Methods and bioinformatic tools to study polymorphic inversions in complex diseases

Carlos Ruiz Arenas

TESI DOCTORAL UPF / 2019

DIRECTOR DE LA TESI

Dr. Juan Ramón González Ruiz

BARCELONA INSTITUTE FOR GLOBAL HEALTH


upf.
Universitat
Pompeu Fabra
Barcelona

A Cecil

Acknowledgements

Esta tesis es la crónica de un fracaso. O, para ser más exactos, es la crónica de cómo un fracaso se convirtió en una oportunidad. Todo empezó en diciembre de 2014. Fue entonces cuando me hice una entrevista para hacer un doctorado industrial en QGenomics, entrevista que no superé con éxito.

Pero en aquella entrevista conocí a los que serían el supervisor de mi tesis, Juan R González, y mi tutor, Luis Pérez Jurado. Es gracias a ellos que esta tesis ha llegado a buen puerto. Agradezco especialmente a JR el haberme dado la oportunidad de incorporarme a su equipo y permitirme descubrir el mundo de la investigación.

Y es en el equipo del BRGE donde he encontrado el apoyo necesario para afrontar estos años de trabajo. Gracias Alejandro por todo el tiempo que te he robado para que ayudaras con mis ideas y por todas las charlas sobre ciencia y la vida. Gracias Jose por estar siempre a nuestro lado, intentándonos sacar una sonrisa. Gracias Dietmar, porque, aunque has sido el último en llegar, te has convertido en una persona muy importante en mi día a día en ISGlobal. Gracias a Marcos, el artista, que siempre ha vigilado porque no me saliera del credo de las inversiones y por la pedazo portada que ha diseñado para mi tesis. También quiero agradecer a todos los compañeros que pasaron por el grupo para después seguir su camino, como José Vargas, Natalia o Ibón, que han contribuido a que el trabajo no fuera tan trabajo. I aquí no em puc oblidar de tu, Carles, el meu gran recolzament durant el temps que vam coincidir fent la tesis. Moltes gràcies per les nostres xerrades, pels

nostres viatges, per les teves idees, pel teu suport i, no menys important, per fer-me descobrir el món dels jocs de taula.

Però la meva estada al CREAL/ISGlobal tampoc es podria explicar sense les noies de la sala C. Gràcies a l'Erica, que abans de ser mare del Quim, ja tenia un nen gran que havia de vigilar que no féssin fora del doctorat per no entregar les coses a temps. Gràcies a l'Ariadna, amb qui he compartit *cotis* i fins i tot aniversaris conjunts. Gràcies a l'Alba, la meva companya de "delegats" i amb qui espero compartir moltes nits de festival. Gracias a Ángela y a Javier, por los buenos momentos compartidos dentro y fuera del trabajo. Finally, thanks to Gosia, Diana and Deborah, who also made my time here more valuable.

Pero no puedo acabar estos agradecimientos sin hablar de vosotros, los *Felise*. Gracias a Alba y a David por todos los viajes y momentos que hemos pasado juntos, y por estar ahí durante los momentos de bajón. Y gracias especialmente a ti, Marta, por estar a mi lado durante todo este tiempo y compartir juntos esta etapa de nuestra vida, la cual espero que sea solo el comienzo de nuestro viaje.

Finalmente, doy gracias a mi familia, por estar ahí siempre que la he necesitado. En especial a mis padres y a mi hermana, que me han apoyado en todo momento y que sin ellos nada de esto sería posible. También a mi primo, por aguantarme en los momentos de bajón y darme su apoyo incondicional.

Pido disculpas de antemano por esta lista que seguro es incompleta. Un trabajo de varios años no se puede completar sin el apoyo de decenas de personas a quienes agradezco su apoyo y su aportación a esta tesis. Y para acabar, solo queda agradecer al AGAUR por haber financiado buena parte de esta tesis.

Abstract

Chromosomal inversions are structural variants where a segment changes its orientation. Chromosomal inversions reduce homologous recombination, producing different haplotypes in standard and inverted chromosomes. As a result, they influence adaptation and selection and play a role in susceptibility to human diseases.

Inversions can be studied using experimental and bioinformatic methods. SNP array data can be used to call inversion genotypes by using haplotype differences between inverted and standard chromosomes. However, these methods are not optimized for large cohorts (thousands of individuals from existing databases such as dbGaP or UK Biobank). Also, current methods can only genotype inversions with two haplotypes and the inversion calling is difficult to be harmonized among cohorts. Finally, it is recognized that chromosomal inversions affect gene expression and DNA methylation. However, there are no accurate methods to globally assess the effect of inversions on local gene expression or DNA methylation.

The main aim of this thesis is to develop new robust and scalable methods and bioinformatic tools to study the phenotypic and functional effects of chromosomal inversions by overcoming the existing limitations. To this end, I have developed a new method to genotype chromosomal inversions that can be used in large cohorts, inversions with multiple haplotypes and that uses reference haplotypes allowing the integrative analysis of multiple cohorts. Second, I have implemented a multivariate method based on redundancy analysis to study the effects of chromosomal inversions on local DNA methylation and gene

expression. Then, I applied both methods to study the role of chromosomal inversions in two groups of complex diseases: neurodevelopmental disorders and cancer. Finally, I developed a new method to study how chromosomal inversions affect recombination patterns. This method is extendable to any genomic regions containing subpopulations with different recombination patterns, allowing associating these subpopulations to phenotypic traits.

Resumen

Las inversiones cromosómicas son variantes estructurales donde un segmento de ADN cambia su orientación. Las inversiones cromosómicas reducen la recombinación homóloga y producen diferentes haplotipos en los cromosomas estándar e invertidos. Como resultado, influyen en la adaptación y la selección y desempeñan un papel en la susceptibilidad a las enfermedades humanas.

Las inversiones se pueden estudiar con métodos experimentales y bioinformáticos. Los datos de SNP array se pueden usar para determinar genotipos de inversión mediante el uso de diferencias de haplotipos entre cromosomas invertidos y estándares. Sin embargo, estos métodos no están optimizados para grandes cohortes (con miles de individuos, como dbGaP o UK Biobank). Además, los métodos actuales solo pueden genotipar las inversiones con dos haplotipos y la clasificación es difícil de armonizar entre cohortes. Finalmente, se conoce que las inversiones cromosómicas afectan la expresión génica y la metilación del ADN. Sin embargo, no existen métodos precisos para evaluar globalmente el efecto de las inversiones en la expresión génica local o la metilación del ADN.

El objetivo principal de esta tesis es desarrollar nuevos métodos robustos y escalables así como herramientas bionformáticas para estudiar los efectos fenotípicos y funcionales de las inversiones cromosómicas, superando las limitaciones existentes. Con este fin, he desarrollado un nuevo método para genotipar las inversiones cromosómicas que se puede usar en grandes cohortes, con inversiones con múltiples haplotipos y que utiliza haplotipos de referencia que permite el análisis

conjunto de múltiples cohortes. En segundo lugar, he implementado un método multivariante basado en el análisis de la redundancia para estudiar los efectos de las inversiones cromosómicas en la metilación del ADN y la expresión génica locales. A continuación, he aplicado ambos métodos para estudiar el papel de las inversiones cromosómicas en dos grupos de enfermedades complejas: trastornos del neurodesarrollo y cáncer. Finalmente, he desarrollado un nuevo método para estudiar cómo las inversiones cromosómicas afectan los patrones de recombinación. Este método es aplicable a cualquier región genómica que contenga subpoblaciones con diferentes patrones de recombinación, lo que permite asociar estas subpoblaciones a rasgos fenotípicos.

Preface

This thesis was written at Barcelona Institute for Global Health (ISGlobal), between September 2015 and January 2019, and it was supervised by Dr. Juan Ramon González. This work consists on a compilation of 4 scientific publications (1 published, 3 under review) co-authored by the PhD candidate. This is in agreement with the procedures of the PhD program in Biomedicine, organized by Department of Experimental and Health Sciences of the Universitat Pompeu Fabra.

The present thesis contributed to: (1) develop new methods to study the phenotypic and functional effects of chromosomal inversions; (2) propose new associations between chromosomal inversions and human diseases; (3) study the effect of chromosomal inversions on recombination patterns; (4) open and discuss future research directions to study the phenotypic and functional effects of chromosomal inversions.

Table of contents

1 General Introduction

Chromosomal inversions

Chromosomal inversions are chromosomal rearrangements were a DNA sequence changes its orientation (Figure 1). They were discovered by Sturtevant in 1921 in *Drosophila* [1] and, since then, they have been found in all genomes, including mammals [2–5], insects [6–8]; plants [9] or bacteria [10]. We can describe inversions using different criteria. A classical criterion classifies inversions with respect the inclusion of the centromere in: pericentric, include the centromere; and paracentric, do not include the centromere (Figure 1). We can also classify inversions with criteria used for genetic variants: (1) polymorphic (common in the population) and sporadic (present in few related individuals); (2) recurrent (commonly appearing) and non-recurrent (rare and ancestral events); (3) germinal (heritable as it is present in germinal cells) and somatic (not heritable as it is not found in germinal cells).

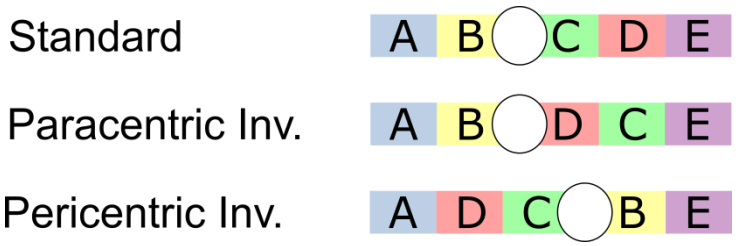

Figure 1: Chromosomal inversion classification with respect to centromere inclusion. First chromosome exemplifies the original sequence. Paracentric inversions do not include the centromere and preserve the chromosome shape. Pericentric inversions include the centromere and change the chromosome shape.

Chromosomal inversions are commonly generated by Non-Allelic Homologous Recombination (NAHR) [11]. NAHR is a subtype of homologous recombination. Homologous recombination consists on the cross-over of homologous chromosomes and the exchange of genetic

content. Thus, homologous recombination generates new chromosomes containing a mixture of parental genetic content, increasing the genetic variability. In NAHR, the cross-over happens between two non-allelic regions. In order to generate a chromosomal inversion, NAHR should happen between two inverted segmental duplications (SD) of the same chromatid. Inverted SDs are regions with more than 90% sequence identity with opposite orientations. When inverted SDs produce a cross-over and the cross-over is solved by NAHR, the region between the SDs changes its orientation generating a chromosomal inversion (Figure 2). This mechanism explains why we find a great number of inversions flanked by inverted SDs [12–14]. Other mechanisms to generate chromosomal inversions are methods to repair DNA such as Non-Homologous End Joining (NHEJ) [15]; or Microhomology-Mediated Break-Induced Replication Model (MMBIR) [16].

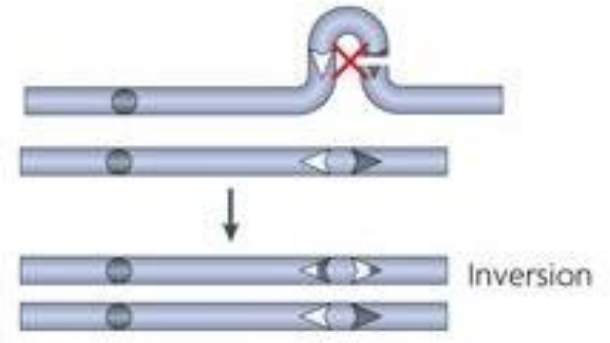

Figure 2: Chromosomal inversion generation by NAHR. Two proximal inverted segmental duplications (SD) can produce a cross-over. The cross-over resolution by NAHR results in the change of orientation of the region between the SD. Adapted from [17]

Chromosomal inversions define different subpopulations

Chromosomal inversions affect homologous recombination in meiosis. During meiosis, sister chromosomes cross-over between homologous regions. If an individual has an inversion in only one chromosome (i.e. it is inversion heterozygous), the inversion region cannot properly pair. Consequently, one chromosome should twist to pair the other chromosome, resulting in the formation of an inversion loop [18]. The inversion loop pairs the inversion region of both chromosomes, but the chromosomes generated by homologous recombination typically contain aberrations. In paracentric inversions, one recombinant has a deletion in the inversion region and the other has a deletion in the terminal region (Figure 3). For pericentric inversions, the recombinant chromosomes have the same terminal region in both extremes (Figure 4). These aberrant chromosomes lead to non-viable zygotes, so we observe a reduction of standard and inverted recombinants in the global population.

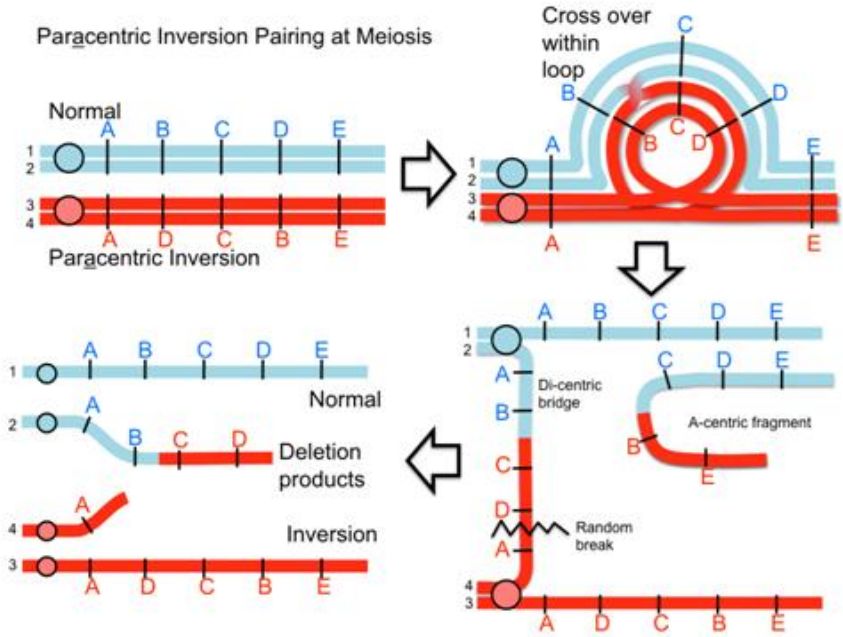

Figure 3: Meiosis in individuals heterozygous for paracentric inversions. Recombinant chromosomes results in a chromosome without centromere (acentric) and a chromosome with two centromeres due to a di-centric bridge. The acentric chromosome is lost during division and the breakage of the dicentric breaks results in two chromosomes containing deletions (Original-Locke-CC:AN).

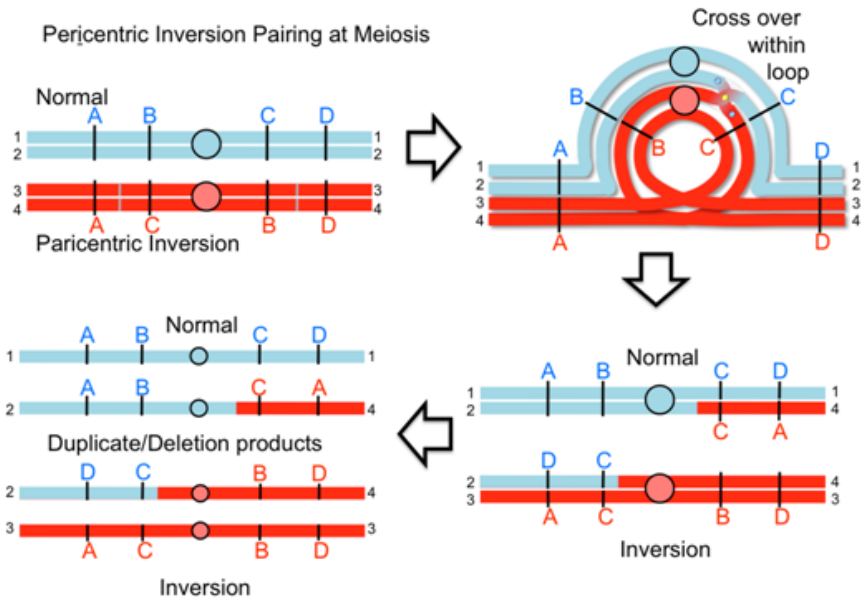

Figure 4 Meiosis in individuals heterozygous for pericentric inversions. Recombinant chromosomes have two copies of one telomeric region (A and D in the figure) and no copies of the other. (Original-Locke-CC:AN)

As inversion recombinants are reduced, there is no genetic exchange between standard and inverted chromosomes. Thus, standard and inverted chromosomes can be considered two isolated populations that evolve independently. Populations that evolve independently accumulate different alleles and have different allele combinations. This hypothesis has been tested in inv8p23.1, a human chromosomal inversion in chromosome 8, where standard and inverted chromosomes have different mutations [19] and different allele combinations reflected in the recombination patterns [20].

Methods to study chromosomal inversions

There are different experimental and bioinformatic methods to study chromosomal inversions, which can be grouped in direct and indirect methods. Direct methods are based on detecting changes in the DNA sequence and follow three main strategies: microscopy, molecular biology and sequencing (Table 1). Indirect methods search signatures on the DNA sequence generated by the inversion, such as extended haplotypes or perturbations of linkage disequilibrium (LD) patterns (Table 2).

Table 1: Summary of direct methods to study chromosomal inversions. General methods can discover new inversions while targeted can only genotype previously known inversions.

Method	Type	Strengths	Limitations
Microscopy			
G-banding	General	Good accuracy	Low throughput Restricted to big inversions
FISH	Targeted	Good accuracy	Low throughput Restricted to big inversions
Molecular biology			
PCR	Targeted	High throughput Good resolution	Fail for inversions flanked by large segmental duplications
Optical mapping	General	Good resolution	Low throughput
Sequencing			
Paired-End	General	High throughput	Restricted to small inversions Fail for inversions flanked by large segmental duplications
Long reads	General	Good resolution	Cost
Strand seq	General	Good resolution	Cost

Table 2: Summary of indirect methods to study chromosomal inversions. General methods can discover new inversions while targeted can only genotype previously known inversions.

Method	Type	Strengths	Limitations
Haplotypes			
Tag SNPs	Targeted	Reuse data Enable association studies	Restricted to polymorphic and non-recurrent inversions Sensitive to genotyping errors
Clustering	Targeted	Reuse data Enable association studies	Restricted to polymorphic and non-recurrent inversions Classification is not standard
LD patterns alteration			
inveRsion	General	Discovery of candidate regions harboring inversions	Restricted to polymorphic and non-recurrent inversions Low accuracy of individual classification

Direct methods

MICROSCOPY

In microscopy methods, chromosomes are stained and observed with the microscope. Then, the presence of an inversion is detected due to changes in the expected staining pattern. There are two main staining methods: Giemsa and fluorescence in situ hybridization (FISH). Staining by Giemsa generates bands on the chromosomes, called G-banding karyotype, which are specific of each chromosome. Large chromosomal inversions affect G-banding karyotype, so alterations in the G-banding karyotype can be used to detect chromosomal inversions [21, 22]. FISH consists on designing fluorescent probes complementary to a genomic region. FISH can be used to detect chromosomal inversions using two

probes with two different colors in two alternative configurations (Figure 5): (1) both probes are placed inside the inversion region and we check their orientation with respect to the chromosome [23]; (2) one probe is placed inside and the other outside the inversion region and we check the distance between the probes [24].

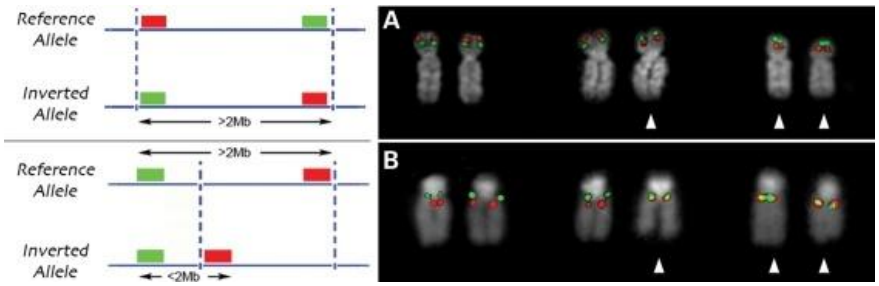

Figure 5: Design of FISH probes to detect chromosomal inversions. Chromosomes containing the inversion are marked with an arrow. A) Two probes placed inside the inversion region. We observe a change of order of the probes in the inverted chromosomes. B) One probe inside the inversion. We observe that probes are separated in standard chromosomes and close in inverted chromosomes. Adapted from [2]

These techniques differ in their application, as G-banding karyotype can discover new inversions while FISH can only test the presence of a previously defined inversion in an individual. Thus, these techniques are complementary, using G-banding to define the inversion region and FISH to confirm the inversion status of the individuals. These methods are very accurate to genotype inversions so they are used in the clinical setting to diagnose inversions in patients [23, 24]. However, they have very low resolution and they are restricted to large inversions, in the order of megabases. In addition, these techniques are not scalable to infer chromosomal inversions in cohorts, preventing their use in association studies.

MOLECULAR BIOLOGY

Molecular biology techniques are used to manipulate and study DNA molecules. Two molecular biology techniques can be used to infer the inversion status of an individual: polymerase chain reaction (PCR) and optical mapping. PCR amplifies a DNA segment enclosed by primers, short sequences designed by the user. PCR can be used to detect inversions by designing primers around an inversion breakpoint for standard and inverted chromosomes. PCR only amplifies short fragments (around 1Kb), so primers designed for standard chromosomes only work in standard chromosomes and the same apply for inverted chromosomes. Thus, homozygous standard individuals only amplify the fragment of standard primers, homozygous inverted individuals only amplify the fragment of inverted primers and inversion heterozygous amplify both fragments [8, 25, 26]. Inversion PCR (iPCR), a modification of basic PCR, was designed to improve chromosomal inversions detection [27]. In optical mapping, DNA strands are stretched and cut using a restriction enzyme. Then, the length of the resulting DNA fragments is measured to create a restriction map (Figure 6). Inversions produce the reversion of fragment lengths in a region, so this alteration in fragment lengths can be used to detect inversions. This approach has been extensively used in bacteria [10, 28, 29].

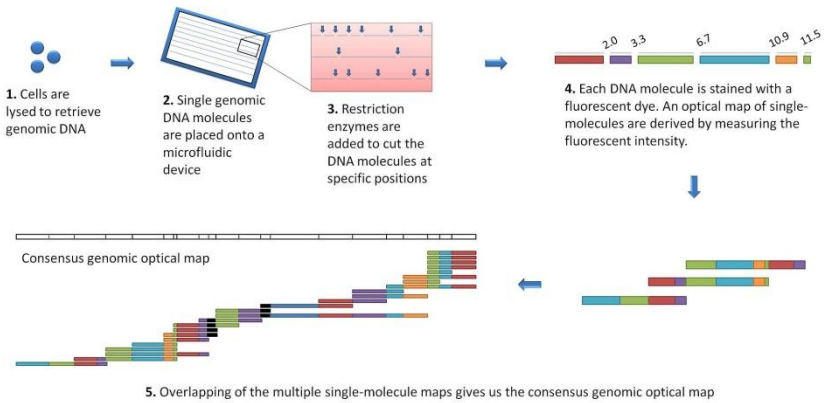

Figure 6: Optical mapping explanation. From Fong Chun Chan and Kendric Wang [30]

Molecular biology techniques have higher resolution and can detect short inversions, in the order of kilobases. PCR is cheap so it is potentially scalable to genotype inversions with known breakpoints. However, PCR does not work for inversions surrounded by segmental duplications larger than the segment amplified by PCR. On the other hand, optical mapping can be used for inversion discovery and it is not affected by segmental duplications but it is not scalable to its use in large cohorts.

SEQUENCING

Finally, some sequencing methods, i.e. techniques to get the DNA sequence of an individual, can also be used to detect inversions. The most common sequencing technique is paired-end sequencing. In pair-end sequencing, the DNA is broken in short fragments of 350-500 bp (base pairs) and we get the sequence of the fragments ends. These short sequences, called reads, measure between 100-150 bp and are mapped to a reference genome. In regions where our sample genome is equal to the reference genome, reads are mapped in a distance close to the DNA fragment and in opposite strands. If this region contains an inversion,

both reads will be mapped to the same strand and to a distance higher than the DNA fragment. Different software tools detect anomalies in read mapping to infer structural variants including inversions. For instance, *Delly* [31], *Pindel* [32] and *SoftSV* [33] have been used to detect inversions in humans [34, 35], highlighting the structural variant calling in the 1000 Genomes Project [36]; and in other species [37, 38]. In the recent years, a new generation of sequencing methods fragments the DNA in bigger fragments (50-80 Kb). These methods improve the detection of structural variants [39] and a algorithm (VALOR: **va**riation using **lo**ng **r**ange information) has been specifically designed to detect chromosomal inversions from this data [40]. Finally, a technique based on sequencing single chromatid strands (strand-seq or single-cell DNA template strand sequencing) has also been applied to detect human chromosomal inversions [41].

Paired-end sequencing can be used to genotype large cohorts and this data is available for a big number of public studies. However, paired-end sequencing can only detect short inversions, as it is limited by the short DNA fragments. In addition, paired-end sequencing cannot detect inversions flanked by inverted repeats larger the DNA fragments. The other two methods seem promising tools, but their cost might be still high for their application in association studies.

Indirect methods

Indirect methods use the effects of chromosomal inversions on genetic sequence to genotype chromosomal inversions. Indirect methods follow two strategies: search differences in mutation content or perturbation of linkage disequilibrium patterns.

MUTATION CONTENT

Different bioinformatics methods exploit differences in mutation content to genotype chromosomal inversions. The simplest approach is searching for a variant having different alleles in standard and inverted chromosomes. The more commonly used variants are Single Nucleotide Polymorphism (SNPs), variants of one base with a frequency higher than 1% in the population. A SNP that differentiates standard and inverted chromosomes is called a tag SNP for this inversion. Tag SNPs have been used to study the human inversions inv17q21.31 [42, 43], inv8p23.1 [23], inv19p12 [44] and inv16q23.1 [45], as well as to infer inversion frequencies in pool sequencing in *Drosophila melanogaster* [46, 47]. Nonetheless, this approach has two main drawbacks: (1) the correlation between the inversion and the tag SNP might be weaker when more individuals are considered [2]; (2) tag SNPs might be only tag the inversion in a given population.

Clustering methods are an extension of tag SNPs. Clustering methods run a reduction of dimensionality technique (e.g. Principal Component Analysis -PCA- or MultiDimensional Scaling -MDS-) on the all the SNPs included in the inversion. If standard and inverted chromosomes differ in their mutation content, a PCA/MDS generates three clusters in the first two components of the inversion region SNPs. These clusters map to inversion genotypes, with the side clusters having inverted and standard homozygous and the middle cluster the inversion heterozygous, as they are a 1:1 mixture of inverted and standard homozygous. This method was theoretically defined by Ma and Amos [48] and implemented in a set of R scripts called *PFIDO* (phase free inversion detection operator) [19]. *PFIDO* was optimized for the human inversion inv8p23.1 and included all the steps, from running a MDS on SNP genotypes to apply the clustering.

A similar algorithm was later implemented in the R package *invClust* [49], which can also account for individuals' genetic ancestry, a factor that confounds the clustering.

Clustering methods are readily applicable to hundreds of samples and can reuse SNP data from previous Genome Wide Association Studies (GWAS), i.e. studies associating hundreds of thousands of SNPs to a disease or phenotype. Therefore, clustering methods have been successfully used in different association studies in humans [49, 50]. Nonetheless, clustering methods have some limitations. First, clustering methods are restricted to polymorphic and non-recurrent inversions with differentiated haplotypes. Second, some inversions generate more than three clusters, such as the human inversion inv16p11.2 [51] or the zebra finch inversion TguZ [52], violating one assumption of the clustering methods. Third, clustering methods require an external validation to map the individual clusters to the real inversion genotypes, requiring further harmonization steps in multi-centric studies. Finally, existing clustering methods are not computationally efficient to be used in cohorts with thousands of individuals.

LINKAGE DISEQUILIBRIUM PATTERNS ALTERATION

Another approach to detect inversions is searching for perturbations in the linkage disequilibrium (LD) patterns. LD is the non-random association between the alleles of a pair of SNPs. LD is inversely associated with recombination rates, as homologous recombination shuffles the alleles, and with the distance between SNPs, as closer SNPs have less points of homologous recombination. If there is an inversion, close SNPs in the reference genome are distant in the inverted chromosomes while distant SNPs in the reference genome are close in

the inverted. Thus, if we define four SNP blocks around the inversion breakpoints (Figure 7):

- Standard chromosomes: LD between blocks around the same breakpoint is high and low between blocks of different breakpoints
- Inverted chromosomes: the LD between blocks around the same breakpoint is low and high between blocks of different breakpoints.

The inversion model [53] states that, if a population contains a polymorphic inversion, the population LD between blocks can be modeled as a mixture of standard and inverted subpopulations. The inversion model was implemented in *inveRsion*, a R package that scans the genome to detect regions potentially containing inversions [54] and used to define chromosomal inversions in cod [55–57]. However, *inveRsion* has low accuracy genotyping individuals. Finally, notice that no method uses differences in recombination patterns between inverted and standard chromosomes to detect inversions.

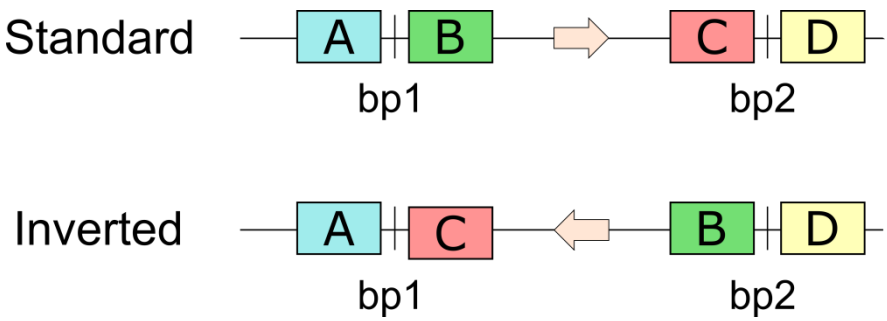

Figure 7: inversion model based on linkage disequilibrium alterations. bp1 and bp2 represent the inversion breakpoints. The arrows indicate the orientation of the inversion fragment with respect the reference genome. A, B, C and D are the four blocks of SNPs defined by the inversion model. In the standard chromosomes, A and B are close and A and C are distant. Thus, LD between A and B is high and while LD between A and C is low. In inverted chromosomes, A and B are distant while A and C are close. Thus, LD between A and B is low while LD between A and C is high.

Functional impact of chromosomal inversions

Evolutionary biology

Chromosomal inversions have been traditionally considered as recombination modifiers in evolutionary biology as they reduce heterozygous recombinants. Chromosomal inversions influence three evolutionary phenomena (reviewed by Kirkpatrick in [58] and [59]): (1) adaptation, (2) speciation and (3) formation of sexual chromosomes.

Chromosomal inversions encapsulate alleles associated with adaptation in the different chromosome configurations. Thus, good allelic combinations fostering adaptation to different environments are preserved from recombination. The role of chromosomal inversions in adaptation has been described in honeybees [60] and monekyflowers [61] and in the climatic clines of *Drosophila melanogaster* [47].

Chromosomal inversions can capture and generate genes producing reproductive isolation and they prevent the breakage of gene combinations producing reproduction isolation. Eventually, this process leads to the formation of different species. Speciation due to chromosomal inversions has been proposed as a theoretical model [62], and as the mechanism of the differentiation between *Drosophila persimilis* and *Drosophila pseudoobscura* [63] and the reproductive isolation in *Boechera stricta* [64].

Finally, mammal sexual chromosomes were generated due to chromosomal inversions [65]. Ancestrally, sex was not defined by chromosome content. One chromosome accumulated male-specific genes (proto Y). Different chromosomal inversions in proto Y captured these genes to reduce recombination with proto X. Finally, there was almost no recombination between proto X and proto Y and proto Y accumulated mutations and degraded to form current Y chromosome.

Human diseases

Chromosomal inversions can participate in human diseases through three mechanisms: (1) generation of new DNA sequences; (2) influencing NAHR; (3) or through haplotypic differences.

A chromosomal inversion can break an existing gene or create a new fusion gene. In both cases, chromosomal inversions have a big impact on fitness and inversions that produce these changes tend to be rare. On one hand, chromosomal inversions that disrupt genes are associated with Duchenne muscular dystrophy and mental retardation [66], Hermansky-Pudlak syndrome [67], haemophilia A [26, 68], phosphoglucomutase 1 deficiency [69] or intellectual disability [70]. On the other hand, chromosomal inversions that create new fusion genes are associated with cancer susceptibility and progression [71–77],

highlighting a recurrent inversion in chromosome 16 that defines a subtype of leukemia (see a revision in [78]).

Chromosomal inversions modify NAHR in the inversion region. On one hand, inversion heterozygous might favor some rearrangements between inverted and standard chromosomes during meiosis that are not possible in homozygous standard or inverted individuals. These rearrangements typically consist on deletions that are transmitted from inversion heterozygous parents to their children. Different chromosomal inversions are associated with diseases through this mechanism, such as inv17q21.31 and 17q21.31 microdeletion syndrome [79], 7q11.21 and Williams syndrome [80, 81], 15q11.13 and Angelman syndrome [82], 5q35 and Sotos syndrome [83] or inv8p23.1 and developmental delay and congenital heart effects [84]. On the other hand, chromosomal inversions block NAHR between sister chromosomes in mitosis. In mitosis, NAHR between sister chromosomes can produce a region of loss of heterozygosity (LOH), where both chromosomes have the same genetic content. This mechanism can recover a mutation causing a disease that is only present in one of the chromosomes. Mutations producing severe congenital ichthyosis can be reverted through this mechanism, thus, inversions heterozygous have worse prognosis for this disease [24].

Finally, inverted and standard chromosomes have different alleles which modify the susceptibility of suffering a disease. For instance, haplotypes in inv8p23.1 are associated with system systematic lupus [85, 86], neuroticism [50], autism [87], schizophrenia [87] and underweight [49]; haplotypes in inv17q21.31 are associated with Parkinson [88–91], neurodegenerative tauopathies [42, 92], Alzheimer [93], neuroticism [50], autism [87], schizophrenia [87] and response to corticosteroids [94];

haplotypes in inv16p11.2 are associated with asthma and obesity [51] and haplotypes in inv16q23.1 are associated with chronic pancreatitis [45].

Functional genomics effects

Omic data consists on biological measures reflecting the state of a tissue. Omic data is useful to propose mechanisms to explain the effect of genetic variants on phenotypic traits. Two common omic data are becoming routinely assessed in genetic studies: (1) transcriptomic that measures the expression of all the genes in a tissue or cell; (2) epigenomics that measures DNA methylation, that is, whether cytosines in CG pairs (CpG) contain a methylation group. DNA methylation is an epigenetic mechanism, i.e. chemical DNA modifications that influence gene regulation.

Polymorphic chromosomal inversions affect both transcriptome (gene expression) and epigenome (DNA methylation). Different polymorphic inversions have been associated with changes in gene expression in humans. The clearest example is inv17q21.31, which affects the expression of genes located in the inversion region in blood [42, 95, 96] and brain [89, 95, 97–100]. This inversion has also *trans*-effects in blood [42]. Other inversions affecting local gene expression in blood or lymphoblastoid lines are inv8p23.1 [23], inv16p11.2 [51] and inv19p12 [44]. Inversion inv16q23.1 affects regional gene expression in pancreatic tissue or acinar cells. Regarding DNA methylation, inv17q21.31 has been associated with changes in regional and global DNA methylation in blood [42].

Studies associating chromosomal inversions with omic data have some limitations. First, most studies genotyped inversions using tag SNPs, a technique sensitive to discrepancies between a single allele and the

inversion status. Second, the analyses were conducted associating each gene or CpG to inversion genotypes and they do not provide an overall estimate of the effect of chromosomal inversions in the inversion region. Therefore, a new approach is required to overcome these limitations.

2 Datasets

This thesis has been carried out using data from publicly available projects. In the following lines, I will briefly describe these projects.

1000 GENOMES

The goal of the 1000 Genomes project was finding the genetic variation present in different human worldwide populations [101]. To this end, they included 2,504 individuals from 26 populations from 5 continental groups (Africa, Europe, East Asia, South Asia and America). They included these 26 populations to have a good representation of global human variation, to have power to detect rare variants and prioritizing populations already included in genetic studies. They extracted blood samples to produce lymphoblastoid cell lines and they applied three genotyping techniques: (1) low coverage whole genome sequencing; (2) deep exome sequencing and (3) SNP arrays. From these data, they detected 88 million variants, including 84.7 million SNPs, 3.6 million indels (short insertion and deletions) and 60 thousand structural variants. All these variants were phased using data from first degree-relatives. The 1000 Genomes project produced a global overview of human genetic variation and has provided reference haplotype to impute SNPs, which has increased the power in GWAS. In addition, the project cell lines are available to apply other genotyping techniques, such as those previously described to study chromosomal inversions.

THE CANCER GENOME ATLAS

The goal of The Cancer Genome Atlas (TCGA) was to characterize the genetic and molecular changes associated with cancer. TCGA was a collaborative project between two US institutions: National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). TCGA studied 33 types of cancer from more than 11,000 patients from

USA and Canada. Samples were provided by collaborating hospital and research centers. Inclusion and exclusion criteria varied depending on the cancer study but all individuals had, at least, one tumor tissue sample and one normal tissue sample, either from blood or from the same tumor tissue. TCGA generated six different omic data types: (1) SNP arrays, used to measure SNPs, CNVs (Copy Number Variants) and LOH (Loss of Heterozygosity); (2) DNAseq, DNA sequencing to detect mutations and structural variants; (3) RNAseq, gene expression measured using sequencing; (4) miRNAseq, miRNA measured using sequencing; (5) DNA methylation using microarray; (6) Reverse-phase protein array (RPPA), measure of protein expression. TCGA project has provided new insights in cancer biology and a better definition of cancer subtypes using molecular markers [102].

THE GENOME-TISSUE EXPRESSION

The Genome-Tissue Expression (GTEx) project, lead by the US National Institute of Health (NIH), aims to associate genetic variants with gene expression in different tissues. GTEx includes data from 714 donors, recently died people who donated tissues for the project, recruited in different American centers [103]. Donors having cancer, drug abuse, recent infectious diseases or other medical conditions were excluded from the analyses. Each donor contributed with 53 different tissues that resulted in 10,361 total tissue samples, after removing samples with low RNA quality or histological alterations. Genotypes and gene expression data from GTEx is publicly available allowing novel analyses of the genetic regulation of gene expression.

PUBLIC REPOSITORIES

I have also used other genetic data from two public repositories: dbGAP (Database of Genotypes and Phenotypes) and EGA (European Genome Archive). Both repositories are databases of genetic and phenotypic data, containing data from previously published GWAS. Although their main focus is on genetic data (SNP array or sequencing data), they also contain gene expression or DNA methylation data. dbGAP belongs to the US NIH and contains 1,178 studies, while EGA belong to EBI (European Bioinformatic Institute) and CRG (Centre for Genomic Regulation) and hosts 1,735 studies.

3 Hypotheses

Chromosomal inversions inhibit recombination in heterozygous individuals. Consequently, standard and inverted chromosomes evolve independently and accumulate genetic differences. We can use these differences to infer inversion status in large cohorts. In addition, standard and inverted chromosomes also differ in their recombination patterns, information that can also be used to call inversion genotypes.

Chromosomal inversions influence gene regulation in the inversion region. In particular, they modify gene expression and DNA methylation in different tissues. Changes in the transcriptome and epigenome are associated with phenotypic variability. Therefore, chromosomal inversions can play a role in complex diseases such as cancer or neurodevelopmental disorders among others.

4 Objectives

The aim of the thesis is to develop new robust and scalable methods and bioinformatic tools to investigate the phenotypic and functional consequences of chromosomal inversions. I also aim to elucidate the role of chromosomal inversions in complex diseases by using these new tools. Therefore, the specific objectives of this thesis are:

- **Objective 1**: to improve detection of chromosomal inversions using genotype data allowing the analysis of inversions with multiple haplotypes and the analysis of multiple studies on the same disease.
- **Objective 2**: to develop a method to study how chromosomal inversions changes methylation or gene expression in the inversion region
- **Objective 3:** to associate chromosomal inversions with complex diseases, in particular, neurodevelopmental disorders and cancer
- **Objective 4**: to study the effect of chromosomal inversions on recombination patterns

5 scoreInvHap: new method to genotype inversions

In this chapter, I describe scoreInvHap, a new method to genotype chromosomal inversion based on SNP data. scoreInvHap can use data from different sources (Whole genome sequencing or WGS, SNP microarray, exome sequencing) and easily harmonizes classification between datasets. Consequently, scoreInvHap can be applied in multi-center studies to elucidate the effect of chromosomal inversions on complex diseases.

Ruiz-Arenas C, Cáceres A, López-Sánchez M, Tolosana I, Pérez-Jurado L, González JR. scoreInvHap: Inversion genotyping for genome-wide association studies. PLoS Genet. 2019 Jul 1;15(7). DOI: 10.1371/journal.pgen.1008203

Supplementary Dataset

scoreInvHap_Sup_Dataset.csv: Inversion genotypes of the 20 inversions included in *scoreInvHap* for the European individuals of 1000 Genomes. Available under request.

6 Redundancy Analysis in omic datasets

In this chapter, I describe the implementation of Redundancy Analysis to omic data in *MEAL*. With *MEAL*, we can obtain a global estimate of the association between DNA methylation or gene expression and chromosomal inversions in the inversion region.

Redundancy Analysis allows improved detection of methylation changes in large genomic regions

Ruiz-Arenas C, González JR. Redundancy analysis allows improved detection of methylation changes in large genomic regions. BMC Bioinformatics. 2017 Dec 14;18(1). DOI: 10.1186/s12859-017-1986-0

7 Inversions and cancer prognosis

In this chapter, I associate two common inversions (inv17q21.31 and inv8p23.1) to cancer prognosis in TCGA and I use omic data to propose a biological mechanism linking the inversion to cancer prognosis.

Common polymorphic inversions at 17q21.31 and 8p23.1 associate with cancer prognosis

Ruiz-Arenas C, Cáceres A, Moreno V, González JR. Common polymorphic inversions at 17q21.31 and 8p23.1 associate with cancer prognosis. Hum Genomics. 2019 Nov 21;13(1). DOI: 10.1186/s40246-019-0242-2

8 *recombClust*: inversions and recombination patterns

In this chapter, I present *recombClust*, a method to partition chromosomes by recombination patterns. I apply *recombClust* to chromosomal inversions to study whether recombination patterns differ between inverted and standard chromosomes, both in simulated inversions and in two common human inversions (inv17q21.31 and inv8p23.1).

Identification of chromosome subpopulations by recombination differences

Authors: Carlos Ruiz-Arenas, Alejandro Cáceres, Marcos López, Josefa González, Juan R. González

(In preparation)

Abstract

Chromosomal subpopulations are characterized by specific mutation content or by singular allelic combinations of common variants. While mutation-based methods are typically used to infer the genetic substructure in a population, methods based on allelic combinations remain to be proposed. We developed *recombClust*, a method that uses SNP data to classify chromosomes according to their recombination patterns within a genomic region. We showed that *recombClust* is able to identify chromosomes with an inverted region under multiple ancestries. We used *recombClust* to detect and validate a recombination substructure leading to four distinct chromosome groups within 1q21.1, the largest genomic region associated to breast cancer. We observed that the four chromosome alleles associated with the gene expression of multiple genes across numerous tissues, and that one chromosome allele associated stronger to breast cancer than any of the SNPs in the region. Our results showed that the chromosome substructure that is associated to differences in allele combinations can help to explain functional and phenotypic differences between individuals.

Introduction

Chromosomal subpopulations emerge when differences in mutation content or in allele combinations accumulate. Both types of genetic differences have been widely observed in species from different taxa, including humans [101, 176], *Drosophila melanogaster* [177, 178] or maize [179, 180]. Content differences are acquired by subpopulations' divergence in mutation quantity and frequency [101] while differences in allele combinations are derived by variations in the recombination patterns between common mutations [176]. Specific mutations and allele combinations can segregate within chromosomal segments that are affected by structural variants, like translocations and inversions, as they block recombination when heterozygous [181]. As such, differences in both mutation content and allele combinations can be intertwined as shown, for instance, for the human inversion at 8p23.1 [19, 20], challenging the assessment of their relative contributions to subpopulation differences. A problem that is primarily challenged by the lack of specific methods to infer chromosome subpopulations from differences in allele combinations.

Chromosomal subpopulations generate large phenotype diversity helping adaptation [55, 56, 104, 182, 183]. Subpopulations are typically detected by their differences in mutation content. A common approach is to perform reduction of dimensionality analyses, such as multidimensional scaling (MDS) or principal component analysis (PCA), on single nucleotide polymorphisms (SNPs) to identify clusters in the population. The approach is extensively used on genome-wide data to define human ancestry [184] as well as to infer chromosomal subpopulations associated with the inversion of a genomic region [48]. By contrast,

detection of population substructure based on differences in recombination patterns has not yet been performed. Therefore, in this work we propose a method lo leverage differences in recombination patterns within a genomic region to classify chromosomes into groups. The method, *recombClust*, comprises two steps. First, it detects the presence of points in a chromosome segment where only a fraction of chromosomes recombined and, second, it tests whether the observed chromosome groups are consistent across the detected points. We thus demonstrate that chromosomal subpopulations that differ in allelic combinations can be identified from differences in recombination patterns, given by different recombination points, and offer the method as a computationally efficient tool, compatible with *Bioconductor*'s packages and implemented for usual data formats, such as the variant call format (VCF).

We tested the performance of the method with numerous simulations. In particular, we applied the method on simulated regions with chromosomal inversions, using the coalescent simulator *invertFREGENE* [109], to demonstrate that *recombClust* correctly detects inversion status from recombination patterns. We then used *recombClust*, on SNP data from the 1000 Genomes Project [185], to compare the calling from recombination patterns and from mutation differences of the well characterized human inversions at 8p23.1 and 17q21.31. Finally, we also applied *recombClust* to reveal a recombination substructure within 1q21.1, the largest region associated with breast cancer observed in a large genome-wide association study (GWAS) meta-analysis [186], and we studied the functional and phenotypic associations of the underlying chromosomal subpopulation.

Material and Methods

recombClust description

We proposed a method to classify chromosomes according to the allele combinations that are allowed by different recombination patterns. Consider a situation where two recombination patterns are latent in the chromosome population generating two chromosome subpopulations in a given genomic region (Figure 1). A first subpopulation of chromosomes comprises those that have recombined at any of three given points within the region, and a second subpopulation comprises those that have recombined at any of two other points. In this case, we can see, for instance, that while two specific haplotypes G1 and H1 are compatible with the recombination pattern 1, they are maximally different in mutation content at each variant. In addition, H1 is more similar in mutation content to H2 than G1 is to H1, despite H1 and H2 belonging to different recombination subpopulations. In this work, we proposed the method *recombClust* that first identifies points in a genomic segment in which only a fraction of chromosomes have recombined and, second, it computes a consensus classification of chromosomes across all detected points, separating the population of chromosomes according to different recombination patterns along the segment.

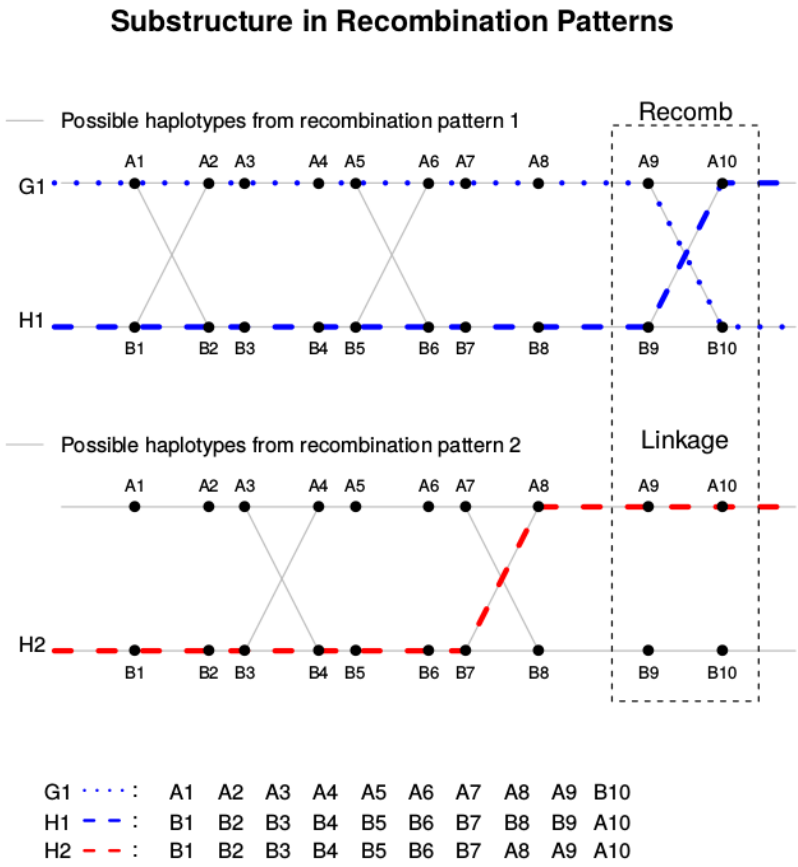

Figure 1: Representation of two chromosomal subpopulations with different recombination patterns in a genomic segment. Lines represent the possible chromosomes present in population 1 (blue) and population 2 (red). Each SNP has two alleles (A and B) and is labeled with a number. Recombination points are placed between SNPs where A and B alleles are joined by a line. G1 and H1 are two possible chromosomes from population 1 and H2 is one of the possible chromosomes from population 2. The dotted box contains a recombination point present in population 1 but not in population 2.

Detection of recombination points in chromosome subpopulations.

The first step of *recombClust* is the identification of points where only a fraction of chromosomes have recombined. Recombination breaks the linkage disequilibrium (LD) between a pair of genetic markers. However, recombination may be restricted between subpopulations of chromosomes in a given genomic region, for instance, by the presence of a chromosomal rearrangement that suppresses the recombination with the chromosomes that do not have it. In those situations, different recombination points may arise at different locations in each subpopulation, where the LD of variants across the point may be completely broken in one subpopulation while remaining high for the other. Therefore the entire population across such recombination points is a mixture of chromosomes, some with very high LD and others with very low LD.

Using SNP phased data, we can detect the presence of a recombination point for only a subpopulation of chromosomes, modeling the likelihood that the subpopulation highly recombined at the point (*recomb*) while the rest remained in complete LD (*linkage*) (Figure 2A). The likelihood is given by a mixture of two latent chromosome subpopulations (A and B). In the first subpopulation, we model the existence of a recombination point that lies in the sequence interval between a pair of SNP blocks (*i=1, 2*) of length L. Phased SNP alleles are encoded by 0 or 1, the haplotype of a chromosome at block *i* is a random variable denoted $X_i \in \{0,1\}^L$ and the haplotype of the joint blocks is the random variable given by the concatenation of the block variables $X_{12} = X_1 \circ X_2$. Under this model, the recombination completely breaks the LD between the

SNP blocks (r^2 = 0) in the *recomb* subpopulation and therefore X_1 and X_2 are statistically independent. Therefore, the probability that a chromosome is observed with haplotype x_{12} in a population under recombination is:

$$P_{recomb}(X_{12} = x_{12} \vee n_1, n_2) = P(X_1 = x_1 \vee n_1)P(X_2 = x_2 \vee n_2) \quad (1)$$

given the haplotype frequencies n_1 and n_2.

For the second chromosome subpopulation, we consider that there is no recombination and we model the SNP blocks to be in complete LD (r^2 = 1). For the chromosomes in the *linkage* subpopulation, X_1 and X_2 are completely linked. X_2 can be unambiguously mapped to X_1 ($f: X_2 \rightarrow X_1$). Under this model, the probability of observing haplotype x_{12} is:

$$P_{linkage}(X_{12} = x_{12} \vee d, f) = \begin{cases} P(X_1 = x_1 \vee d), x_1 = f(x_2) \\ 0, otherwise \end{cases} \quad (2)$$

where d are the frequencies of X_1.

We define the *mixture* model with two components, following equations (1) and (2). The model represents a chromosome population with a mixture of *recomb* and *linkage* subpopulations with proportion π. We therefore assume that the probability of observing a chromosome with haplotype x_{12} is

$$P_{mixture}(X_{12} = x_{12} \vee r_{1,}r_{2,}l_{1,}g, \pi) = \pi P_{recomb}(X_{12} = x_{12} \vee r_1, r_2) + (1 - \pi)P_{linkage}(X_{12} = x_{12} \vee l_1, g) \quad (3)$$

where r_1 and r_2 are the frequencies of haplotypes X_1 and X_2 in the *recomb* subpopulation, l_1 is the haplotype frequencies of X_1 in the *linkage* subpopulation, where g is the function linking X_2 to X_1.

Given a set of m independent chromosomes ($k = 1, \ldots m$), we denote the random variable for the joint blocks over all chromosomes as

$Y_{12} = (X_{12}^1, X_{12}^2, \ldots X_{12}^m)$ and therefore the likelihoods of observing the data y_{12} under the *recomb, linkage* and *mixture* models are:

$$L_{recomb}(n_1, n_2 | y_{12}) = \prod_{k=1}^{m} P_{recomb}\left(X_{12}^k = x_{12}^k \vee n_1, n_2\right) \quad (4)$$

$$L_{linkage}(d, f | y_{12}) = \prod_{k=1}^{m} P_{linkage}\left(X_{12}^k = x_{12}^k \vee d_1, f\right) \quad (5)$$

$$L_{mixture}\left(r_1, r_{2,} l_1, g, \pi \middle| y_{12}\right) = \prod_{k=1}^{m} P_{mixture}\left(X_{12}^k = x_{12}^k \vee r_1, r_{2,} l_1, g, \pi\right) \quad (6)$$

We use the estimated models' likelihoods to test whether the *mixture* model is the best fit to the data and thus to detect the presence of a recombination point in a subset of chromosomes. The parameters for the *recomb* and *linkage* models are estimated by a maximum likelihood. In particular, n_1, n_2 and d are determined by their empirical frequencies. The function f is defined using a greedy algorithm which sequentially pairs each observed x_2, in decreasing order by their frequency, with the x_1 for which the observed frequency of x_{12} is maximum and has not been previously paired.

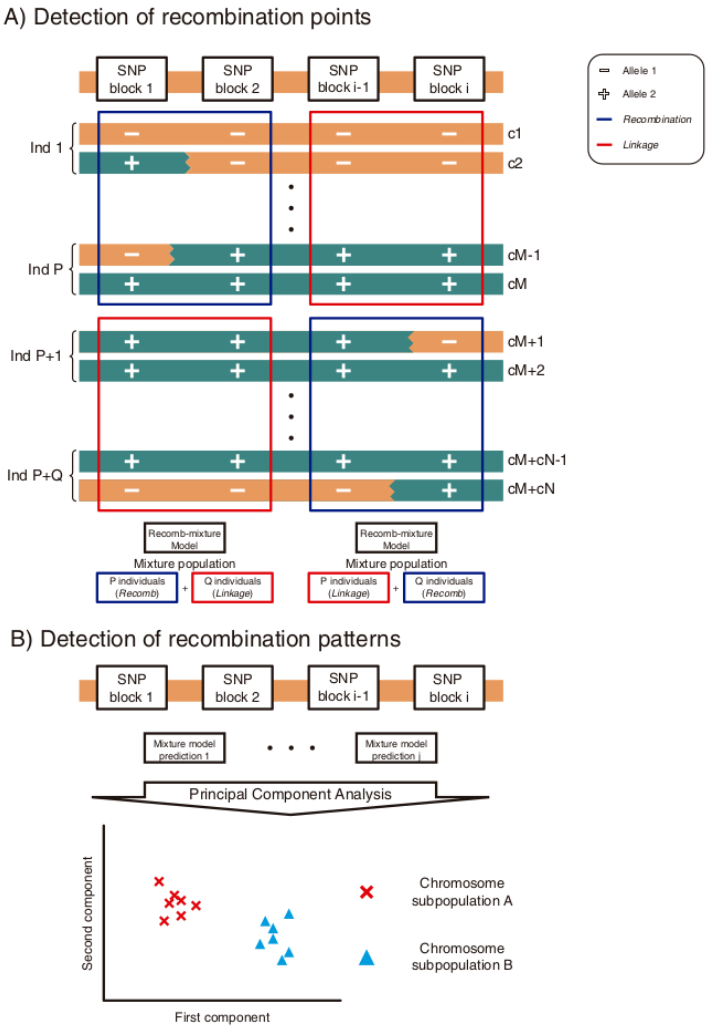

Figure 2: *RecombClust* definition. A) Recomb-mixture model definition. Colors represent two main haplotypes. In the chromosome subpopulation on the left, there is recombination between the blocks, and chromosomes have all possible allele combinations between the SNP blocks (illustrated in geometrical figures). Whereas, for the population on the right, there is no recombination and the chromosomes have only two possible combinations between SNP blocks. Ind P denotes the chromosome P of the population. The Recomb-mixture model contains a mixture of the two models. B) Chromosome classification. Recomb-mixture model is applied to all pairs of SNP blocks in a given region. A consensus classification between all selected Recomb-mixture models is computed by running a PCA on the classification probabilities. Finally, chromosomes are clustered on groups with similar recombination patterns.

The *mixture* model parameters are determined using an Expectation-Maximization (EM) algorithm. For each chromosome, we define a hidden variable $z_k \epsilon \{0,1\}$. This variable indicates if the chromosome belongs to the *recomb* or the *linkage* subpopulations. The EM algorithm updates the model parameters iteratively maximizing the expectation of the data. Given the parameters of the model ω, $\omega = (r_1, r_2, l_1, g, \pi)$, we define the probability that chromosome *k* belongs to the *linkage* subpopulation, $s_{0,k}(\omega) = P\big(z_k = 0 \vee x_{12}^k, \omega\big)$. Similarly, the probability that individual k belongs to the *recomb* subpopulation given ω is $s_{1,k}(\omega) = P\big(z_k = 1 \vee x_{12}^k, \omega\big)$. For each k the probability of belonging to any subpopulation is 1 and, therefore, $s_{0,k}(\omega) + s_{1,k}(\omega) = 1$. In each step of the EM algorithm, we find the value of ω' that maximizes:

$$\omega' = \underset{\omega}{argmax} \sum_{k=1}^{m} \Big[log\left((1-\pi')P_{link}\big(x_{12}^k \vee l_1', g'\big)\right) s_{0,k}(\omega) + log\big(\pi' P_{rec}(x_{12}^k \vee r_1', r_2')\big) s_{1,k}(\omega) \Big] \quad (7)$$

We therefore update the mixture likelihood by ω'given by:

$$\pi' = \underset{\pi}{argmax} \big[log\big((1-\pi')s_0(\omega)\big) + log\left(\pi' s_1(\omega)\right) \big] \quad (8)$$

$$r_1' = \underset{r_1}{argmax} \sum_{k=1}^{m} log\big[P\big(x_{12}^k \vee r_1\big)\big] s_{1,k}(\omega) \quad (9)$$

$$r_2' = \underset{r_2}{argmax} \sum_{k=1}^{m} log\big[P\big(x_{12}^k \vee r_2\big)\big] s_{1,k}(\omega) \quad (10)$$

$$l_1' = \underset{l_1}{argmax} \sum_{k=1}^{m} log\big[P\big(x_{12}^k \vee l_1\big)\big] s_{0,k}(\omega) \quad (11)$$

We estimate haplotype frequencies r_1, r_2, and l_1 in close form using Lagrange multipliers, following Sindi *et al.* [53]. In particular, we obtain

$$\pi' = \frac{s_1(\omega)}{s_0(\omega) + s_1(\omega)} \quad (12)$$

Where $s_0(\omega)$ and $s_1(\omega)$ are the probabilities that a chromosome in the population belongs to the *linkage* or the *recomb* subpopulations ($s_0(\omega) = \sum_{k=1}^{m} s_{0,k}(\omega)$; $s_1(\omega) = \sum_{k=1}^{m} s_{1,k}(\omega)$). We determine the function g' with the actualized parameters using the same procedure than for f'. The final ω' is such that its square root difference with the previous estimate is lower than machine precision. In addition, for numerical stability we set the zero in equation 2 to 10^{-5}.

We then assess if the point, flanked by the SNP blocks, has a credible recombination in a subpopulation of chromosomes by testing whether the *mixture* model is the best model for the data as compared with the *recomb* and the *linkage* models separately. We compare all three models using the Bayesian Information Criteria (BIC) and choose the *mixture* model when [187]:

$$BIC_{mixture} + 10 < min\left(BIC_{recomb}, BIC_{linkage}\right) \quad (13)$$

In addition, we considered that the most robust *mixture* models for a positive detection of a recombination point and for chromosome classification were those for which their estimated frequencies r_1, r_2, and l_1 converged near the observed ones and, therefore, the chi-squared test between observed and estimated frequencies was not significant. When a robust model was observed, a recombination point was called and the classification of chromosome *k* in the *recomb* subpopulation was given by $s_{1,k} > 0.5$.

Clustering of chromosomes into different recombination patterns

A recombination pattern is a set of recombination points in which only a fraction of chromosomes have recombined. In the second step of *recombClust*, a consensus clustering is performed on all the recombination points detected over a genomic region to determine whether individual chromosomes are consistently classified into different recombination patterns given by multiple recombination points. Therefore, to detect recombination points for a subpopulation of chromosomes across the region, *recombClust* first extensively fits the mixture model between numerous 2-SNP blocks, which do not overlap and are at a maximum distance of 100 Kb [188]. For each model with substantial evidence of having a recombination point in a subset of chromosomes, the method computes the probability that the chromosomes belong to a recombination group. Finally, *recombClust* produces a consensus classification of the chromosomes by clustering the first principal component of the group probabilities across all recombination points (Figure 2B). The method is implemented in the R package, *recombClust* (https://github.com/isglobal-brge/recombClust), which accepts haplotype data from VCF files.

Performance of *recombClust* to detect recombination points and recombination patterns

We evaluated the performance of the mixture model to detect a single recombination point in a chromosome subpopulation using simulated datasets, computing the ability to detect the point and the accuracy in chromosome classification. We produced diverse synthetic datasets. The

reference data consisted on pairs of SNP-blocks flanking a recombination point, with 2 SNPs per block on 1000 chromosomes. The blocks had intermediate LD between the SNPs, and SNP allele frequencies were selected at random between 0.55 and 0.95. To evaluate the performance of the mixture model to correctly classify a subpopulation of chromosomes with a recombination point, we simulated a population (R_A+L_B) with a mixture of subpopulations A and B with high recombination (R_A) and high linkage (L_B), respectively. Under this scenario, we changed the mixture frequency, the linkage between the SNPs within a block, and the frequency of the reference SNP alleles to evaluate the model robustness.

To evaluate the performance of the mixture model to detect a recombination point in a subpopulation of chromosomes, we produced various population mixtures (Table 1). In the first set, no mixture was present and chromosomes belong to a single population A with high recombination (R_A), intermediate linkage (ML_A) and high linkage between blocks (L_A). In the second set, we produced populations with different types of mixtures that included: 1) the target scenario, where one subpopulation was under high recombination and the other with high linkage between blocks (R_A+L_B); 2) both subpopulations under high recombination (R_A+R_B); 3) both subpopulations with high LD (L_A+L_B); 4) both subpopulations with intermediate linkage (ML_A+ML_B) and 5) one subpopulation in intermediate and the other in high linkage (ML_A+L_B). Note that populations A and B differ in allele frequency, so populations with a mixture of two populations with the same model (R_A+R_B, L_A+L_B and ML_A+ML_B) are different from a single population.

Table 1: Simulated scenarios for the mixture of two chromosome groups A and B with differences in recombination history at a given point. The chromosomal subpopulations differ in the linkage between two blocks of two SNPs flanking the recombination point. Scenarios correspond to high recombination for population A and B (R_A and R_B), intermediate linkage (ML_A and ML_B) and high linkage (L_A and LB) and their possible mixtures. The mixture model targets scenario R_A+L_B in bold face.

Scenario	Chromosome mixture	Between Block LD Population A	Between Block LD Population B
R_A	No	$r^2 = 0$, D' = 0	-
ML_A	No	$r^2 < 1$, D' = 1	-
L_A	No	$r^2 = 1$, D' = 1	-
R_A+L_B	**Yes**	**$r^2 = 0$, D' = 0**	**$r^2 = 1$, D' = 1**
R_A+R_B	Yes	$r^2 = 0$, D' = 0	$r^2 = 0$, D' = 0
L_A+L_B	Yes	$r^2 = 1$, D' = 1	$r^2 = 1$, D' = 1
ML_A+ML_B	Yes	$r^2 < 1$, D' = 1	$r^2 < 1$, D' = 1
ML_A+L_B	Yes	$r^2 < 1$, D' = 1	$r^2 = 1$, D' = 1

We also evaluated the performance of classifying the chromosomes under different recombination patters using simulated inversions. As inversion polymorphisms produce chromosomal subpopulations that differ in their recombination patterns, we tested the ability of *recombClust* to detect inversion status in simulated inversions. We simulated an inversion of 800 Kb and a frequency of 20% using *invertFREGENE* [109] to evaluate the mixture model at different recombination points. We varied the inversion length (from 50 Kb to 1 Mb) and inversion frequency (from 0.1 to 0.9) to evaluate the overall *recombClust* performance to call the inversion status of the chromosomes. Each combination of frequency and length was run 100

times. In all simulations, we used the default values of *invertFREGENE* parameters (recombination rate: 1.25×10^{-7}, mutation rate: 2.3×10^{-7}).

Human inversions

We studied the extent to which inversion polymorphisms can be better characterized by allele combinations rather than mutation differences among the SNPs within the inversion. We therefore used *recombClust* to classify inversion status of chromosomes for the two best characterized human inversions, which are found at 8p23.1 (chr8:8055789-11980649, hg19) and 17q21.31 (chr17:43661775-44372665, hg19). We used SNP phased data from the 1000 Genomes project [185]. We inferred the individuals' inversion genotypes with *recombClust* and compared it with those obtained with *invClust* [49], a reference method to determine the inversion status based on mutation differences. We compared both inversion callings with experimental inversion genotypes available in the inversion repository invFEST [107].

Recombination substructure in a susceptibility locus for breast cancer

We ran *recombClust* in the 0.2Mb region 1q21.1 between chr1:145.55Mb-145.75Mb (hg19) which contains the largest LD block with significant SNPs associated with breast cancer, as observed in a large GWAS meta-analysis [186]. The GWAS results do not indicate clearly towards a causal SNP and, in addition, no inversion or structural variation has been reported that could explain the association of the entire block with breast cancer. We, therefore, tested whether the chromosomes in the region presented a recombination substructure that could underlie the associations.

We analyzed four independent studies to validate the chromosome substructure in the region: The breast cancer case-control GWAS study CGEMS (dbGAP accession: phs000147.v3.p1)[110], the breast cancer samples from The Cancer Genome Atlas (TCGA) project [189, 190], the European individuals from the 1000 Genomes project, and the Genotype-Tissue Expression (GTEx) project [103]. For CGEMS data and TCGA, we imputed SNPs from chromosome 1 on Michigan server [118], using HRC r1.1 2016 as reference panel and EAGLE v.2.3 to phase which returned phased haplotypes, and computed the LD (R^2) between SNPs in region and the chromosome subpopulation genotypes using *snpStats* [191]. In TCGA, we selected the individuals classified as European by *peddy* [118] with a probability higher than 0.9. In GTEx, we phased chromosome with *SHAPEIT* [192] and we also selected European individuals with *peddy.* In the *recombClust* analysis, we included SNPs with a MAF > 0.05 and performed the consensus clustering across the detected points with a hierarchical clustering on the first two PCs of the chromosome subpopulation probabilities. Chromosome subpopulation genotypes were computed by coding chromosome subpopulations as alleles and testing Hardy-Weinberg equilibrium using *SNPassoc* [193].

We studied whether the chromosome genotypes were associated with gene expression and phenotype differences between individuals. We evaluated the association with gene expression in multiple tissues using GTEx data, using the gene raw counts from *recount2* [194]. For each tissue, we removed genes with less than 10 counts in more than 90% of the samples. We tested the association between the chromosome alleles and gene expression, applying *limma* [128] to $\log_2$CPM values obtained with *voom* [157]. We included sex, platform, top three genome-wide principal components and variables from *PEER* as covariates. We also

tested the transcriptomic effect of the interaction between the chromosome subpopulation genotypes and 26 SNPs (MAF > 0.1) in the region. Finally, in the case-control study, CGEMS (1,145 cases and 1,142 controls), we studied the association between the chromosome subpopulation genotypes and cancer status, adjusting for age and genome-wide PCs.

Results

We developed and implemented *recombClust* to classify chromosomes into different recombining groups across multiple recombination points within a genomic region. The method comprises two steps. First, it detects recombination points where only a fraction of chromosomes have recombined, and then it classifies chromosomes into subpopulations based on a consensus clustering across the detected points.

Detection of recombination points in a chromosome subpopulation

The detection of a recombination point in a subpopulation of chromosomes was given by the observation of a robust mixture model, as defined in the Methods section. The mixture model tests whether the chromosomes of a population can be classified in two subpopulations (A and B) defined by the presence of a recombination point in one subpopulation but not in the other. We first evaluated the accuracy of the mixture model to classify individual chromosomes in a target scenario, (R_A+L_B) in Table 1, where one subpopulation had maximum recombination and the other maximum linkage. We observed that the

mixture model was more accurate in calling chromosomes from the recombining subpopulation than those at high linkage (Supplementary Note). We also observed that the model was robust under different parameters of initialization, LD between the SNPs composing the blocks or the genetic divergence between the subpopulations (Supplementary Note).

We evaluated the performance of the mixture model to detect a recombination point in a subpopulation of chromosomes between two SNP blocks, using the synthetic datasets described in Table 1. In particular, we tested whether the model correctly identified the scenarios in which only one chromosome subpopulation had a recombination point between the SNP blocks. Scenarios with detectable recombination points were selected from those mixture models with lowest BIC and robust fitting, and compared with their simulated values. As expected, we confirmed that the model was optimal for the target scenario (R_A+L_B) (Figure 3). In addition, the model completely discarded points with no mixture and whose chromosomes where all in complete recombination (R_A) or high linkage (R_L). Nonetheless, we observed that the model detected points with no mixture under intermediate linkage (ML_A) or under more complex mixtures. This observation can be explained by the fact that the mixture model considers the ideal case where two subpopulations are either in complete linkage or full recombination. In reality, intermediate linkage values and mixtures are expected, where the model still detects a significant mixture signal. Note, however, that the final aim of *recombClust* is to consistently classify chromosomes across the several points that constitute a recombination pattern within an extended genomic region and, therefore, inaccuracies

at individual points are expected to cancel out when considering strong consistency over large regions.

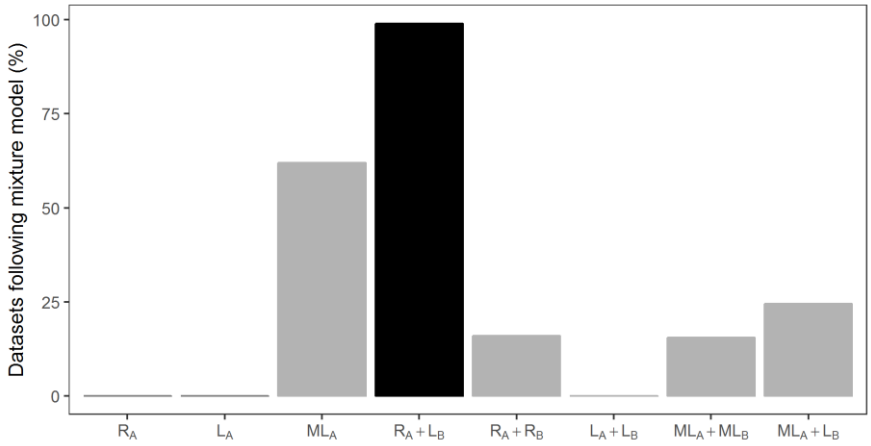

Figure 3: Performance of the *Recomb*-mixture model in simulated scenarios. Fitting of the mixture model that assumes a mixture of chromosomal subpopulations A and B, where chromosomes in A are under maximum recombination (R) and chromosomes in B under maximum linkage (L) in a single recombination point. Scenarios are described in Table 1, where chromosomal populations where created with and without mixture and chromosomal subpopulations were characterized by maximum recombination, intermediate linkage (ML) and maximum linkage. The figure shows the proportion of scenarios detected by the model with significant mixture of the type $R_A + L_B$ under all scenarios. The mixture model is clearly optimal for the correct scenario, while it also detects some positive signal in others.

Classification of chromosome subpopulations across multiple recombination points

recombClust performs a consensus clustering of multiple mixture models at numerous recombination points. To test the accuracy of the method to identify subpopulations of chromosomes with different recombination patterns across multiple points, we simulated two subpopulations of 1000 chromosomes, each with five different recombination points under a R_A+L_B scenario (Table 1). From the identified recombination points, we performed consensus clustering using a k-means algorithm on the first PC of the chromosome classification probabilities across all mixture models. We thus observed a neat separation of the chromosome subpopulations (Figure 4A), which we did not observed for the first two PCs across all simulated SNPs, confirming that the model selected recombination substructure and allele combinations rather than mutation differences between chromosomes.

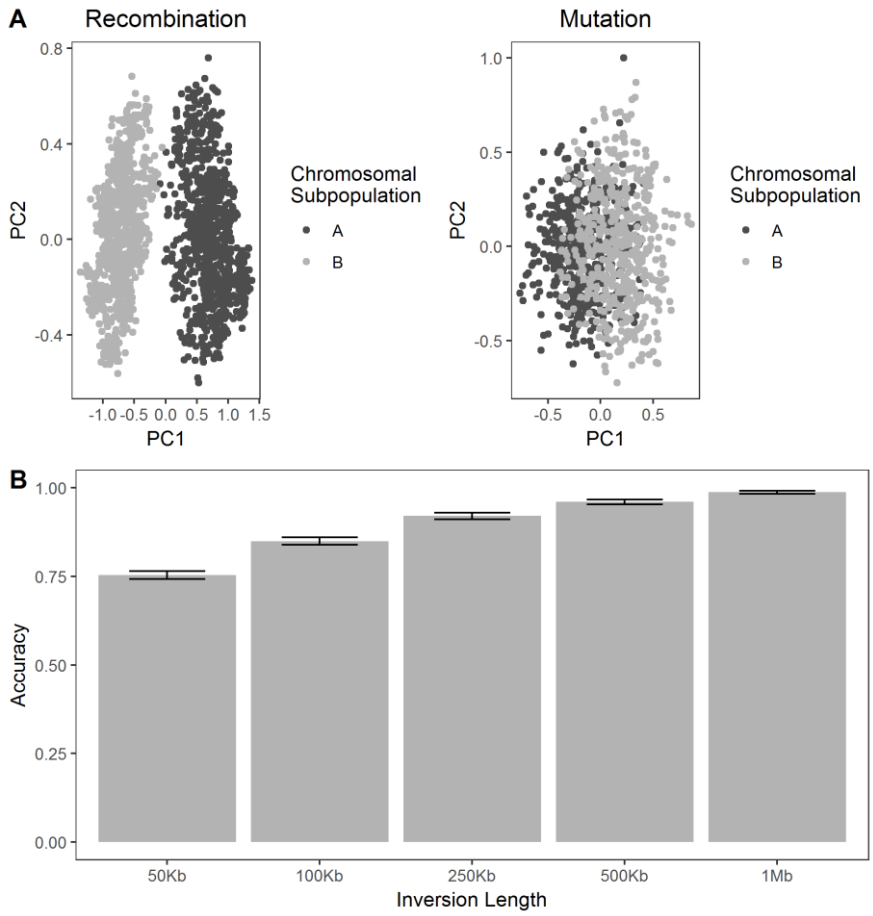

Figure 4: *recombClust* accuracy for detecting subpopulations with different recombination patterns. A) Comparison between using recombination or mutation data in a simulated population. The simulated population contained a mixture of two subpopulations (A and B) with different recombination patterns; 5 different recombination points were simulated for each subpopulation under a $R_A(B) + L_B(A)$ scenario. A-left) PCA for the chromosome classification at 10 different recombination points of a simulated chromosome population. The figure shows a clear separation of the subpopulations based on different recombination patterns at multiple recombination points. A-right) PCA for the genotype values showing that the separation based on mutation differences between the subpopulations is not neat. B) The figure shows the accuracy for the predicted chromosome classification into inversion status as obtained by *recombClust* for 9,000 simulated inversions at a given size (1000 simulations at 9 different inversion frequencies). The figure shows the mean accuracy and standard error. The accuracy is high but drops with the inversion length for values lower than 100Kb.

Classification of inversion status based on recombination differences

Inversion polymorphisms differ in the recombination patterns inside the inverted region. We therefore asked the extent to which the inversion status of chromosomes can be inferred by recombination differences, using *recombClust*. We evaluated the performance of the method using the coalescent simulator of inversions *invertFREGENE*. We first tested the performance of the mixture model to classify chromosomes into recombining groups at different recombination points across the inversion and then we tested the accuracy of *recombClust* to call inversion status.

To first test the accuracy of chromosome classification at single recombination points, we simulated an 800 Kb inversion at 20% frequency. We fitted the mixture model across multiple points in the inverted region, detecting the points were recombination occurred only in inverted or standard chromosomes. We evaluated the accuracy of each robust model, corresponding to a detected recombination point, to classify chromosomes into the inverted status. We observed that the mixture models had median specificity of 1 and a median sensitivity of 0.89 across all the recombination points detected. We observed that LD-based SNP pruning and SNP block size can affect the accuracy in chromosome classification (Figures S1-S2). However, the chromosomes were clearly separated by inversion status under all conditions (Figures S3-S4).

We then evaluated the accuracy of *recombClust* to classify the inversion status across all recombination points by performing a consensus clustering based on the k-means clustering of the first PC of all mixture

model classifications. We simulated inversions with different lengths and frequencies. We observed high accuracy in inversion calling (Figure 4B), in particular, greater than 90% for inversions larger than 0.25MB. As expected, accuracy for short inversions was lowered as they presented fewer recombination points. *recombClust*'s accuracy was stable for inversion frequencies within the range (0.2, 0.8) (Figure S5) and did not correlate with inversion's age (r = 0.02, p-value = 0.19) (Figure S6). Overall, we confirmed that *recombClust* was able to call the inversion status of chromosomes from their recombination patterns.

Recombination differences improves classification of human inversions across multiple ancestries

We compared the *recombClust* calling of human inversions at 8p23.1 and 17q21.31, based on recombination differences, with the calling produced by mutation differences. We first observed that within inv-8p23.1, most SNP pairs closer than 100Kb showed high differences in LD between inversion and standard chromosomes (Figure S7). The observation is in line with reported differences in recombination patterns between inversion status [20], suggesting that recombination differences can be used to infer inv-8p23.1 genotypes.

Using *recombClust* in the European samples of the 1000 Genomes Project, we searched for recombination points in either the standard or inverted chromosomes. We tested if the inferred chromosome subpopulations matched experimental inversion status as reported in invFEST and compared the results with *invClust*, a standard method for inversion calling based on mutation differences between chromosomes. We found that *recombClust* separated inverted and standard chromosomes, based on the k-means clustering on the first PC of the

mixture model's probabilities of detected recombination points (Figure S8). As the first PC, clearly showed two clusters, individuals' k-means classification was highly accurate with respect to the experimental inversion-alleles (Table 2). In inversion 8p23.1, some chromosomes from heterozygous samples lied between the clusters likely affected by phasing errors. We observed that *invClust* also classified inversion genotypes accurately (Table 2). These results show that recombination substructure can be reliably used to call inversion status at 8p23.1 and 17q21.31 in Europeans in addition to mutation differences, as detected by *invClust.*

Table 2: Classification accuracy of two human inversions in European individuals of the 1000 Genomes project. recombClust shows that experimental inversion status of chromosomes can be classified from differences in recombination patterns, in addition to differences in mutations, mainly detected by invClust.

	Inversion status	*recombClust*	*invClust*
inv-8p23.1	All	1.000	1.000
	Inv/Inv	1.000	1.000
	Std/Inv	1.000	1.000
	Std/Std	1.000	1.000
inv-17q21.31	All	0.993	0.998
	Inv/Inv	1.000	1.000
	Std/Inv	0.987	1.000
	Std/Std	0.996	0.996

We then compared the inversion calling between *recombClust* and *invClust* for all the individuals in the 1000 Genomes Project, testing the performance of *recombClust* under multiple ancestries [49]. We first observed a lower number of detected recombination points than for the European individuals, which could be caused by the large variability in the ancestry of the data (Table S1). For the two inversions, we observed that classification accuracy was strongly affected by individuals' ancestry

(Table 3). Thus, both methods had high accuracy for all ancestries except African (Table 3). Asians showed moderate accuracy for inv-8p23.1.

Table 3: Classification accuracy of recombClust and invClust for two inversions on multiple ancestries, from the 1000 Genomes Project. Accuracy estimates are reported in each 1000 Genomes superpopulation. recombClust classification showed equal or greater accuracy than invClust for the chromosome classification of experimentally validated inversion alleles. None of the AMR individuals had inversion genotypes for inversion inv-8p23.1 so they are not reported. EUR: European, AFR: African, AMR: American, EAS: East asia.

	Ancestry	N	recombClust	invClust
inv-8p23.1	All	80	0.700	0.688
	EUR	34	1.000	1.000
	AFR	20	0.250	0.250
	EAS	26	0.654	0.615
inv-17q21.31	All	1142	0.885	0.800
	EUR	425	0.988	0.988
	AMR	190	0.916	0.911
	AFR	238	0.538	0.134
	EAS	289	1.000	1.000

We inspected the first two PC components of the mixture model prediction across inv-8p23.1 and inv-17q21.31, for all ancestries (Figure 5), and observed multiple clusters, in which chromosomes segregated both by inversion status and ancestry. Similar clustering has been observed for mutation differences in these inverted regions [49]. However, differences in allele combinations revealed differences between the chromosomal substructure of the inversions. For inv-17q21.31, we observed multiple clusters that mapped to the inversions status but not to ancestral differences, while for inv-8p23.1, ancestry subgroups were observed within each inversion status. These observations confirmed that that clusters identified in the first PCs of the mixture model predictions can be interpreted as non-recombining chromosome groups that differ in ancestry or inversion status, or other

unobserved factors that suppress recombination between the groups, such as copy number variants likely segregating the standard chromosomes at 17q21.31 [113].

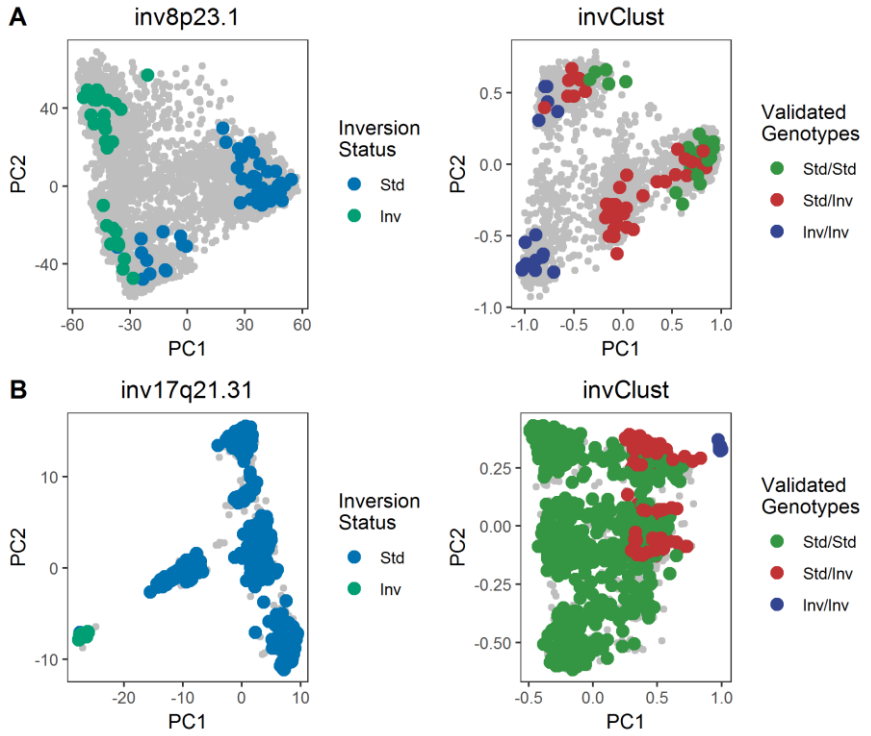

Figure 5: Identification of chromosomal subpopulations from different ancestries in two inverted regions. The figures show the first two PCA components for the all mixture model predictions at numerous recombination points across inv-8p23.1 and inv-17q21.31, computed for all 1000 Genomes ancestries. Chromosomes are clearly separated by inversion status (Std, Inv) and ancestry. For inv-8p23.1 clear ancestral groups are identified within inversion status whereas ancestry is mixed within each inv-17q21.31 status. Colored points indicate experimentally validated observations of inversion status and ancestry.

A recombination substructure in 1q21 strongly associates with breast cancer susceptibility

We applied *recombClust* in a 0.2Mb region at 1q21.1 containing numerous SNPs associated with breast cancer. While no structural variation has been detected in the region that can account for the large association across the block, we aimed to determine if chromosomes in the region could be classified in different recombination groups and if the groups conferred a higher risk compared with any of the individual SNPs. We therefore run *recombClust* across the region in four independent SNP datasets and observed a reproducible 4-cluster pattern in the first two PCs of the chromosome probabilities (Figure 6A). We performed a hierarchical clustering on the first two PCs and defined chromosome alleles by allele frequency (freq allele 1: 0.39, freq allele 2: 0.36, freq allele 3: 0.20, freq allele 4: 0.05), observing that allele 2 was not in Hardy-Weinberg equilibrium (p-value: $3.7x10^{-3}$). We also applied a MDS analysis to the breast cancer dataset from the case-control CGEMS study to partition the population using mutation differences. Although we observed clusters that correlated with *recombClust* classification, several groups emerged (Figure S9) showing that mutation differences substantially differed from the classification derived from recombination patterns. We run *recombClust* in 10q26.13, another susceptibility locus with a causal SNP candidate for breast cancer risk where no chromosome substructure was expected. As expected, we could not identify a recombination substructure in 10q26.13 (Figure S10); suggesting the peculiarity of the recombination pattern of 1q21.1 against other susceptibility locus of breast cancer.

We then studied whether the identified chromosomal subpopulations affected the expression of nearby genes using the GTEx data for multiple tissues. While we did not find any differential expression of genes with the chromosome alleles in breast tissue, we observed that all chromosome subpopulations changed the gene expression in multiple genes across different tissues (Figure 6B). In particular, chromosome subpopulations 1 and 2 had effects on numerous genes and tissues. We also tested whether the chromosome alleles modulated the eQTL effect of SNPs in the region. We observed that the interaction between rs56110740 and chromosome allele 2 was significantly associated with the expression of *PDKZ1* in transverse colon (Figure 6C). These observations suggest a functional role of the recombination substructure at 1q21.1 and a possible modulation of eQTL effects.

We finally tested the associations between the different chromosome alleles and breast cancer status, and compared them with the associations for all the SNPs in the region. We therefore analyzed the case-control CGEMS study. We found a significant association between chromosome allele 4 and cancer (OR: 1.62, p-value: $5x10^{-4}$) that was more significant than any other SNP association in the region (Table 4, Figure 6D). The significance of the association was of the same order of magnitude than those found for rs144778858 and rs116124754. Interestingly, these SNPs were not included in the recombination patterns detected by *recombClust* and were not in LD with allele 4 (R^2= 0.27). Consequently, rs144778858 and rs116124754 are unlikely to drive the association between chromosome allele 4 and breast cancer. In addition, we did not see any tag SNPs for the allele 4 that suggests the existence of an unobserved chromosome rearrangement or process that suppresses recombination between the allele 4 and the other

chromosomes. While none of the structural variant reported in the 1000 Genomes project was associated with the chromosome alleles, we finally noted that this region has suffered numerous reconstructions between different builds of the reference genome.

Table 4: Top SNP associations with breast cancer for the CGEMS study in the 1q21.1 locus. SNPs in bold face were used to derive the recombination substructure of the region.

SNP	Chr	Position	Effect Allele	OR (95% CI)	p-value
Allele 4	*1*	-	-	*1.62 (1.23-2.15)*	*5.34 x10^{-4}*
rs144778858	1	145573536	C	1.93 (1.32-2.82)	7.27x10^{-4}
rs116124754	1	145574977	T	1.93 (1.32-2.82)	7.27x10^{-4}
rs10797655	**1**	**145633187**	**T**	**1.19 (1.05-1.33)**	**4.70x10^{-3}**
rs6424379	**1**	**145555653**	**C**	**1.18 (1.05-1.33)**	**6.45x10^{-3}**
rs882210	1	145732946	C	1.43 (1.09-1.89)	1.14x10^{-2}
rs10399658	**1**	**145686437**	**A**	**1.16 (1.03-1.30)**	**1.49x10^{-2}**
rs28841052	**1**	**145616867**	**T**	**1.15 (1.02-1.29)**	**1.88x10^{-2}**
rs116746633	1	145679352	G	2.21 (1.14-4.29)	1.93x10^{-2}
rs4970860	**1**	**145619411**	**A**	**1.15 (1.02-1.29)**	**2.04x10^{-2}**
rs12735609	**1**	**145643945**	**G**	**1.15 (1.02-1.29)**	**2.04x10^{-2}**

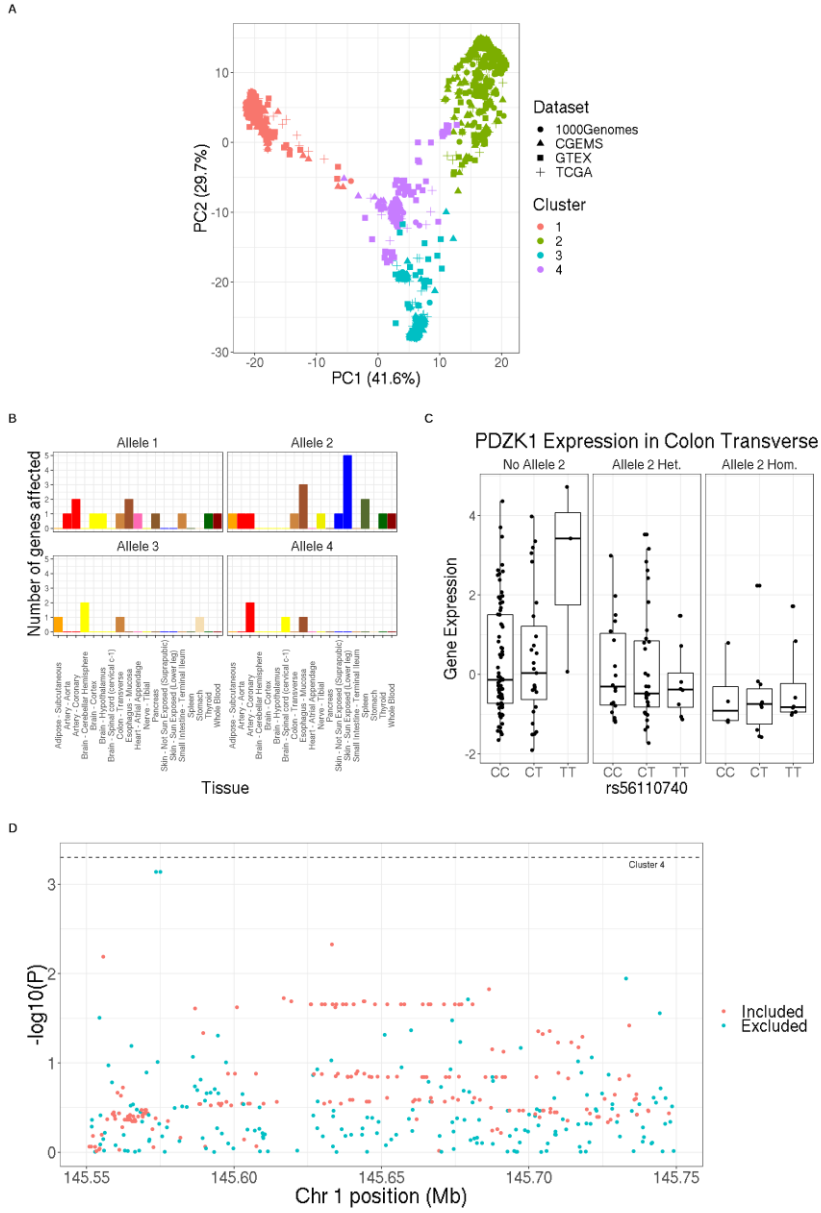

Figure 6: Chromosome subpopulations identified from recombination differences within 1q21.1, a susceptibility locus for breast cancer. A) PCA for the chromosome classification at multiple recombination points across the region, as implemented in *recombClust*. Hierarchical clustering was applied to identify four chromosome clusters, which were validated across four independent studies. B) Effects of chromosome alleles on gene expression across multiple tissues. The figure shows numerous eQTL effects of the chromosome alleles, particular for the most common alleles 1 and 2.

C) Modulation of the effect of rs56110740 on PDK1 expression by chromosome allele 2. Allele 2 Het.: Individuals heterozygous for allele 2. Allele 2 Hom.: Individuals homozygous for allele 2. D) Chromosomal allele 4, defined by middle cluster in figure A was tested for association with breast cancer. The dotted line shows p-value for the association. Points show individual SNP associations with breast cancer across the region. SNPs in red were selected by recombClust for classifying the chromosomes into different recombining groups; SNPs in blue were excluded. The figure shows that the association of chromosome allele 4 with breast cancer is more significant than any of the SNP associations.

Discussion

We proposed an analysis method based on SNP data that identifies subpopulations of chromosomes that differ in their recombination patterns. Recombination differences, given by the distribution of recombination points along a chromosome segment, are evident in subpopulations of different ancestry [176] or in subpopulations that have rearrangements that suppress recombination between chromosomal groups [20]. Our method addresses the, yet unasked, question of whether chromosomal subpopulations can be identified from latent differences in recombination patterns within a population. The existence of such chromosomal subpopulations may thus provide evidence about unobserved recombination modifiers, such as chromosomal rearrangements.

Population substructure is commonly detected from mutation differences between the chromosomes, that is, the amount of their allelic mismatch [48], typically obtained by the clustering of the first principal components of SNP genotypes [48, 49, 184]. By contrast, differences in recombination patterns are based on the differences in the combination of common alleles that are present in each chromosomal subpopulation and, as such, they constitute a different source of genetic

divergence. We observed that meaningful differences in recombination patterns between chromosomes can be inferred but are limited to regions greater than 0.1Mb, as shorter regions are not likely to contain enough recombination points [117]. While errors in the phasing of chromosomes can also limit our observations, we found two cases in which the chromosome subpopulations derived from mutation and recombination differences differed. First, we observed that chromosome subpopulations derived from recombination differences improved the mutation-based calling of human inversions at 17q21.31. The increased accuracy can be explained because the appearance of new allele combinations is more frequent than the emergence of new alleles [195]. Second, we observed that four robust chromosome subpopulations were detected in the 1q21.1 susceptibility locus for breast cancer [186], where large number of subpopulations were found based on mutation differences. In particular, we observed that a chromosome subpopulation showed higher association with breast cancer that any of the SNPs in the 1q21.1, suggesting that the causal variant could be other than a SNP [196, 197]. These results indicate that singular combination of alleles could either be an important source for explaining the associations or it could signal the presence of an unobserved causal process that acts as a recombination modifier. In this context, our method, *recombClust,* can be used to further investigate the extent to which recombination substructure can help to explain genome-wide associations.

Other important questions that follows are how common is the presence of chromosomal subpopulations derived from recombination substructure and whether it correlates with selection signals. Some methods scan the genome to detect regions of significant population

differentiation due to genetic structure. A common approach is to identify genomic regions subjected to selection based on the genetic distance (Fst) between populations [198, 199] which has been successfully applied to different species and taxa [200–202]. In terms of inversions, another approach is to detect LD differences between chromosomal subpopulations across breakpoints to infer regions where inversions may be present [54]. Similarly, *recombClust* can be extended to scan the genome and test whether the detected signal correlates with inversion or Fst genomic signals. In particular, regions with strong chromosome clustering given by *recombClust* could shade light into regions with early genetic divergence.

recombClust is the first method to detect chromosome subpopulations based on allele combinations. It can be used to call inversion polymorphisms yet its main advantage is the detection of regions where suppression of recombination is likely at place. In particular, the method can provide evidence of unobserved structural variants that may underlie the associations of numerous SNPs in large LD blocks, as typically reported in GWASs. As such, the method can be applied to large amounts of public GWAS data, available from repositories like dbGAP.

Availability

The datasets analyzed were derived from the following public domain resources:

- 1000 Genomes project phase 3: project web page (http://www.internationalgenome.org/)
- The Cancer Genome Atlas (TCGA): dbGAP authorized access (accession code: phs000178.v10.p8)

- The Genotype-Tissue Expression (GTEx) Project: dbGAP authorized access (accession code: phs000424.v7.p2)
- Cancer Genetic Markers of Susceptibility (CGEMS): dbGAP authorized access (accession code: accession number phs000147.v3.p1)

recombClust is available in Github (https://github.com/isglobal-brge/recombClust).

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness [MTM2015-68140-R]; and the Catalan Government [#016FI_B 00272 to CR-A]. Funding for open access charge: Spanish Ministry of Economy and Competitiveness.

Acknowledgments

The authors would like to express their gratitude to the Supercomputing and Bioinnovation Center (SCBI) of the University of Malaga (Spain) for their support and resources. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. GTEx data were obtained from: the GTEx Portal on 06/07/2018 and dbGaP accession number phs000424.v7.p2 on 12/05/2017. CGEMS data was obtained from dbGaP (accession number phs000147.v3.p1). The results shown here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Conflict of interest

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Supplementary Figures

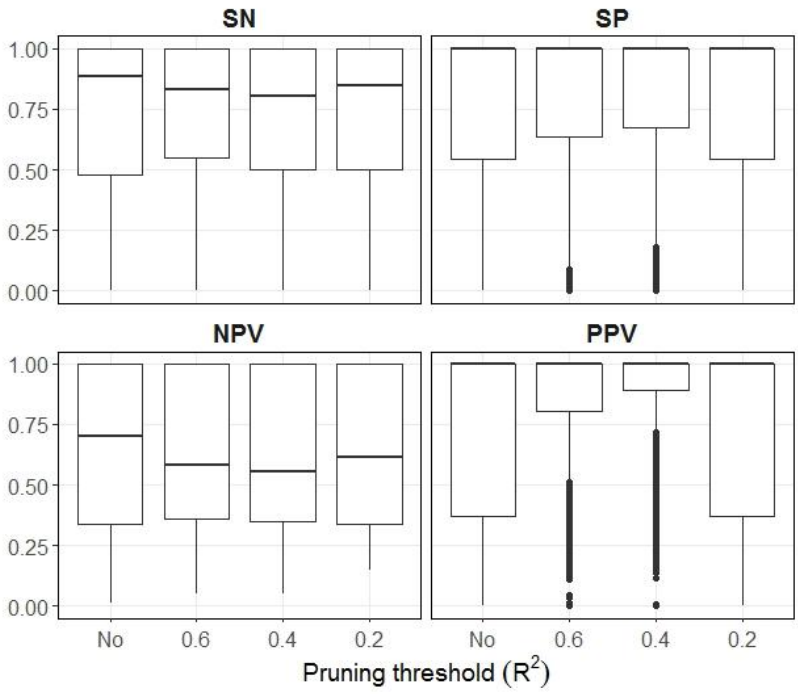

Figure S1: Models accuracy for different SNPs pruning. Each boxplot includes only those block-pairs belonging to mixture population. SN: sensitivity; SP: specificity; NPV: negative predictive value; PPV: positive predictive value.

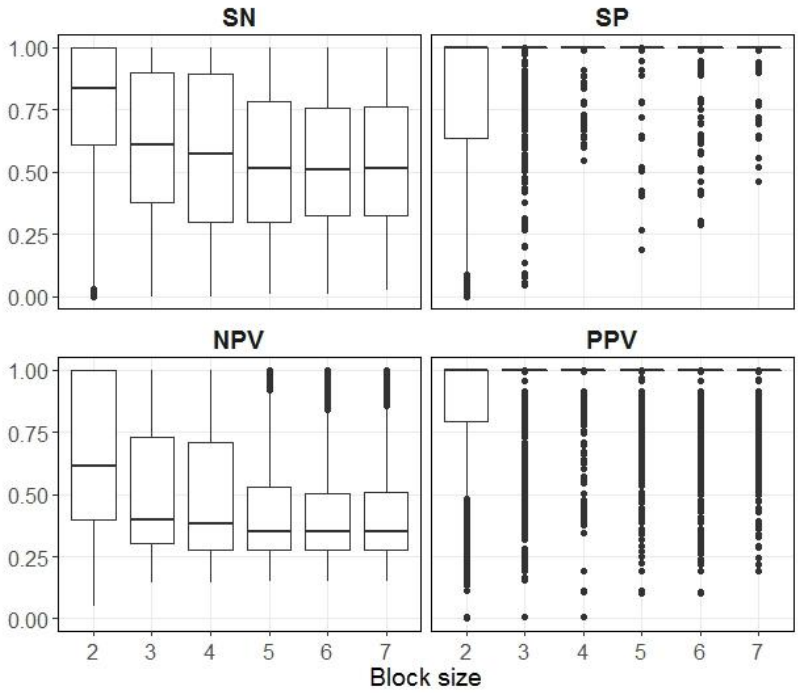

Figure S2: Models accuracy for different block sizes. Each boxplot includes only those block-pairs belonging to mixture population. SN: sensitivity; SP: specificity; NPV: negative predictive value; PPV: positive predictive value.

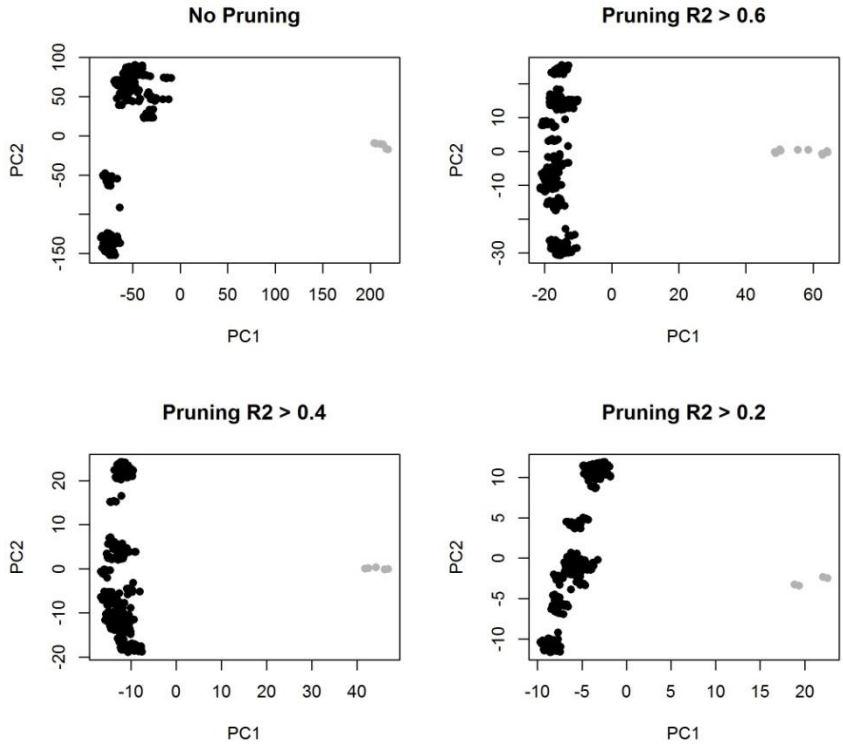

Figure S3: PCA of chromosome probabilities for different SNP prunings. Each point represents a phased chromosome. Phased chromosomes are colored based on inversion status (black: standard, grey: inverted).

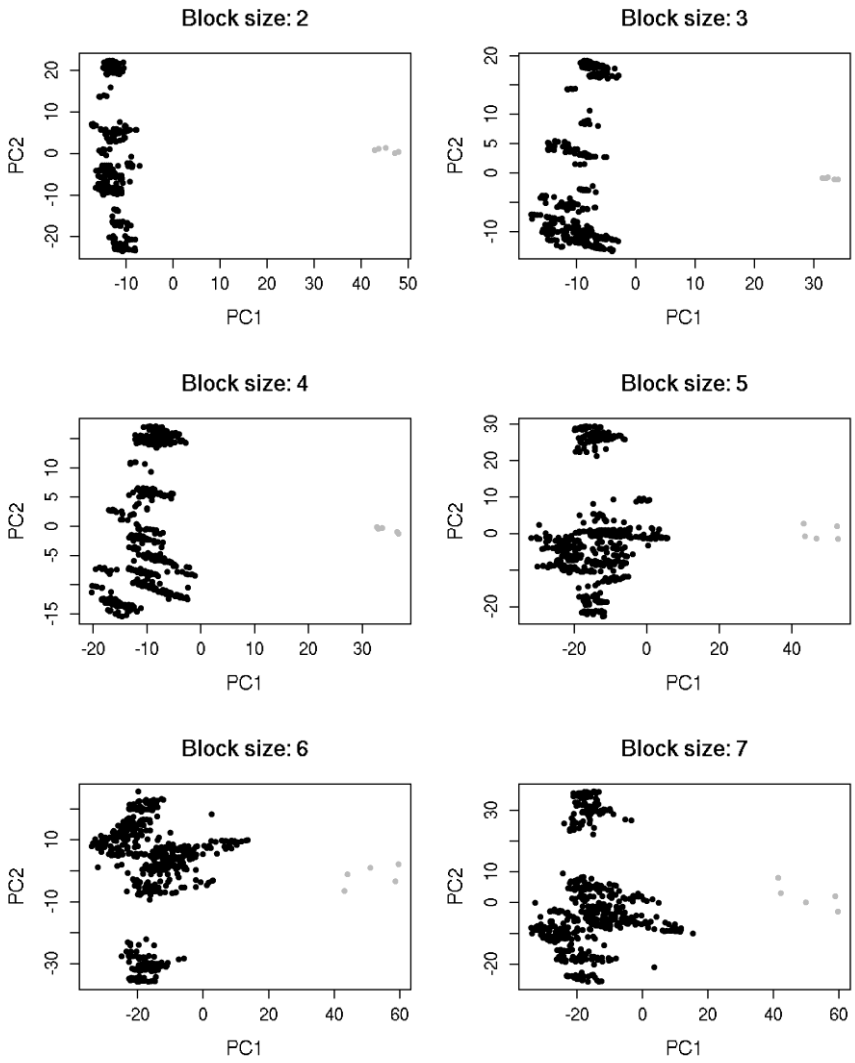

Figure S4: PCA of chromosome probabilities for different block sizes. Each point represents a phased chromosome. Phased chromosomes are colored based on inversion status (black: standard, grey: inverted).

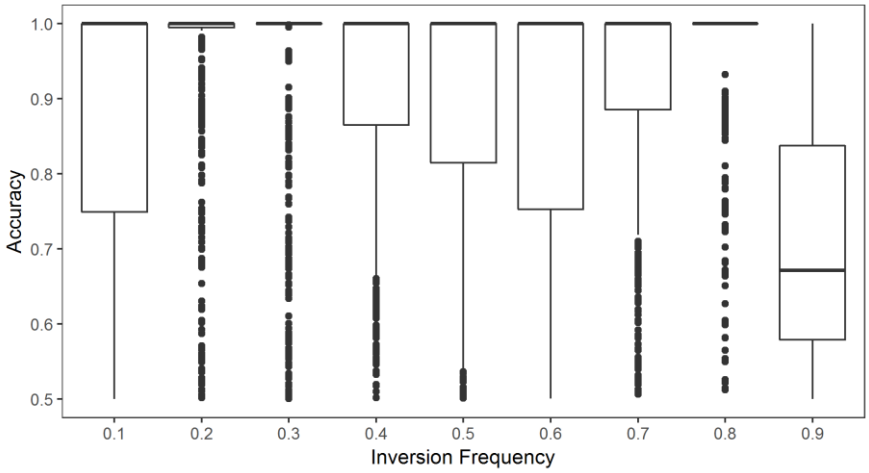

Figure S5: *recombClust* accuracy for different inversion frequencies. Accuracy is the proportion of phased chromosomes correctly classified. Each boxplot includes 500 simulations.

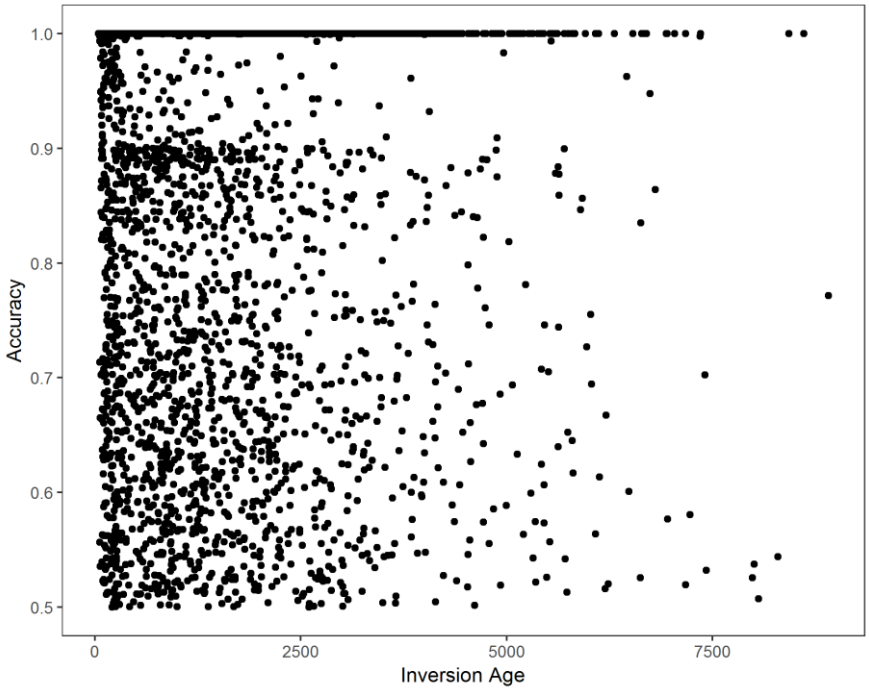

Figure S6: *recombClust* accuracy for different inversion ages. Accuracy is the proportion of phased chromosomes correctly classified. Each point is the accuracy of an independent simulation.

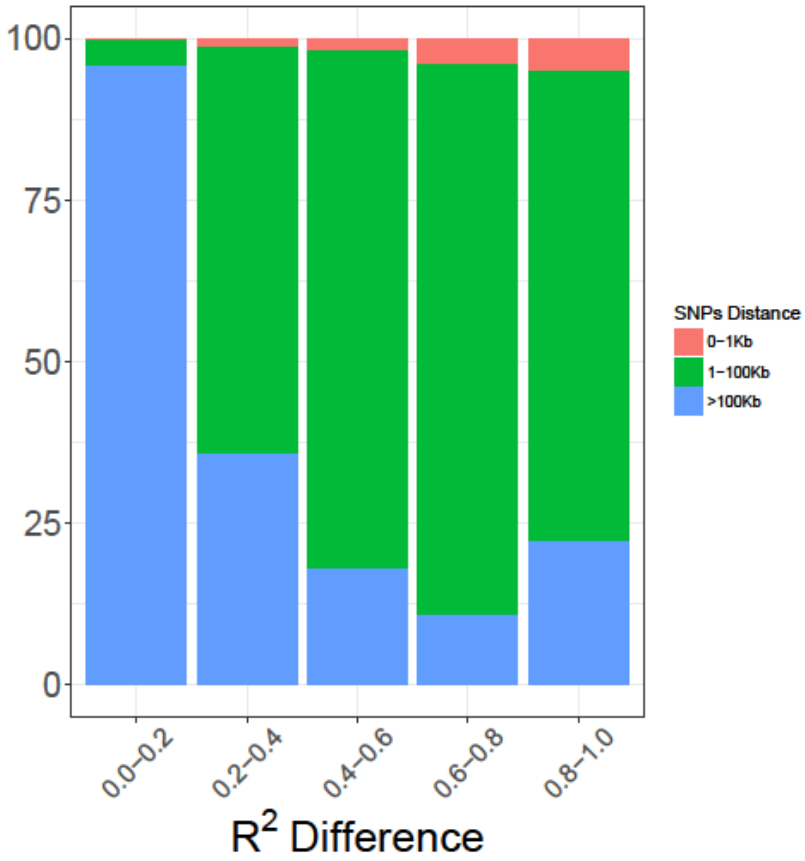

Figure S7: Distribution of LD difference between inverted and standard chromosomes and SNPs-distance in inv8p23.1. X-axis is the absolute difference between the LD of a SNP pair in inverted chromosomes versus standard chromosomes. SNP pairs are binned by their R2 difference and by the distance between SNPs. Columns show the distance distribution of pairs with similar LD difference.

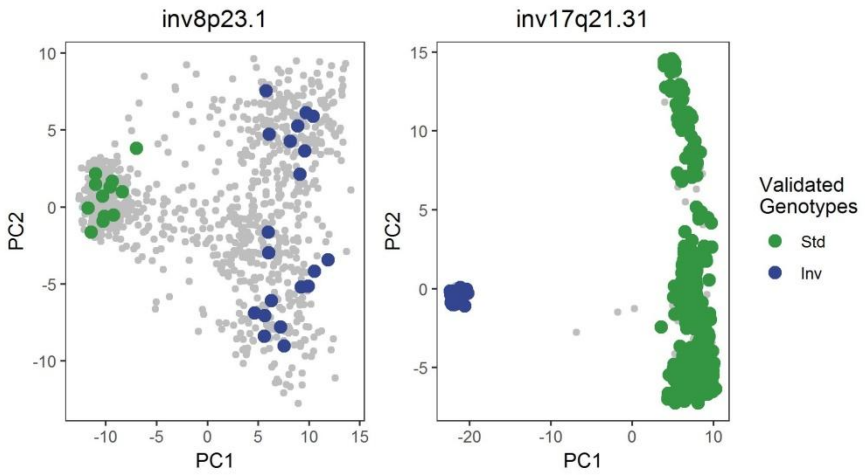

Figure S8: PCAs of chromosome probabilities for inv8p23.1 and inv17q21.31 in European samples of 1000 Genome Project. Chromosomes with known inversion genotype are colored (green: standard, blue: inverted).

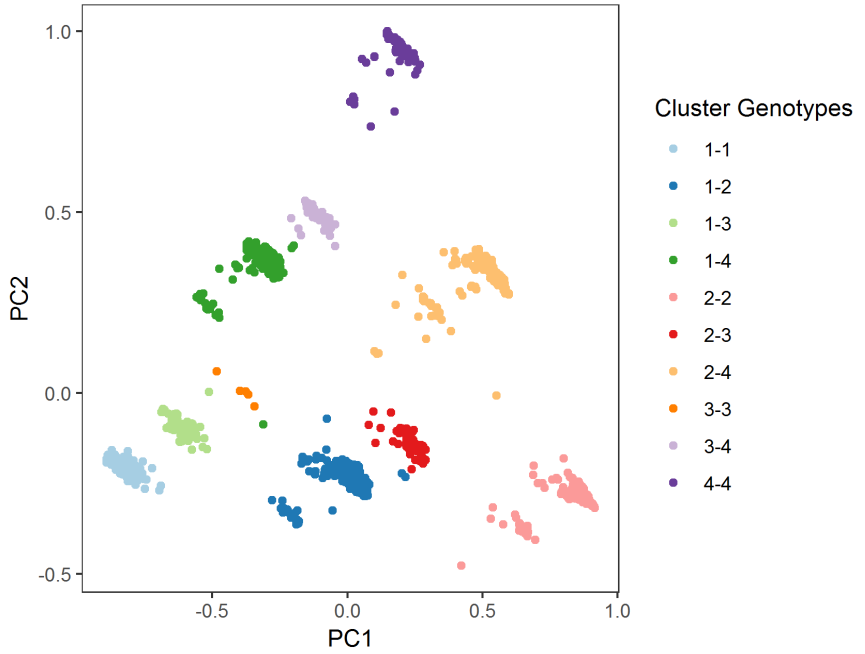

Figure S9: Genotypes' MDS in target region 1q21.1 from CGEMS study. Individuals are colored based on the cluster classification from *recombClust*.

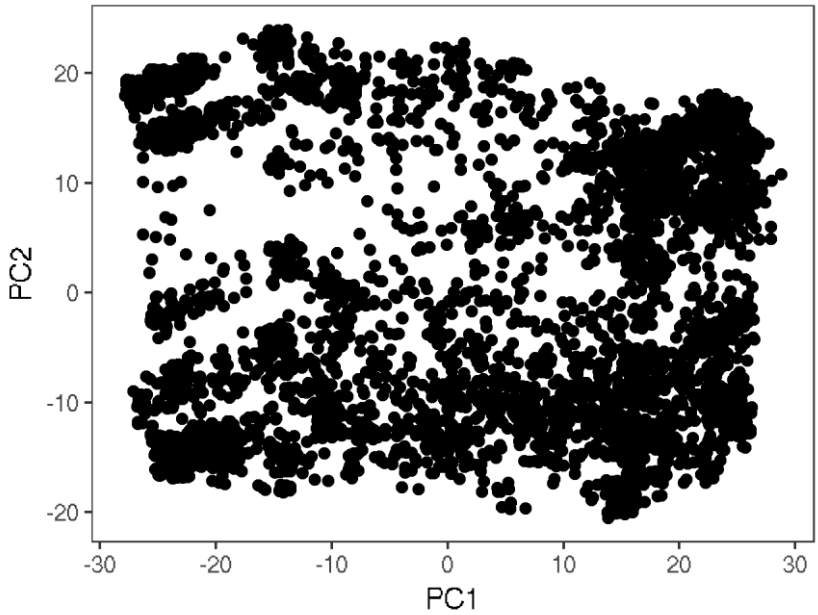

Figure S10: PCAs of chromosome probabilities in target region 10q26.13 in CGEMS study. Individuals are not colored as they do not form clear clusters.

Supplementary Tables

Table S1: Summary of models selected

		Selected Models	Total Model tested	Percentage
Simulation	No Pruning	248123	360527	68.82
	R2 > 0.6	16373	30039	54.51
	R2 > 0.4	9172	17384	52.76
	R2 > 0.2	2584	5555	46.52
European Samples	inv8p23.1	5118	20479	24.99
	inv17q21.31	1961	4664	42.05
All Samples	inv8p23.1	81465	555616	14.66
	inv17q21.31	3550	17904	19.83

Supplementary Note: Evaluation of LD-mixture model on simulated data

First, we evaluated the LD-mixture model performance in a general setting. LD-mixture model had higher specificity (mean SP: 0.85) and predictive positive value (mean PPV: 0.90) than sensitivity (mean SN: 0.64) and negative predictive value (mean NPV: 0.74) (Figure N1). Thus, LD-mixture model has few false positives: a chromosome detected as *recomb* is very likely to belong to *recomb* population. We also tested the effect of initializing the EM algorithm with two π values (π_0 = 0.05 and π_0 = 0.95) to account for converge problems. Low π_0 had higher SP (Δmean = 0.08, p-value < $2.2 \cdot 10^{-16}$) and PPV (Δmean = 0.04, p-value < $2.2 \cdot 10^{-16}$) and lower SN (Δmean = -0.07, p-value < $2.2 \cdot 10^{-16}$) and NPV (Δmean = -0.03, p-value < $2.2 \cdot 10^{-16}$) (Figure N1).

Second, we evaluated the performance of LD-mixture model using simulated datasets with different features. We investigated if the LD between the SNPs that compose the blocks (intrablock LD) affected the model's performance. Model's performance was penalized by the number of blocks in intermediate linkage (SP: -0.05, p-value < $2 \cdot 10^{-16}$; SN: -0.02, p-value < $2 \cdot 10^{-16}$; PPV: -0.03, p-value < $2 \cdot 10^{-16}$; NPV: -0.02, p-value < $2 \cdot 10^{-16}$) and the number of blocks in linkage (SP: -0.10, p-value < $2 \cdot 10^{-16}$; SN: -0.01, p-value < $2 \cdot 10^{-16}$; PPV: -0.07, p-value < $2 \cdot 10^{-16}$; NPV: -0.02, p-value < $2 \cdot 10^{-16}$). (Figure N2). Blocks in intermediate linkage or linkage contain less SNP combinations than blocks in *recomb*, so the LD-mixture model has less information to differentiate the mixture populations. We next tested how genetic

divergence between the populations affected the LD-mixture model performance (Figure N3). In general, median values of the SN, SP, NPV and PPV where higher than 0.93. However, simulations without genetic divergence had lower SN (median: 0.54) and NPV (median: 0.68) and simulations with the maximum divergence had lower SP (median: 0.85) and PPV (median: 0.86). A possible explanation is that when both populations have the same genetic content, the most frequent SNP-blocks for *recomb* and *linkage* population are the same. In this situation, LD-mixture model will assign chromosomes having these combinations of blocks to *linkge* population, increasing the number of false negatives (chromosomes from *recomb* population classified as *linkage*) and reducing SN and NPV. Finally, we tested how the mixture proportions affected the LD-mixture model's performance. Proportion of *recomb* population was positively correlated with SP ($r=0.56$, p-value $< 2 \cdot 10^{-16}$) and NPV ($r=0.47$, p-value $< 2 \cdot 10^{-16}$) and negatively correlated with SN ($r=-0.59$, p-value $< 2 \cdot 10^{-16}$) (Figure N4). Median PPV was independent of the proportion of *recomb* population and was higher than 0.9 in all simulations. These results also support that LD-mixture model has a low rate of chromosomes from *linkage* population classified as *recomb*, independent of mixture proportion.

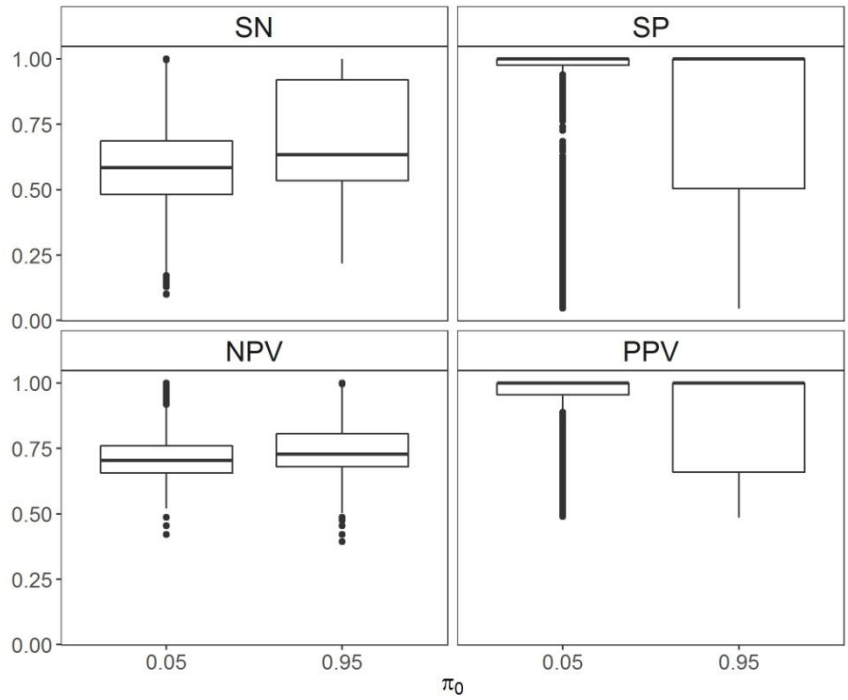

Figure N1: Model accuracy for different values of π_0. π_0 is the initial value of the mixture parameter π in the LD-mixture model. Each boxplot contains 5400 simulations. SN: sensitivity; SP: specificity; NPV: negative predictive value; PPV: positive predictive value.

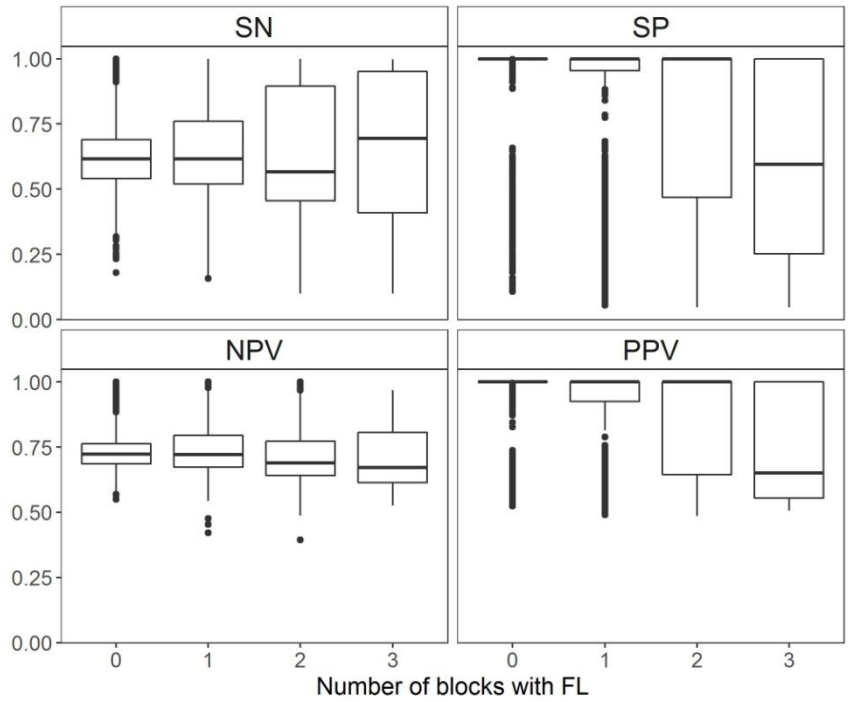

Figure N2: Model accuracy for different intrablock LD. X-axis is the number of blocks with linkage, i.e. one SNP can be mapped to the other. SN: sensitivity; SP: specificity; NPV: negative predictive value; PPV: positive predictive value.

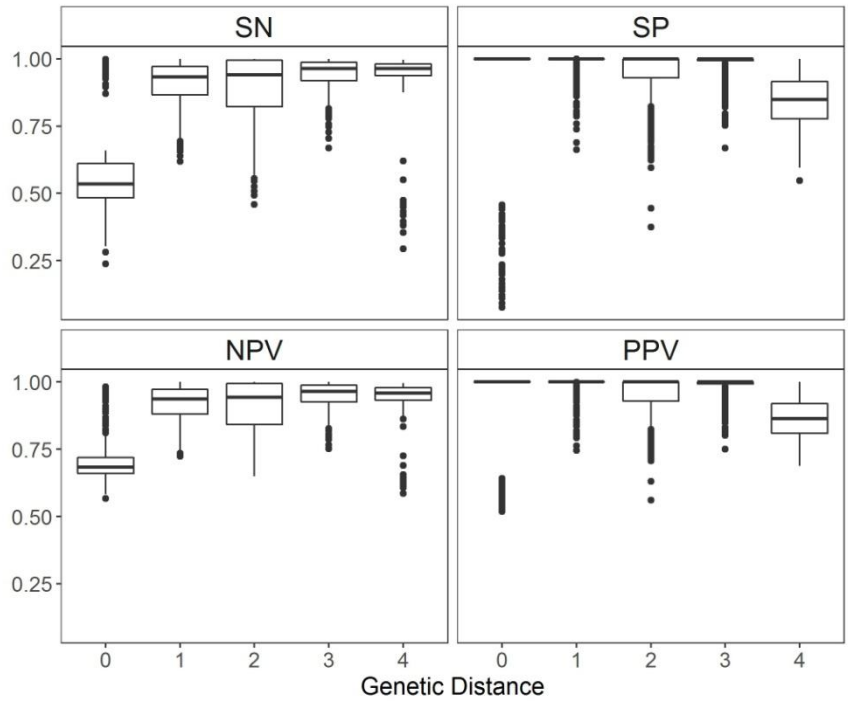

Figure N3: Model accuracy for different genetic distances between populations. Genetic distance is the number of SNPs having a different major allele between the populations. SN: sensitivity; SP: specificity; NPV: negative predictive value; PPV: positive predictive value.

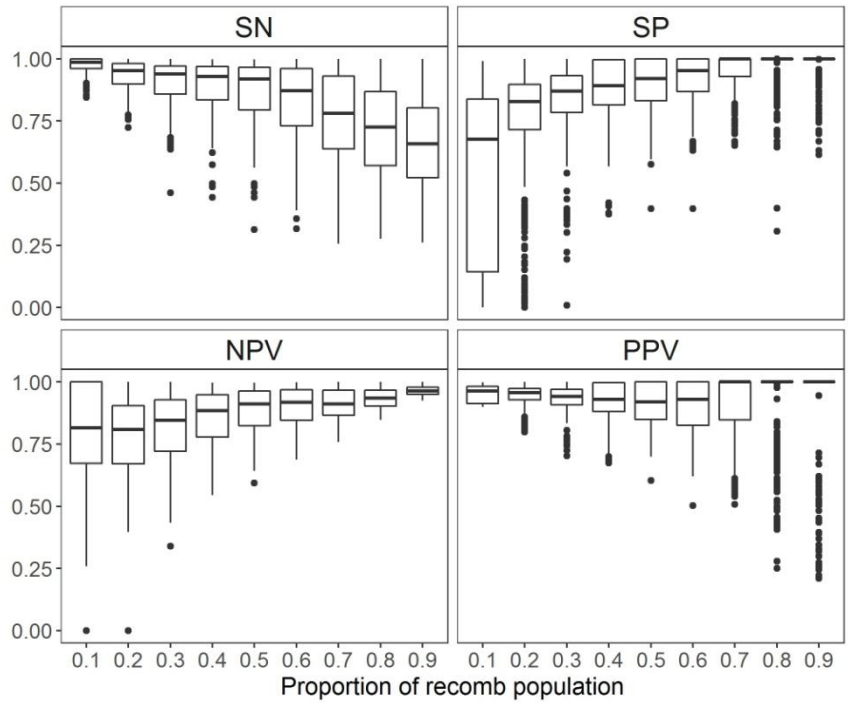

Figure N4: Models accuracy for difference in LD population proportion. X-axis is the proportion of chromosomes in the mixed dataset that belong to the *recomb* population. Each proportion was simulated 400 times. SN: sensitivity; SP: specificity; NPV: negative predictive value; PPV: positive predictive value.

9 General Discussion

In the following lines, I will discuss the main topics addressed in this doctoral thesis that includes: the new methods to study chromosomal inversions, the inversions studied in this thesis, the effects of those inversions in complex diseases and the role of inversions on recombination.

New methods to study chromosomal inversions

In chapters 5 and 6, I propose two bioinformatic methods to study the phenotypic and functional effects of chromosomal inversions. These methods contribute to two different fields: inversion association studies and regional omic analyses.

scoreInvHap framework

In chapter 5, I propose a new framework to run association studies with chromosomal inversions based on *scoreInvHap*. Two current approaches have been used to perform association studies: one fully bioinformatic and another including experimental data along with tag SNPs.

In the fully bioinformatic approach, the inversion genotypes are inferred using clustering methods, such as *invClust* [49], in the same dataset used to perform the association. Clusters are mapped to inversion status by frequency (i.e. the most frequent cluster is the standard) or by comparing the cluster haplotypes with the reference genome. Then, inferred inversion genotypes are used to perform the association study. The main limitation of this approach is the lack of a clear link between the haplotypes defined by the clusters and the inversion status. Thus, this approach does not guarantee that the inferred inversion genotypes correlate with the true inversion status, as discussed in an association study of inversions and psoriasis [112] or in the association study of region 15q24.2 [203]. In addition, this approach has been applied in

regions where an inversion was only predicted by bioinformatic methods, such as *inveRsion* [54] or the method of Ma and Amos [48], but not confirmed with any direct method. Consequently, the correlation between the cluster classification and inversion status is more unclear.

The second approach has three main steps. First, inversion genotypes are inferred in a discovery dataset using direct methods such as PCR. Second, inversion genotypes are used to find tag SNPs. Third, tag SNPs are used to infer inversion genotypes in a big dataset to perform association studies. This approach was used to study the human inversion inv19p12 [44]. The main limitation of this approach is the use of tag SNPs to perform association studies as: (1) inversions with multiple haplotypes might not have tag SNPs; (2) tag SNPs are specific of a population [204], potentially leading to wrong classifications; (3) tag SNPs are not included in many microarrays, preventing the reutilization of GWAS data [204].

I propose a new framework to genotype inversions based on *scoreInvHap* by combining the strengths of both approaches. In *scoreInvHap* framework, inversion genotypes are obtained from direct methods, such as PCR or sequencing. Then, we use the inversion-haplotype model to define the haplotypes that map to the inversion status. The inversion-haplotype model helps us building *scoreInvHap* references to infer inversion genotypes in other datasets. Consequently, *scoreInvHap* incorporates the strengths of both methods: association between haplotypes and true inversion status in a discovery dataset along with high-throughput inversion genotyping using all inversion SNP.

Although the inversion-haplotype model is not essential to build *scoreInvHap* references, it plays a key role in the *scoreInvHap* framework. The inversion-haplotype model determines whether an inversion contains differentiated haplotypes. Thus, the inversion-haplotype model

can prioritize those inversions that could be genotyped using indirect methods, to focus the genotyping using direct methods. Experimental inversion genotypes will further discard those inversions do not comply with the inversion-haplotype model. The inversion-haplotype model helped us discarding some inversions described by Antonnaci and colleagues [2] that have experimental inversion genotypes but not a clear haplotype structure. Nonetheless, the inversion-haplotype model considers haplotypes ranging the whole inversion and may miss shorter haplotypes. For instance, the inversion-haplotype model did not detect the previously described subhaplotypes of inv8p23.1 [19] and inv17q21.31 [205], which were defined using few SNPs. The inversion-haplotype model might also discard inversions that can be inferred using tag SNPs but that do not contain extended haplotypes.

scoreInvHap limitations come from its dependence on inversion haplotypes. Thus, *scoreInvHap* can only classify polymorphic and non-recurrent chromosomal inversions. Rare inversions do not produce enough changes in the genetic sequence and recurrent inversions do not have haplotypes specific of an inversion status. *scoreInvHap* also works better for long inversions, as they contain more SNPs that form distinct haplotypes. Finally, *scoreInvHap* references can be limited to one population, requiring their recomputation before applying *scoreInvHap* to a different population. Note that a chromosomal inversion might only comply with the inversion-haplotype model in a given population, so a chromosomal inversion might only be inferred using *scoreInvHap* in a this population but not in others.

Current version of *scoreInvHap* includes references for 20 human inversions, which can be genotyped in any dataset from a European population. Nonetheless, *scoreInvHap* can be extended to genotype

more human inversions in more human populations. To this end, two different lines of research are proposed. First, *scoreInvHap* framework can be applied to the inversions reported in invFEST database [107] and by Sanders and colleagues [41] to increase the number of chromosomal inversions genotyped. Sanders and colleagues reported several polymorphic inversions longer than 10 Kb, which seem good candidates to contain differentiated haplotypes. Second, *scoreInvHap* framework should be applied to our set of 20 inversions in populations other than European, extending the study of chromosomal inversions to these populations. Finally, *scoreInvHap* framework can also be applied to other genomes, such as *Drosophila Melanogaster*, to foster the study of chromosomal inversions in other species.

In this work, I used the European individuals of 1000 Genomes project to check the inversion-haplotype model and to build the references. I highly recommend using 1000 Genomes project data to build *scoreInvHap* references for European population by two reasons. First, individuals from 1000 Genomes project have cell lines available in public repositories, which can be used to genotype the inversions using direct methods. Second, European individuals from 1000 Genomes project come from 5 different populations, so we expect that they represent most of the genetic variance in Europe. Indeed, 1000 Genomes has been used to define the genetic traits of European populations, so *scoreInvHap* references should also be applicable to any European population. However, none of the European populations from 1000 Genomes project comes from Eastern Europe. Genetic differences between western and eastern European populations have been previously reported [206], although they only explained 0.15% of the total genetic variance. Thus, we expect that observed genetic differences between European

individuals do not compromise our *scoreInvHap* references in East Europe populations, but further research is needed to confirm this end.

New applications of *scoreInvHap* framework

scoreInvHap enables new strategies to associate chromosomal inversions to diseases. First, *scoreInvHap* allows reanalyzing most genetic datasets from EGA and dbGAP. Large inversions can be directly genotyped from WGS (Whole Genome Sequencing), WES (Whole Exome Sequencing) or SNP array data, while short inversions can be genotyped from WGS and from after imputation of SNP array data. The reanalysis of these datasets dramatically reduces the multiple tests performed, from millions of SNPs to 20 inversions, increasing the statistical power to find new associations. Second, *scoreInvHap* framework facilitates analyzing datasets with partial access, as *scoreInvHap* only requires few thousand of SNPs to genotype our 20 inversions. Thus, *scoreInvHap* reduces the amount of data shared and eases collaborations with other cohorts. For instance, UK Biobank is a big prospective cohort from United Kingdom [207] that contains 500,000 individuals with genetic and exposure data. Although researchers can access to UK Biobank genetic data under request, they can only access few thousand SNPs at a time, which is enough for *scoreInvHap* to genotype all our inversions.

Third, *scoreInvHap* enables the incorporation of chromosomal inversions in current collaborative meta-analysis in genetic epidemiology. Recently, different consortia of genetic cohorts have been established. Two examples are the Early Growth Genetics (EGG) consortium [208] and the EArly Genetics and Lifecourse Epidemiology (EAGLE) consortium [209], consortia that aim to study the effects of genetic variants on growth and childhood, respectively. In these consortia, an analyst runs the analysis in each cohort and sends the results to a leading group, who meta-analyze

the results from the cohorts. These cohorts usually have genetic data from SNP array, which is imputed to homogenize the cohorts. These meta-analyses largely increase the sample size so they have more power to find new associations. *scoreInvHap* has the potential to be included in these analyses, either alone or in parallel to SNP associations. In the latter, the simultaneous analysis of SNPs and chromosomal inversions help the interpretation of the results. An example of this approach is the association of inversions inv8p23.1 and inv17q21.31 with neuroticism [50]. An initial meta-analysis showed that regions 8p23.1 and 17q21.31 contained a high number of SNPs associated with neuroticism, while additional analyses showed that these associations were due to inv8p23.1 and inv17q21.31, helping to interpret the initial GWAS findings.

Fourth, an alternative to collaborative meta-analyses is DataSHIELD [210], a software infrastructure to run integrative analyses from different cohorts without sharing data. DataSHIELD infrastructure consists on cohort servers, with the cohort data, and a client server, which is accessed by the user to run the analyses. The client server asks the cohort servers to perform different computations. Cohort servers only send summarized data to the client server and never individual data, so the data does never leave the cohort server and the user can never access the private data. The client server integrates all the summarized data and returns the same result that would be obtained if analyzing all the data together, having more statistical power than a meta-analysis. DataSHIELD relies on implementing standard statistical algorithms using summarized data, so *scoreInvHap* could be adapted to be included in this infrastructure. To do so, cohort servers should contain imputed genotyped data and then run *scoreInvHap* in each private server. The

inversion genotypes will be temporary stored in the cohort server to run an association analysis with the desired outcome. Thus, only the results of the association will be returned to the client server, preserving the integrity of the data.

Finally, *scoreInvHap* can be used in a future to design microarrays to diagnose chromosomal inversions. As previously mentioned, *scoreInvHap* needs few thousand SNPs to genotype all our inversions, so a SNP microarray could contain all the SNPs needed to genotype all our inversions with *scoreInvHap*. SNP microarray could be a quick and easy way to screen inversions to detect individual susceptibility to different diseases. In a latter step, chromosomal inversions detected with *scoreInvHap* could be further validated using direct methods, to get a more robust diagnosis.

Regional Omic Analyses

In chapter 6, I described RDA, a new approach to get an overall estimate of the effect of chromosomal inversions in regional gene expression or DNA methylation. As RDA codes the inversion genotypes in a linear model, it can incorporate inversion genotypes from any approach and using any genetic model. RDA provides a standardized measure of the effect, R^2, so it enables comparing the effect between different inversions or between inversions and other factors, such as phenotypic variants (e.g. age or sex) or exposures (e.g. smoking). Finally, RDA also allows testing the combined effect of different variables at the same time and other complex models, such as interactions.

Chromosomal inversions studied in this thesis

The 20 chromosomal inversions included in this thesis are polymorphic and non-recurrent. I obtained the experimental genotypes needed to

build the *scoreInvHap* references from invFEST and 1000 Genomes project. As a result, most inversions are shorter than 10Kb. However, our inversions might not be representative of the human polymorphic inversions. For instance, Giner-Delgado studied 45 polymorphic inversions from invFEST using PCR [204]. Although her inversions had similar length than our subset, she found that most inversions were recurrent so the haplotypes were not specific of an inversion status.

Nonetheless, our 20 inversions are good candidates to explain genetic effects in complex diseases. Non-recurrent chromosomal inversions preserve the allelic combinations, which are associated with adaptation. Thus, non-recurrent chromosomal inversions are potentially associated with adaptation. In addition, genetic variants involved in adaptation are likely to be involved in complex diseases. For instance, components of type 2 immune response pathway are subjected to positive selection in primates but cause asthma in humans [211]. Our chromosomal inversions might also participate in complex diseases through a similar mechanism.

In our subset, most chromosomal inversions generate more than two haplotypes. Several hypotheses can be proposed to explain this phenomenon. First, standard or inverted chromosomes might contain additional structural variants, which would block recombination between chromosomes with the same inversion status. This hypothesis was proposed to explain the three haplotypes of inv16p11.2 [51]. Second, inversions with multiple haplotypes might be complex inversions, where a big inversion event is followed by a short internal inversion event. Thus, we will have three types of chromosomes: fully standard, fully inverted and half standard and half inverted. This hypothesis was proposed to explain the rearrangements in 22q11.2 [212]. Finally, short inversions

might lack recombination points. The average distance between recombination hotspots in the human genome is 122 Kb [117], which is larger than most of our inversions. Thus, inversions shorter than this distance might not have recombination points and the mutations history alone can produce different haplotypes [117]. Further research is needed to elucidate the mechanism generating the different haplotypes in our inversions, which will help to better understand the evolution and structure of our chromosomal inversions.

New possible effects of chromosomal inversions

In this thesis, I have run different association studies with chromosomal inversions. On one hand, I have replicated an association between chromosomal inversions and schizophrenia [87] and I have found new possible associations with breast cancer and cancer prognosis. Some strengths of these studies are the inclusion of exome sequencing data to replicate the results from GWAS and the simultaneous evaluation of 20 chromosomal inversions. On the other hand, I was unable to replicate a previous association of chromosomal inversions with autism [87]. A possible reason is that the studies were run in different populations (UK vs USA).

I tested four different genetic models in our association studies: additive, dominant, recessive and overdominant. Interestingly, I only found some associations when coding the inversions using a genetic model other than the additive. Our approach contrasts with previous studies with chromosomal inversions, which only tested the additive model, suggesting that some effects of chromosomal inversions might have been underreported. One exception is the association between inv17q21.31 and taoupathies by Li and colleagues [42], who tested the dominant and

the recessive models. Our chromosomal inversions can affect phenotypes through all the four genetic models. On one hand, our chromosomal inversions have different alleles between inverted and standard chromosomes, which justifies the use of additive, dominant and recessive genetic models (models commonly evaluated in SNPs) to characterize chromosomal inversions effects. On the other hand, chromosomal inversions are structural variants and they can also affect phenotypes through rearrangements in heterozygous individuals. This mechanism justifies the use of the overdominant model, where heterozygous individuals have different risk than the homozygous. Although there is some evidence supporting this model, such as the association between inversion and ichtyosis prognosis [24] or between inversions and cancer prognosis in chapter 7, further studies are needed to confirm biological mechanism supporting these observations. Finally, the low number of inversions evaluated allowed us testing the four different models, which might not be feasible in a typical GWAS assessing millions of variants.

A limitation of our studies is that most associations were only reported in a single population. Data from TCGA and GTEx comes from North American individuals, while data from autism, schizophrenia and breast cancer came from the UK. Although I restricted the associations to individuals presenting European ancestry, we do not know whether these associations are valid in other European populations and, in particular, in Spanish individuals. For instance, I observed a lower effect of inv17q21.31 in colorectal progression in the Spanish cohort compared to TCGA, suggesting a different effect of the inversion in different populations. Therefore, further replication of the associations reported in this thesis is required. Another limitation of our studies is the reduced

information about individual features (e.g. socio-economical status) or environmental exposures (e.g. smoking or diet). Although these variables are likely to influence the associations between chromosomal inversions and diseases, I could not include them in our models.

scoreInvHap can return the subhaplotypes in chromosomal inversions with multiple haplotypes, but I run all the associations using the inversion genotypes. By using inversion genotypes, I included positional effects in the association and simplified the analyses, as only one allele is considered at a time. In addition, inversion haplotypes are more likely to be tagged by a SNP than inversion status, so associations between haplotypes and diseases might have been captured by previous GWAS studies. However, specific haplotypes can also cause phenotypic effects, as shown by the association between subhaplotypes of inv17q21.31 and Parkinson [89]. Therefore, in new studies, inversion haplotypes should also be considered to evaluate whether the haplotypes or the inversion cause the phenotypic effects.

I report in this work an association between inv17q21.31 and changes in gene expression and DNA methylation in other chromosomes. Although there is no a clear biological mechanism to explain this effect, other studies have found effects of inv17q21.31 in other chromosomes. inv17q21.31 was observed to affect recombination in other chromosomes [173], and Li and colleagues also found changes of DNA methylation and gene expression in other chromosomes [42]. However, the genes and CpGs modified by the inversion in our study are not consistent with those found by Li and colleagues. These discrepancies can be explained by the differences between the studies: we are using a different population (colorectal patients vs healthy subjects), a different tissue (colorectal tumor tissue vs peripheral blood) and a different

genetic model (overdominant vs dominant and recessive). All in all, bigger samples sizes and more studies are required to confirm the effects of inv17q21.31 in other chromosomes.

recombClust

In chapter 8, I presented *recombClust*, a new method to partition chromosomes by differences in recombination patterns. I applied *recombClust* to chromosomal inversions in simulations and in real data and I found that standard and inverted chromosomes can be differentiated using recombination patterns. These results confirm and extend the previous observation of Alves and colleagues, who showed that individuals from different ancestries but with the same inversion status for inv8p23.1 had similar recombination patterns [20].

As *scoreInvHap*, *recombClust* is limited to polymorphic and non-recurrent inversions. In addition, *recombClust* relies on detecting recombination points, so it only works for inversions larger than 100Kb. All in all, *recombClust* is not well suited for chromosomal inversions classification, as it is only able to classify a small subset of inversions. However, chapter 8 shows that analysis of recombination patterns might lead to different than results analyses based exclusively on genotypes. Consequently, a combination of both approaches might give new insights into inversion history and structure.

Finally, I observed that *recombClust* can detect chromosomal sub-populations not caused by chromosomal inversions. These subpopulations might be produced by other genetic elements (e.g. translocations) and might be linked to adaptation or selection. Sub-populations include different alleles and different combinations of alleles, so they can potentially explain better phenotypic traits than

single SNPs. In addition, some chromosomal sub-populations cannot be tagged by single SNPs, so their effects have not been reported in existing GWAS. All in all, further studies of chromosome sub-populations are needed to know how common they are and their influence in phenotypic traits.

10 Conclusions

The main conclusions of this thesis are outlined below:

- *scoreInvHap* is a new robust and scalable method to genotype chromosomal inversions from genotype data.
- *scoreInvHap* outperforms current methods to genotype chromosomal inversions:
    - *scoreInvHap* classification is easily harmonized between different datasets
    - *scoreInvHap* can be applied in cohorts with thousands of individuals
    - *scoreInvHap* allows running association studies to new inversions and including new data sources
- Redundancy analysis enables studying the overall effect of chromosomal inversions on regional DNA methylation and gene expression.
    - Redundancy analysis allows comparing the effect on DNA methylation and gene expression between different chromosomal inversions and between a chromosomal inversion and other factors.
- Combination of *scoreInvHap* and redundancy analysis allows discovering new functional effects of chromosomal inversions.
    - The inverted haplotype of inv17q21.31 protects from suffering schizophrenia.
    - Inv8p23.1 and inv17q21.31 are new candidate genetic risk factors of cancer prognosis.
    - The effect of inv17q21.31 on colorectal cancer prognosis might be mediated by DNA methylation.
- Chromosomal inversions generate different recombination patterns between standard and inverted chromosomes

- *recombClust* is a new method to partition a population of chromosomes in subpopulations of chromosomes with similar recombination patterns
    - *recombClust* successfully recovered inversion status from recombination patterns.
    - Chromosomal subpopulations detected by *recombClust* can have stronger phenotypic effects than individual SNPs.

List of abbreviations

AGP: Autism Genome Project

BBC: British Birth Cohort

BIC: Bayesian Information Criteria

bp: base pairs

BRCA: breast invasive carcinoma

CGEMS: Cancer Markers of Susceptibility

CNV: Copy Number Variant

CpG: CG pairs

CPMs: Counts Per Million

COAD: colon adenocarcinoma

CRG: Centre for Genomic Regulation

dbGaP: Database of Genotypes and Phenotypes

DEG: Differentially Expressed Genes

DMP: Differentially Methylated Probe

DMR: Differentially Methylated Region

DNAseq: DNA sequencing

EBI: European Bioinformatic Institute

EGA: European Genome Archive

EM: Expectation-Maximization

ER: Estrogen Receptor

eQTL: expression Quantative Trait Loci

FISH: fluorescence in situ hybridization

GDC: Genomic Data Commons

GTEx: The Genome-Tissue Expression project

GWAS: Genome Wide Association Studies

HER2: human epidermal growth factor receptor 2

iPCR: Inversion PCR

LD: linkage disequilibrium

LIHC: liver hepatocellular carcinoma

LOH: loss of heterozygosity

LUAD: lung adenocarcinoma

LUSC: lung squamous cell carcinoma

MAF: Minor Allele Frequency

MDS: MultiDimensional Scaling

miRNAseq: sequencing of miRNA

MMBIR: Microhomology-Mediated Break-Induced Replication Model

NAHR: Non-Allelic Homologous Recombination

NBD: National Blood Service

NCI: National Cancer Institute

NHEJ: Non-Homologous End Joining

NHGRI: National Human Genome Research Institute

NIH: National Institute of Health

NPV: negative predictive value

PCA: Principal Component Analysis

PCR: polymerase chain reaction

PPV: positive predictive value

R^2: R-squared

RDA: redundancy analysis

READ: rectum adenocarcinoma

SN: sensitivity

SP: specificity

STAD: stomach adenocarcinoma

RNAseq: sequencing of RNA

RPPA: Reverse-phase protein array

SD: segmental duplications

SNP: single nucleotide polymorphisms

SSC: Simon Simplex Collection

SVA: Surrogate Variable Analysis

TCGA: The Cancer Genome Atlas

VCF: variant call format

WES: Whole Exome Sequencing

WGS: Whole genome sequencing

References

1. Sturtevant AH: **A Case of Rearrangement of Genes in Drosophila.** *Proc Natl Acad Sci U S A* 1921, **7**:235–7.
2. Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE: **Characterization of six human disease-associated inversion polymorphisms**. *Hum Mol Genet* 2009, **18**:2555–2566.
3. Kehrer-Sawatzki H, Szamalek JM, Tänzer S, Platzer M, Hameister H: **Molecular characterization of the pericentric inversion of chimpanzee chromosome 11 homologous to human chromosome 9**. *Genomics* 2005, **85**:542–550.
4. Brooks SA, Lear TL, Adelson DL, Bailey E: **A chromosome inversion near the <i>KIT</i> gene and the Tobiano spotting pattern in horses**. *Cytogenet Genome Res* 2008, **119**:225–230.
5. Haase B, Jude R, Brooks SA, Leeb T: **An equine chromosome 3 inversion is associated with the tobiano spotting pattern in German horse breeds**. *Anim Genet* 2008, **39**:306–309.
6. Kenig B, Kurbalija Novičić Z, Patenković A, Stamenković-Radak M, Anđelković M: **Adaptive Role of Inversion Polymorphism of Drosophila subobscura in Lead Stressed Environment**. *PLoS One* 2015, **10**:e0131270.
7. Cirulli ET, Noor MAF: **Localization and Characterization of X Chromosome Inversion Breakpoints Separating Drosophila mojavensis and Drosophila arizonae**. *J Hered* 2007, **98**:111–114.
8. Matoke-Muhia D, Gimnig JE, Kamau L, Shililu J, Bayoh MN, Walker ED: **Decline in frequency of the 2La chromosomal inversion in Anopheles gambiae (s.s.) in Western Kenya: correlation with increase in ownership of insecticide-treated bed nets**. *Parasit Vectors* 2016, **9**:334.
9. Ma J, Gao S, Stiller J, Jiang Q-T, Lan X-J, Liu Y-X, Pu Z-E, Wang J, Wei Y, Zheng Y-L, Gustafson JP: **Identification of genes bordering breakpoints of the pericentric inversions on 2B, 4B, and 5A in bread wheat (*Triticum aestivum* L.)**. *Genome* 2015, **58**:385–390.
10. Ferreira AC, Dias R, de Sá MIC, Tenreiro R: **Whole-genome mapping reveals a large chromosomal inversion on Iberian Brucella suis biovar 2 strains**. *Vet Microbiol* 2016, **192**:220–225.
11. Pang AWC, Migita O, MacDonald JR, Feuk L, Scherer SW: **Mechanisms of Formation of Structural Variation in a Fully Sequenced Human Genome**. *Hum Mutat* 2013, **34**:345–354.
12. Wang D, Li S, Guo F, Ning K, Wang L: **Core-genome scaffold comparison reveals the prevalence that inversion events are associated with pairs of inverted repeats**. *BMC Genomics* 2017, **18**:268.

13. Ranz JM, Maurin D, Chan YS, von Grotthuss M, Hillier LW, Roote J, Ashburner M, Bergman CM: **Principles of genome evolution in the Drosophila melanogaster species group.** *PLoS Biol* 2007, **5**:e152.
14. Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X: **Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements**. *Hum Mol Genet* 2003, **12**:2201–2208.
15. Cartwright IM, Kato TA: **Role of various DNA repair pathways in chromosomal inversion formation in CHO mutants**. *Int J Radiat Biol* 2015, **91**:925–933.
16. Hastings PJ, Lupski JR, Rosenberg SM, Ira G: **Mechanisms of change in gene copy number**. *Nat Rev Genet* 2009, **10**:551–564.
17. Sasaki M, Lange J, Keeney S: **Genome destabilization by homologous recombination in the germ line**. *Nat Rev Mol Cell Biol* 2010, **11**:182–195.
18. Griffiths A, Gelbart W, Miller J, Lewontin R: **Chromosomal Rearrangements**. In *Modern Genetic Analysis*. New York: W. H. Freeman; 1999.
19. Salm MPA, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, Schadt EE, Cookson WO, Wierzbicki AS, Naoumova RP, Shoulders CC: **The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism**. *Genome Res* 2012, **22**:1144–1153.
20. Alves JM, Chikhi L, Amorim A, Lopes AM: **The 8p23 Inversion Polymorphism Determines Local Recombination Heterogeneity across Human Populations**. *Genome Biol Evol* 2014, **6**:921–930.
21. Partida-Pérez M, Domínguez MG, Neira VA, Figuera LE, Rivera H: **De novo inv(17)(p11.2q21.3) in an intellectually disabled girl: appraisal of 21 inv(17) constitutional instances.** *J Genet* 2012, **91**:241–4.
22. Gorello P, Nofrini V, Brandimarte L, Pierini V, Crescenzi B, Nozza F, Daniele G, Storlazzi CT, Di Giacomo D, Matteucci C, La Starza R, Mecucci C: **Inv(11)(p15q22)/NUP98-DDX10 fusion and isoforms in a new case of de novo acute myeloid leukemia**. *Cancer Genet* 2013, **206**:92–96.
23. Bosch N, Morell M, Ponsa I, Mercader JM, Armengol L, Estivill X: **Nucleotide, Cytogenetic and Expression Impact of the Human Chromosome 8p23.1 Inversion Polymorphism**. *PLoS One* 2009, **4**:e8269.
24. Nomura T, Suzuki S, Miyauchi T, Takeda M, Shinkuma S, Fujita Y, Nishie W, Akiyama M, Shimizu H: **Chromosomal inversions as a hidden disease-modifying factor for somatic recombination phenotypes.** *JCI insight* 2018, **3**.
25. Coulibaly MB, Pombi M, Caputo B, Nwakanma D, Jawara M, Konate L, Dia I, Fofana A, Kern M, Simard F, Conway DJ, Petrarca V, Torre A, Traoré S, Besansky NJ: **PCR-based karyotyping of Anopheles gambiae inversion 2Rj identifies the BAMAKO chromosomal form**. *Malar J* 2007, **6**:133.

26. BAGNALL RD, GIANNELLI F, GREEN PM: **Int22h-related inversions causing hemophilia A: a novel insight into their origin and a new more discriminant PCR test for their detection**. *J Thromb Haemost* 2006, **4**:591–598.
27. Aguado C, Gayà-Vidal M, Villatoro S, Oliva M, Izquierdo D, Giner-Delgado C, Montalvo V, García-González J, Martínez-Fundichely A, Capilla L, Ruiz-Herrera A, Estivill X, Puig M, Cáceres M: **Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence.** *PLoS Genet* 2014, **10**:e1004208.
28. Mariano DCB, Sousa T de J, Pereira FL, Aburjaile F, Barh D, Rocha F, Pinto AC, Hassan SS, Saraiva TDL, Dorella FA, de Carvalho AF, Leal CAG, Figueiredo HCP, Silva A, Ramos RTJ, Azevedo VAC: **Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002.** *BMC Genomics* 2016, **17**:315.
29. Shukla SK, Kislow J, Briska A, Henkhaus J, Dykes C: **Optical Mapping Reveals a Large Genetic Inversion between Two Methicillin-Resistant Staphylococcus aureus Strains**. *J Bacteriol* 2009, **191**:5717–5723.
30. **File:Optical mapping.jpg --- Wikimedia Commons{,} the free media repository**
[https://commons.wikimedia.org/w/index.php?title=File:Optical_mapping.jpg&oldid=173202957]
31. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO: **DELLY: structural variant discovery by integrated paired-end and split-read analysis**. *Bioinformatics* 2012, **28**:i333–i339.
32. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z: **Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads**. *Bioinformatics* 2009, **25**:2865–2871.
33. Bartenhagen C, Dugas M: **Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms.** *Brief Bioinform* 2015, **17**:bbv028-.
34. Coccé MC, Mardin BR, Bens S, Stütz AM, Lubieniecki F, Vater I, Korbel JO, Siebert R, Alonso CN, Gallego MS: **Identification of *ZCCHC8* as fusion partner of *ROS1* in a case of congenital glioblastoma multiforme with a t(6;12)(q21;q24.3)**. *Genes, Chromosom Cancer* 2016, **55**:677–687.
35. Jaratlerdsiri W, Chan EKF, Petersen DC, Yang C, Croucher PI, Bornman MSR, Sheth P, Hayes VM: **Next generation mapping reveals novel large genomic rearrangements in prostate cancer**. *Oncotarget* 2017, **8**:23588.
36. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, Konkel MK, Malhotra A,

Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Mu XJ, Alkan C, Antaki D, et al.: **An integrated map of structural variation in 2,504 human genomes.** *Nature* 2015, **526**:75–81.
37. Liu C, Ran X, Wang J, Li S, Liu J: **Detection of genomic structural variations in Guizhou indigenous pigs and the comparison with other breeds**. *PLoS One* 2018, **13**:e0194282.
38. Lu L, Chen J, Robb SMC, Okumoto Y, Stajich JE, Wessler SR: **Tracking the genome-wide outcomes of a transposable element burst over decades of amplification**. *Proc Natl Acad Sci* 2017, **114**:E10550–E10559.
39. Collins RL, Brand H, Redin CE, Hanscom C, Antolik C, Stone MR, Glessner JT, Mason T, Pregno G, Dorrani N, Mandrile G, Giachino D, Perrin D, Walsh C, Cipicchio M, Costello M, Stortchevoi A, An J-Y, Currall BB, Seabra CM, Ragavendran A, Margolin L, Martinez-Agosto JA, Lucente D, Levy B, Sanders SJ, Wapner RJ, Quintero-Rivera F, Kloosterman W, Talkowski ME: **Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome**. *Genome Biol* 2017, **18**:36.
40. Eslami Rasekh M, Chiatante G, Miroballo M, Tang J, Ventura M, Amemiya CT, Eichler EE, Antonacci F, Alkan C: **Discovery of large genomic inversions using long range information**. *BMC Genomics* 2017, **18**:65.
41. Sanders AD, Hills M, Porubský D, Guryev V, Falconer E, Lansdorp PM: **Characterizing polymorphic inversions in human genomes by single cell sequencing**. *Genome Res* 2016:gr.201160.115.
42. Li Y, Chen JA, Sears RL, Gao F, Klein ED, Karydas A, Geschwind MD, Rosen HJ, Boxer AL, Guo W, Pellegrini M, Horvath S, Miller BL, Geschwind DH, Coppola G: **An epigenetic signature in peripheral blood associated with the haplotype on 17q21.31, a risk factor for neurodegenerative tauopathy.** *PLoS Genet* 2014, **10**:e1004211.
43. Zabetian CP, Hutter CM, Factor SA, Nutt JG, Higgins DS, Griffith A, Roberts JW, Leis BC, Kay DM, Yearout D, Montimurro JS, Edwards KL, Samii A, Payami H: **Association analysis of MAPT H1 haplotype and subhaplotypes in Parkinson's disease**. *Ann Neurol* 2007, **62**:137–144.
44. Puig M, Castellano D, Pantano L, Giner-Delgado C, Izquierdo D, Gayà-Vidal M, Lucas-Lledó JI, Esko T, Terao C, Matsuda F, Cáceres M: **Functional Impact and Evolution of a Novel Human Polymorphic Inversion That Disrupts a Gene and Creates a Fusion Transcript.** *PLoS Genet* 2015, **11**:e1005495.
45. Rosendahl J, Kirsten H, Hegyi E, Kovacs P, Weiss FU, Laumen H, Lichtner P, Ruffert C, Chen J-M, Masson E, Beer S, Zimmer C, Seltsam K, Algül H, Bühler F, Bruno MJ, Bugert P, Burkhardt R, Cavestro GM, Cichoz-Lach H, Farré A, Frank J, Gambaro G, Gimpfl S, Grallert H, Griesmann H,

Grützmann R, Hellerbrand C, Hegyi P, Hollenbach M, et al.: **Genome-wide association study identifies inversion in the *CTRB1-CTRB2* locus to modify risk for alcoholic and non-alcoholic chronic pancreatitis**. *Gut* 2018, **67**:1855–1863.
46. Kapun M, van Schalkwyk H, McAllister B, Flatt T, Schlötterer C: **Inference of chromosomal inversion dynamics from Pool-Seq data in natural and laboratory populations of *Drosophila melanogaster***. *Mol Ecol* 2014, **23**:1813–1827.
47. Kapun M, Fabian DK, Goudet J, Flatt T: **Genomic Evidence for Adaptive Inversion Clines in *Drosophila melanogaster***. *Mol Biol Evol* 2016, **33**:1317–1336.
48. Ma J, Amos CI: **Investigation of inversion polymorphisms in the human genome using principal components analysis.** *PLoS One* 2012, **7**:e40224.
49. Cáceres A, González JR: **Following the footprints of polymorphic inversions on SNP data: from detection to association tests.** *Nucleic Acids Res* 2015:1–11.
50. Okbay A, Baselmans BML, De Neve J-E, Turley P, Nivard MG, Fontana MA, Meddens SFW, Linnér RK, Rietveld CA, Derringer J, Gratten J, Lee JJ, Liu JZ, de Vlaming R, Ahluwalia TS, Buchwald J, Cavadino A, Frazier-Wood AC, Furlotte NA, Garfield V, Geisel MH, Gonzalez JR, Haitjema S, Karlsson R, van der Laan SW, Ladwig K-H, Lahti J, van der Lee SJ, Lind PA, Liu T, et al.: **Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses.** *Nat Genet* 2016, **48**:624–33.
51. González JR, Cáceres A, Esko T, Cuscó I, Puig M, Esnaola M, Reina J, Siroux V, Bouzigon E, Nadif R, Reinmaa E, Milani L, Bustamante M, Jarvis D, Antó JM, Sunyer J, Demenais F, Kogevinas M, Metspalu A, Cáceres M, Pérez-Jurado LA: **A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity.** *Am J Hum Genet* 2014, **94**:361–72.
52. Knief U, Hemmrich-Stanisak G, Wittig M, Franke A, Griffith SC, Kempenaers B, Forstmeier W: **Fitness consequences of polymorphic inversions in the zebra finch genome**. *Genome Biol* 2016, **17**:199.
53. Sindi SS, Raphael BJ: **Identification and Frequency Estimation of Inversion**. *J Comput Biol* 2010, **17**:517–531.
54. Cáceres A, Sindi SS, Raphael BJ, Cáceres M, González JR: **Identification of polymorphic inversions from genotypes**. *BMC Bioinformatics* 2012, **13**:28.
55. Berg PR, Star B, Pampoulie C, Sodeland M, Barth JMI, Knutsen H, Jakobsen KS, Jentoft S: **Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod**. *Sci Rep* 2016, **6**:23246.

56. Sodeland M, Jorde PE, Lien S, Jentoft S, Berg PR, Grove H, Kent MP, Arnyasi M, Olsen EM, Knutsen H: **"Islands of Divergence" in the Atlantic Cod Genome Represent Polymorphic Chromosomal Rearrangements.** *Genome Biol Evol* 2016, **8**:1012–22.
57. Berg PR, Star B, Pampoulie C, Bradbury IR, Bentzen P, Hutchings JA, Jentoft S, Jakobsen KS: **Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions**. *Heredity (Edinb)* 2017, **119**:418–428.
58. Kirkpatrick M: **How and Why Chromosome Inversions Evolve**. *PLoS Biol* 2010, **8**:e1000501.
59. Kirkpatrick M: **The Evolution of Genome Structure by Natural and Sexual Selection**. *J Hered* 2017, **108**:3–11.
60. Christmas MJ, Wallberg A, Bunikis I, Olsson A, Wallerman O, Webster MT: **Chromosomal inversions associated with environmental adaptation in honeybees**. *Mol Ecol* 2018.
61. Gould BA, Chen Y, Lowry DB: **Pooled ecotype sequencing reveals candidate genetic mechanisms for adaptive differentiation and reproductive isolation**. *Mol Ecol* 2017, **26**:163–177.
62. Dagilis AJ, Kirkpatrick M: **Prezygotic isolation, mating preferences, and the evolution of chromosomal inversions**. *Evolution (N Y)* 2016, **70**:1465–1472.
63. Fuller ZL, Leonard CJ, Young RE, Schaeffer SW, Phadnis N: **Ancestral polymorphisms explain the role of chromosomal inversions in speciation**. *PLOS Genet* 2018, **14**:e1007526.
64. Lee C-R, Wang B, Mojica JP, Mandáková T, Prasad KVSK, Goicoechea JL, Perera N, Hellsten U, Hundley HN, Johnson J, Grimwood J, Barry K, Fairclough S, Jenkins JW, Yu Y, Kudrna D, Zhang J, Talag J, Golser W, Ghattas K, Schranz ME, Wing R, Lysak MA, Schmutz J, Rokhsar DS, Mitchell-Olds T: **Young inversion with multiple linked QTLs under selection in a hybrid zone**. *Nat Ecol Evol* 2017, **1**:0119.
65. Lahn BT, Page DC: **Four evolutionary strata on the human X chromosome.** *Science* 1999, **286**:964–7.
66. Thu Tran TH, Zhang Z, Yagi M, Lee T, Awano H, Nishida A, Okinaga T, Takeshima Y, Matsuo M: **Molecular characterization of an X(p21.2;q28) chromosomal inversion in a Duchenne muscular dystrophy patient with mental retardation reveals a novel long non-coding gene on Xq28**. *J Hum Genet* 2013, **58**:33–39.
67. Jones ML, Murden SL, Brooks C, Maloney V, Manning RA, Gilmour KC, Bharadwaj V, de la Fuente J, Chakravorty S, Mumford AD: **Disruption of AP3B1by a chromosome 5 inversion: a new disease mechanism in Hermansky-Pudlak syndrome type 2**. *BMC Med Genet* 2013, **14**:42.
68. Xin Y, Zhou J, Ding Q, Chen C, Wu X, Wang X, Wang H, Jiang X: **A**

pericentric inversion of chromosome X disrupting *F8* and resulting in haemophilia A. *J Clin Pathol* 2017, **70**:656–661.
69. Yokoi K, Nakajima Y, Ohye T, Inagaki H, Wada Y, Fukuda T, Sugie H, Yuasa I, Ito T, Kurahashi H: **Disruption of the Responsible Gene in a Phosphoglucomutase 1 Deficiency Patient by Homozygous Chromosomal Inversion**. In *JIMD reports*; 2018.
70. Rigola MA, Baena N, Català V, Lozano I, Gabau E, Guitart M, Fuster C: **A 11.7-Mb paracentric inversion in chromosome 1q detected in prenatal diagnosis associated with familial intellectual disability**. *Cytogenet Genome Res* 2015, **146**:109–114.
71. Yamazaki H, Suzuki M, Otsuki A, Shimizu R, Bresnick EH, Engel JD, Yamamoto M: **A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression.** *Cancer Cell* 2014, **25**.
72. Rhees J, Arnold M, Boland CR: **Inversion of exons 1-7 of the MSH2 gene is a frequent cause of unexplained Lynch syndrome in one local population.** *Fam Cancer* 2014, **13**:219–25.
73. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, Bando M, Ohno S, Ishikawa Y, Aburatani H, Niki T, Sohara Y, Sugiyama Y, Mano H: **Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer**. *Nature* 2007, **448**:561–566.
74. Argani P, Zhang L, Reuter VE, Tickoo SK, Antonescu CR: **RBM10-TFE3 Renal Cell Carcinoma**. *Am J Surg Pathol* 2017, **41**:655–662.
75. Agnihotri S, Jalali S, Wilson MR, Danesh A, Li M, Klironomos G, Krieger JR, Mansouri A, Khan O, Mamatjan Y, Landon-Brace N, Tung T, Dowar M, Li T, Bruce JP, Burrell KE, Tonge PD, Alamsahebpour A, Krischek B, Agarwalla PK, Bi WL, Dunn IF, Beroukhim R, Fehlings MG, Bril V, Pagnotta SM, Iavarone A, Pugh TJ, Aldape KD, Zadeh G: **The genomic landscape of schwannoma**. *Nat Genet* 2016, **48**:1339–1348.
76. Fraser M, Sabelnykova VY, Yamaguchi TN, Heisler LE, Livingstone J, Huang V, Shiah Y-J, Yousif F, Lin X, Masella AP, Fox NS, Xie M, Prokopec SD, Berlin A, Lalonde E, Ahmed M, Trudel D, Luo X, Beck TA, Meng A, Zhang J, D'Costa A, Denroche RE, Kong H, Espiritu SMG, Chua MLK, Wong A, Chong T, Sam M, Johns J, et al.: **Genomic hallmarks of localized, non-indolent prostate cancer**. *Nature* 2017, **541**:359–364.
77. Gruber TA, Larson Gedman A, Zhang J, Koss CS, Marada S, Ta HQ, Chen S-C, Su X, Ogden SK, Dang J, Wu G, Gupta V, Andersson AK, Pounds S, Shi L, Easton J, Barbato MI, Mulder HL, Manne J, Wang J, Rusch M, Ranade S, Ganti R, Parker M, Ma J, Radtke I, Ding L, Cazzaniga G, Biondi A, Kornblau SM, et al.: **An Inv(16)(p13.3q24.3)-Encoded CBFA2T3-GLIS2 Fusion Protein Defines an Aggressive Subtype of Pediatric Acute**

Megakaryoblastic Leukemia. *Cancer Cell* 2012, **22**:683–697.
78. Pulikkan JA, Castilla LH: **Preleukemia and Leukemia-Initiating Cell Activity in inv(16) Acute Myeloid Leukemia**. *Front Oncol* 2018, **8**:129.
79. Koolen DA, Vissers LELM, Pfundt R, de Leeuw N, Knight SJ, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M, Schinzel A, Baumer A, Anderlid B-M, Schoumans J, Knoers N V, van Kessel AG, Sistermans EA, Veltman JA, Brunner HG, de Vries BBA: **A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism**. *Nat Genet* 2006, **38**:999–1001.
80. Bayés M, Magano LF, Rivera N, Flores R, Pérez Jurado LA: **Mutational mechanisms of Williams-Beuren syndrome deletions.** *Am J Hum Genet* 2003, **73**:131–51.
81. Hobart HH, Morris CA, Mervis CB, Pani AM, Kistler DJ, Rios CM, Kimberley KW, Gregg RG, Bray-Ward P: **Inversion of the Williams syndrome region is a common polymorphism found more frequently in parents of children with Williams syndrome.** *Am J Med Genet C Semin Med Genet* 2010, **154C**:220–8.
82. Gimelli G, Pujana MA, Patricelli MG, Russo S, Giardino D, Larizza L, Cheung J, Armengol L, Schinzel A, Estivill X, Zuffardi O: **Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions**. *Hum Mol Genet* 2003, **12**:849–858.
83. Visser R, Shimokawa O, Harada N, Kinoshita A, Ohta T, Niikawa N, Matsumoto N: **Identification of a 3.0-kb Major Recombination Hotspot in Patients with Sotos Syndrome Who Carry a Common 1.9-Mb Microdeletion**. *Am J Hum Genet* 2005, **76**:52–67.
84. Mohajeri K, Cantsilieris S, Huddleston J, Nelson BJ, Coe BP, Campbell CD, Baker C, Harshman L, Munson KM, Kronenberg ZN, Kremitzki M, Raja A, Catacchio CR, Graves TA, Wilson RK, Ventura M, Eichler EE: **Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region**. *Genome Res* 2016.
85. Namjou B, Ni Y, Harley ITW, Chepelev I, Cobb B, Kottyan LC, Gaffney PM, Guthridge JM, Kaufman K, Harley JB: **The effect of inversion at 8p23 on BLK association with lupus in Caucasian population.** *PLoS One* 2014, **9**:e115614.
86. Demirci FY, Wang X, Morris DL, Feingold E, Bernatsky S, Pineau C, Clarke A, Ramsey-Goldman R, Manzi S, Vyse TJ, Ilyas Kamboh M: **Multiple signals at the extended 8p23 locus are associated with susceptibility to systemic lupus erythematosus**. *J Med Genet* 2017, **54**:381–389.
87. Gutiérrez Arumi A: **Ancestral genomic submicroscopic inversions of human genome and their relation with multifactorial human diseases**.

Univ Pompeu Fabra 2015.
88. Vandrovcova J, Pittman AM, Malzer E, Abou-Sleiman PM, Lees AJ, Wood NW, de Silva R: **Association of MAPT haplotype-tagging SNPs with sporadic Parkinson's disease**. *Neurobiol Aging* 2009, **30**:1477–1482.
89. Tobin JE, Latourelle JC, Lew MF, Klein C, Suchowersky O, Shill HA, Golbe LI, Mark MH, Growdon JH, Wooten GF, Racette BA, Perlmutter JS, Watts R, Guttman M, Baker KB, Goldwurm S, Pezzoli G, Singer C, Saint-Hilaire MH, Hendricks AE, Williamson S, Nagle MW, Wilk JB, Massood T, Laramie JM, DeStefano AL, Litvan I, Nicholson G, Corbett A, Isaacson S, et al.: **Haplotypes and gene expression implicate the MAPT region for Parkinson disease: the GenePD Study.** *Neurology* 2008, **71**:28–34.
90. Setó-Salvia N, Clarimón J, Pagonabarraga J, Pascual-Sedano B, Campolongo A, Combarros O, Mateo JI, Regaña D, Martínez-Corral M, Marquié M, Alcolea D, Suárez-Calvet M, Molina-Porcel L, Dols O, Gómez-Isla T, Blesa R, Lleó A, Kulisevsky J: **Dementia Risk in Parkinson Disease**. *Arch Neurol* 2011, **68**:359–64.
91. Goris A, Williams-Gray CH, Clark GR, Foltynie T, Lewis SJG, Brown J, Ban M, Spillantini MG, Compston A, Burn DJ, Chinnery PF, Barker RA, Sawcer SJ: **Tau and α-synuclein in susceptibility to, and dementia in, Parkinson's disease**. *Ann Neurol* 2007, **62**:145–153.
92. Webb A, Miller B, Bonasera S, Boxer A, Karydas A, Wilhelmsen KC: **Role of the tau gene region chromosome inversion in progressive supranuclear palsy, corticobasal degeneration, and related disorders.** *Arch Neurol* 2008, **65**:1473–8.
93. Myers AJ, Kaleem M, Marlowe L, Pittman AM, Lees AJ, Fung HC, Duckworth J, Leung D, Gibson A, Morris CM, de Silva R, Hardy J: **The H1c haplotype at the MAPT locus is associated with Alzheimer's disease**. *Hum Mol Genet* 2005, **14**:2399–2404.
94. Tantisira KG, Lazarus R, Litonjua AA, Klanderman B, Weiss ST: **Chromosome 17: Association of a large inversion polymorphism with corticosteroid response in asthma**. *Pharmacogenet Genomics* 2008, **18**.
95. de Jong S, Chepelev I, Janson E, Strengman E, van den Berg LH, Veldink JH, Ophoff RA: **Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner.** *BMC Genomics* 2012, **13**:458.
96. Bekpen C, Tastekin I, Siswara P, Akdis CA, Eichler EE: **Primate segmental duplication creates novel promoters for the LRRC37 gene family within the 17q21.31 inversion polymorphism region**. *Genome Res* 2012, **22**:1050–1058.
97. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson J V, Holmans P, Heward CB,

Reiman EM, Stephan D, Hardy J: **A survey of genetic human cortical gene expression**. *Nat Genet* 2007, **39**:1494–1499.
98. Yuan T, Jiao Y, de Jong S, Ophoff RA, Beck S, Teschendorff AE, Maegawa S, Hinkal G, Kim H, Shen L, Zhang L, Teschendorff A, Menon U, Gentry-Maharaj A, Ramus S, Weisenberger D, Rakyan V, Down T, Maslau S, Andrew T, Yang T, Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Horvath S, Beerman I, Bock C, Garrison B, et al.: **An Integrative Multi-scale Analysis of the Dynamic DNA Methylation Landscape in Aging**. *PLOS Genet* 2015, **11**:e1004996.
99. Allen M, Kachadoorian M, Quicksall Z, Zou F, Chai H, Younkin C, Crook JE, Pankratz V, Carrasquillo MM, Krishnan S, Nguyen T, Ma L, Malphrus K, Lincoln S, Bisceglio G, Kolbert CP, Jen J, Mukherjee S, Kauwe JK, Crane PK, Haines JL, Mayeux R, Pericak-Vance MA, Farrer LA, Schellenberg GD, Parisi JE, Petersen RC, Graff-Radford NR, Dickson DW, Younkin SG, et al.: **Association of MAPT haplotypes with Alzheimer's disease risk and MAPT brain gene expression levels**. *Alzheimers Res Ther* 2014, **6**:39.
100. International Parkinson Disease Genomics Consortium, Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin U-M, Saad M, Simón-Sánchez J, Schulte C, Lesage S, Sveinbjörnsdóttir S, Stefánsson K, Martinez M, Hardy J, Heutink P, Brice A, Gasser T, Singleton AB, Wood NW: **Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies.** *Lancet (London, England)* 2011, **377**:641–9.
101. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Gibbs RA, Boerwinkle E, et al.: **A global reference for human genetic variation**. *Nature* 2015, **526**:68–74.
102. Tomczak K, Czerwińska P, Wiznerowicz M: **The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.** *Contemp Oncol (Poznan, Poland)* 2015, **19**:A68-77.
103. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, et al.: **The Genotype-Tissue Expression (GTEx) project**. *Nat Genet* 2013, **45**:580–585.
104. Kirkpatrick M, Barton N: **Chromosome inversions, local adaptation and speciation.** *Genetics* 2006, **173**:419–34.
105. Puig M, Casillas S, Villatoro S, Cáceres M: **Human inversions and**

their functional consequences. *Brief Funct Genomics* 2015, **14**:369–79.
106. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N, Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM, Sainz J, Agnarsson K, Birgisdottir B, Ghosh S, Olafsdottir A, Cazier J-B, Kristjansson K, Frigge ML, Thorgeirsson TE, Gulcher JR, Kong A, Stefansson K: **A common inversion under selection in Europeans**. *Nat Genet* 2005, **37**:129–137.
107. Martínez-Fundichely A, Casillas S, Egea R, Ràmia M, Barbadilla A, Pantano L, Puig M, Cáceres M: **InvFEST, a database integrating information of polymorphic inversions in the human genome**. *Nucleic Acids Res* 2014, **42**:D1027–D1032.
108. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061–73.
109. O'Reilly PF, Coin LJM, Hoggart CJ: **invertFREGENE: software for simulating inversions in population genetic data**. *Bioinformatics* 2010, **26**:838–840.
110. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF, Hoover RN, Thomas G, Chanock SJ: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer**. *Nat Genet* 2007, **39**:870–874.
111. Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC, Wang X, Ademuyiwa F, Ahmed S, Ambrosone CB, Baglietto L, Balleine R, Bandera E V, Beckmann MW, Berg CD, Bernstein L, Blomqvist C, Blot WJ, Brauch H, Buring JE, Carey LA, Carpenter JE, Chang-Claude J, Chanock SJ, Chasman DI, Clarke CL, Cox A, Cross SS, Deming SL, Diasio RB, et al.: **A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor–negative breast cancer**. *Nat Genet* 2011, **43**:1210–1214.
112. Ma J, Xiong M, You M, Lozano G, Amos CI: **Genome-wide association tests of inversions with application to psoriasis.** *Hum Genet* 2014, **133**:967–74.
113. Steinberg KM, Antonacci F, Sudmant PH, Kidd JM, Campbell CD, Vives L, Malig M, Scheinfeldt L, Beggs W, Ibrahim M, Lema G, Nyambo TB, Omar SA, Bodo J-M, Froment A, Donnelly MP, Kidd KK, Tishkoff SA, Eichler EE: **Structural diversity and African origin of the 17q21.31 inversion polymorphism.** *Nat Genet* 2012, **44**:872–80.
114. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu X-

Q, Vincent JB, Skaug JL, Thompson AP, Senman L, Feuk L, Qian C, Bryson SE, Jones MB, Marshall CR, Scherer SW, Vieland VJ, Bartlett C, Mangin LV, Goedken R, Segre A, Pericak-Vance MA, Cuccaro ML, Gilbert JR, Wright HH, Abramson RK, Betancur C, Bourgeron T, Gillberg C, Leboyer M, et al.: **Mapping autism risk loci using genetic linkage and chromosomal rearrangements**. *Nat Genet* 2007, **39**:319–328.
115. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, Murtha MT, Bal VH, Bishop SL, Dong S, Goldberg AP, Jinlu C, Keaney JF, Klei L, Mandell JD, Moreno-De-Luca D, Poultney CS, Robinson EB, Smith L, Solli-Nowlan T, Su MY, Teran NA, Walker MF, Werling DM, Beaudet AL, Cantor RM, Fombonne E, Geschwind DH, Grice DE, Lord C, et al.: **Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci**. *Neuron* 2015, **87**:1215–1233.
116. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF: **Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America**. *Hum Mutat* 2009, **30**:69–78.
117. International HapMap Consortium TIH: **A haplotype map of the human genome.** *Nature* 2005, **437**:1299–320.
118. Pedersen BS, Quinlan AR: **Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy.** *Am J Hum Genet* 2017, **100**:406–413.
119. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh P-R, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C: **Next-generation genotype imputation service and methods**. *Nat Genet* 2016, **48**:1284–1287.
120. UK IBD Genetics Consortium JC, Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, Wesley E, Parnell K, Zhang H, Drummond H, Nimmo ER, Massey D, Blaszczyk K, Elliott T, Cotterill L, Dallal H, Lobo AJ, Mowat C, Sanderson JD, Jewell DP, Newman WG, Edwards C, Ahmad T, Mansfield JC, Satsangi J, Parkes M, Mathew CG, Wellcome Trust Case Control Consortium 2 L, Donnelly P, et al.: **Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region.** *Nat Genet* 2009, **41**:1330–4.
121. Kulis M, Esteller M: **DNA Methylation and Cancer**. In *Advances in genetics*. *Volume 70*; 2010:27–56.
122. Hassler MR, Egger G: **Epigenomics of cancer - emerging new concepts.** *Biochimie* 2012, **94**:2219–30.
123. María Martín-Núñez G, Rubio-Martín E, Cabrera-Mulero R, Rojo-Martínez G, Olveira G, Valdés S, Soriguer F, Castaño L, Morcillo S: **Type 2**

diabetes mellitus in relation to global LINE-1 DNA methylation in peripheral blood: A cohort study. *Epigenetics* 2014, **9**:1322–1328.
124. Enquobahrie DA, Moore A, Muhie S, Tadesse MG, Lin S, Williams MA: **Early Pregnancy Maternal Blood DNA Methylation in Repeat Pregnancies and Change in Gestational Diabetes Mellitus Status—A Pilot Study**. *Reprod Sci* 2015, **22**:904–910.
125. Steenaard R V, Ligthart S, Stolk L, Peters MJ, van Meurs JB, Uitterlinden AG, Hofman A, Franco OH, Dehghan A: **Tobacco smoking is associated with methylation of genes related to coronary artery disease.** *Clin Epigenetics* 2015, **7**:54.
126. Simpkin AJ, Hemani G, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, Ring SM, Sharp GC, Tilling K, Horvath S, Kunze S, Peters A, Waldenberger M, Ward-Caviness C, Nohr EA, Sørensen TIA, Relton CL, Smith GD: **Prenatal and early life influences on epigenetic age in children: A study of mother-offspring pairs from two cohort studies.** *Hum Mol Genet* 2015, **25**:191–201.
127. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, Reese SE, Markunas CA, Richmond RC, Xu C-J, Küpers LK, Oh SS, Hoyo C, Gruzieva O, Söderhäll C, Salas LA, Baïz N, Zhang H, Lepeule J, Ruiz C, Ligthart S, Wang T, Taylor JA, Duijts L, Sharp GC, Jankipersadsing SA, Nilsen RM, Vaez A, Fallin MD, Hu D, et al.: **DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis**. *Am J Hum Genet* 2016, **98**:680–696.
128. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Res* 2015, **43**:e47.
129. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash JB, Sabunciyan S, Feinberg AP: **The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.** *Nat Genet* 2009, **41**:178–86.
130. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, Wei C-L: **Dynamic changes in the human methylome during differentiation.** *Genome Res* 2010, **20**:320–31.
131. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA: **Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies.** *Int J Epidemiol* 2012, **41**:200–9.
132. Peters T, Buckley M, Statham A, Pidsley R, Samaras K, Lord R, Clark S, Molloy P: **De novo identification of differentially methylated regions in the human genome**. *Epigenetics Chromatin* 2015, **8**:6.
133. Butcher LM, Beck S: **Probe Lasso: A novel method to rope in differentially methylated regions with 450K DNA methylation data**.

Methods 2015, **72**(C):21–28.
134. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, Johnson CS, Smiraglia DJ, Liu S: **IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data.** *Bioinformatics* 2012, **28**:729–30.
135. Du P, Bourgon R: **methyAnalysis: DNA methylation data analysis and visualization.** .
136. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S: **ChAMP: 450k Chip Analysis Methylation Pipeline**. *Bioinformatics* 2014, **30**:428–430.
137. Limbach M, Saare M, Tserel L, Kisand K, Eglit T, Sauer S, Axelsson T, Syvänen A-C, Metspalu A, Milani L, Peterson P: **Epigenetic profiling in CD4+ and CD8+ T cells from Graves' disease patients reveals changes in genes associated with T cell receptor signaling.** *J Autoimmun* 2016, **67**:46–56.
138. Binder AM, LaRocca J, Lesseur C, Marsit CJ, Michels KB: **Epigenome-wide and transcriptome-wide analyses reveal gestational diabetes is associated with alterations in the human leukocyte antigen complex.** *Clin Epigenetics* 2015, **7**:79.
139. Lando M, Fjeldbo CS, Wilting SM, C Snoek B, Aarnes E-K, Forsberg MF, Kristensen GB, Steenbergen RD, Lyng H: **Interplay between promoter methylation and chromosomal loss in gene silencing at 3p11-p14 in cervical cancer.** *Epigenetics* 2015, **10**:970–80.
140. Li D, Xie Z, Pape M Le, Dye T: **An evaluation of statistical methods for DNA methylation microarray data analysis.** *BMC Bioinformatics* 2015, **16**:217.
141. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA: **Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.** *Bioinformatics* 2014, **30**:1363–9.
142. Braak CJF Ter: **Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis**. *Ecology* 1986, **67**:1167–1179.
143. Ruiz C, Hernandez-Ferrer C, González J: **MEAL: Perform methylation analysis. R package version 1.10.0**. 2016.
144. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H, T.C.G.A.R. N, M.K. S, P.W. L, G. R, J.N. W, D.J. B, K.A. H, M. C, B.E. B, R.M. M, L. Y, L. C, M. G, T. H, H. N, Y. Z, M.K. S, et al.: ***TCGAbiolinks*: an R/Bioconductor package for integrative analysis of TCGA data**. *Nucleic Acids Res* 2016, **44**:e71–e71.
145. van Dongen J, Nivard MG, Willemsen G, Hottenga J-J, Helmer Q,

Dolan C V., Ehli EA, Davies GE, van Iterson M, Breeze CE, Beck S, Hoen PAC'., Pool R, van Greevenbroek MMJ, Stehouwer CDA, Kallen CJH van der, Schalkwijk CG, Wijmenga C, Zhernakova S, Tigchelaar EF, Beekman M, Deelen J, van Heemst D, Veldink JH, van den Berg LH, van Duijn CM, Hofman BA, Uitterlinden AG, Jhamai PM, Verbiest M, et al.: **Genetic and environmental influences interact with age and sex in shaping the human methylome**. *Nat Commun* 2016, **7**:11115.
146. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD: **sva: Surrogate Variable Analysis**. *R package version 3.20.0.* 2016.
147. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *J R Stat Soc Ser B* 1995, **57**:289–300.
148. Ruiz-Arenas C, Cáceres A, López-Sánchez M, Tolosana I, Pérez-Jurado L, González JR: **scoreInvHap: inversion genotyping for genome-wide association studies**. *Rev* .
149. Permuth-Wey J, Lawrenson K, Shen HC, Velkova A, Tyrer JP, Chen Z, Lin H-Y, Ann Chen Y, Tsai Y-Y, Qu X, Ramus SJ, Karevan R, Lee J, Lee N, Larson MC, Aben KK, Anton-Culver H, Antonenkova N, Antoniou AC, Armasu SM, Bacot F, Baglietto L, Bandera E V., Barnholtz-Sloan J, Beckmann MW, Birrer MJ, Bloom G, Bogdanova N, Brinton LA, Brooks-Wilson A, et al.: **Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31**. *Nat Commun* 2013, **4**:1627.
150. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project**. *Nat Genet* 2013, **45**:1113–1120.
151. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM: **Toward a Shared Vision for Cancer Genomic Data**. *N Engl J Med* 2016, **375**:1109–1112.
152. **birdseed2vcf**
153. **scoreInvHap Bioconductor version** [https://bioconductor.org/packages/release/bioc/html/scoreInvHap.html]
154. **WHO Cancer Fact Sheets**
155. Ramos M, Waldron L, Schiffer L, Obenchain V MM: **curatedTCGAData: Curated Data From The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects.** 2018.
156. Viechtbauer W: **Conducting Meta-Analyses in *R* with the** metafor **Package**. *J Stat Softw* 2010, **36**:1–48.
157. Law CW, Chen Y, Shi W, Smyth GK: **voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.** *Genome Biol* 2014, **15**:R29.

158. Ruiz-Arenas C, González JR: **Redundancy analysis allows improved detection of methylation changes in large genomic regions**. *BMC Bioinformatics* 2017, **18**:553.
159. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis**. *Bioinformatics* 2005, **21**:3439–3440.
160. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nat Protoc* 2009, **4**:1184–91.
161. Chen Y, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R: **Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray.** *Epigenetics* 2013, **8**:203–9.
162. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K: mediation **: *R* Package for Causal Mediation Analysis**. *J Stat Softw* 2014, **59**:1–38.
163. Itsara A, Vissers LELM, Steinberg KM, Meyer KJ, Zody MC, Koolen DA, de Ligt J, Cuppen E, Baker C, Lee C, Graves TA, Wilson RK, Jenkins RB, Veltman JA, Eichler EE: **Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing.** *Am J Hum Genet* 2012, **90**:599–613.
164. Zhang N, Huo Q, Wang X, Chen X, Long L, Guan X, Jiang L, Ma T, Hu W, Yang Q: **A genetic variant in p63 (rs17506395) is associated with breast cancer susceptibility and prognosis**. *Gene* 2014, **535**:170–176.
165. Rafiq S, Tapper W, Collins A, Khan S, Politopoulos I, Gerty S, Blomqvist C, Couch FJ, Nevanlinna H, Liu J, Eccles D: **Identification of inherited genetic variations influencing prognosis in early-onset breast cancer.** *Cancer Res* 2013, **73**:1883–91.
166. Rafiq S, Khan S, Tapper W, Collins A, Upstill-Goddard R, Gerty S, Blomqvist C, Aittomäki K, Couch FJ, Liu J, Nevanlinna H, Eccles D: **A genome wide meta-analysis study for identification of common variation associated with breast cancer prognosis.** *PLoS One* 2014, **9**:e101488.
167. Wang X, Lin Y, Lan F, Yu Y, Ouyang X, Wang X, Huang Q, Wang L, Tan J, Zheng F: **A GG allele of 3'-side AKT1 SNP is associated with decreased AKT1 activation and better prognosis of gastric cancer**. *J Cancer Res Clin Oncol* 2014, **140**:1399–1411.
168. Tahara T, Okubo M, Shibata T, Kawamura T, Sumi K, Ishizuka T, Nagasaka M, Nakagawa Y, Arisawa T, Ohmiya N, Hirata I: **Association between common genetic variants in pre-microRNAs and prognosis of advanced gastric cancer treated with chemotherapy.** *Anticancer Res* 2014, **34**:5199–204.

169. Kim JG, Chae YS, Lee SJ, Kang BW, Park JY, Lee E-J, Jeon H-S, Park JS, Choi GS: **Genetic variation in microRNA-binding site and prognosis of patients with colorectal cancer**. *J Cancer Res Clin Oncol* 2015, **141**:35–41.
170. Lee SJ, Kang BW, Chae YS, Kim HJ, Park SY, Park JS, Choi GS, Jeon H-S, Lee WK, Kim JG: **Genetic Variations in STK11, PRKAA1, and TSC1 Associated with Prognosis for Patients with Colorectal Cancer**. *Ann Surg Oncol* 2014, **21**:634–639.
171. Haja Mohideen AMS, Hyde A, Squires J, Wang J, Dicks E, Younghusband B, Parfrey P, Green R, Savas S: **Examining the polymorphisms in the hypoxia pathway genes in relation to outcome in colorectal cancer.** *PLoS One* 2014, **9**:e113513.
172. Roehl AC, Vogt J, Mussotter T, Zickler AN, Spöti H, Högel J, Chuzhanova NA, Wimmer K, Kluwe L, Mautner V-F, Cooper DN, Kehrer-Sawatzki H: **Intrachromosomal mitotic nonallelic homologous recombination is the major molecular mechanism underlying type-2 NF1 deletions**. *Hum Mutat* 2010, **31**:1163–1173.
173. Chowdhury R, Bois PRJ, Feingold E, Sherman SL, Cheung VG: **Genetic Analysis of Variation in Human Meiotic Recombination**. *PLoS Genet* 2009, **5**:e1000648.
174. Dallol A, Al-Maghrabi J, Buhmeida A, Gari MA, Chaudhary AG, Schulten H-J, Abuzenadah AM, Al-Ahwal MS, Sibiany A, Al-Qahtani MH: **Methylation of the polycomb group target genes is a possible biomarker for favorable prognosis in colorectal cancer.** *Cancer Epidemiol Biomarkers Prev* 2012, **21**:2069–75.
175. Park SJ, Kim S, Hong YS, Lee J-L, Kim J-E, Kim K, Hong S-M, Jin D-H, Kim CW, Yoon YS, Park IJ, Lim S-B, Yu CS, Kim JC, Kim TW: **TFAP2E Methylation Status and Prognosis of Patients with Radically Resected Colorectal Cancer**. *Oncology* 2015, **88**:122–132.
176. Laayouni H, Montanucci L, Sikora M, Melé M, Dall'Olio GM, Lorente-Galdos B, McGee KM, Graffelman J, Awadalla P, Bosch E, Comas D, Navarro A, Calafell F, Casals F, Bertranpetit J: **Similarity in Recombination Rate Estimates Highly Correlates with Genetic Differentiation in Humans**. *PLoS One* 2011, **6**:e17913.
177. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, Richardson MF, Anholt RRH, Barrón M, Bess C, Blankenburg KP, Carbone MA, Castellano D, Chaboub L, Duncan L, Harris Z, Javaid M, Jayaseelan JC, Jhangiani SN, Jordan KW, Lara F, Lawrence F, Lee SL, Librado P, Linheiro RS, Lyman RF, et al.: **The Drosophila melanogaster Genetic Reference Panel**. *Nature* 2012, **482**:173–178.
178. Hunter CM, Huang W, Mackay TFC, Singh ND: **The Genetic Architecture of Natural Variation in Recombination Rate in Drosophila**

melanogaster. *PLOS Genet* 2016, **12**:e1005951.
179. Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, Campo L, Flament P, Melchinger AE, Menz M, Meyer N, Moreau L, Moreno-González J, Ouzunova M, Pausch H, Ranc N, Schipprack W, Schönleben M, Walter H, Charcosset A, Schön C-C: **Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction.** *Genetics* 2014, **198**:3–16.
180. Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Rincent R, Schipprack W, Altmann T, Flament P, Melchinger AE, Menz M, Moreno-González J, Ouzunova M, Revilla P, Charcosset A, Martin OC, Schön C-C: **Intraspecific variation of recombination rate in maize**. *Genome Biol* 2013, **14**:R103.
181. Ishii K, Charlesworth B: **Associations between allozyme loci and gene arrangements due to hitch-hiking effects of new inversions**. *Genet Res* 1977, **30**:93.
182. Schaeffer SW: **SELECTION IN HETEROGENEOUS ENVIRONMENTS MAINTAINS THE GENE ARRANGEMENT POLYMORPHISM OF *DROSOPHILA PSEUDOOBSCURA***. *Evolution (N Y)* 2008, **62**:3082–3099.
183. Lowry DB, Willis JH: **A Widespread Chromosomal Inversion Polymorphism Contributes to a Major Life-History Transition, Local Adaptation, and Reproductive Isolation**. *PLoS Biol* 2010, **8**:e1000500.
184. Patterson N, Price AL, Reich D, Reich D, Daly M: **Population Structure and Eigenanalysis**. *PLoS Genet* 2006, **2**:e190.
185. McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Gibbs (Principal Investigator) RA, Dinh H, Kovar C, et al.: **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012, **491**:56–65.
186. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, Maranian MJ, Bolla MK, Wang Q, Shah M, Perkins BJ, Czene K, Eriksson M, Darabi H, Brand JS, Bojesen SE, Nordestgaard BG, Flyger H, Nielsen SF, Rahman N, Turnbull C, Fletcher O, Peto J, Gibson L, dos-Santos-Silva I, Chang-Claude J, Flesch-Janys D, Rudolph A, Eilber U, Behrens S, et al.: **Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer**. *Nat Genet* 2015, **47**:373–380.
187. Kass RE, Raftery AE: **Bayes Factors**. *J Am Stat Assoc* 1995, **90**:773–795.
188. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome**.

Science (80-) 2005, **310**:321–324.
189. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu C-J, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, et al.: **The somatic genomic landscape of glioblastoma.** *Cell* 2013, **155**:462–77.
190. Cancer Genome Atlas Research Network, Brat DJ, Verhaak RGW, Aldape KD, Yung WKA, Salama SR, Cooper LAD, Rheinbay E, Miller CR, Vitucci M, Morozova O, Robertson AG, Noushmehr H, Laird PW, Cherniack AD, Akbani R, Huse JT, Ciriello G, Poisson LM, Barnholtz-Sloan JS, Berger MS, Brennan C, Colen RR, Colman H, Flanders AE, Giannini C, Grifford M, Iavarone A, Jain R, Joseph I, et al.: **Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas**. *N Engl J Med* 2015, **372**:2481–2498.
191. Clayton D: **snpStats: SnpMatrix and XSnpMatrix classes and methods**. 2017.
192. Delaneau O, Zagury J-F, Marchini J: **Improved whole-chromosome phasing for disease and population genetic studies**. *Nat Methods* 2013, **10**:5–6.
193. Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, Estivill X, Moreno V: **SNPassoc: an R package to perform whole genome association studies**. *Bioinformatics* 2007, **23**:654–655.
194. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT: **Reproducible RNA-seq analysis using recount2**. *Nat Biotechnol* 2017, **35**:319–321.
195. Stevison LS, Woerner AE, Kidd JM, Kelley JL, Veeramah KR, McManus KF, Bustamante CD, Hammer MF, Wall JD: **The Time Scale of Recombination Rate Evolution in Great Apes**. *Mol Biol Evol* 2016, **33**:928–945.
196. Moyerbrailean GA, Kalita CA, Harvey CT, Wen X, Luca F, Pique-Regi R: **Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding?** *PLOS Genet* 2016, **12**:e1005875.
197. Hormozdiari F, Kichaev G, Yang W-Y, Pasaniuc B, Eskin E: **Identification of causal genes for complex traits**. *Bioinformatics* 2015, **31**:i206–i213.
198. Beaumont MA, Nichols RA: **Evaluating Loci for Use in the Genetic Analysis of Population Structure**. *Proc R Soc B Biol Sci* 1996, **263**:1619–1626.
199. Foll M, Gaggiotti O: **A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian**

perspective. *Genetics* 2008, **180**:977–93.
200. Bradbury IR, Hubert S, Higgins B, Bowman S, Borza T, Paterson IG, Snelgrove PVR, Morris CJ, Gregory RS, Hardie D, Hutchings JA, Ruzzante DE, Taggart CT, Bentzen P: **Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish**. *Evol Appl* 2013, **6**:450–461.
201. Van Wyngaarden M, Snelgrove PVR, DiBacco C, Hamilton LC, Rodríguez-Ezpeleta N, Zhan L, Beiko RG, Bradbury IR: **Oceanographic variation influences spatial genomic structure in the sea scallop, *Placopecten magellanicus***. *Ecol Evol* 2018, **8**:2824–2841.
202. Shih K-M, Chang C-T, Chung J-D, Chiang Y-C, Hwang S-Y: **Adaptive Genetic Divergence Despite Significant Isolation-by-Distance in Populations of Taiwan Cow-Tail Fir (Keteleeria davidiana var. formosana)**. *Front Plant Sci* 2018, **9**:92.
203. Cáceres A, Esko T, Pappa I, Gutiérrez A, Lopez-Espinosa M-J, Llop S, Bustamante M, Tiemeier H, Metspalu A, Joshi PK, Wilsonx JF, Reina-Castillón J, Shin J, Pausova Z, Paus T, Sunyer J, Pérez-Jurado LA, González JR: **Ancient Haplotypes at the 15q24.2 Microdeletion Region Are Linked to Brain Expression of MAN2C1 and Children's Intelligence**. *PLoS One* 2016, **11**:e0157739.
204. Giner Delgado C: **Large-scale evolutionary analysis of polymorphic inversions in the human genome**. *TDX (Tesis Dr en Xarxa)* 2017.
205. Boettger LM, Handsaker RE, Zody MC, McCarroll SA: **Structural haplotypes and recent evolution of the human 17q21.31 region.** *Nat Genet* 2012, **44**:881–5.
206. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD: **Genes mirror geography within Europe**. *Nature* 2008, **456**:98–101.
207. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R: **UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age**. *PLOS Med* 2015, **12**:e1001779.
208. **Early Growth Genetics Consortium** [https://egg-consortium.org/]
209. **The EArly Genetics and Lifecourse Epidemiology Consortium** [https://www.wikigenes.org/e/art/e/348.html]
210. **DataSHIELD** [http://www.datashield.ac.uk/]
211. Barber MF, Lee EM, Griffin H, Elde NC: **Rapid Evolution of Primate Type 2 Immune Response Factors Linked to Asthma Susceptibility.** *Genome Biol Evol* 2017, **9**:1757–1765.
212. Demaerel W, Hestand MS, Vergaelen E, Swillen A, López-Sánchez M, Pérez-Jurado LA, McDonald-McGinn DM, Zackai E, Emanuel BS, Morrow

BE, Breckpot J, Devriendt K, Vermeesch JR, Antshel K, Arango C, Armando M, Bassett A, Bearden C, Boot E, Bravo-Sanchez M, Breetvelt E, Busa T, Butcher N, Campbell L, Carmel M, Chow E, Crowley TB, Cubells J, Cutler D, Demaerel W, et al.: **Nested Inversion Polymorphisms Predispose Chromosome 22q11.2 to Meiotic Rearrangements**. *Am J Hum Genet* 2017, **101**:616–622.

Annex I: PhD portfolio

Other merits

- AGAUR FI Grant to develop a PhD thesis
- Finalist in Rin4, a competition to explain your PhD thesis in 4 minutes.

Other tasks developed during the PhD

- Analyst of INMA cohort in PACE and GoDMC, two consortia that analyze changes in DNA methylation in children.
- Quality control of DNA methylation data from HELIX and INMA projects
- Development of three R packages accepted in Bioconductor: scoreInvHap, MultiDataSet and MEAL.
- ISGlobal PhD representative
- Jury in Premi PRBB (Treballs de Recerca de Batxillerat)
- BRGE seminars' organization

Other publications

Reese SE, Xu C-J, den Dekker HT, Lee MK, Sikdar S, **Ruiz-Arenas C**, Merid SK, Rezwan FI, Page CM, Ullemar V, Melton PE, Oh SS, Yang IV, Burrows K, Söderhäll C, Jima DD, Gao L, Arathimos R, Küpers LK, Wielscher M, Rzehak P, Lahti J, Laprise C, Madore A-M, Ward J, Bennett BD, Wang T, Bell DA, The BIOS Consortium, Vonk JM, Håberg SE, Zhao S, Karlsson R, Hollams E, Hu D, Richards AJ, Bergström A, Sharp GC, Felix JF, Bustamante M, Gruzieva O, Maguire RL, Gilliland F, Baïz N, Nohr EA,

Corpeleijn E, Sebert S, Karmaus W, Grote V, Kajantie E, Magnus MC, Örtqvist AK, Eng C, Liu AH, Kull I, Jaddoe VWV, Sunyer J, Kere J, Hoyo C, Annesi-Maesano I, Arshad SH, Koletzko B, Brunekreef B, Binder EB, Räikkönen K, Reischl E, Holloway JW, Jarvelin M-R, Snieder H, Kazmi N, Breton CV, Murphy SK, Pershagen G, Anto JM, Relton CL, Schwartz DA, Burchard EG, Huang R-C, Nystad W, Almqvist C, Henderson AJ, Melén E, Duijts L, Koppelman GH, London SJ, Epigenome-wide Meta-analysis of DNA Methylation and Childhood Asthma, *Journal of Allergy and Clinical Immunology (2019)*, doi: https://doi.org/10.1016/j.jaci.2018.11.043

Gemma C Sharp, Lucas A Salas, Claire Monnereau, Catherine Allard, Paul Yousefi, Todd M Everson, Jon Bohlin, Zongli Xu, Rae-Chi Huang, Sarah E Reese, Cheng-Jian Xu, Nour Baïz, Cathrine Hoyo, Golareh Agha, Ritu Roy, John W Holloway, Akram Ghantous, Simon K Merid, Kelly M Bakulski, Leanne K Küpers, Hongmei Zhang, Rebecca C Richmond, Christian M Page, Liesbeth Duijts, Rolv T Lie, Phillip E Melton, Judith M Vonk, Ellen A Nohr, ClarLynda Williams-DeVane, Karen Huen, Sheryl L Rifas-Shiman, **Carlos Ruiz-Arenas**, Semira Gonseth, Faisal I Rezwan, Zdenko Herceg, Sandra Ekström, Lisa Croen, Fahimeh Falahi, Patrice Perron, Margaret R Karagas, Bilal M Quraishi, Matthew Suderman, Maria C Magnus, Vincent W V Jaddoe, Jack A Taylor, Denise Anderson, Shanshan Zhao, Henriette A Smit, Michele J Josey, Asa Bradman, Andrea A Baccarelli, Mariona Bustamante, Siri E Håberg, Göran Pershagen, Irva Hertz-Picciotto, Craig Newschaffer, Eva Corpeleijn, Luigi Bouchard, Debbie A Lawlor, Rachel L Maguire, Lisa F Barcellos, George Davey Smith, Brenda Eskenazi, Wilfried Karmaus, Carmen J Marsit, Marie-France Hivert, Harold Snieder, M Daniele Fallin, Erik Melén, Monica C Munthe-Kaas, Hasan Arshad, Joseph L Wiemels, Isabella Annesi-Maesano, Martine Vrijheid, Emily Oken, Nina Holland, Susan K Murphy, Thorkild I A Sørensen, Gerard H Koppelman,

John P Newnham, Allen J Wilcox, Wenche Nystad, Stephanie J London, Janine F Felix, Caroline L Relton; Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: findings from the pregnancy and childhood epigenetics (PACE) consortium, *Human Molecular Genetics*, Volume 26, Issue 20, 15 October 2017, Pages 4067–4085, https://doi.org/10.1093/hmg/ddx290

Hernandez-Ferrer C, **Ruiz-Arenas C**, Beltran-Gomila A, González JR. MultiDataSet: an R package for encapsulating multiple data sets with application to omic data integration. *BMC Bioinformatics*. 2017;18(1):36. Published 2017 Jan 17. doi:10.1186/s12859-016-1455-1

Joubert, B. R., Felix, J. F., Yousefi, P., Bakulski, K. M., Just, A. C., Breton, C., Reese, S. E., Markunas, C. A., Richmond, R. C., Xu, C. J., Küpers, L. K., Oh, S. S., Hoyo, C., Gruzieva, O., Söderhäll, C., Salas, L. A., Baïz, N., Zhang, H., Lepeule, J., **Ruiz, C**., Ligthart, S., Wang, T., Taylor, J. A., Duijts, L., Sharp, G. C., Jankipersadsing, S. A., Nilsen, R. M., Vaez, A., Fallin, M. D., Hu, D., Litonjua, A. A., Fuemmeler, B. F., Huen, K., Kere, J., Kull, I., Munthe-Kaas, M. C., Gehring, U., Bustamante, M., Saurel-Coubizolles, M. J., Quraishi, B. M., Ren, J., Tost, J., Gonzalez, J. R., Peters, M. J., Håberg, S. E., Xu, Z., van Meurs, J. B., Gaunt, T. R., Kerkhof, M., Corpeleijn, E., Feinberg, A. P., Eng, C., Baccarelli, A. A., Benjamin Neelon, S. E., Bradman, A., Merid, S. K., Bergström, A., Herceg, Z., Hernandez-Vargas, H., Brunekreef, B., Pinart, M., Heude, B., Ewart, S., Yao, J., Lemonnier, N., Franco, O. H., Wu, M. C., Hofman, A., McArdle, W., Van der Vlies, P., Falahi, F., Gillman, M. W., Barcellos, L. F., Kumar, A., Wickman, M., Guerra, S., Charles, M. A., Holloway, J., Auffray, C., Tiemeier, H. W., Smith, G. D., Postma, D., Hivert, M. F., Eskenazi, B., Vrijheid, M., Arshad, H., Antó, J. M., Dehghan, A., Karmaus, W., Annesi-Maesano, I., Sunyer, J., Ghantous, A., Pershagen, G., Holland, N., Murphy, S. K., DeMeo, D. L., Burchard, E. G., Ladd-Acosta, C.,

Snieder, H., Nystad, W., Koppelman, G. H., Relton, C. L., Jaddoe, V. W., Wilcox, A., Melén, E., … London, S. J. (2016). DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *American journal of human genetics*, *98*(4), 680-96.

Presentations in congresses:

- BioC 2017: Where Software and Biology Connect (26-28th July 2017) – Oral presentation. Boston, USA

  Title: MEAL "2": story of software package
- 2017 Congress of the European Society for Evolutionary Biology (20-25th August 2017) – Poster presentation. Groningen, the Netherlands.

  Title: Genotyping large microscopic inversions from recombination patterns in SNP data
- HELIX Symposium (30-31st October 2017) – poster presentation. Barcelona, Spain.

  Title: Association between DNA methylation and gene expression in HELIX subcohort (pilot)
- 4th ISGlobal-CREAL PhD Symposium (28th November 2017) – oral presentation. Barcelona, Spain.

  Title: Mapping polymorphic inversions to haplotypes
- 3rd ISGlobal-CREAL PhD Symposium (28th November 2016) - poster presentation. Barcelona, Spain.

  Title: Robust method to genotype large ancestral chromosomal inversions using SNPs data

- ISCB NGS-Barcelona - Structural Variation and Population Genomics Conference (3rd - 4th April 2017) - poster presentation. Barcelona, Spain.

  Title: MultiDataSet: an R package for encapsulating multiple data sets with application to omic data integration

- V Bioinformatics Student Symposium (12th May) - oral presentation. Barcelona, Spain.

  Title: MultiDataSet: an R package for encapsulating multiple data sets with application to omic data integration

- 2nd ISGlobal-CREAL PhD Symposium (4th November 2015) - poster presentation. Barcelona, Spain.

  MEAL: Methylation and Expression AnaLyzer

- European Bioconductor Developers Meeting (7th - 8th December 2015)- oral presentation. Cambridge, UK.

  Title: MEAL: Methylation and Expression AnaLyzer

- 4th Jornada d'Investigadors Predoctorals Interdisciplinària (2nd February) - oral presentation. Barcelona, Spain.

  Title: Expanding Genomic Knowledge in Neurodevelopment disorders

- Finalist in the competition Rin4, Universitat Pompeu Fabra (21st April 2016). Barcelona, Spain.

Title of the presentation: Expandiendo el conocimiento del efecto de variables estructurales y variables genéticas comunes en enfermedades complejas. Aplicación práctica en trastornos del neurodesarrollo.

- XIII Symposium on Bioinformatics (10th - 13th May 2016) - poster. Valencia, Spain.

  Title: MEAL: Methylation and Expression AnaLyzer

- The EUROPEAN HUMAN GENETICS CONFERENCE 2016 (21^{th} - 24^{th} May 2016) - poster. Barcelona, Spain.

  Title: Robust method to genotype large ancestral chromosomal inversions using NGS data

Courses:

- UPF course: "The craft of the scientific research article"
- Epidemiology II and Epidemiology III from Master in Public Health (UPF)
- Intervals course at PRBB: "How to design a visually stunning scientific poster"
- UPF course: "Oral Communication of Scientific Content. Keys for Communicating your Research Effectively"
- Intervals course at PRBB: "Putting the Why? before the How?"
- Seminar "How to perform and interpret a systematic review" (CREAL)
- Interval course at PRBB: "Técnicas actorales para la comunicación científica"