are 146 outliers from 13 different tumour types of which the highest number of samples are from Eso-AdenoCA (48.6%), followed by Lymph-BNHL (19.9%) and Stomach-AdenoCA (13.7%). This corresponds for Eso-AdenoCA to 73.2% of the samples, 27.1% for Lymph-BNHL and 29.4% for Stomach-AdenoCA.

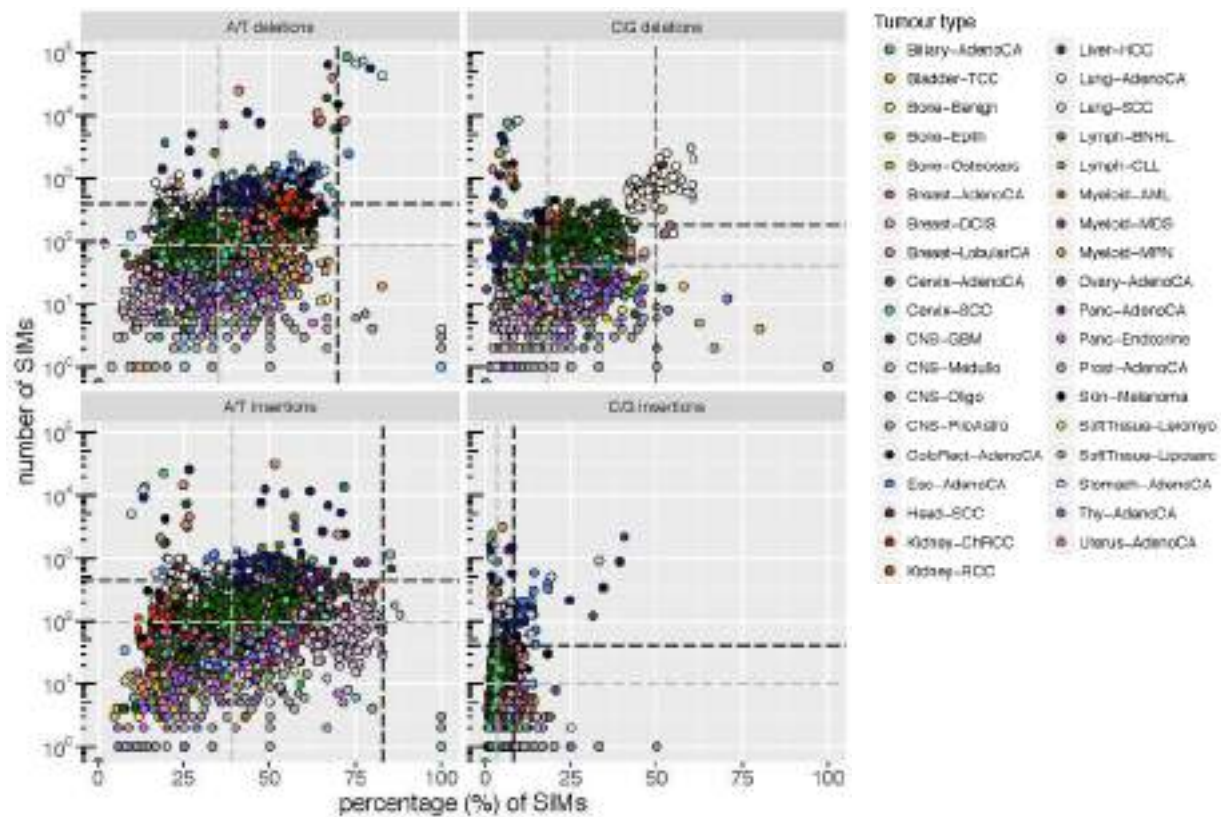


Fig D. Absolute and relative number of 1 bp SIMs across the four subtypes.

Shown for each sample are the percentage of SIMs of the indicated subtype and the corresponding absolute number. Per sample the four percentages sum up to 100%. The grey lines indicate the medians and the black lines the $Q3+1.5 \times IQR$ based on, for the vertical lines, the percentage of SIMs of the particular subtype, and for the horizontal lines, the absolute numbers. The median percentage across the entire cohort is highest for 1 bp A/T insertions (38.8%), followed by 1 bp A/T deletions (35.2%), 1 bp C/G deletions (18.2%), and 1 bp C/G insertions (3.7%). Due to the large range of percentages for the 1 bp A/T deletions and insertions there are only 7 and 2 outliers, respectively, in terms of percentage and absolute number. There are 405 samples for which at least 50% of the 1 bp SIMs are A/T deletions. For three tumour types this holds for half or more of their samples: Kidney-RCC (71.3%), Skin-Melanoma (51.4%) and Lymph-CLL (50.0%). For 1 bp A/T insertions there are 630 samples for which this subtype makes up at least 50% of their 1 bp SIMs. For four tumour types this holds for half or more of their samples: Cervix-AdenoCA (100%, 2 samples), CNS-Medullo (87.2%), Cervix-SCC (72.2%) and Panc-AdenoCA (69.0%). For the 1 bp C/G deletions there are 23 outliers in terms of percentage and absolute number of which 11 are from Lung-AdenoCA, 10 from Lung-SCC, 1 each from Blader-TCC and Head-SCC. Interestingly, for these outliers 1 bp C/G deletions are the majority of their 1 bp SIMs. For 1 bp C/G insertions there are 39 outliers of which 16 are from Eso-AdenoCA, 10 from ColoRect-AdenoCA, 6 from Stomach-AdenoCA, 4 from Panc-AdenoCA and 3 from Skin-Melanoma.

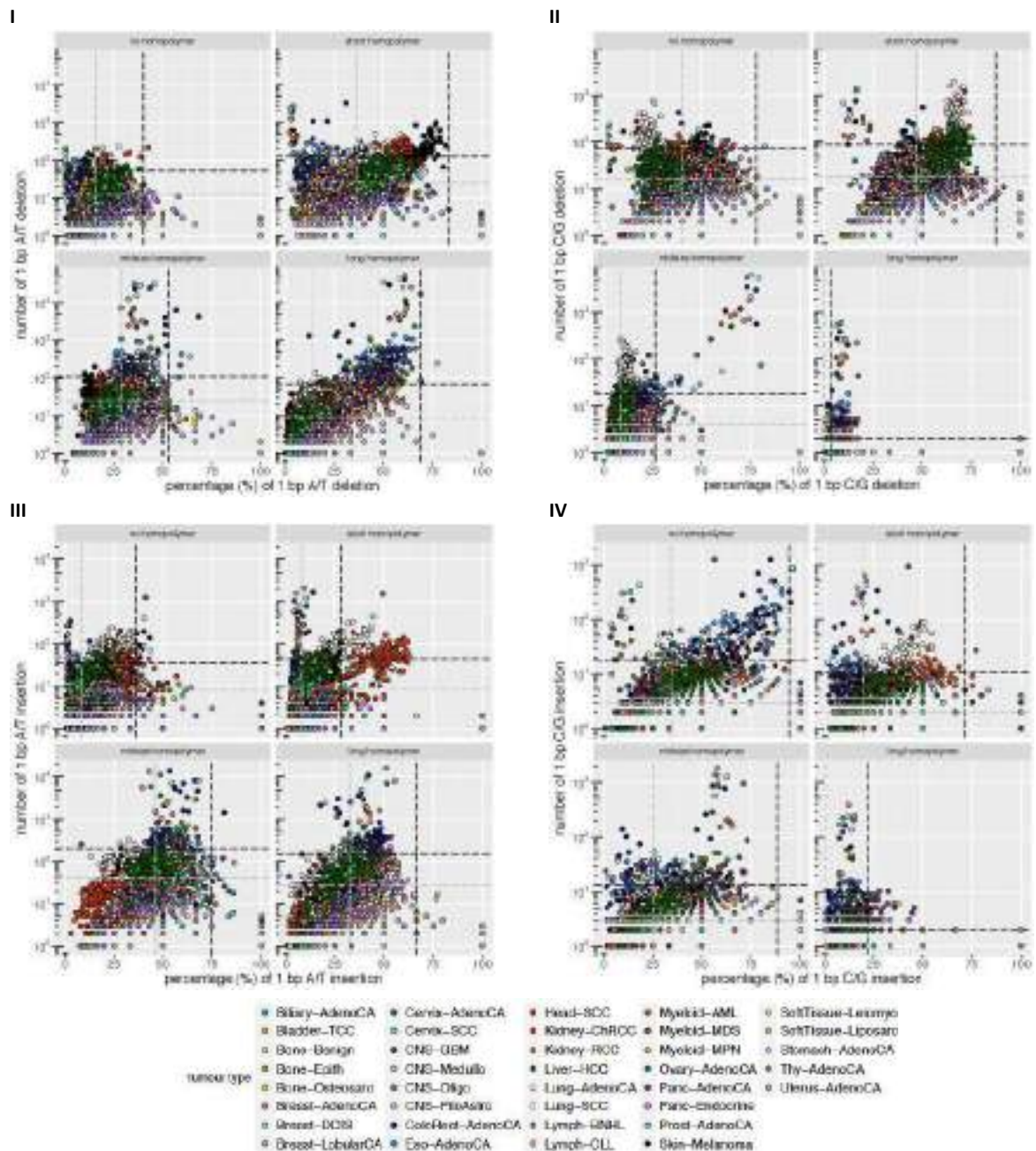


Fig E. Homopolymer context of 1 bp SIMs.

For each of the four SIM subtypes we computed per sample the percentage of 1 bp SIMs in the four homopolymer contexts (see **Main Text**). The grey lines indicate the medians and the black lines the $Q3+1.5 \times IQR$ based on, for the vertical lines, the percentage of SIMs in the particular homopolymer context, and for the horizontal lines, the absolute numbers. For most contexts, there are few outliers (12 or less) in terms of percentage and absolute number. Exceptions are the midsize and long homopolymer context for 1 bp C/G deletions (33 and 161 cases, respectively), short homopolymer context for 1 bp A/T insertions (102 cases) and long homopolymer context for 1 bp C/G insertions (40 cases). For a number of samples more than 50% of a particular SIM subtype is in one of the four homopolymer contexts. These are for (I) 1 bp A/T deletions: 13 samples in no, 487 samples in a short, 77 samples in a midsize, and 174 samples in a long homopolymer context; (II) 1 bp C/G deletions: 507 samples in no, 1,013 samples in a short, 22 samples in a midsize and 3 samples in a long homopolymer context; (III) 1 bp A/T insertions: 18 samples in no, 66 samples in a short, 852 samples in a midsize and 100 samples in a long homopolymer context; (IV) 1 bp C/G insertions: 608 samples in no, 165 samples in a short, 321 samples in a midsize and 9 samples in a long homopolymer context.

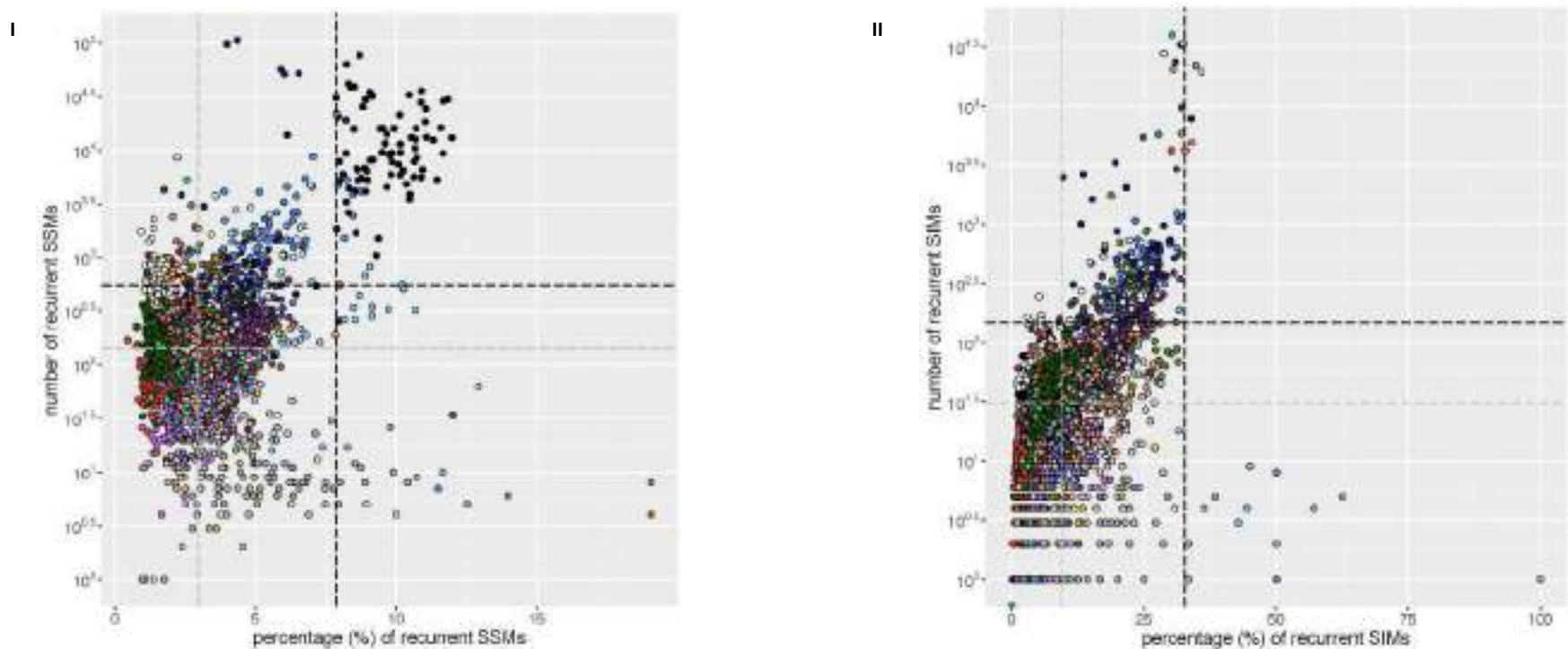
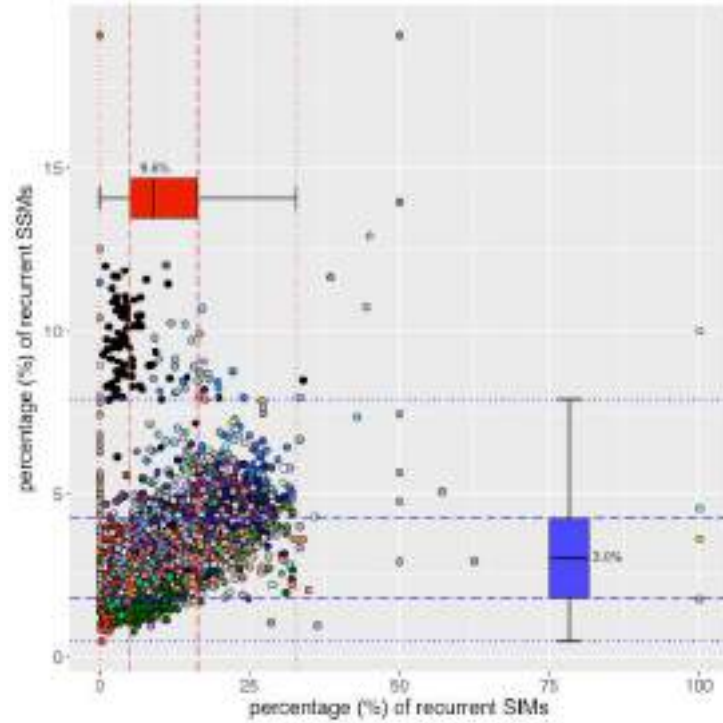


Fig F. Overall level of recurrence in terms of SSMs and SIMs per sample.

(I) The percentage versus the absolute number of recurrent SSMs. The grey lines indicate the medians and the black lines the $Q3+1.5 \times IQR$ based on, for the vertical lines, the percentage of recurrent SSMs and, for the horizontal lines, the absolute numbers. There are 89 samples that are outliers in both relative and absolute terms of which 77 are Skin-Melanoma samples. Only based on absolute number, there are 333 outliers of which 24.6% are Skin-Melanoma samples, followed by 22.2% Eso-AdenoCA samples. Lung-SCC samples have a high absolute number of recurrent SSMs, but the percentage that is recurrent is below the median. (II) The percentage versus the absolute number of recurrent SIMs. The grey lines indicate the medians and the black lines the $Q3+1.5 \times IQR$ based on, for the vertical lines, the percentage of recurrent SIMs and, for the horizontal lines, the absolute numbers. There are only 4 outliers for both measurements and 295 if we instead base it only on absolute number of recurrent SIMs of which the largest percentage are Eso-AdenoCA samples (25.4%), followed by Panc-AdenoCA (19.0%) and ColoRect-AdenoCA (15.9%). Noticeable is the group of eight samples from four different tumour types, each of which has over 19,000 recurrent SIMs and at least 28.7% of the SIMs are recurrent.

(continues below)

III



(continued from above)

(III) Percentage of recurrent SIMs versus recurrent SSMs. The red boxplot corresponds to the recurrent SIMs and the blue boxplot to the recurrent SSMs. There are 344 samples from 21 different tumour types for which the percentage of recurrent SSMs and SIMs are both above the third quartile. The four tumour types for which half or more of their samples are in this set: Eso-AdenoCA (54 out of 97), ColoRect-AdenoCA (28 out of 52), Panc-AdenoCA (116 out of 232) and Cervix-AdenoCA (1 out of 2). There are 381 samples from 18 different tumour types for which both percentages are below the first quartile. For Kidney-RCC 88.8% of the samples are in this set. This is followed by Lung-AdenoCA with 48.6%, Ovary-AdenoCA with 45.5% and Lung-SCC with 44.7%.

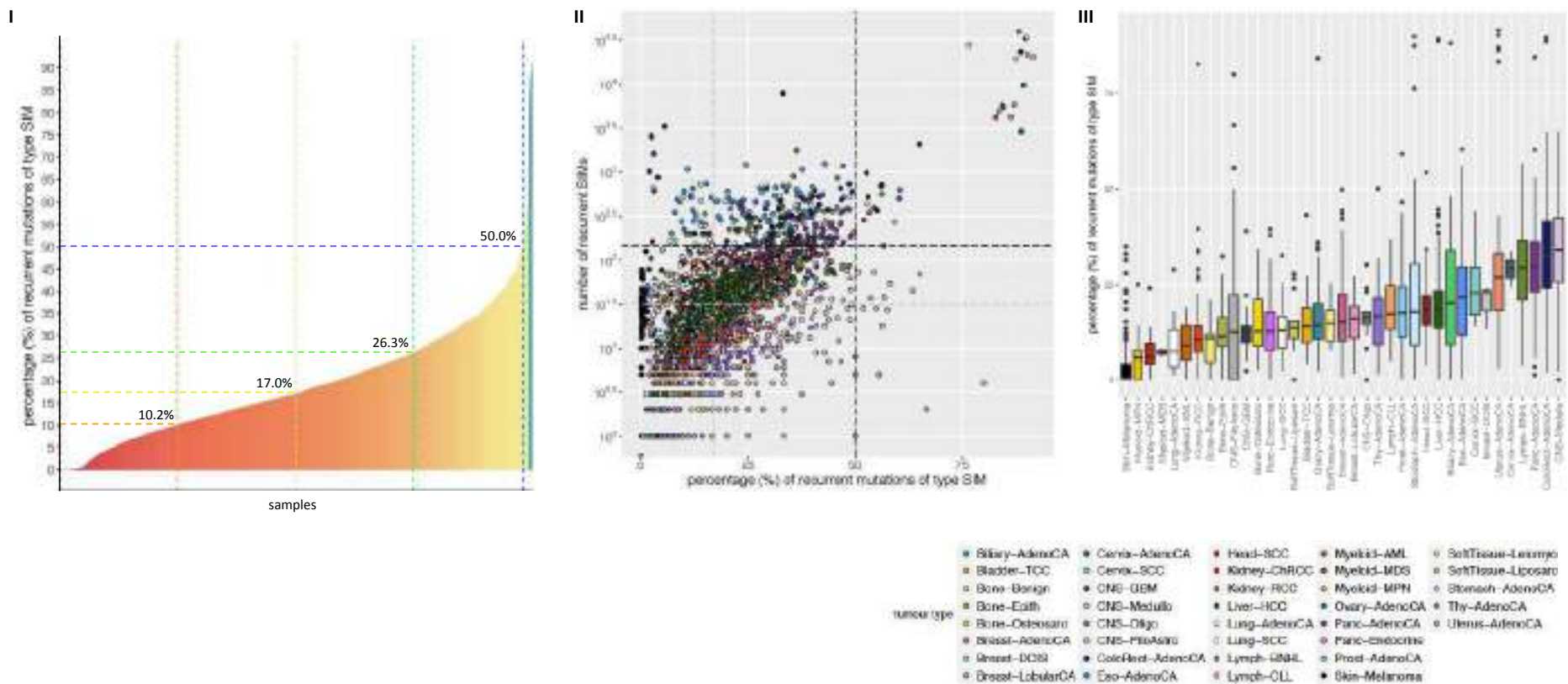


Fig G. The percentage of recurrent mutations of type SIM.

(I) Recurrent mutations show a higher percentage of type SIM than mutations overall. The yellow line indicates the median percentage of mutations of type SIM across the dataset (17%). To the right of the vertical yellow line the samples have a percentage above the median. The orange (10.2%) and green (26.3%) lines indicate the first and third quartile, respectively. The Q1-1.5xIQR is equal to 0% and is not shown. The blue line (50%) indicates the Q3+1.5xIQR, to the right of which samples are outliers. There are 45 samples with more recurrent SIMs than SSMS. (II) The percentage of recurrent mutations of type SIM versus the number of recurrent SIMs per sample. The grey lines indicate the medians and the black lines indicate the Q3+1.5xIQR. There are 30 samples from 12 different tumour types that are outliers in terms of percentage and absolute number. This includes 7 samples from ColoRect-AdenoCA, 5 samples from Uterus-AdenoCA and 4 each from Panc-AdenoCA and Stomach-AdenoCA. (III) The boxplots per tumour type representing the percentage of recurrent mutations of type SIM, which show a considerable variability within and between tumour types. They are ordered according to the median percentage.

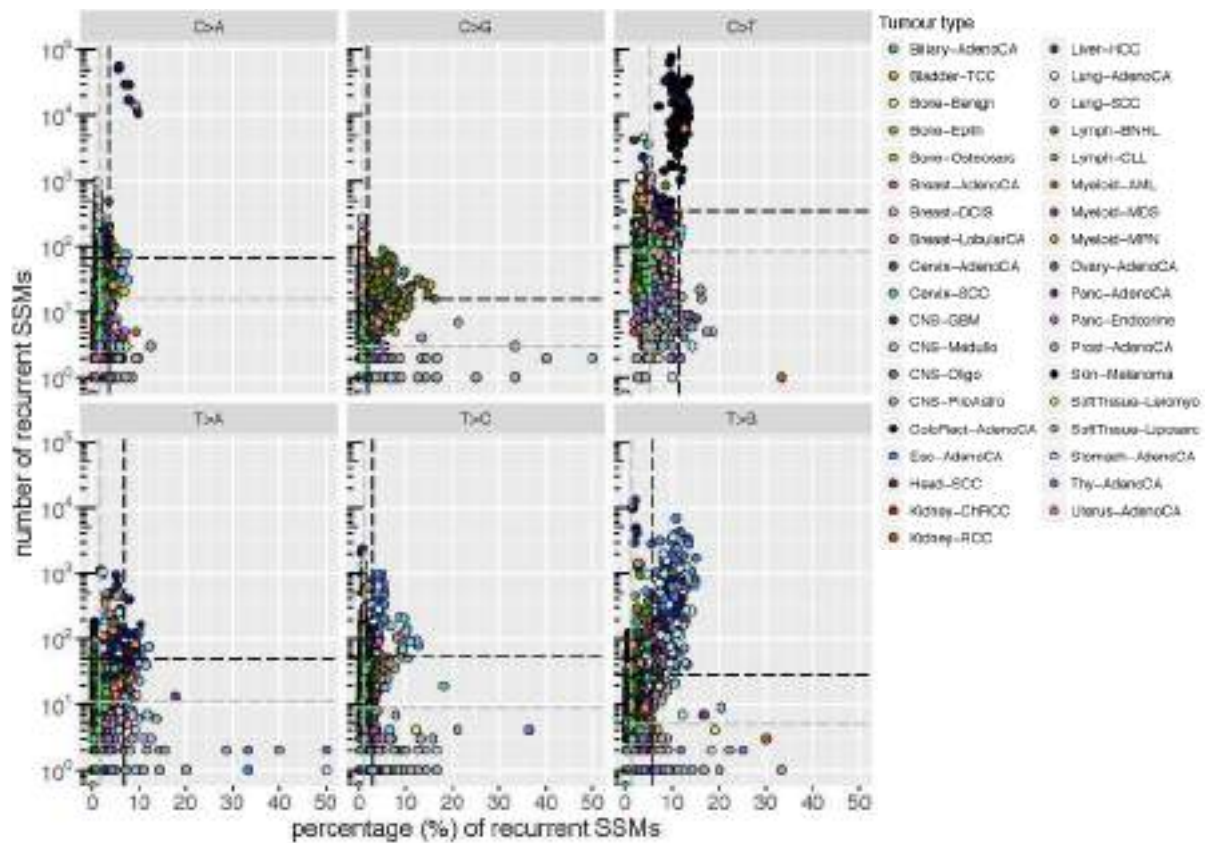


Fig H. Absolute and relative numbers of recurrent SSMs across the six subtypes.

For each sample the percentage and absolute number of recurrent SSMs per subtype is shown. The grey lines indicate the medians and the black lines the $Q3+1.5 \times IQR$ based on, for the vertical lines, the percentage of SSMs of the particular subtype that is recurrent and, for the horizontal lines, the absolute numbers. For C>A SSMs there are 12 samples that are outliers in terms of percentage and number of recurrent SSMs. Of these 12 there are seven ColoRect-AdenoCA samples and one Uterus-AdenoCa sample that particularly stand out. Each has over 10,000 recurrent C>A SSMs and at least 5.6% are recurrent. For C>G SSMs there are 82 outliers of which 62 are from Lymph-BNHL. There are 37 outliers for the C>T SSMs of which 33 are Skin-Melanoma samples. For T>A SSMs there are 17 outliers of which 7 are from ColoRect-AdenoCA and 5 from Prost_AdenoCA. For T>C SSMs there are 99 outliers of which 58 are from Eso-AdenoCA and 17 from Stomach-AdenoCA. Finally, for T>G SSMs there are 187 outliers of which again Eso-AdenoCA and Stomach-AdenoCA form the majority with 83 and 42 samples, respectively.

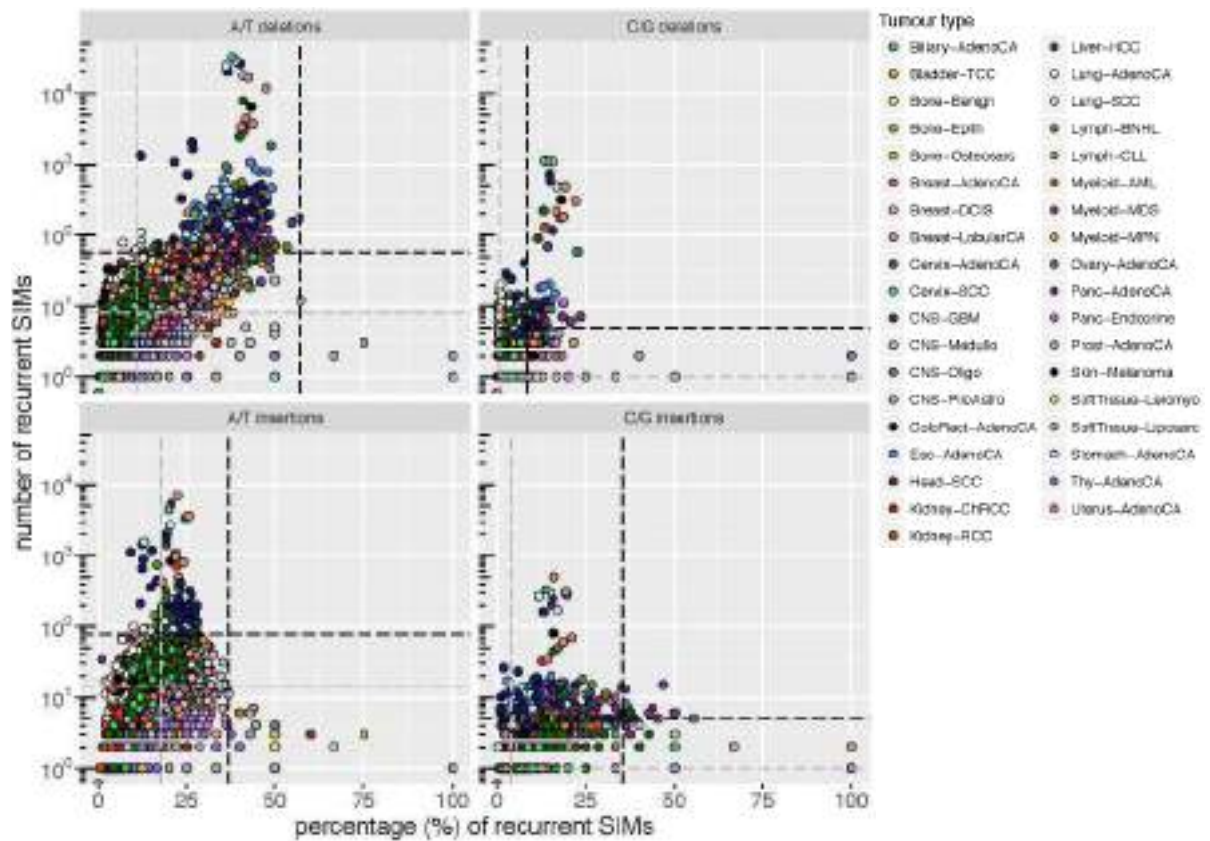
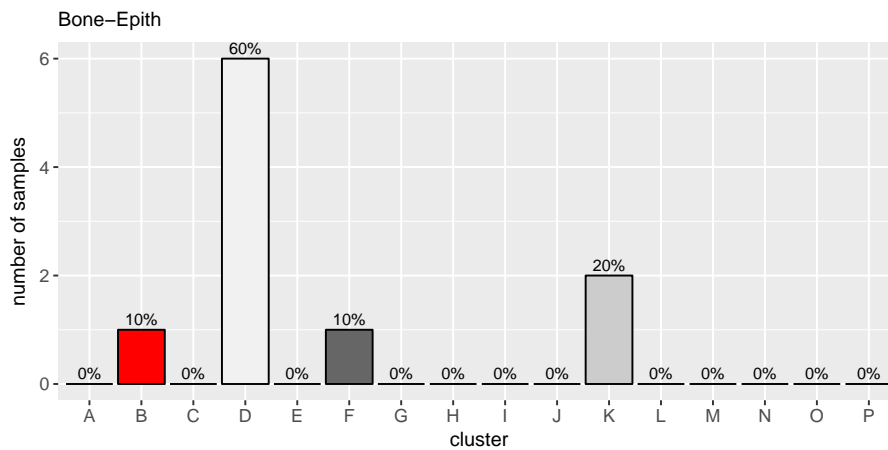
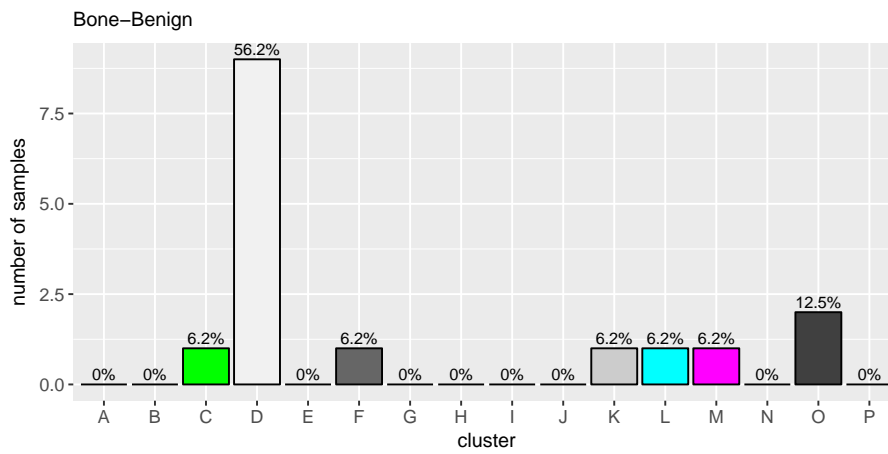
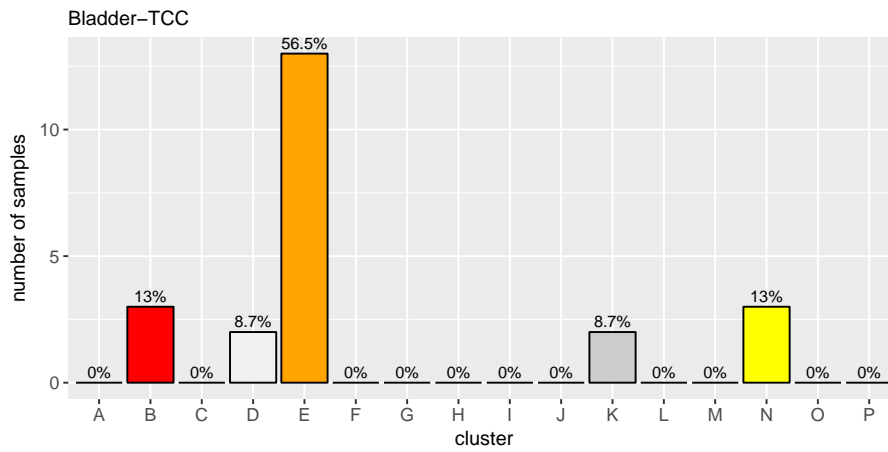
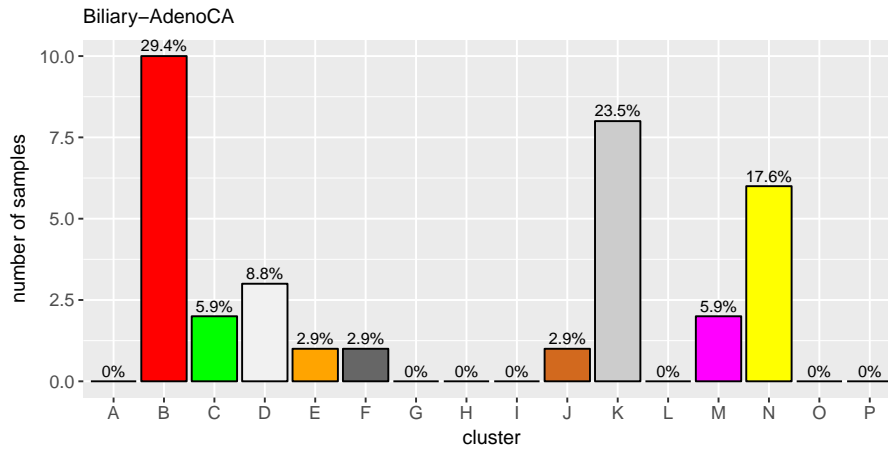
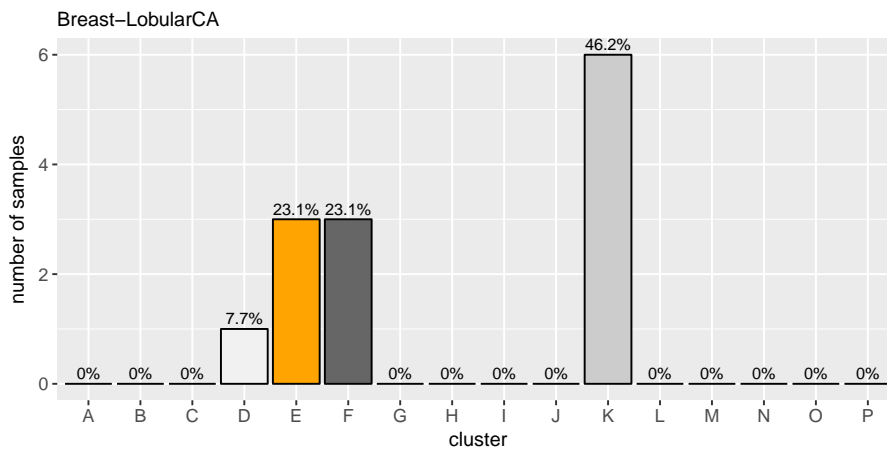
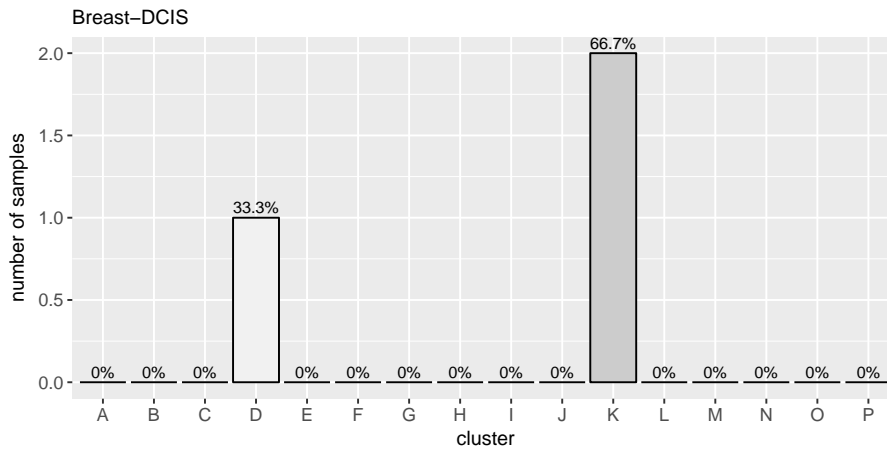
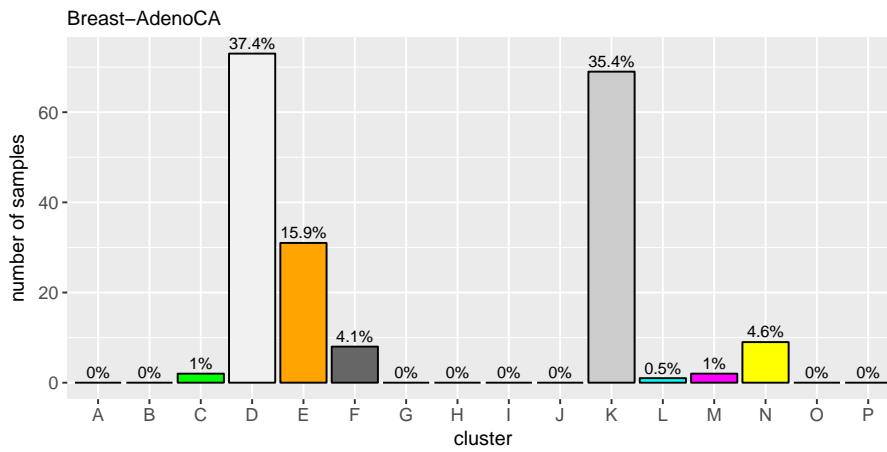
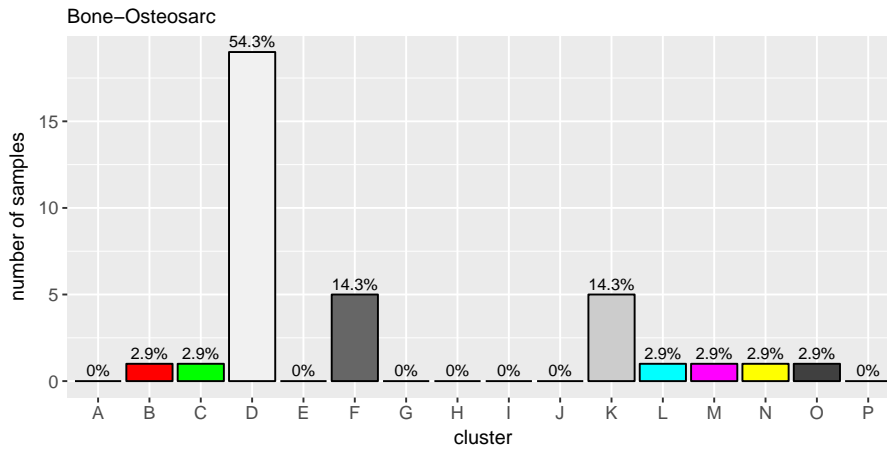


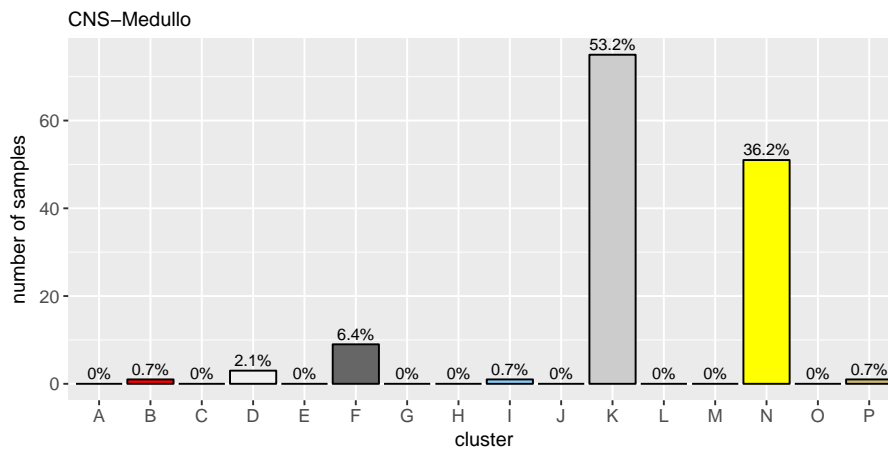
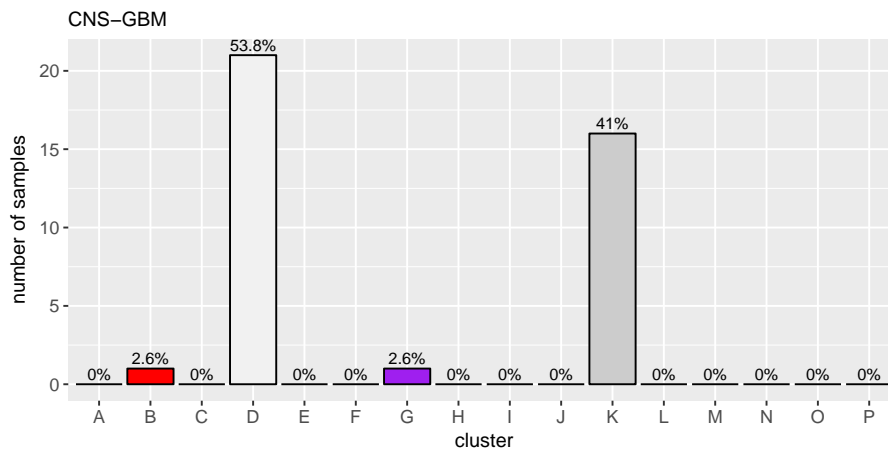
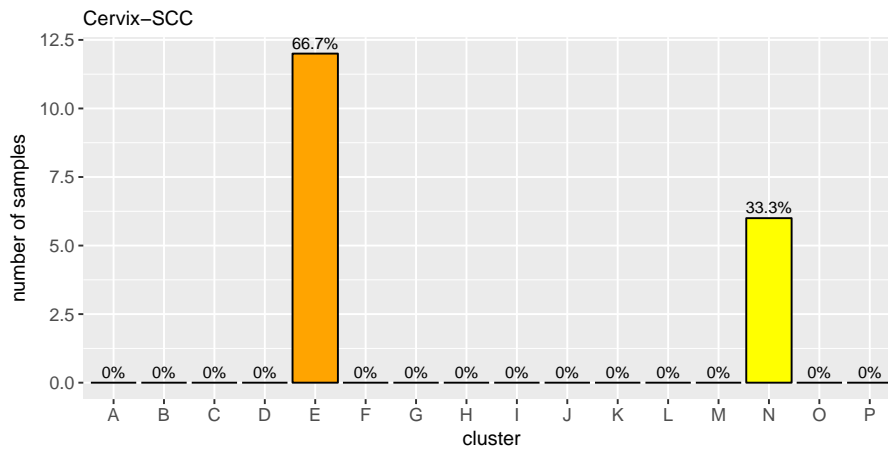
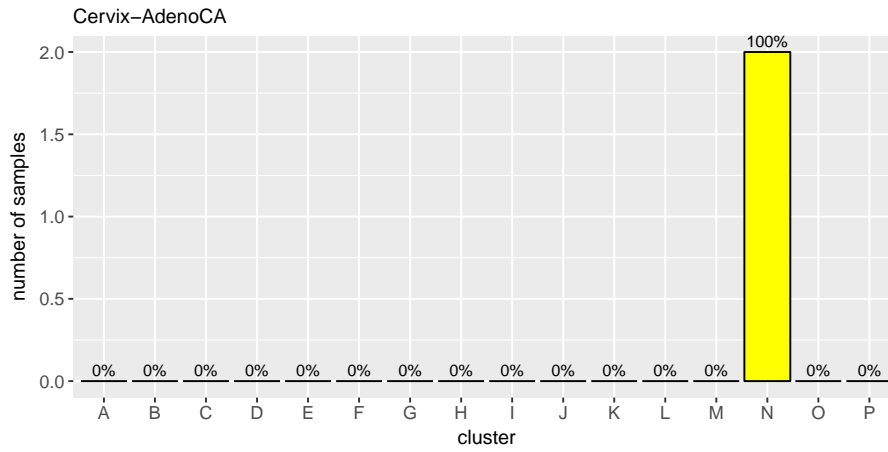
Fig 1. Absolute and relative numbers of recurrent 1 bp SIMs across the four subtypes.

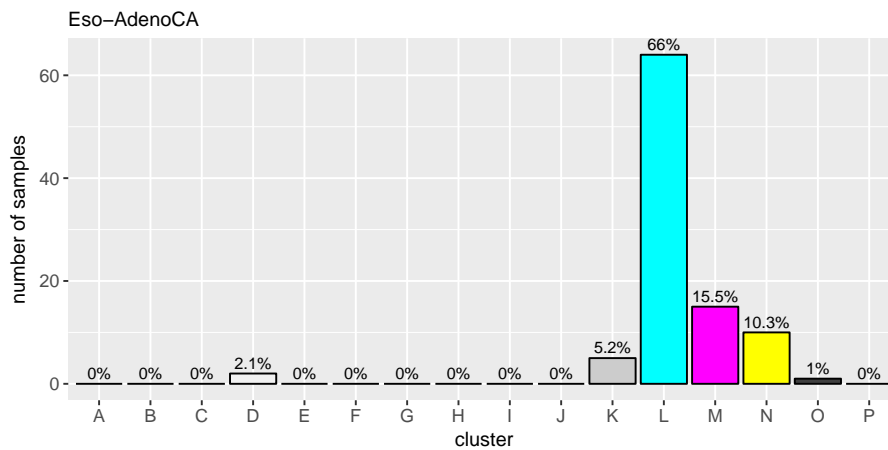
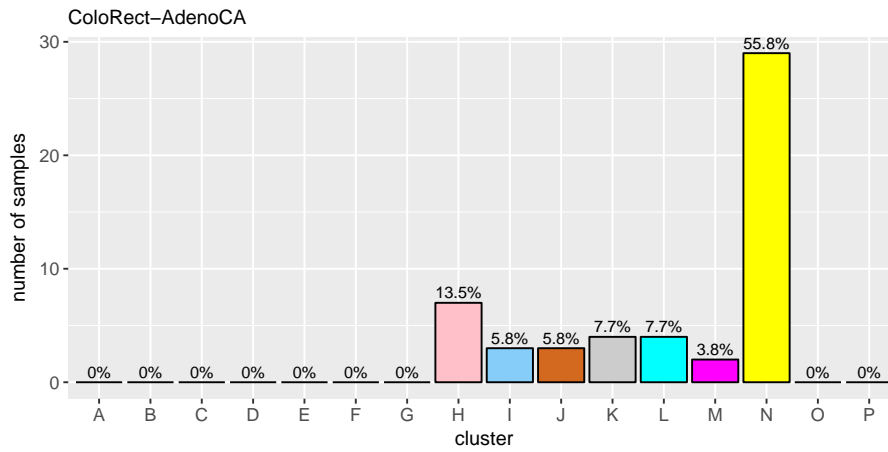
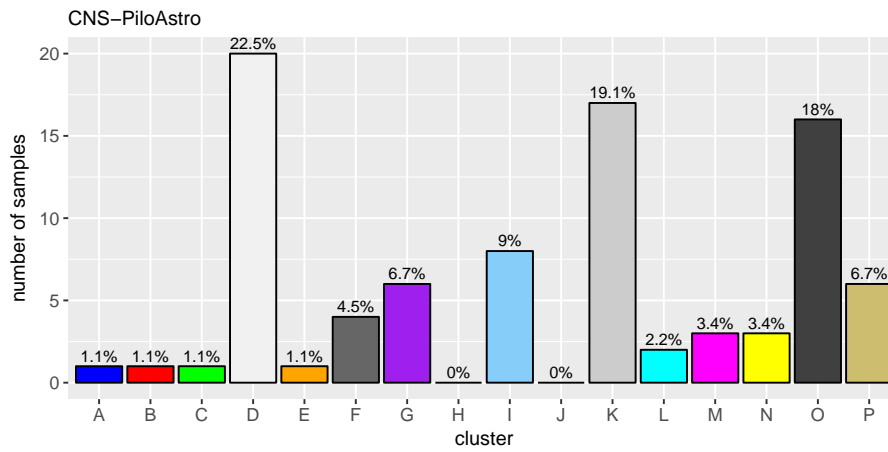
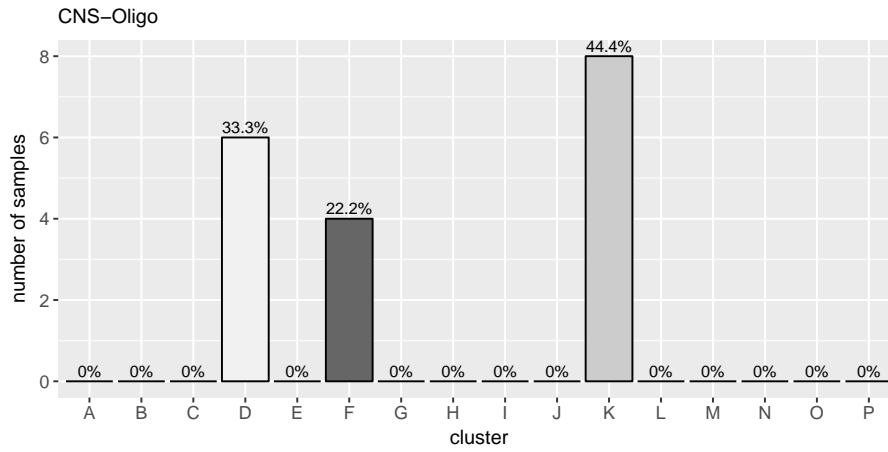
For each sample the percentage and absolute number of recurrent 1 bp SIMs per subtype is shown. The grey lines indicate the medians and the black lines the $Q3+1.5 \times IQR$ based on, for the vertical lines, the percentage of SIMs of the particular subtype that is recurrent and, for the horizontal lines, the absolute numbers. There is a large spread of the percentages for 1 bp A/T deletions and insertions and therefore there are no outliers in terms of percentage and absolute number. There are 352 outliers in terms of absolute number of recurrent 1 bp A/T deletions of which Eso-AdenoCA constitutes the largest percentage (22.4%), followed by Panc-AdenoCA (20.2%) and Lymph-BNHL (18.8%). For the number of recurrent 1 bp A/T insertions there are 236 outliers of which again Eso-AdenoCA contributes the highest percentage of samples (26.7%), followed by ColoRect-AdenoCA (19.5%) and Panc-AdenoCA (16.1%). For recurrent 1 bp C/G deletions there are 58 outliers in terms of percentage and absolute number of which 29.3% are from Eso-AdenoCA and 19.0% from ColoRect-AdenoCA. For recurrent 1 bp C/G insertions there are 9 outliers in terms of percentages and absolute numbers of which 7 are from Panc-AdenoCA and 1 each from Eso-AdenoCA and Liver-HCC.

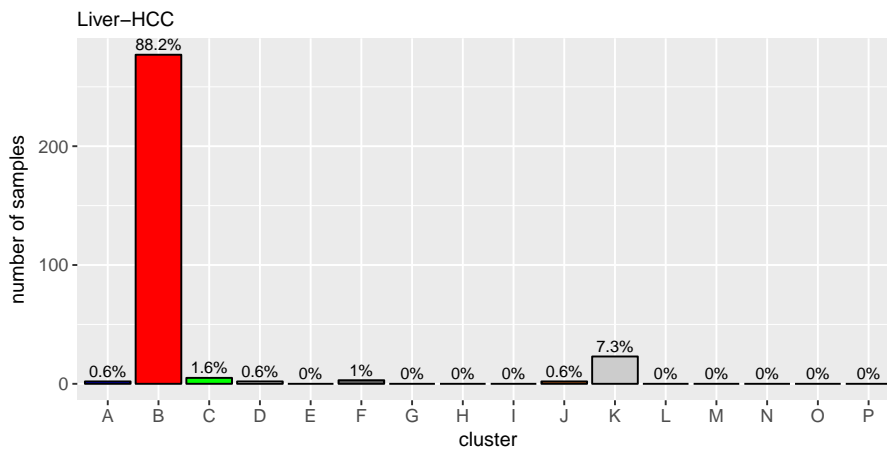
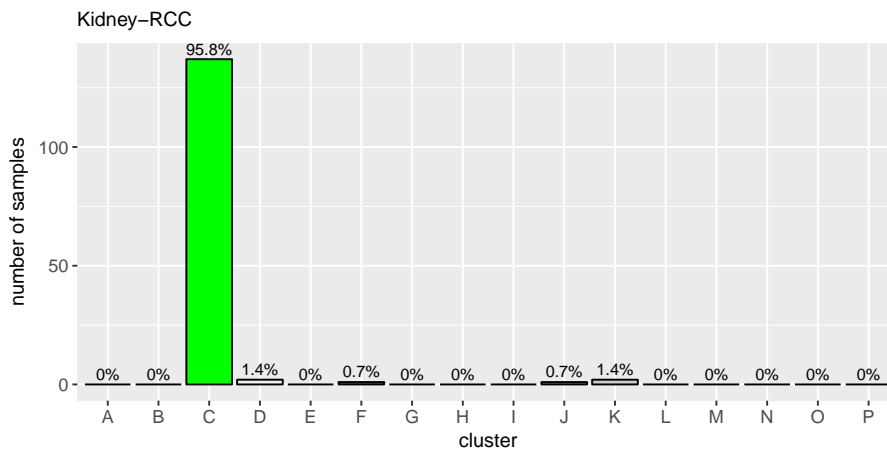
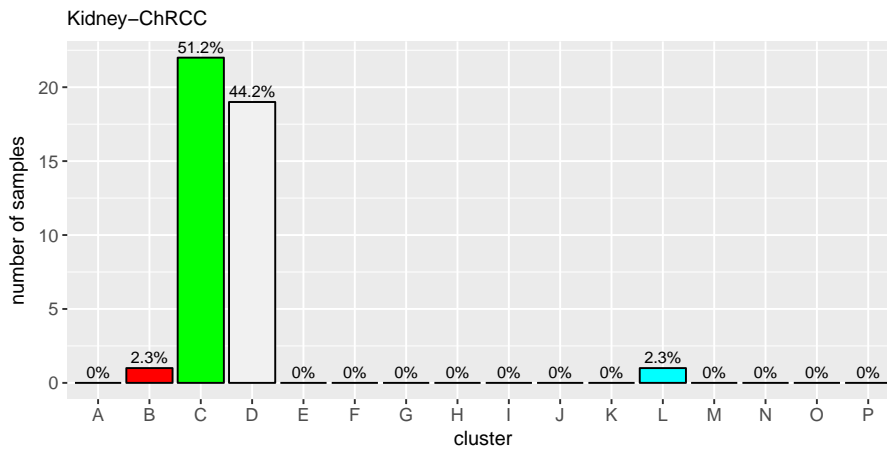
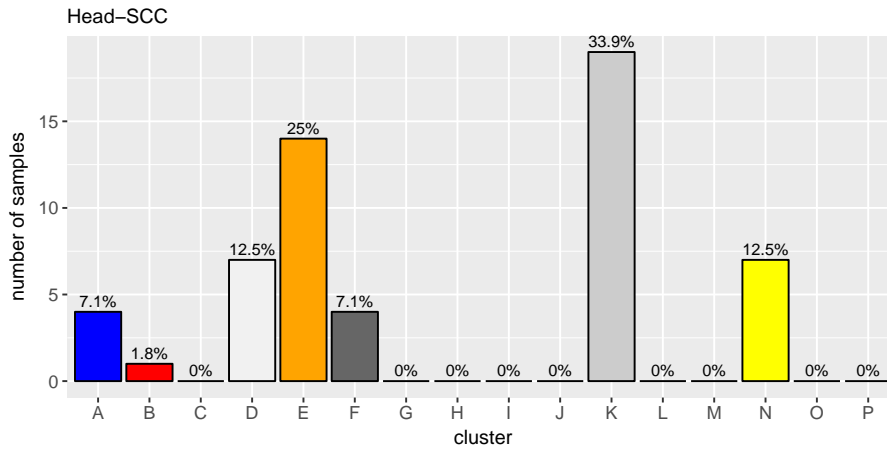
S2 File. Sample distribution per tumour type across the 16 clusters.

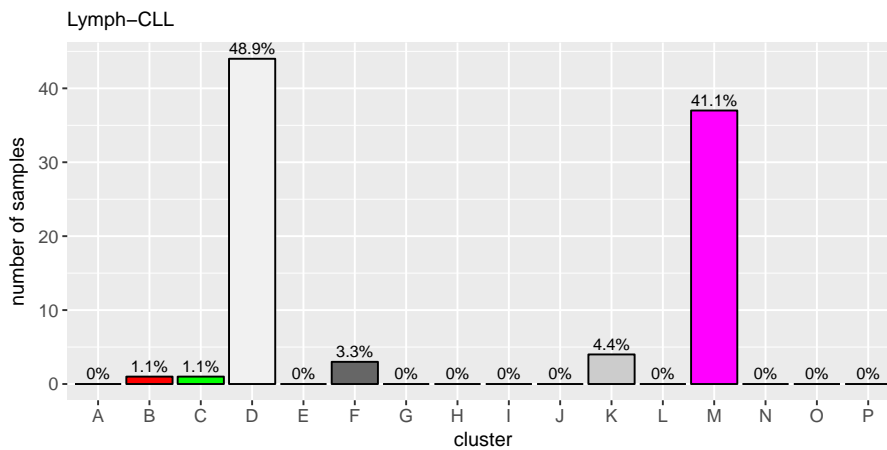
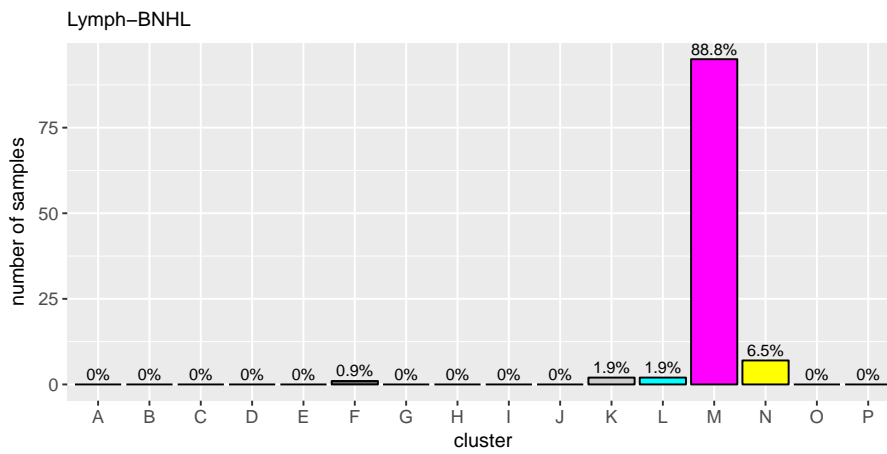
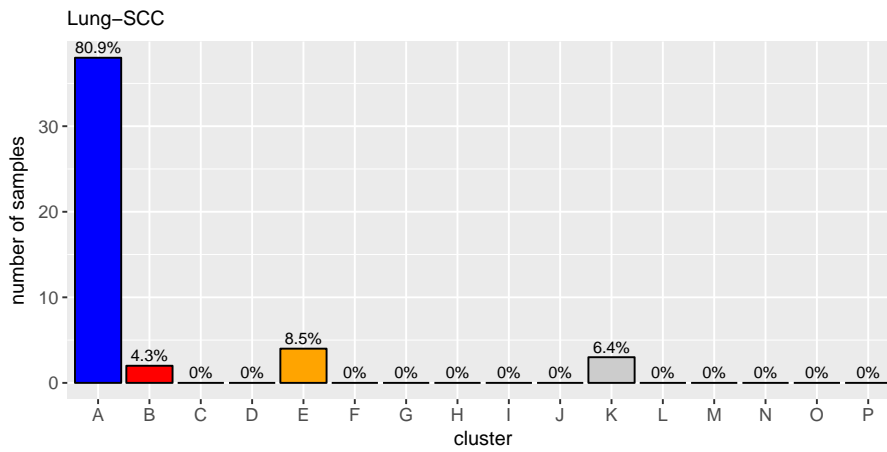
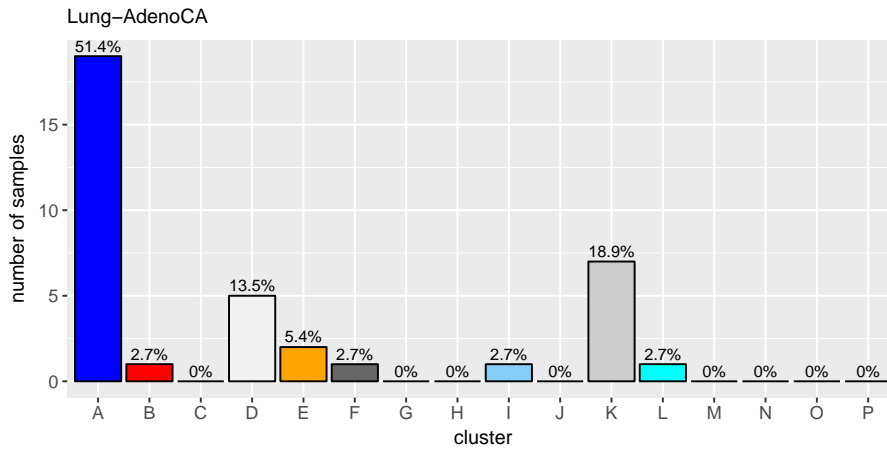


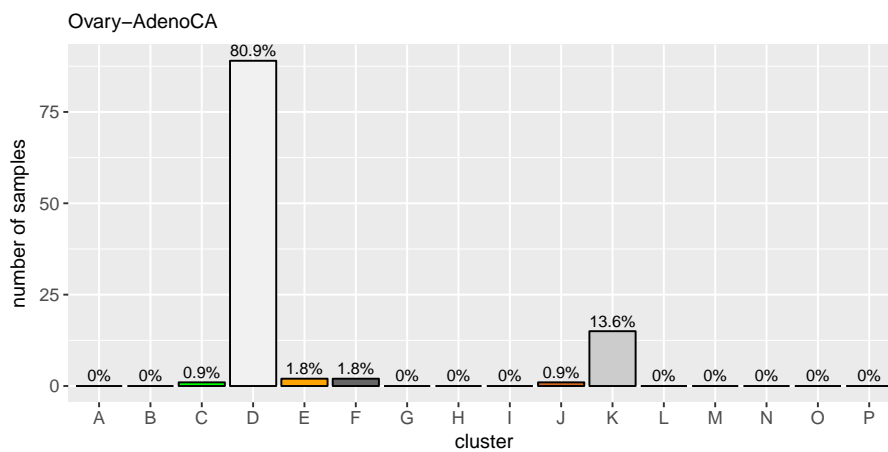
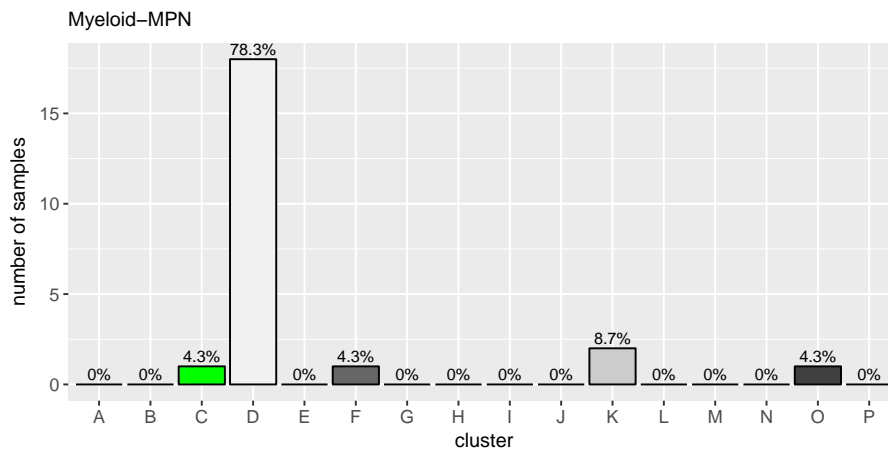
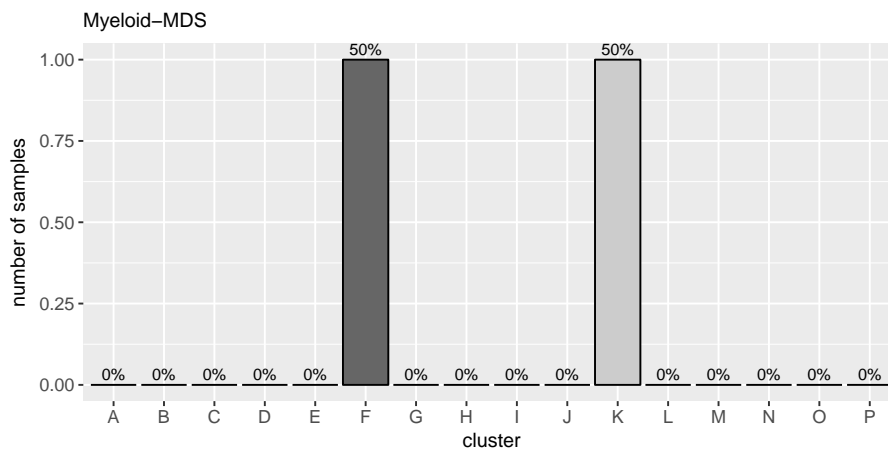
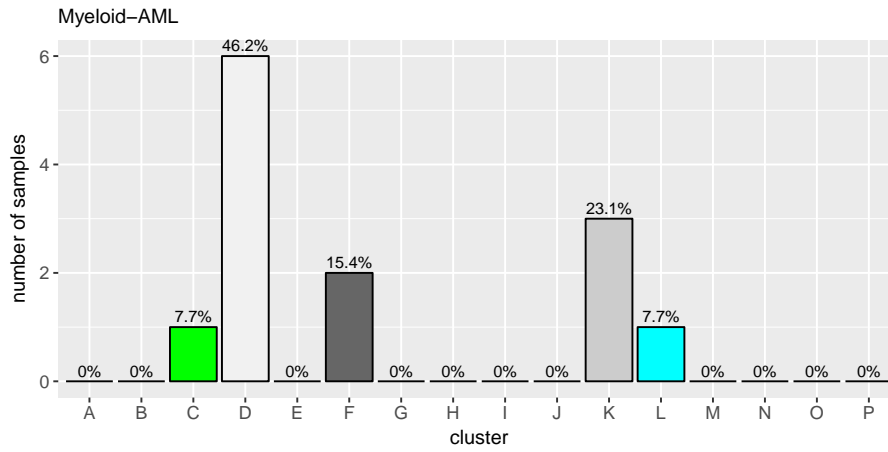


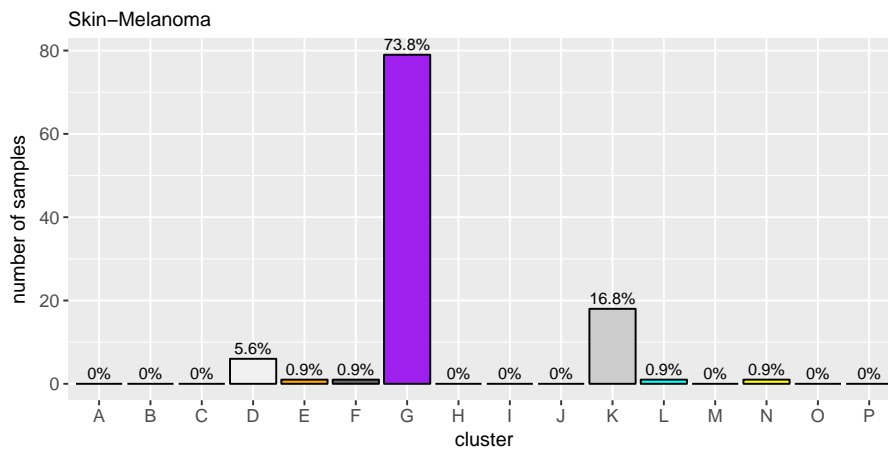
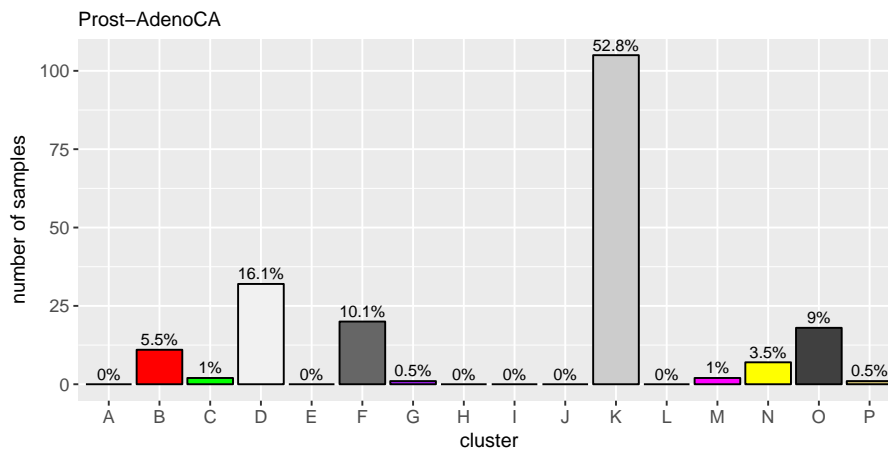
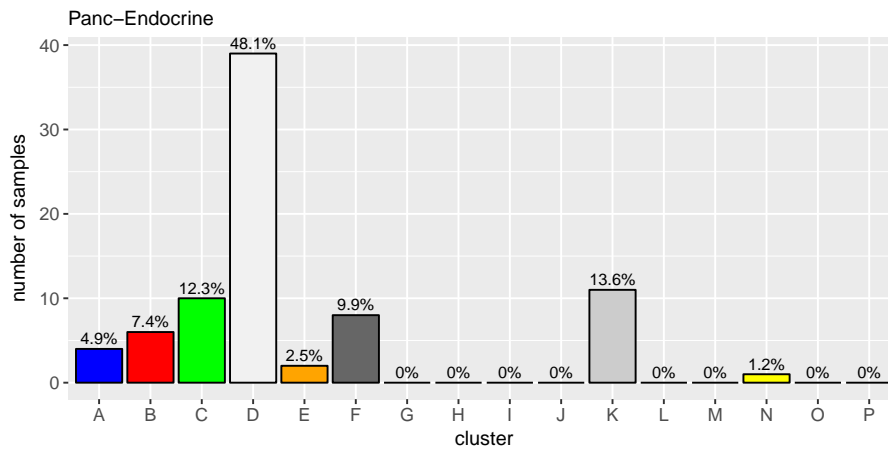
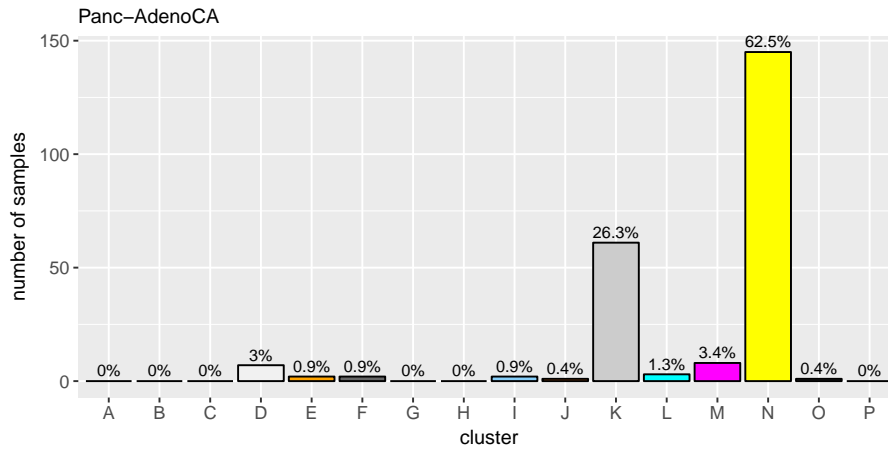


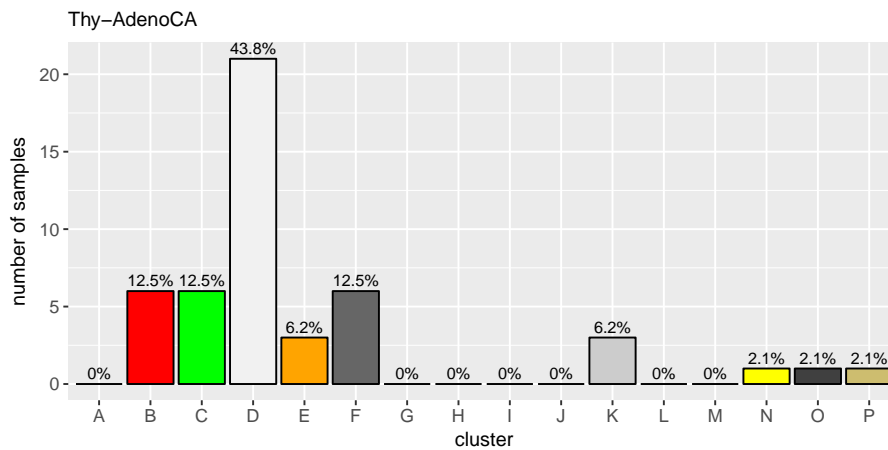
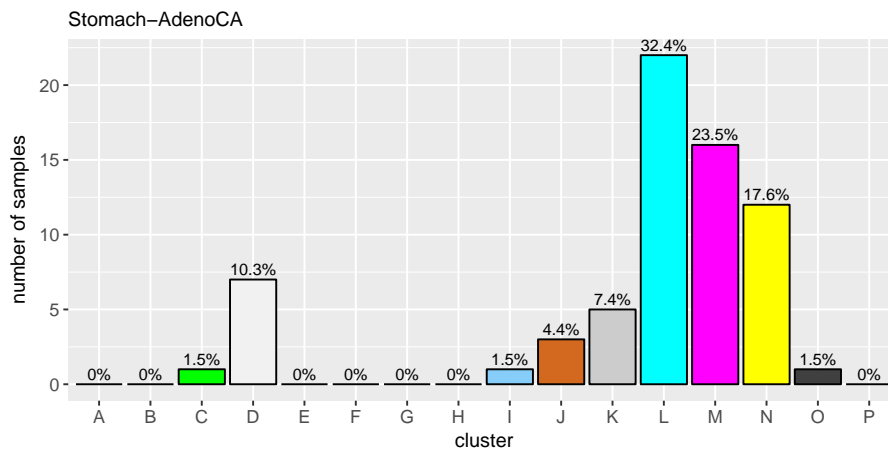
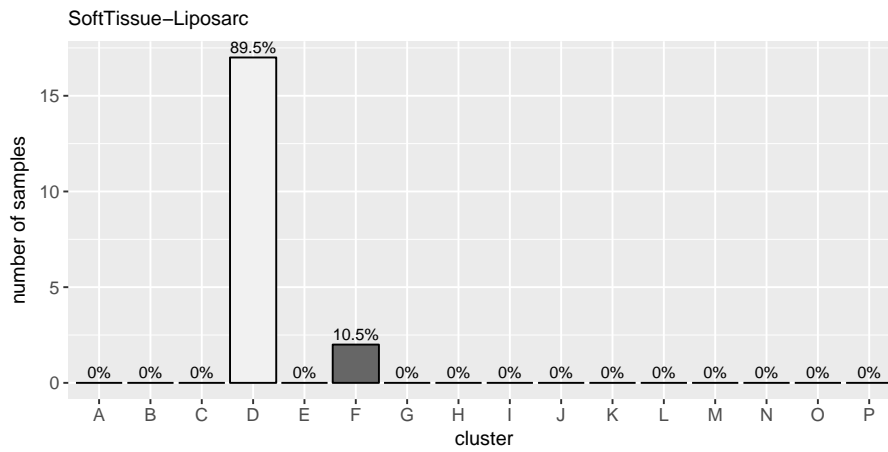
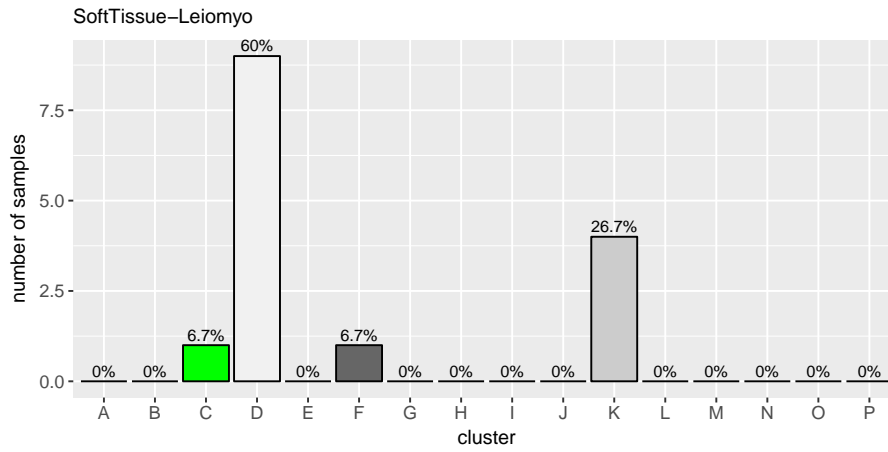


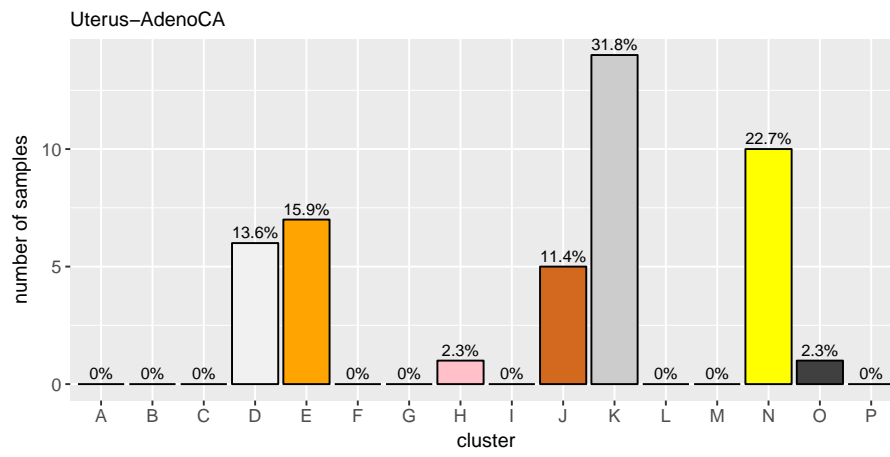






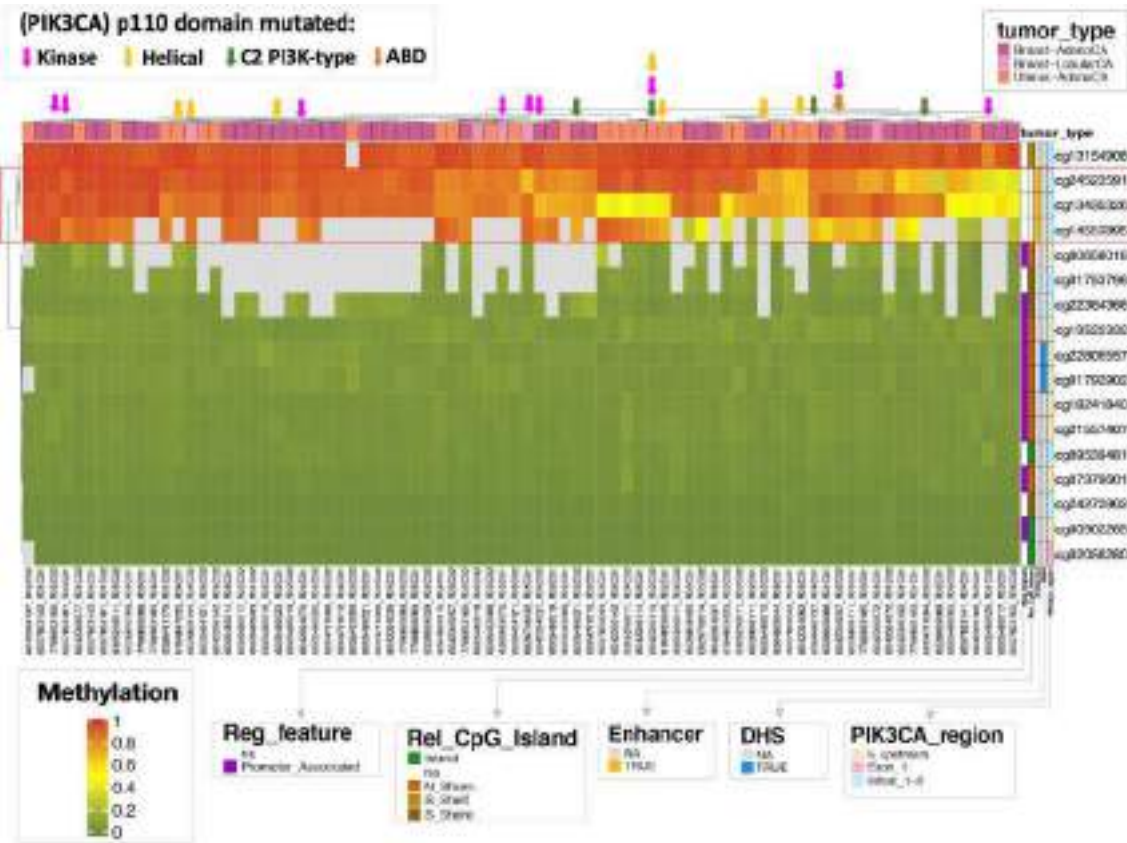




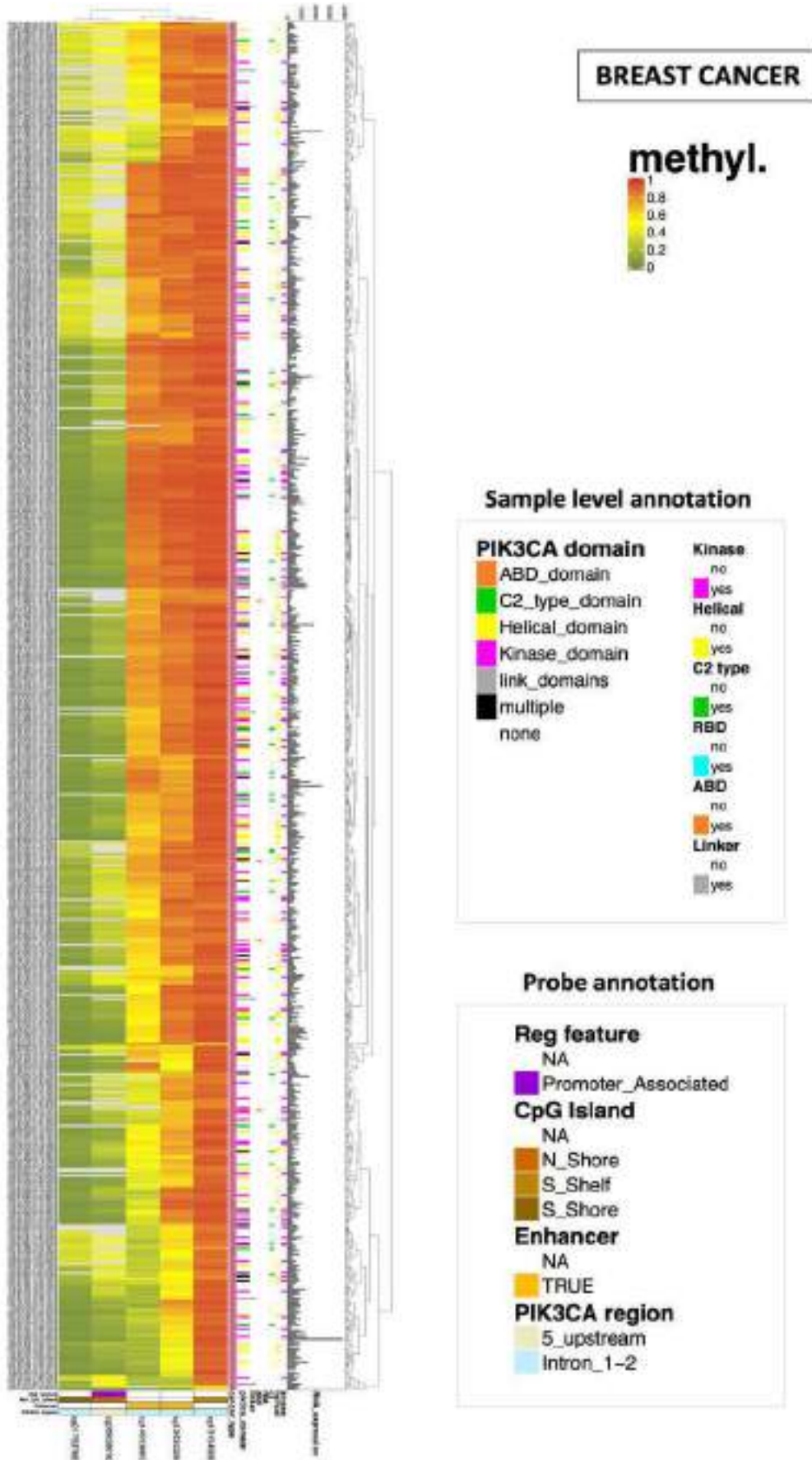


APPENDIX 2.

(A) Methylation (450K array) across breast and uterus cancer samples in PCAWG dataset.



(B) Methylation (450K array) across breast cancer samples in TCGA dataset.



(C) Methylation (450K array) across uterus cancer samples in TCGA dataset.

