






Universitat Autònoma de Barcelona

Integrative analysis of the functional consequences of inversions in the human genome

Jon Lerga Jaso

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

Tesis doctoral

Integrative analysis of the functional consequences
of inversions in the human genome

JON LERGA JASO

Director:

MARIO CÁCERES AGUILAR



Departamento de Genética y de Microbiología
Facultad de Biociencias
Universitat Autònoma de Barcelona
Septiembre 2019

Integrative analysis of the functional consequences
of inversions in the human genome

Memoria presentada por Jon Lerga Jaso
para optar al grado de Doctor en Genética
por la Universitat Autònoma de Barcelona

Autor:

Jon Lerga Jaso

Director:

Mario Cáceres Aguilar

Cerdanyola del Vallès, Septiembre de 2019

Abstract

Structural variation contributes substantially to the genetic diversity, but its association with complex traits and diseases is not well understood and deserves detailed characterisation. This is particularly true for chromosomal inversions, whose functional consequences have remained elusive in humans, with very few notable exceptions. Despite the rising interest in identifying all types of genomic variants, inversions have been often set aside due to the presence of repetitive regions in their breakpoints along with their balanced nature. The InvFEST Project has tried to overcome this technical challenge by developing unique methods for inversion genotyping. Thanks to this effort, a total of 111 polymorphic human inversions have been accurately genotyped in a large number of individuals from diverse populations, becoming the most complete and reliable resource of this type of variation available to date. In the current era of precision medicine, quantitative trait loci (QTL) analysis has emerged as a key approach to determine how genetic polymorphisms influence gene expression and, in turn, phenotypic traits. Thus, this thesis makes the most of the great amount of data generated to perform for the first time a systematic quantification of the functional impact of human polymorphic inversions. The results show that inversions can affect gene expression by maintaining differentiated haplotypes, disrupting or reorganizing gene structures, creating novel fusion transcripts or acting through changes in epigenetic patterns. Strikingly, half of the inversions analysed act as lead QTLs or are in high linkage disequilibrium (LD) with top QTLs for gene expression and epigenetic changes across different tissues and cell lines, which suggests that inversions are enriched for functional effects. In particular, this influence on molecular phenotypes is even stronger for long inversions (>100kb), which are involved in 80% of the QTL associations. Although structural variants are known to have a higher chance to be associated with expression levels and complex traits, the detected inversion effects may reflect a trade-off between beneficial expression changes and potential negative costs on fertility due to production of unbalanced gametes by recombination. Furthermore, inversions present an enrichment of genome-

wide association studies (GWAS) signals in their surrounding area, and 14 of them are in high LD with trait-associated variants, supporting their potential implication in human phenotypes. Finally, the phenotypic consequences of two interesting inversions have been investigated in detail. HsInv0102 inverts an alternative non-coding exon from the *RHOH* gene. Interestingly, this inversion is also associated with RhoH protein levels and the inverted allele could act as a moderate protective locus on blood cancer susceptibility. On the other hand, HsInv0124 regulates the expression of several genes in the *IFITM* region, including *IFITM2* and *IFITM3*, through changes in histone modification patterns. Moreover, under infection conditions, this inversion has a pervasive effect on the expression of genes related with immune response, indicating that it may play an important role in defense against viral infections. All together, these findings illustrate the potential functional impact of inversions on the human genome and help to uncover previously missing variants related to phenotype variability.

Index

1	Introduction	1
1.1	Why chromosomal inversions? A brief overview	4
1.1.1	It started in <i>Drosophila</i> : The story of a genetic variant	4
1.1.2	Molecular mechanisms of inversion origin	7
1.1.3	Functional consequences of polymorphic inversions .	9
1.1.4	Role of inversions in local adaptation and complex phenotypes	23
1.1.5	Current strategies for inversion discovery and genotyping at large scale	27
1.1.6	The InvFEST project	31
1.2	Human genome function	33
1.2.1	Genetic determinants of gene expression variation .	34
1.2.2	Gene activity is regulated by interaction with epigenetic states	42
1.2.3	Genetic regulation of protein levels	49
1.2.4	Current methods for QTL analysis	55

1.3	Objectives	57
2	Materials and Methods	61
2.1	Methods for “Functional and phenotypic effects of a human inversion regulating <i>RHOH</i> isoforms and RhoH protein levels”	61
2.2	Methods for “A human inversion influences antiviral response through different regulatory effects on interferon response genes”	67
2.3	Methods for “Comprehensive analysis of the influence of human inversions on gene expression, epigenetic changes and phenotypic variation.”	71
3	Results	79
3.1	Evolutionary and functional impact of common polymorphic inversions in the human genome	80
3.2	Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR	125
3.3	Functional and phenotypic effects of a human inversion regulating <i>RHOH</i> isoforms and RhoH protein levels	165
3.3.1	HsInv0102 impact on <i>RHOH</i> gene structure and expression in LCLs	167
3.3.2	Additional effects on gene and protein expression	171
3.3.3	HsInv0102 association with blood cancer	173
3.4	A human inversion influences antiviral response through different regulatory effects on interferon response genes	179
3.4.1	HsInv0124 frequency and distribution	182

3.4.2	HsInv0124 impact on gene expression in LCLs	183
3.4.3	HsInv0124 and epigenetic changes	186
3.4.4	HsInv0124 effect on gene expression under infection	191
3.5	Comprehensive analysis of the influence of human inver- sions on gene expression, epigenetic changes and phenotypic variation	196
3.5.1	Inversion set and genotype calling	196
3.5.2	The impact of human inversions on gene expression and epigenetics	198
3.5.3	Inversions and phenotypes	203
3.5.4	Detailed characterization of HsInv0014 and HsInv0030 functional effects	207
4	Discussion	211
4.1	Methodology, data and limitations of this study	212
4.1.1	Gene expression quantification and QTL mapping	213
4.1.2	Tissues assayed and cell-specificity	216
4.1.3	Sample size, inversion frequency and genotype im- putation	219
4.1.4	Inversions and phenotypic effects	222
4.2	Significance of findings. The functional impact of human inversions	226
4.2.1	The influence of human inversions on gene expres- sion, epigenetic changes and phenotypic traits	228

Index

4.2.2	HsInv0102 association with Rhoh protein levels and blood cancer susceptibility	239
4.2.3	HsInv0124 is associated with the expression of immunologically-related genes	243
4.2.4	Do inversions have more functional effects than expected?	249
4.3	Future perspectives	249
5	Conclusions	253
6	Bibliography	255

List of Figures

1.1	Inversion detected in a <i>Drosophila</i> chromosome.	6
1.2	Suppression of recombination by inversions.	13
1.3	Recurrent inversions are the most common cause of haemophilia.	16
1.4	Position-effect variegation in <i>Drosophila</i>	19
1.5	Inversions predisposing to Koolen-de Vries and Williams-Beuren syndromes.	22
1.6	Inversion effects on mating systems in the white-throated sparrow and the ruff.	26
1.7	Identification of regulatory variants through eQTL analysis.	37
1.8	Outline of our integrative analysis of omics data exploring the contribution of polymorphic inversions to human traits.	58
3.1	HsInv0102 impact on <i>RHOH</i> expression in LCLs	171
3.2	Potential missed phenotypic associations of HsInv0102 by common genotyping arrays.	177
3.3	Possible association of HsInv0102 with blood cancer susceptibility.	178

List of Figures

3.4	HsInv0124 LD patterns and imputation accuracy.	183
3.5	HsInv0124 global distribution across 26 populations from 1000GP Ph3.	184
3.6	<i>De novo</i> transcriptome annotation around inversion HsInv0124	187
3.7	Regulation of gene expression by HsInv0124 in LCLs.	188
3.8	Differentially-expressed genes associated to HsInv0124 in LCLs stimulated with IFN.	189
3.9	Association of HsInv0124 to a <i>cis</i> -regulatory domain (CRD) activity.	190
3.10	HsInv0124 regulates gene expression in <i>IFITM</i> locus through histone modification marks in LCLs.	192
3.11	Detail of Inv0124 affecting histone modification levels.	193
3.12	Potential mechanism by which HsInv0124 is affecting histone modification levels in the surrounding region.	193
3.13	HsInv0124 is associated to <i>IFITM3</i> expression in CD14 ⁺ monocytes.	195
3.14	Pervasive effects of HsInv0124 on gene expression in monocytes stimulated with IAV.	196
3.15	Feasibility of inversion imputation.	199
3.16	Examples of <i>cis</i> INV-eQTLs in different tissues.	203
3.17	Enrichment of previously reported GWAS signals on inversion regions.	204
3.18	Candidate INV-eQTLs associated to GWAS hits.	206
3.19	Molecular consequences of inversion HsInv0030.	208

3.20	HsInv0014 affects <i>AKR1C1</i> and <i>AKR1C2</i> expression in LCLs.	209
3.21	HsInv0014 regulation of <i>AKR1C1</i> and <i>AKR1C2</i> expression is dependent on tissue.	209
4.1	Manhattan plots of cis eQTLs of the genes <i>NLRP6</i> in IAV-exposed monocytes and <i>CTRB2</i> in pancreas with and without inversion data.	223
4.2	HsInv0573 maintains two separated haplotypes with functional implications.	230
4.3	Potential functional impact of HsInv0031.	238
4.4	Interplay between SNP rs12252 and HsInv0124 alleles. . . .	245

List of Tables

- 1.1 Published pQTL studies in human blood derived samples. . . 52

- 3.1 Summary of inversions acting as QTLs of different molecular phenotypes. 202

Chapter 1

Introduction

The finalization of the first essentially complete version of the human genome sequence by the Human Genome Project more than 15 years ago, covering a total of $\sim 99\%$ of the euchromatin (Lander et al. 2001; International Human Genome Sequencing Consortium 2004), was just the starting point of a wide range of large-scale collaborative enterprises aimed at discovering all genetic elements implicated in the design of our phenotypic characteristics. The efforts devoted to generate a comprehensive description of human genetic variation by the International HapMap Project in the first place (The International HapMap Consortium 2003; The International HapMap Consortium 2005; International HapMap 3 Consortium 2010) and the 1000 Genomes Project (1000GP) afterwards (The 1000 Genomes Project Consortium 2012; The 1000 Genomes Project Consortium 2015) have provided a systematic catalogue of DNA sequence variation from the entire spectrum of allele frequencies across several populations. In addition, this has allowed us to discover the correlation patterns between nearby variants in linkage disequilibrium (LD) blocks or haplotypes, becoming a valuable public genomic resource and greatly enhancing our molecular understanding of the global genetic composition of the human genome and the processes that shape such diversity. Likewise, other recent sequencing projects have also offered an insight into the landscape of human genome variation, either by focusing on particular populations

Chapter 1. Introduction

(Wong et al. 2013; The Genome of the Netherlands Consortium 2014; The UK10K Consortium 2015; Besenbacher et al. 2015; Nagasaki et al. 2015; Ameer et al. 2017; Choudhury et al. 2017; Chheda et al. 2017; Kim et al. 2018) or on individuals from diverse groups (Gurdasani et al. 2015; Mallick et al. 2016).

Propelled by these projects and the commercial development of microarray technology for efficient genotyping of variation at individual nucleotides, known as single nucleotide polymorphisms (SNPs), genome-wide association studies (GWAS) became rapidly popular. By testing allele frequencies between healthy controls and affected patients, they have been widely used as a systematic approach to discover those genetic markers associated with the risk of suffering a particular disease. So far, GWAS analyses, through the generation of many international consortia, have identified thousands of robust associations with human traits and complex diseases (as for example The Wellcome Trust Case Control Consortium 2007; Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Day et al. 2017; Elliott et al. 2018; Pulit et al. 2019; Jansen et al. 2019). Despite this success, most associated loci confer relatively small increments in disease risk and they explain just a small part of the heritability estimated from familial clustering, which was called as the “missing heritability” problem (Manolio et al. 2009). Moreover, other issues present an obstacle to the interpretation of GWAS findings, including strong LD patterns in some loci, incomplete tagging of rare variants, the assessing almost exclusively of polymorphisms from European populations, or the fact that the vast majority (>90%) of GWAS signals reside in non-coding regions of the genome (Hindorff et al. 2009; Edwards et al. 2013). Establishing which gene functions are affected by associated variants has been thereby delayed. However, comprehensive functional annotation of the genome by other large projects such as the ENCODE project (Birney et al. 2007) has improved our biological knowledge of the genetic associations, and GWAS results have been translated into clinical application, such as in risk prediction and classification (Natarajan et al. 2017; Khera et al. 2018), diagnostic procedures and drug development (Manolio 2013; Okada et al. 2014; Finan et al. 2017; Fang et al. 2019).

Although GWAS arrays are efficient at genotyping SNPs across the whole genome, structural variants (SV) have frequently been ignored. SVs account for merely $\sim 0.2\%$ of genomic variation and their detection and genotyping are more challenging (Alkan et al. 2011; Lappalainen et al. 2019). Despite this group is less abundant than SNPs or small (<10 bp) insertions or deletions (indels), owing to their much larger size, SVs affect a higher fraction of nucleotide differences in the human genome (Pang et al. 2010; Alkan et al. 2011; The 1000 Genomes Project Consortium 2015). Structural variation has been found to be disproportionately enriched in functional associations; i.e. SVs are more likely to be associated to GWAS hits or affect gene expression with larger effect sizes compared to SNPs (Sudmant et al. 2015 Chiang et al. 2017). For instance, SVs have been identified as notable contributors to several conditions, such as copy number variation in autism (Pinto et al. 2010) or schizophrenia (Marshall et al. 2017). Therefore, this type of variants probably play a very important role in human pathology.

Nonetheless, not all SVs have been studied at the same level. This is particularly true for inversions, which could be considered the least understood of all genetic variants. Compared to other SVs, like copy number variants (CNVs), they still constitute a challenge because its balanced nature and the complex repetitive regions in which their breakpoints usually appear (Puig et al. 2015a). In spite of being discovered nearly a century ago in *Drosophila melanogaster* (Sturtevant 1917; Sturtevant 1921a), only few human polymorphic inversion have been studied in some detail so far, and their potential clinical relevance is almost unknown. However, inversions have been implicated in a broad range of phenotypic traits and adaptation in other species (Wellenreuther and Bernatchez 2018). Besides, it has been found that a notable fraction of human inversions are recurrent and are likely missed by GWAS arrays (Aguado et al. 2014). Therefore, we may be missing the functional role of this kind of variants, which makes the study of human inversions one of the areas with greater potential within the genomic structural variation field.

1.1 Why chromosomal inversions? A brief overview

1.1.1 It started in *Drosophila*: The story of a genetic variant

Already in the first quarter of the twentieth century, Alfred H. Sturtevant postulated the existence of inverted segments in the genetic material as an event that can suppress recombination between chromosomes of *Drosophila* species (Sturtevant 1917; Sturtevant 1921a). In 1913, as an undergraduate, Sturtevant published the first genetic map based on the proportion of crossing over events between normal and mutant alleles of six genes on the chromosome X of *D. melanogaster*. Crossover rates were useful to arrange genes in a linear series, but he also realized that some chromosomes segregating in natural populations of flies carried crossing-over “modifiers”, after observing an unusually low percentage of crossovers affecting particular loci in some specimens (Sturtevant 1917). By comparing the genetic linkage maps on chromosome X of *D. melanogaster* and *D. simulans*, Sturtevant demonstrated that five genes controlling similar characters were located in the same order, which suggested some degree of conservation between closely related species (1921b). However, while comparing the third chromosome of the same species, Sturtevant found that in this case three identical loci were arranged in a different order, which was consistent with having a block of genes rotated by 180° and constitute the first published evidence of an inversion (Sturtevant 1921a). Moreover, this inverted gene order explained the observed “crossover suppression” phenomenon for those samples that were inversion heterozygotes.

Initially, laborious observations on the genetic linkage in crosses between different strains of the same *Drosophila* species were mainly used to detect inversions. The posterior discovery of the polytene salivary gland chromosomes in fly larvae (known as “giant chromosomes” at that time) allowed the direct observation of the karyotypes and inversion breakpoints by simple microscopic analysis (Dobzhansky and Sturtevant 1938; Krimbas and

Powell 1992; Kirkpatrick 2010). These dipteran polytene chromosomes are interphase chromosomes formed by the intimate association of thousand of DNA strands resulting from numerous S-phase rounds of the cell cycle that result in one large size chromosome with easy to visualize structural features. Thus, heterozygous inversion could be seen clearly during this somatic chromosome pairing as a loop-like configuration (Figure 1.1).

The substantial data on inversion polymorphisms identified within species and fixed inversions between different species of the *Drosophila* genus were used as the first genetic markers to infer phylogenetic trees (Dobzhansky and Sturtevant 1937). However, the improvement of molecular biology techniques in the 1970s promoted the decline in popularity of research on inversions in the following years (Kirkpatrick 2010; Wellenreuther and Bernatchez 2018). More recently, the new sequencing and genomic techniques and the pervasive discovery of other structural variants made their study more attractive thanks to the development of array-based strategies that allowed to identify large unbalanced changes provoked by gains or losses of DNA segments. These technologies supposed indeed a big advance in the knowledge of structural variation, but, at the same time, the balanced nature of inversions also meant a delay in their knowledge compared to deletions, insertions or copy number variants (Alkan et al. 2011).

Until a few years ago, in humans, inversion discovery was limited to traditional cytogenetics, like classical G-banded karyotypes, but only large-scale rearrangements -of several megabases in size- were perceptible under the microscope, whereas submicroscopic changes remained largely ignored (Feuk 2010). In this regard, several pericentric inversion variants (i.e. including the centromere) affecting constitutive heterochromatic DNA have been reported by cytogenetic analysis in some chromosomes with no apparent phenotypic effects. An example is the inverted polymorphism in chromosome 9 with incidences around 0-4% across human populations (Hsu et al. 1987; Thomas et al. 2008), about which a few subsequent studies have shown no adverse influence on fertility and pregnancy (Mozdarani et al. 2007; Dana and Stoian 2012; Nonaka et al. 2019). Furthermore, inversion

rearrangements can appear *de novo* in patients or families and be routinely discovered by cytogenetic analysis in hospitals, often associated to abnormal manifestations (Partida-Perez et al. 2012; Sotoudeh et al. 2017), or appear in cancer (Le Beau et al. 1983; Gorello et al. 2013; Kaneko et al. 2014). Inversion analysis was helped later on by the development of other techniques. These labour-intense and target-based approaches resulted in the identification of only a few human balanced structural rearrangements (see below), but the absence of high-throughput discovery and genotyping methods for inversions restricted their characterization to a modest catalogue of polymorphisms (Feuk 2010) and despite their initial finding about a century ago, inversions were set aside in favor of the rest of genetic variants.

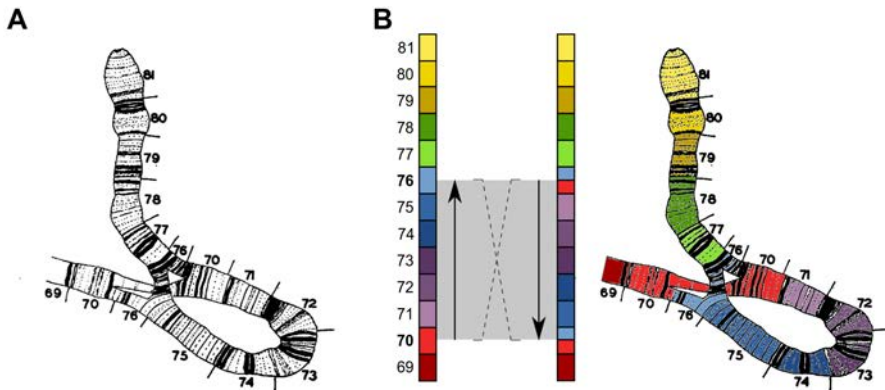


Figure 1.1 – Inversion detected in a *Drosophila* chromosome. (A) Drawing of a polytene chromosome of *D. pseudoobscura* showing its division into sections, obtained by Dobzhansky and Sturtevant in 1938 (Dobzhansky and Sturtevant 1938). Multiples copies of each parental third chromosome of this species are tightly associated with its homologue in somatic synapsis. The loop configuration observed is due to the presence of an inversion in heterozygosis. (B) Drawing modified with a colored pattern and a schema of the inversion for easier visualization. Sections 70 and 76 are highlighted to indicate the sections in which inversion breakpoints occur (red and blue sections), leading to an inverted region.

1.1.2 Molecular mechanisms of inversion origin

An inversion arises when a chromosomal segment breaks at two points, frequently within repetitive sequences, rotates 180 degrees, and is inserted at the same location, reversing the DNA sequence contained in it. Classically, inversions are classified with respect to the inclusion of the centromere in the inverted region, termed pericentric if they do involve the centromere, and paracentric otherwise. Besides, inversions fall into distinct classes based on the sequence similarity found at the breakpoints, which reflects the mechanism by which these variants have been generated (Escaramís et al. 2015; Weckselblatt and Rudd 2015). Many of them have very similar sequences in inverted orientation at the breakpoints and are the result of a recombination-based process, termed non-allelic homologous recombination (NAHR). Nonetheless, two other general mechanisms can lead to the formation of inversions: pathways aimed to repair DNA breaks, such as non-homologous end-joining (NHEJ) and microhomology mediated end-joining (MMEJ), or problems produced during replication, like fork stalling and template switching (FoSTeS), microhomology-mediated break-induced replication (MMBIR), serial replication slippage (SRS) and break-induced SRS (BISRS). These repair-based and replicative-based examples -jointly referred here as non-homologous processes- are, in this manner, mostly caused by a variety of different mechanisms (Lee et al. 2007) and give rise to unique events.

Long sequences of homology provide the substrate for NAHR, which occur by illegitimate recombination of paralogous sequences that have misaligned in mitosis or meiosis. Hence, segmental duplications, also known as low-copy repeats, constitute the common catalysts of such events, although this process can also occur at other repetitive elements of high sequence similarity such as SINEs (Short interspersed nuclear elements), LINEs (Long interspersed nuclear elements) or LTRs (Long terminal repeat elements) (Startek et al. 2015). The resulting SV is determined by the location and relative orientation of the homologous sequences: within the same chromosome, (i), a deletion appears when sequences are in direct orientation and (ii) an inversion when such sequences are placed in inverted

Chapter 1. Introduction

orientation; (iii) deletions and duplications occur when the sequences implicated are from homologous chromosomes in direct orientation; and (iv) translocations, on the other hand, happen when regions with high similarity are located on different chromosomes. The factors behind the frequency of NAHR still remain obscure, but it is clear that non-allelic crossovers are associated with longer sequences of extremely high identity and negatively correlated with the distance at which repetitive elements are situated (Liu et al, 2011). Importantly, repeat-mediated NAHR inversions, as well as other rearrangements, have been shown to happen recurrently (Aguado et al. 2014; Weckselblatt and Rudd 2015), and some of them have been related to genomic disorders (Fawcett and Innan 2013).

DNA double-strand break and repair mechanisms can also lead to the generation of inversions. Together with homologous recombination, NHEJ is a major process of break-induced repair in mammals, where the break ends are directly ligated with almost no homology requirements (<4 bp). Although NHEJ can provoke gain or loss of nucleotides, it usually takes place with high fidelity, whereas the alternative pathway, MMEJ, is more error-prone. MMEJ accounts for a minor proportion of double strand break repair and, although it was believed that generally occurs as a back-up mechanism when NHEJ is suppressed or a sister chromatid is not available for homologous recombination, increasing evidence indicates that MMEJ may play a more important role in DNA repair (McVey and Lee 2008). The distinguishing element of MMEJ is the use of substantial microhomology (5-25 bp) during rejoining of broken ends and it is considered a source of genomic instability (Iliakis et al. 2004). Double-strand breaks are generated randomly in the genome by both endogenous and exogenous genotoxic agents and inversions that arise from them are often distributed through in the genome without a clear pattern and are accompanied by short insertions or deletions at the breakpoint junctions due to these error-prone mechanisms.

Replication-based processes have been also proposed as origin of complex rearrangements, including inversions and translocations (Lee et al. 2007). The stalling of an active replication fork can cause the drifting of the DNA

polymerase, which may invade a second fork where the DNA synthesis is re-initiated (Zhang et al. 2009; Weckselblatt and Rudd 2015). This is defined as FoSTeS and the template exchange is usually mediated by microhomology to any nearby single-stranded DNA. Serial template switching may occur several times before resumption of replication on the original template, causing different types of rearrangements that can range from some kilobases to few megabases. Other replication-based models causing smaller complex rearrangements include SRS, which produces forward and backward replication slippage due to the presence of short repeats in direct orientation, provoking a dissociation followed by improper reassociation of the replisome (Escaramís et al. 2015).

Pang and colleagues (2013) attempted to clarify the underlying mechanisms that give rise to all structural variation between two genomes, including inversions. A slightly larger number of the inversions in the analysed dataset were generated by NAHR rather than by non-homologous processes that require little or no sequence homology, but each procedure accounted for almost half of the total inversions. It is worth noting that these relative proportions are presumably biased by the selected dataset and the techniques available to detect inversions. Most common sequencing strategies are blind to the detection of inversions originated at segmental duplications and repetitive regions (Lucas-Lledó and Cáceres 2013), which points out NAHR as the dominant generation mechanism. Actually, segmental duplications account for about 5% of the human genome (Bailey et al. 2002), and at least half of the genome sequence is composed by repetitive elements (International Human Genome Sequencing Consortium 2001), although some studies suggest that is over two-thirds (de Koning et al. 2011). Therefore, there is considerable raw material for NAHR events.

1.1.3 Functional consequences of polymorphic inversions

We have already done a brief overview of how chromosomal inversions were discovered and the mechanisms underlying their origin. But, why

inversions? Given that in many cases genetic content within inverted and uninverted chromosomes remains the same, what type of impact could just a change in the DNA linear order have? Two main mechanisms have been hypothesized since early on (Hoffmann and Rieseberg 2008; Kirkpatrick 2010; Wellenreuther and Bernatchez 2018). First, chromosomal inversions have intrigued biologists for a long time due to their unique properties in recombination suppression between the two orientations. Thus, inversions can capture and maintain together an advantageous combination of alleles that will not recombine. Further, it is possible that adaptive mutations are accumulated within the inversion and protected over time. Second, another potential consequences of inversions is what has been traditionally called as position effects due to the change of position of genes and regulatory sequences. Moreover, inversions could cause significant functional alterations by disrupting gene sequences that span the breakpoints. Finding the association between inversion polymorphisms and distinct functional elements across the genome is an essential task to understand how they could remodel the regulatory machinery and explain their impact on gene expression. In this section, we review in detail the potential implications derived from inverting a genomic segment, including: (i) inhibition of recombination; (ii) mutational events at inversion breakpoints; (iii) the so-called position effects; and (iv) predisposition to further genomic rearrangements (Puig et al. 2015a).

Inhibition of recombination

One of the most characteristic effects of inversions is the inhibition of recombination in heterozygotes with both orientations, which is key to figure out their evolutionary and functional consequences. Inversions have for long been known to affect the normal intimate pairing between homologous chromosomes during the first meiotic division of diploid organisms, including humans (Hoffmann and Rieseberg 2008; Kirkpatrick 2010). Given the loss of linear homology at inverted regions in heterozygotes, inversions are known to generate a chromosomal loop that allow chromosomes to synapse with different orientation along the inverted segment, as it was observed

initially in *Drosophila* polytene chromosomes (Figure 1.1). Still, recombination within the inversion will generate unbalanced products, which consist of acentric fragments and dicentric chromosomes if the crossover occurs within a paracentric inversion, or chromosomes with deletions and duplications of the whole arm fraction outside the inversion if a pericentric inversion is implicated (Figure 1.2). Therefore, unless a physical mechanism that blocks crossovers exist, inversions are likely to have an impact on fertility since unbalanced chromosomes cannot give rise to viable progeny (Anton et al, 2005). Also in the case that two crossovers happen within the inversion (actually, an even number), the resulting chromosomes will be balanced. The exception are the species of *Drosophila* and other diptera, since males do not recombine and in females the aberrant chromosomes are relegated to the polar bodies.

Given that recombination allows the formation of new allelic combinations, its impairment has important evolutionary consequences. For example, the X and Y chromosomes do not recombine along almost their full length and inversions have been an efficient mechanism for recombination suppression during sex chromosome evolution. In fact, the establishment of sex chromosomes during an early stage of mammal evolution involved the appearance of a group of inversions that gradually increased the non-recombining region, giving rise to separated sex-determinant alleles (Ming and Moore 2007; Hoffmann and Rieseberg 2008; Kirkpatrick 2010; Bachtrog 2013). Moreover, the fact that inversions inhibit recombination is an effective way to trap a favourable combination of independent alleles and preserve linkage between them. We will comment in the next section several cases in distinct species of animals and plants with long inversions acting as “supergenes”, which can lead to complex phenotypes, environmental adaptation and speciation. In addition, further adaptive mutations can arise and establish within the inversion, giving rise to divergent haplotypes associated with different functional profiles, which will be protected over time.

In humans, the most notable example is the 589-kb inversion 17q21.31 (Stefansson et al. 2005). As a result of local recombination suppression, a

strong LD spanning all the region has been observed with two highly divergent lineages, H1 and H2 (Boettger et al. 2012; Steinberg et al. 2012; Alves et al. 2015). H2 haplotype is at low frequency in Africans and East Asians, but is found at higher levels in Europe. Remarkably, positive selection in one of the H2 inversion subhaplotypes has been reported in European individuals due to the association of this inversion with an increase in fertility of the female carriers in the Icelandic population (Stefansson et al. 2005). As we will see later, this inversion is also associated to a microdeletion syndrome and also many gene expression changes (de Jong et al. 2012) and clinical conditions, such as Alzheimer’s disease (Myers et al. 2005), Parkinson’s disease (Skipper et al. 2004; Zabetian et al. 2007; Tobin et al. 2008; Setó-Salvia et al. 2011), neuroticism (Okbay et al. 2016), progressive supranuclear palsy and corticobasal degeneration (Baker et al. 1999; Webb et al. 2008).

Inversions causing direct mutational events

Among the most evident molecular consequences of inversions are direct mutations at the gene structure level. In this regard, inversions can disrupt genes that span one of the breakpoints, impairing or completely inhibiting the transcription of the resulting truncated copy. There are many cases of pathological inversions that disrupt genes (Puig et al. 2015a). For instance, a high fraction of the patients suffering the X-linked disorders haemophilia A and Hunter syndrome are known to carry inversions that appear recurrently in the population. In particular, 45% of severe cases of haemophilia A in humans, a coagulopathy caused by a genetic deficiency in the clotting factor VIII (encoded by the *F8* gene), is due to inversions generated by NAHR between a 9.5 kb repeated element called “int22h”, which is located flanking the exon 22 of the *F8* gene, and one of the two copies located ~565 kb away closer to the telomere in inverted orientation (Figure 1.3). This misalignment leads to inversions of the whole region and the disruption of the coding capacity of the *F8* gene (Lakich et al. 1993; Antonarakis et al. 1995; Bagnall et al. 2005; Gouw et al. 2012). Another inversion mediated by homologous recombination between repeats

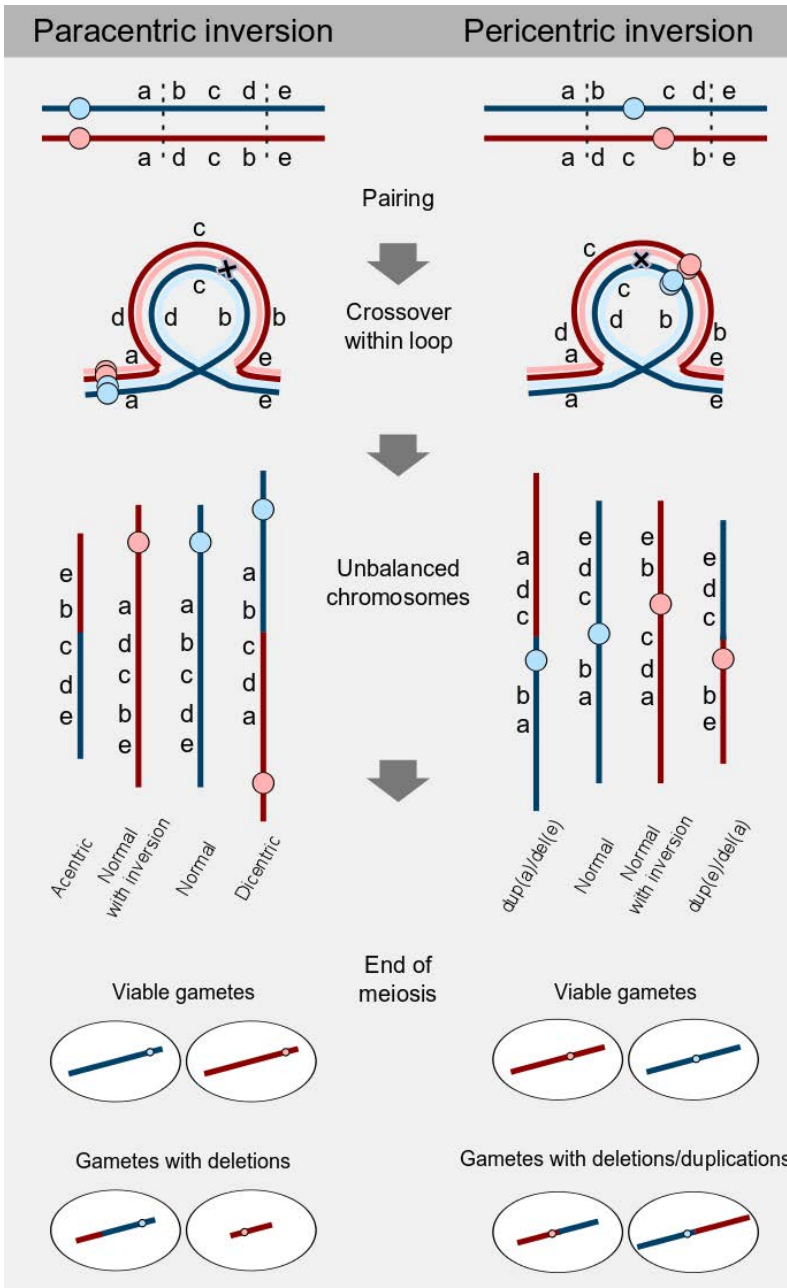


Figure 1.2 – Suppression of recombination by inversions. Caption on the following page.

Figure 1.2 – Suppression of recombination by inversions. Paracentric (left) and pericentric (right) inversions form a loop structure during the pairing of homologous chromosomes. A single crossing over event within the inverted segment results in unbalanced chromosomes. Letters represent loci order and circles indicate centromeres. In paracentric inversions, one acentric fragment and another dicentric chromosome are produced. The acentric fragment is lost because it cannot be drawn to either pole in chromosome segregation, and the dicentric fragment forms a bridge between spindle poles in the telophase of Meiosis I and is broken randomly, leading to two products with deletions. In pericentric inversions two chromosomes with both duplications and deletions are originated. The outcome of this is the selection of just non-crossover chromosomes in viable offspring.

accounts for around 2-4.8% of haemophilia A cases and generates hybrid transcripts between the first exon of *F8* and the gene *VBP1* (Bagnall et al. 2002; Gouw et al. 2012; Oldenburg et al. 2014). Hunter syndrome, or mucopolysaccharidosis type II, is a rare disease caused by the accumulation of glycosaminoglycans in body tissues due to a deficiency of the lysosomal enzyme iduronate 2-sulfatase (encoded by the *IDS* gene), which is implicated in the catabolism of these molecules. Patients with the Hunter syndrome frequently present an inversion caused by an homologous recombination event between the *IDS* gene and an adjacent partial pseudogene (*IDS2*), resulting in the disruption of the *IDS* gene in intron 7 (Bondeson et al. 1995). Also, inversions have been shown to break coding genes or create fusion genes in cancer (Soda et al. 2007; Gruber et al. 2012; Gao et al. 2013; Rhees et al. 2014; Hamatani et al. 2014; Argani et al. 2017).

One of the best examples of a human polymorphic inversion with these effects is HsInv0379, which was already described by Puig and colleagues (Puig et al. 2015b). HsInv0379 is located in chromosome 19 and is among the longest polymorphic inversions described in humans, with a length of 415 kb. Transcription factor *ZNF257* is disrupted by one inversion break-

point located in an intron of the gene, and, as expected its transcription level is reduced in heterozygous individuals and no expression was detected in the unique analysed individual carrying both inverted alleles. Interestingly, the inversion relocates the promoter and the first two exons of this gene, inducing the transcription of a new fusion transcript in the new location that incorporates a novel 3' exon made up of fragments from repetitive elements. Although the inversion generates a defective *ZNF257* copy and a novel transcript isoform, no clear phenotypic traits were associated to the inversion, but its limited global distribution -it is almost exclusively present in individuals with East Asian ancestry with an average frequency of the inverted allele of $\sim 5\%$ - suggests that if anything, it may have detrimental effects.

Moreover, there are also some other inversions with less drastic effects. Several examples of polymorphic inversions with two gene sequences placed at both breakpoints have been described (Puig et al. 2015a). This is frequent among rearrangements generated by NAHR, where two gene copies located within highly homologous segmental duplications can exchange sequences when the inversion occurs. If final transcripts generated in the inverted conformation do not possess significant nucleotidic differences, expression changes are unlikely. However, if rearranged genes were divergent enough, hybrid transcripts are created and expression or function may vary. This is the case of HsInv0030, where the first exon and promoter of the two chymotrypsinogen precursor genes *CTRB1* and *CTRB2* are exchanged and hybrid transcript sequences have been documented (Pang et al. 2013). In fact, HsInv0030 has been associated to alcoholic chronic pancreatitis risk and changes in the *CTRB1/CTRB2* expression ratio (Rosendahl et al. 2018). Additionally, when both breakpoints are placed within the same gene, inversion can reverse exons and presumably affect splicing, like in inversion HsInv0102, which inverts an alternative non-coding exon of the protein-coding gene *RHOH* (Puig et al. 2015a; Sudmant et al. 2015). In such cases, a more detailed analysis of the consequences on expression of the different isoforms are required to disentangle the potential functional role. Altogether, these examples highlight the importance of these variants in human pathology.

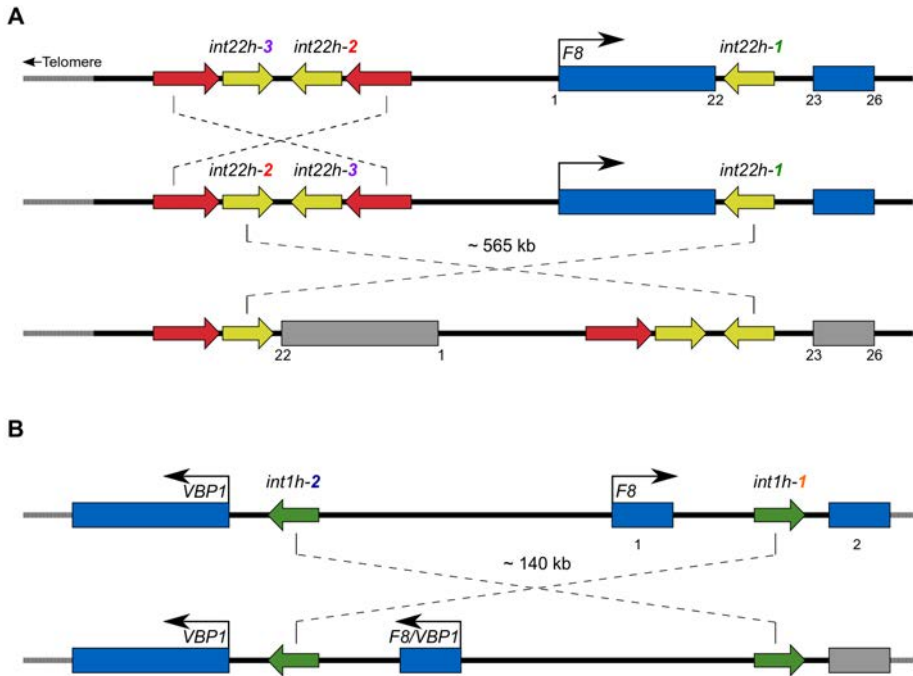


Figure 1.3 – Recurrent inversions are the most common cause of haemophilia. (A) In chromosome X, two repeated copies of the element *int22h* (yellow arrows), which is within intron 22 of the factor VIII encoding gene *F8* (the coding sequence is depicted in blue and the numbers indicate the exons), are located approximately 565 kb away towards the telomere. These copies are flanked by highly identical segmental duplications (red arrows) that can give rise to a polymorphic inversion that reverses the orientation of both repeating elements. A misalignment and NAHR with the other element located in opposite orientation lead to an inversion of the whole region, disrupting in turn the *F8* protein-coding gene. (B) Another inversion mediated by the repeated elements 'int1h' (green arrows), one of the located at the first intron of *F8*, can results in this gene disruption. These two inversions account for half of haemophilia A cases (Gouw et al. 2012). Non-functional parts of genes after inversion disruption are indicated in grey.

Position effect: Relocation of genomic segments by inversions

Although mutational effects on gene sequences could be considered as a direct consequence of the inversion position, we decided to dedicate a separate section for those potential inversions that can influence mRNA levels without mutating directly the gene sequence. Indeed, for some authors, a “position effect” has been defined as an alteration of gene-expression patterns provoked by relocating the whole gene from its original place to another genomic location, but this change should not affect the gene itself and the transcription unit is supposed to remain intact (Wilson et al. 1990; Kleinjan and van Heyningen 1998). Thus, it is worth noting that inversions can reshape genomic regions by modifying the position of diverse functional elements with respect to their ordinary chromosomal environment.

Hypothetically, it is possible to imagine various mechanisms that could be responsible for expression changes (Wilson et al. 1990; Kleinjan and van Heyningen 1998; Sharp et al. 2006; Puig et al, 2015a). For instance, an inversion could separate a *cis*-acting regulatory element from the gene on which such element exerts its influence. Depending on the type of regulatory element, the consequences could vary. For example, removing an enhancer may lead to a reduction of the gene transcription in the inverted allele, while disconnecting a silencer may give rise to inappropriate activity of the gene in certain tissues or cell types. Alternatively, if the regulatory elements from this gene are situated closer to another transcript unit because of the rearrangement, the novel regulatory sequences may increase or impair the expression of this second gene by providing novel regulatory sequences. Also, a competition for the interaction with these regulatory sequences between both genes may occur, resulting in aberrant expression levels as well. Although these cases differ from the reorganization of gene sequences in the previous section, the consequences of some of these position effects resemble the gene fusions that frequently occur in lymphoid leukaemias and lymphomas (Wang et al. 2017). A good example is Burkitt lymphoma, which is classically characterized by a translocation that juxtaposes the *MYC* gene to regulatory elements of the immunoglob-

ulin heavy or light chain. As a consequence, *MYC* becomes constitutively activated owing to the control of the immunoglobulin enhancers (Boerma et al. 2009). Thus, inversions could act in the same manner by bringing elements together. Moreover, such position effect of inversions might also result in gene expression alteration due to changes in local chromatin structure. Indeed, the classical example is that of a gene sequence that is relocated from a euchromatic region into the heterochromatin, or vice versa, as it happens in position-effect variegation in *Drosophila* (Figure 1.4). In this case, an inversion places the eye colour gene *white* into the pericentric heterochromatin, which results in the methylation of the promoter and the subsequent down-regulation of gene expression (Vogel et al. 2009).

Position effects can be also extended to insulators or boundary components, affecting the chromatin context. As will be seen below, the three-dimensional structure of mammalian genomes is organized into topologically associated domains (TADs) (Dixon et al. 2012). These domains have been proposed to define regions flanked by CCCTC binding factor (CTCF) sites that insulate the regulatory activity within the TAD and prevent interactions with non-target genes (ref). In this regard, recent studies highlighted the disruption of TADs by distinct structural variants as a cause of gene misexpression in various human diseases (Lupianez et al. 2015; Flavahan et al. 2016; Franke et al. 2016; Hnisz et al. 2016). For instance, the split of the mammalian *HoxD* gene cluster in mice by an induced inversion rearrangement led to artificial repositioning of the genes relative to flanking enhancer regions and dysregulated patterns of gene activity during limb development (Spitz et al. 2005). Moreover, Kraft and colleagues induced a series of inversions to place an active limb enhancer cluster into the neighbouring gene-dense region and explored their impact on gene expression *in vivo* in mice (Kraft et al. 2019). These inversions broke the previous TAD configuration, leading to aberrant gene activation. These alterations contributed to congenital malformations like severe polydactyly (Kraft et al. 2019). Consequently, inversions that are not breaking gene sequences cannot be discarded as functionally neutral; rather, those inverted regions encompassing non-coding segments could be

affecting the regulatory machinery due to these positional effects. Even the slightest modification, such as the inversion of an enhancer, which are known to be theoretically flexible in terms of orientation, could have consequences on gene expression, since a few studies have described enhancers with orientation-dependent activity (Nishimura et al. 2000; Swamynathan and Piatigorsky 2002). Nonetheless, no polymorphic inversions affecting gene expression through any of these mechanisms have been reported in humans so far.

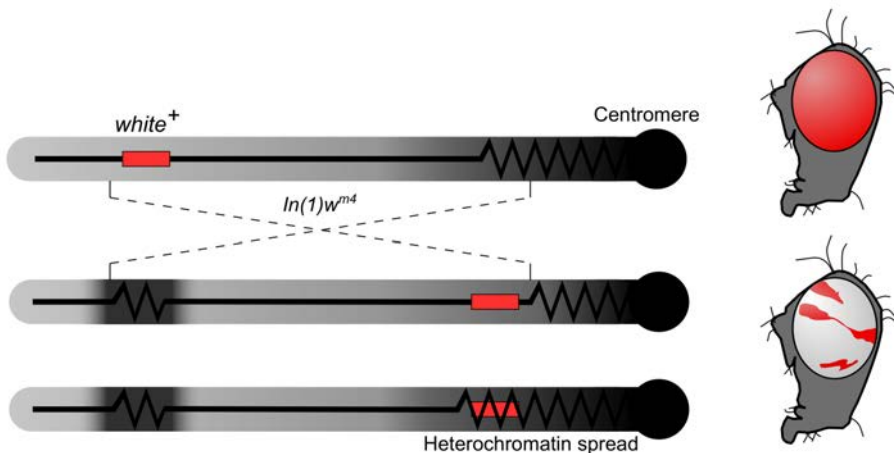


Figure 1.4 – Position-effect variegation in *Drosophila*. Normally, every cell from *Drosophila* eyes expresses the white gene, which results in a red-eye phenotype. If the inversion places the white gene close to the pericentric heterochromatin, this gene could be silenced by the abnormal spread of the heterochromatic marks into the euchromatin. Flies that inherit this rearrangement possess variegated eyes (mosaic of red and white eye color), since there is variation in heterochromatin spread and some cells express the gene while others do not.

Inversions predisposing to further rearrangements

Under certain circumstances, inversions may not have an obvious functional impact as in the examples described above, but they could rather confer a predisposition to further rearrangements that lead to several disor-

ders. These secondary rearrangements, which are deletions in most cases, are presumably catalyzed by the presence of large and highly identical direct repeats. In these cases, inversions are characterized by switching such repetitive elements and expand the number of segments of sequence present in the same orientation, allowing their subsequent misalignment during synapsis and hence facilitating illegitimate recombination (Sharp et al. 2006; Feuk 2010; Puig et al. 2015a). In addition, it has been proposed that the presence of an inversion in heterozygosis, and probably the recombination problems associated, somehow facilitate this process (Puig et al. 2015a).

Accordingly, higher inverted allele frequencies have been reported in the parents of the patients suffering from some of these disorders compared with the general population. Individuals affected by Angelman syndrome frequently display deletions of chromosome 15q11-q13 and an inversion allele involving this region was detected in 67% of mothers of patients, while only 9% of control subjects were carriers (Gimelli et al. 2003). Moreover, in another study it was found that all fathers of children affected by Sotos syndrome with a deletion in the paternally derived chromosome were heterozygous for an inversion in the 5q35 region (Visser et al. 2005). Similarly, heterozygous inversions involving the critical region were detected in transmitting parents of Williams-Beuren (Osborne et al. 2001; Hobart et al. 2010) and Koolen-de Vries (Koolen et al. 2006) syndromes with higher frequency than in the normal population.

Koolen-de Vries syndrome constitutes an interesting example in which the 17q21.31 inversion polymorphism facilitates the misalignment and abnormal recombination between flanking segmental duplications of the disease-critical region (Koolen et al. 2006). This disorder, also known as 17q21.31 microdeletion syndrome, is characterized by developmental delay, intellectual disability, hypotonia and recognizable facial features. The 17q21.31 locus is a genomically complex region with diverse structural configurations associated to direct H1 and inverted H2 divergent inversion alleles (Stefansson et al. 2005; Boettger et al. 2012; Steinberg et al. 2012). One of the potential inversion subhaplotypes, H2D, bears two duplications of

~150 kb placed in direct orientation and only carriers of this configuration are predisposed to the deletion of the entire region (500-650 kb) by NAHR (Steinberg et al. 2012) (Figure 1.5). This particular subhaplotype is rare in Asians and Africans, while it is at high frequency (~18%) in European populations, indicating that individuals with European ancestry are at much higher risk (Boettger et al. 2012). Consequently, virtually all 17q21.31 microdeletion syndrome cases reported in the scientific literature have been developed in individuals of European origin (Mefford et al. 2009; Cooper et al. 2011). The estimated prevalence of 17q21.31 microdeletion syndrome is around 0.64% of the individuals with idiopathic mental retardation (Koolen et al. 2008). Likewise, Williams-Beuren syndrome is a rare disorder caused by a hemizygous deletion of ~1.5-1.8 Mb on chromosome region 7q11.23, which encompasses 28 genes (Bayes et al. 2003; Shubert et al. 2009; Merla et al. 2010). Typical symptoms are mental retardation, vascular stenosis, visual problems, overfriendliness, short stature and specific facial characteristics. The deleted region is flanked by blocks of segmental duplications, which gave rise to a known polymorphic inversion of this region by NAHR. About 30% of transmitting parents are carriers of the inverted chromosome, whereas the inversion is only in 5% of the general population individuals. As already mentioned, it has been proposed that interchromosomal pairing in meiosis of the non-inverted and inverted chromosome would produce a loop structure, where an unequal crossover between segmental duplications could lead to the subsequent deletion and a reciprocal duplication (Figure 1.5). In general, both syndromes are assumed to be caused by haploinsufficiency for the deleted genes. For instance, elastin gene haploinsufficiency causes arterial stenosis, which is a component of Williams-Beuren syndrome (Metcalf et al. 2000). In the case of the 17q21.31 deletion phenotype, the implicated region includes five known protein-coding genes, but affected individuals with *de novo* loss-of-function mutations in *KANSL1* display most of the clinical signs, suggesting that the deletion of one copy of this gene is sufficient to cause the disease (Koolen et al. 2012; Zollino et al. 2012). However, we cannot discard that the alteration of the other elements residing within the deleted region contribute to exacerbate the phenotype.

Further examples of chromosomal rearrangements promoted by inversions include translocations. For instance, Y chromosomes carrying a 4 Mb inversion are more prone to becoming involved in a translocation between

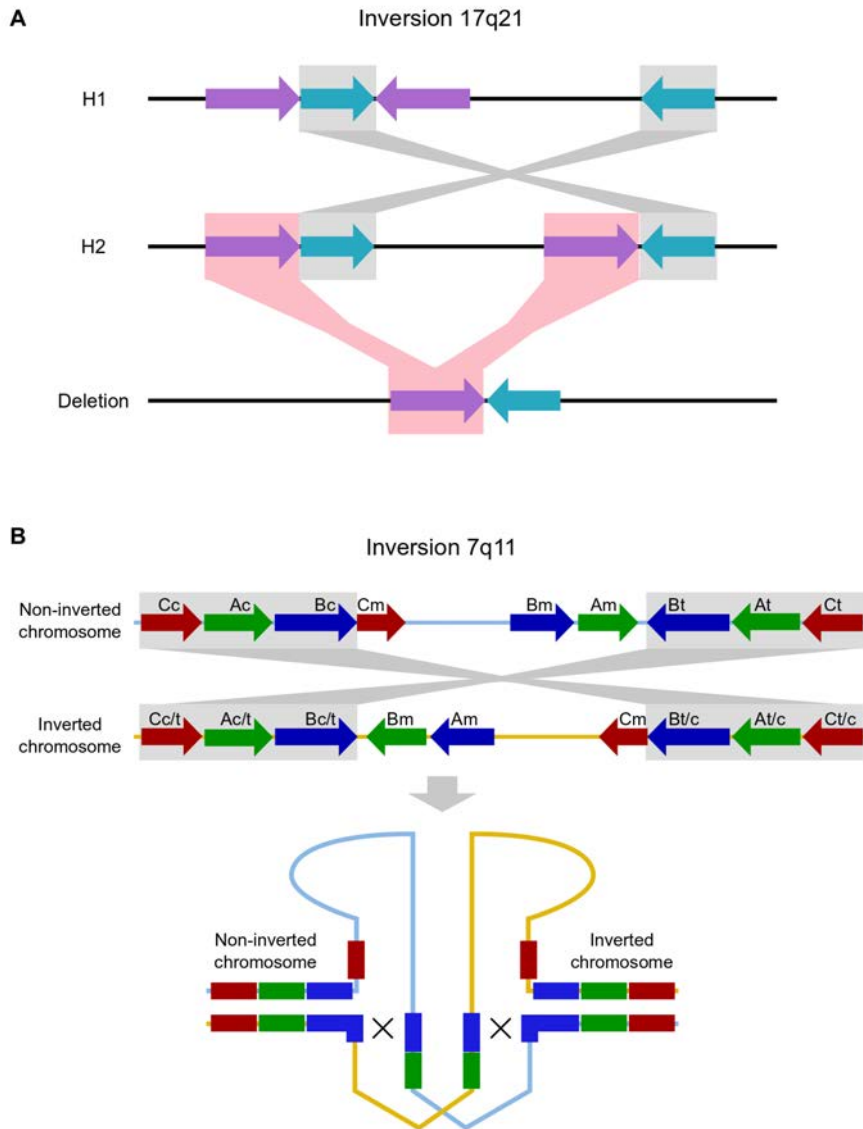


Figure 1.5 – Inversions predisposing to Koolen-de Vries and Williams-Beuren syndromes. Caption on the following page.

Figure 1.5 – Inversions predisposing to Koolen-de Vries and Williams-Beuren syndromes. (A) Simplified scheme of the genomic architecture at 17q21.31 (actual 17q21.31 inversion rearrangement is more complex). The H2D subhaplotype is present in $\sim 20\%$ of the European population, in which flanking segmental duplications are directly oriented and can undergo a deletion rearrangement via NAHR. The deletion in 17q21.31 may be responsible for up to 1% of cases of mental retardation. (B) Diagram of the pairing of an inverted and non-inverted chromosome that produce a loop in the 7q11 region, where a misalignment of segmental duplication blocks and an unequal crossover would lead to gametes with deletions and duplications. Segmental duplications are depicted as arrows, and distinct units are showed in different colours.

homologous genes *PRKX* and *PRKY* on chromosome X and Y, respectively, leading to sex reversal and infertility (Jobling et al. 1998). Among other disorders caused by secondary rearrangements associated to an inversion are the 3q29 microdeletion syndrome, Wolf-Hirschhorn syndrome, 15q13 microdeletion syndrome, 15q24 microdeletion syndrome, Emery-Dreifuss muscular dystrophy and RCAD syndrome (Puig et al. 2015a). Nonetheless, these findings must be taken with extreme care given the high frequency of many of these inversion polymorphisms in the population and the low prevalence of such disorders, suggesting a very low increase in the probability of having affected offspring by an inversion carrier.

1.1.4 Role of inversions in local adaptation and complex phenotypes

Substantial empirical evidence shows that inversions are common across many species, and are associated with fascinating phenotypes, such as diverse mating systems or social behaviours, as well as play a role in climate adaptation and ultimately speciation (Hoffmann and Rieseberg 2008; Wellenreuther and Bernatchez 2018). Although most human inversions

are small (<10 kb) (Feuk et al. 2005; Puig et al. 2015a), this is not the rule in other species of animals and plants, where large inversions spanning several megabases are common. Consequently, such inversions may cover a large proportion of the genome and affect a great number of genes. As already mentioned, by limiting local recombination inversions can act effectively as supergenes (Thompson and Jiggins 2014). A supergene is considered a group of two or more genes and other functional elements at a single chromosomal region that are jointly inherited and that simultaneously affect multiple complex characteristics. Supergenes segregate in a simple Mendelian manner and have been reported to generate intricate phenotypic systems and underlie adaptation in various species. Indeed, the change in the activity of a single protein-coding gene is unlikely to regulate many phenotypic traits. Since reduced recombination within supergenes is crucial to maintain the block of multiple genes as a single unit, inversions are a central element in evolution of supergenes and have been reported to harbour supergene complexes (Thompson and Jiggins 2014; Wellenreuther and Bernatchez 2018).

Well-known examples of supergenes come from mimetic wing forms of butterfly species, where inversions have been observed to control colour patterns. Mimicry in *Papilio polytes* and *Heliconius numata* has been shown to be caused by an inversion that includes the gene *doublesex*, key in insect sexual dimorphism (Kunte et al. 2014), and different inversion rearrangements at the supergene locus *P* (Joron et al. 2011), respectively. Likewise, inversions are associated to a broad range of complex phenotypes such as migratory ecotypes and social forms. Supergene examples related to migratory phenotypes include the Atlantic cod *Gadus morhua* (Berg et al. 2016; Berg et al. 2017), the salmonid *Oncorhynchus mykiss* (Pearse et al. 2014), and the willow warbler *Phylloscopus trochilus*, where large haplotype blocks are consistent with the presence of inversions (Lundberg et al. 2017). On the other hand, a large non-recombining chromosomal region in the Alpine silver ant *Formica selysi* genome has been linked to social organization, which can derive in different social forms with variation in queen number. While ant colonies are typically headed by one queen, the workers with a specific inversion orientation can tolerate several

fertile queens, which has been observed that can be advantageous. This nonrecombining “social chromosome” has convergently evolved in a similar separate evolutionary lineage for the red imported fire ant *Solenopsis invicta* (Purcell et al. 2014; Linnrechy and Kronauer 2014). Finally, reproductive behaviours are also susceptible to be influenced by chromosomal inversions. Two remarkable examples involve the ruff *Philomachus pugnax* (Küpper et al. 2016; Lamichhaney et al. 2016), a medium-sized and lek-breeding wading bird that breeds in wetlands across northern Eurasia, and the white-throated sparrow *Zonotrichia albicollis* (Tuttle et al. 2016), a common songbird in North America (Figure 1.6).

On the other hand, a large body of literature supports that inversions are strongly involved in climatic adaptation. Canonical examples come from geographical clines (i.e. adaptation to altitudinal and latitudinal gradients) in multiple species of *Drosophila*. A prominent case is the inversion 3RP in *D. melanogaster*, with environmental clines on several continents that shift in parallel according to climatic change (Krimbas et al. 1992; Anderson et al. 2005). Other cases of inversion clines can be found in *Anopheles* mosquito species (Ayala et al. 2017) and maize (*Zea mays*) (Fang et al. 2012). Another typical example of adaptive response to climate change is the association between frequency changes of *D. subobscura* chromosomal inversions and increased temperature due to global warming (Balanyá et al. 2006; Rodríguez-Trelles and Rodríguez 2007). While during environmental adaptation mechanisms, reproductive barriers may emerge, leading to speciation. In this regard, a chromosomal inversion distinguishes alternative annual and perennial ecotypes in the yellow monkeyflower *Mimulus guttatus*, with adaptive effects on flowering time and growth-related traits that give rise in turn to isolating barriers (Lowry and Willis 2010).

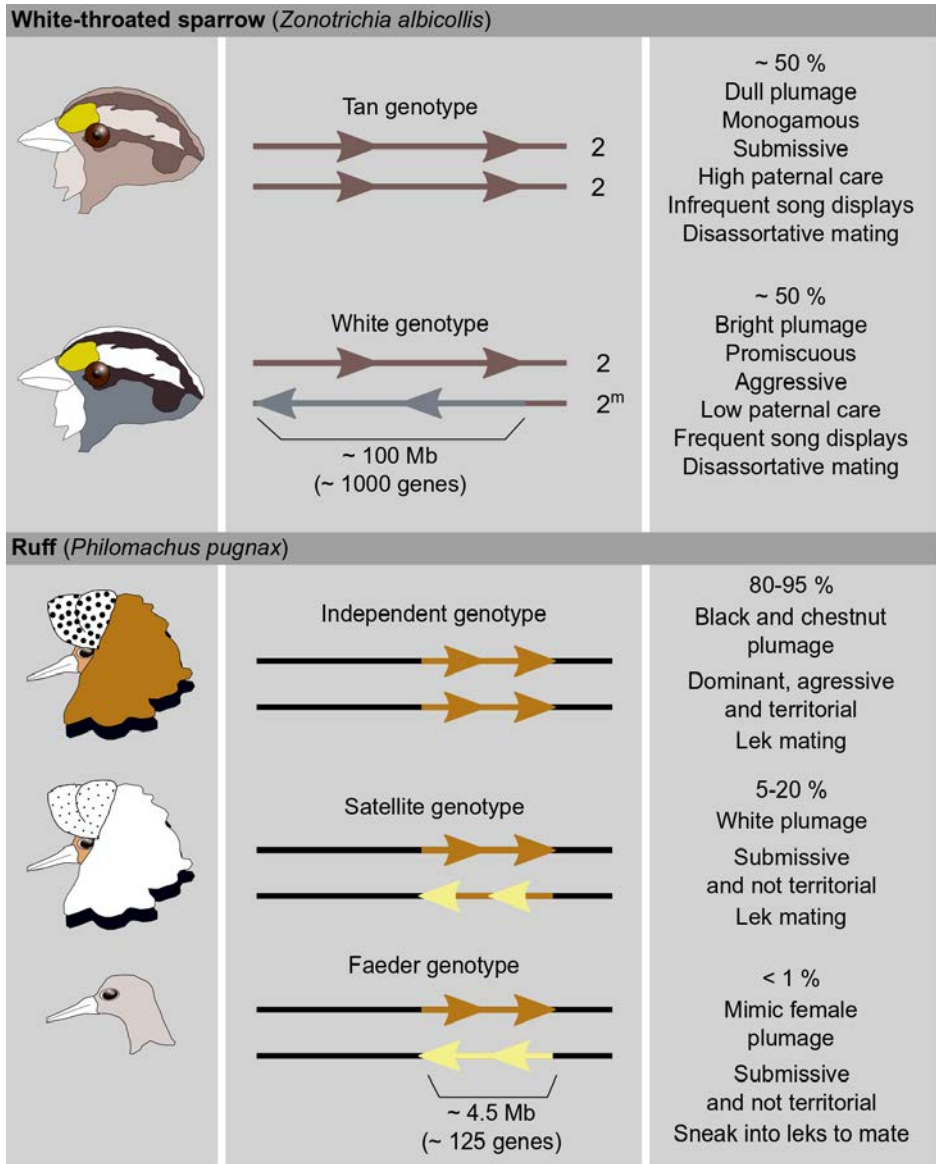


Figure 1.6 – Inversion effects on mating systems in the white-throated sparrow and the ruff. Caption on the following page.

Figure 1.6 – Inversion effects on mating systems in the white-throated sparrow and the ruff. Well described effects in the two colour morphs of the white-throated sparrow and the three alternative male mating types of the ruff. In the white-throated sparrow a ~ 100 Mb pericentric inversion in chromosome 2 ($\sim 10\%$ of its genome) controls behaviour and two color morphs: white-striped and tan-striped plumage. White males are heterozygous for the inverted form and tend to have more mates, but tan males, which carry only non-inverted chromosomes, are monogamous and invest in parental care. A similar situation can be found in the ruff, in which a ~ 4.5 Mb inversion on chromosome 11 encodes for three alternative male forms -independents, satellites and faeders-, which differ from each other in mating strategy, size, plumage, physiology, testis size, and behaviour. Independent males (most common in the population) with ornamented dark or chestnut plumage are dominant and aggressive, and carry two copies of the ancestral, noninverted chromosome. Satellite males (less common) with white ruffs are submissive, while faeder males (rare), mimic females in their size and plumage and sneak into the leks of independent males to steal copulations. Satellites and faeders carry distinct supergenes: the faeder supergene is a product of an inversion of the ancestral chromosome, whereas a rare recombination between the ancestral and the faeder supergene originated the satellite version (Lamichhaney et al. 2016). Each inverted region contains from 125 (ruff) to over 1000 (sparrow) genes, all of which are highly differentiated from their respective non-inverted haplotypes and divergent alleles apparently drive phenotypic differences between morphs. Satellite and faeder alleles in the ruff and chromosome 2^m in the white-throated sparrow are lethal or highly deleterious when homozygous.

1.1.5 Current strategies for inversion discovery and genotyping at large scale

The recent introduction of sequencing techniques has demonstrated that they are suitable genomic approaches to identify inversions at a large scale. Strategies aimed to discovery inversions genome-wide have been originally

based on the comparison of the reference human genome with the genome of non-human primates (Feuk et al. 2005; Kronenberg et al 2018; Catachio et al. 2018) or with other *de novo* assembled human samples (Levy et al. 2007). This was initially a useful approach to identify inverted segments between two organisms or individuals, and is one of the most exhaustive and precise strategies to obtain detailed information of structural variation along the whole genome. Nevertheless, it is quite expensive since it involves whole-genome sequencing and large segmental duplications at inversion breakpoints often complicate the *de novo* reconstruction of the genome assembly. Perhaps, one of the most successful methods for inversion detection developed so far is based on the systematic comparison of the human reference genome to a second genome represented by a set of paired-end sequences (Tuzun et al. 2005). The resulting sequenced reads from the two ends of a cloned fragment coming from an individual sample are expected to align to the reference genome in an appropriate orientation (e.g. plus(+)/minus(-)) and at a specific distance; if not, this could be indicating the presence of a structural variation between both genomes. This general protocol was adapted to next-generation sequencing platforms, and is currently known as paired-end mapping (PEM), resulting in a major progress in the field of inversion identification (Korbel et al. 2007; Kidd et al. 2008; Ahn et al. 2009; Sudmant et al. 2015). However, PEM is not free of limitations, since it relies heavily on the reference genome, which might contain miss-assembled regions that could lead to erroneous calling of structural variants (Vicente-Salvador et al. 2017). Another problem that should be outlined is that PEM often fails when inversions are flanked by long inverted repeats of high identity, as happened when reconstructing *de novo* assemblies, since reads from these regions cannot be uniquely mapped (Lucas-Lledó and Cáceres 2013). For these inversions, target-based assays are thereby needed (Feuk, 2010; Puig et al, 2015a).

Novel systems aimed to identify structural variation in the human genome are continuously being developed, and some promising studies published recently offered the ability to handle balanced events, such as inversions. For example, the invention of optical mapping allowed the detection of SVs as differences in restriction maps from single molecules of labeled

DNA scanned by fluorescence microscopy (Teague et al. 2010). In a similar manner, BioNano Genomics technology uses fluorescence imaging in single DNA molecules to construct complete genomic maps on nanochannel arrays (Lam et al. 2012), and it has been used to predict many inversions (Li et al. 2017a, Levy-Sakin et al. 2019). On the other hand, long-read sequencing can provide insights into inversion polymorphism as well by sequencing through longer repetitive elements (Chaisson et al. 2015; Huddleston et al. 2017; Shao et al. 2018; Audano et al. 2019). Also, Strand-seq is a single-cell technique based on the sequencing of a particular strand of DNA inherited by daughter cells that have incorporated 5-bromo-2'-deoxyuridine (BrdU) during DNA replication to allow the selection of the BrdU-negative template strand (Falconer et al. 2012). Sanders and colleagues (2016) applied this method to detect more than 100 human polymorphic inversions, which can be identified as segments of strand switches in Strand-seq libraries. Although this is an interesting application, it is quite labor intensive and the size of inversions detected are usually big, resulting in that small inversions would be still hidden. Finally, the combination of several of these techniques can increase the power and overcome the inherent limitations of each of them to maximise the capacity of discovering inversions. In particular, a recent study integrated diverse genomic methods, including short- and long-read, strand-specific sequencing technologies and optical mapping for SV detection in three parent-child trios, and was able to identify 308 inversions with relatively good accuracy (Chaisson et al. 2019).

Bioinformatic solutions for inversion discovery have been developed as well. The existence of strong LD blocks in the genome, sequences of several kilobases in which variants are strongly correlated with each other, allow that many genetic variants can be tagged with only a fraction of genotyped markers. This phenomenon is employed by imputation softwares like IMPUTE2 (Howie et al. 2008) to infer the state of many polymorphisms not typed in GWAS arrays, for instance. In addition, specific programs for inversion identification and genotype calling based on signatures in nucleotide variation patterns (Bansal et al. 2007; Sindi and Raphael 2010; Ma and Amos 2012) have been also proposed, such as PFIDO (Salm et

al. 2012), *invClust* (Cáceres et al. 2015) or *scoreInvHap* (Ruiz-Arenas et al. 2019). In this last case, inversion status is inferred from similarity to orientation specific haplotypes determined experimentally. However, all these methods have limitations and it is unclear how accurate are these genotypes are, especially for recurrent inversions, which are unlikely to be captured by any combination of linked genetic variants.

It is important to note that inversion discovery is just the first step and despite all these efforts, it is still not clear how many common inversions exist in the human genome, what the size distribution of inversions variants is, and to what extent inversions are associated with human disorders (Puig et al. 2015a). In addition, there is still very little information about population frequencies and worldwide distribution of human polymorphic inversions. Therefore, it is necessary to have methods to validate the genome-wide predictions and genotype them in multiple individuals. For instance, one of the first techniques used for inversion validation was fluorescence *in situ* hybridization (FISH), which is a molecular cytogenetic approach that uses fluorescent probes designed to bind complementarily particular genomic regions. This is useful to detect chromosomal inversions by using two or three probes labeled with different colors placed inside and outside the inversion. Thus, depending on the inversion orientation, the distance and order of the probes will vary (Antonacci et al. 2009; Bosch et al. 2009; Hobart et al. 2010; Nomure et al. 2018). Other traditional molecular approaches that have been used to detect inversions also comprise southern blot hybridization (Small et al. 1997) or pulsed-field gel electrophoresis (Osborne et al. 2001). All these techniques can be very accurate to test variants in a particular genomic localization, but neither of them is scalable to genotype inversions on a population scale or in the large cohorts necessary to perform association studies. Besides, the low resolution of FISH limits the analysis to large inversions (>1 Mb). Other molecular biology techniques like PCR can amplify DNA segments enclosed by primers around inversion breakpoints, which makes possible to design orientation-specific primer combinations. Also, a modified PCR protocol can be applied to inversions with long inverted repeats at the breakpoints, called inverse PCR (iPCR). Therefore, targeted inversions

can be assayed by PCR in larger number of samples (Aguado et al. 2014; Puig et al. 2015a; Vicente-Salvador et al. 2017). In the near future, with the introduction of new powerful high-throughput techniques to screen for inversions in an unbiased manner, a complete picture of human inversion polymorphism is expected to be finally available.

1.1.6 The InvFEST project

As a consequence of the rising interest in structural variants and the challenge that inversions represent among them, the InvFEST (INVersion Functional & Evolutionary Studies) Project started in 2010 with the aim of improving our little knowledge about this type of variation in humans. The project attempts to characterise the population distribution and functional consequences of polymorphic inversions present in human populations and aspires to become an indispensable complement to the Database of Genomic Variants (dgv.tcag.ca) (MacDonald et al. 2014).

We have already commented that early studies predicted few hundred inversions by using genome assembly comparison and PEM (Korbel et al. 2007; Levy et al. 2007; Kidd et al. 2008). Thus, these initial inversion predictions were combined into a unified non-redundant set and classified depending on their reliability according to an intense curation effort (Martínez-Fundichely et al. 2014; Martínez-Fundichely et al. in prep.). Nonetheless, the repetitive nature of the human genome gives rise to a high proportion of false positives when using these methods (Martínez-Fundichely et al. 2014). Indeed, an exhaustive sequence analysis and experimental validation using optimized PCR-based techniques for genotyping inversions, such as iPCR, have allowed to determine that a large fraction of predicted inversions are false positives, but also to expand considerably the number of experimentally confirmed polymorphic inversions in humans, to more than 200 nowadays (Martínez-Fundichely et al. 2014; Aguado et al. 2014; Vicente-Salvador et al. 2017). All this information is stored in the InvFEST database (invfestdb.uab.cat) (Martínez-Fundichely et al. 2014), which constitutes an ongoing effort to provide the most re-

liable and largest collection on human inversions by integrating multiple sources of available data. It is worth pointing out that there is a notable variation in terms of size, ranging from ~ 100 bp to dozens of Mb, although many larger inversions (>10 kb) are probable false positives (Puig et al. 2015a). Also, the longer the inversion, the more likely to be flanked by segmental duplications, suggesting differences in the mechanisms of origin (Puig et al. 2015a). Given that PEM strategies have been demonstrated to be unsuccessful at detecting inversions flanked by large segmental duplications (Lucas-Lledó and Cáceres 2013), the database is continuously updated with the new studies covering novel technologies aimed to detect inversions (Sanders et al. 2016; Huddleston et al. 2017; Shao et al. 2018; Levy-Sakin et al. 2019; Audano et al. 2019; Chaisson et al. 2019).

The InvFEST project has already produced a big advance in our knowledge of human inversions. In fact, a subset of validated inversions have been genotyped in a large number of individuals (Aguado et al. 2014; Vicente-Salvador et al. 2017). Specifically, information about mechanisms of origin, breakpoint definitions, worldwide frequency, ancestral state and evolutionary history has allowed us to get a global picture of each inversion (Aguado et al. 2014; Vicente-Salvador et al. 2017). Also, these studies have found that recurrent inversions seem to be common in humans -that is, an inversion event can appear several times across the human lineage due to NAHR between the IRs at the breakpoints- and even a low recurrence rate would reduce LD between inversions and nearby variants, which limits their genotyping with tag SNPs (Antonacci et al. 2009; Aguado et al. 2014; Vicente-Salvador et al. 2017).

The potential functional consequences of inversions in different organisms, including humans, highlight the importance of studying inversions in biomedical research (Puig et al. 2015a). However, only three polymorphic inversions in humans -17q21.31, 8p23.1 and 16p11- had been studied in some detail (Stefansson et al. 2005; Myers et al. 2005; Zabetian et al. 2007; Webb et al. 2008; de Jong et al. 2012; Salm et al. 2012; Okbay et al. 2016; González et al. 2014). Besides the well-studied 17q21.31 inversion, inversion 8p23 has been related to autoimmune diseases and personality

traits (Salm et al. 2012; Okbay et al. 2016) and 16p11 inverted allele appears to protect against the joint susceptibility of asthma and obesity and correlates with expression levels of neighboring genes (González et al. 2014). In this regard, the InvFEST project has given us the opportunity to start assessing the functional role of human inversions in a systematic manner. For instance, the accurate annotation of inversions has made possible to identify several cases in which inversions affect gene sequences in different ways (Aguado et al. 2014; Vicente-Salvador et al. 2017). An interesting example includes the East-Asian specific inversion HsInv0379 that disrupts a zinc-finger protein gene and creates a novel fusion transcript (Puig et al. 2015b). However, the current knowledge about the contribution of human polymorphic inversions to gene expression and phenotypic variation is still limited and most cases have not been investigated in detail yet. For instance, the 1000GP did not find any inversion as lead eQTL of expression changes in LCLs. Additionally, a comprehensive analysis of the effect of structural variants on gene expression in several tissues could only associate a 198-bp inversion (HsInv0191 in InvFEST) with the expression of genes *P4HA1* in skin and *FUT11* in nerve tissues, but these signals were lost in a joint eQTL analysis in favour of other genetic variants. Therefore, inversion effects have remained elusive. For that, it is necessary direct genotyping methods that provide reliable genotypes of human inversions in large cohorts. As we will see in the next chapters of this thesis, new experimental high-throughput assays have been developed, allowing us to carry out an complete characterisation of the functional impact of human inversions for the first time (Giner-Delgado et al. 2019; Puig et al. 2019).

1.2 Human genome function

A central goal of human genomics is to interpret the functional consequences of millions of discovered genetic variants, including inversions. The understanding of these potential effects is key to evaluate their contribution to disease susceptibility, progression or treatment effective-

ness. Much of our initial knowledge derived from the study of monogenic “Mendelian” diseases, where the disease-causing variants and the protein-coding genes in which they reside were successfully identified through linkage mapping in human pedigrees (Botstein and Risch 2003). However, this is not true for most common diseases, which involve contributions from multiple genes as well as environmental and lifestyle factors. In the past few years, the method of choice for the study of complex phenotypes has been to carry out GWAS in large cohorts of cases and controls. Thanks to GWAS efforts, thousands of genetic variants have been linked to human traits and diseases.

Nonetheless, GWAS merely uncover and list statistical associations, but this data does not provide biological insights about the mechanisms underlying the respective phenotype *per se*. Indeed, the majority of these association studies do not go beyond reporting the most significant loci. Moreover, discovered variants only confer relatively small increments in risk and are mostly located at non-coding regions. Therefore, our ability to predict the functional implications of such variants is just speculative and is based on available annotation of genes in the vicinity of GWAS signals. To solve this challenge, a large number of molecular traits can be potentially interrogated, including gene expression, DNA methylation, histone modifications or chromatin conformation. These “intermediate” phenotypes at cellular level allow to establish which gene functions are affected by the variants implicated in disease. This approach helps to elucidate the processes by which a genetic variant has a functional impact and links cellular states with whole-organism phenotypes.

1.2.1 Genetic determinants of gene expression variation

While proteins are the chief actors in carrying out cellular functions, measuring protein levels in a precise and high-throughput manner is technologically challenging, although some progress has been made (see below; Enroth et al. 2014; Battle et al. 2015; Sun et al. 2018). In contrast, mRNA levels are easier to measure genome-wide and most current stud-

ies typically use transcript rather than protein abundance as a proxy for gene expression. Until a decade ago, gene expression profiling had been mainly performed using microarrays that contained one or a few probe sets for each interrogated gene. However, the quantification of previously un-annotated transcripts or alternative isoforms of the same gene was impossible. With the advent of next-generation sequencing methods, the quantification of transcript levels by massive sequencing of RNA molecules (RNA-Seq) has become the preferred method to detect and measure transcriptome-wide gene expression. RNA-Seq can accurately quantify low expressed genes and is not limited to known genes (Marioni et al. 2008; Sun et al. 2013; T’Hoen et al. 2013; Stark et al. 2019). For example, the expression of more than 4,000 un-annotated transcript were reported in a RNA-Seq study of the Nigerian HapMap population (Pickrell et al. 2010). Furthermore, RNA-seq profiles of monocyte-derived dendritic cells uncovered thousands of novel isoforms synthesized in response to influenza infection (Ye et al. 2018).

Thanks to these genomic methods, there has been tremendous progress in identifying genetic variants that affect gene expression, termed expression quantitative trait loci (eQTLs). In other words, an eQTL is a genetic variant that explains part of the expression variance of a gene; that is, the intermediate phenotype tested is the gene expression level, and a direct association between genetic variation and gene expression is measured (Figure 1.7). The most associated variant is usually called lead, top or sentinel eQTL, and is more likely to be the causal variant affecting expression (Brown et al. 2017). However, LD structure makes difficult to identify the true causal variants, since many eQTL signals will be correlated. This association analysis, also termed as eQTL mapping, can involve variants located either in *cis*-proximal to the gene- or in *trans* -distally-. On the one hand, it is common to test variants within 1 Mb of the transcription start site (TSS), since very few variants beyond this distance may influence expression. Indeed, most eQTLs lie either within gene bodies or close to TSSs (Veryerias et al. 2008; Stranger et al. 2012). Also, only 25% of top eQTLs were located at a distance higher than 50 kb from the TSS in a large-scale study in lymphoblastoid cell lines (LCLs) (Lappalainen et al.

2013). On the other hand, *trans*-eQTLs have been much more difficult to detect, because interrogating the whole-genome for potential effects has proven statistically challenging. Nonetheless, when a reasonable sample size is employed, *trans*-eQTLs can be reliably found (Grundberg et al. 2012; Westra et al. 2013; Fairfax et al. 2014, GTEx Consortium 2017; Bonder et al. 2017; Vösa et al. 2018). It is worth mentioning that genes can be regulated by multiple eQTLs. In this sense, conditional analysis, in which the eQTL mapping is conditioned on the lead variant to test any other significant variant, can lead to the discover of secondary independent eQTLs (Jansen et al. 2017).

First eQTL studies in humans explored gene expression in LCLs (Stranger et al. 2007a; Stranger et al. 2007b; Montgomery et al. 2010; Pickrell et al. 2010; Stranger et al. 2012; Stranger et al. 2007a; Stranger et al. 2007b; Stranger et al. 2012; Lappalainen et al. 2013). In this regard, the most well-known work is probably from the GEUVADIS (Genetic EUropean VARIation in DISease) Consortium (Lappalainen et al. 2013). This project contributed to much of the knowledge available to that moment about quantitative and qualitative transcriptome variation in human populations with the creation of one of the most comprehensive transcriptome reference datasets, given the already published genome sequences of these individuals by the 1000 Genomes project (Lappalainen et al. 2013). The authors analysed 462 individuals and reported 8,329 genes with eQTLs, including expression levels and splicing, but also miRNA-mRNA interactions, fusion genes and RNA editing, giving a precise view of the spectrum and functional mechanisms of regulatory variation. Additional early studies also focused on other accessible cell types, such as skin or blood (Grundberg et al. 2012; Battle et al. 2014; Buil et al. 2016).

Importantly, an increasing number of studies have shown that many of GWAS signals are enriched for eQTLs (Nicolae et al. 2010). For example, schizophrenia susceptibility alleles are enriched for eQTLs in brain (Richards et al. 2012), whereas type 2 diabetes SNPs affect gene expression in liver and adipose tissues (Zhong et al. 2010). This enrichment might not be so surprising if we take into account that a GWAS variant

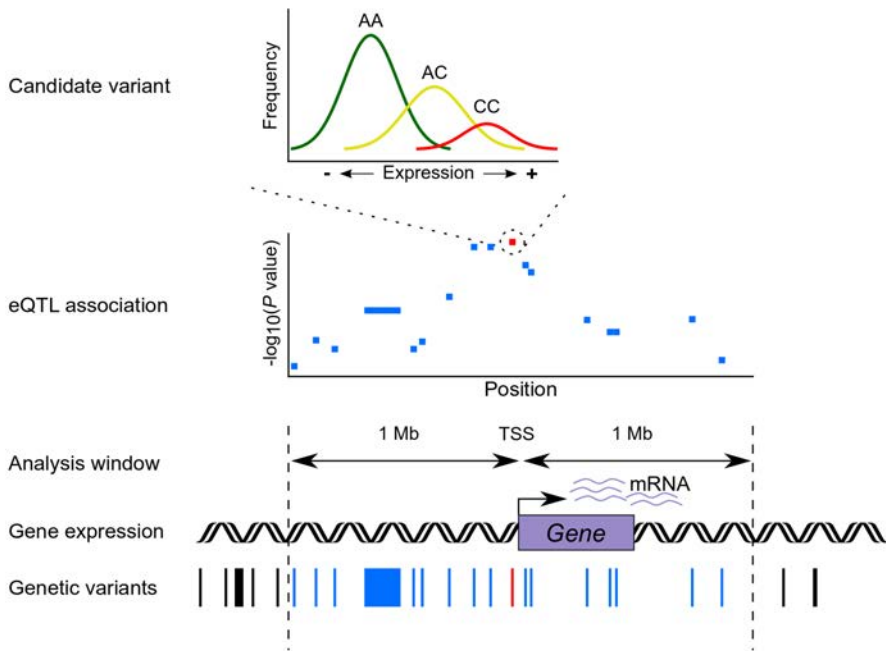


Figure 1.7 – Identification of regulatory variants through eQTL analysis. From bottom to top. Many SNPs and other genetic variants are tested against molecular phenotypes, such as gene-expression levels. Variants of interest that are acting in cis on gene expression are detected by focusing on a defined genomic window around the transcription start site (TSS) (1 Mb in this case). Also, all variants can be tested in a genome-wide manner to discover those acting in trans. Manhattan plot shows the statistical significance of the tested variants along their genomic coordinates. The graph above displays the distribution of gene-expression levels stratified by the three genotypes of the most significant variant.

can act as an eQTL and affect the expression of a gene that in turn influences the disease. For instance, eQTLs discovered in LCLs helped to explain a GWAS association between the 17q21 locus and asthma susceptibility. This region contains at least 15 genes and none of them had evident roles in the disease at that time (Moffatt et al. 2007). However, transcript levels of *ORMDL3* were strongly associated with the top GWAS signal, suggesting the role of this gene as the leading candidate

in the aetiology of the disease. In fact, further functional studies confirmed this finding (Ono et al. 2014). eQTLs found in LCLs have been also useful to determine candidate genes in Crohn’s disease (Libiollé et al. 2007), autism (Nishimura et al. 2007) or bipolar disorder (Iwamoto et al. 2004). Although some variants are expected to act in a tissue- or condition-restricted manner, recent data suggest that many eQTLs are shared across tissues and thus these first findings in LCLs were useful to interpret GWAS results (Nica et al. 2011; Grundberg et al. 2012; Nica and Dermitzakis 2013; GTEx Consortium 2017). However, it is also well known that genes are differentially expressed and regulated across distinct conditions and tissues, and many only show activity in specific cell lines (Grundberg et al. 2012; GTEx consortium 2017), under certain condition such as infection (Nedelec et al. 2016; Quach et al. 2016), or in a particular developmental stage (Cardoso-Moreira et al. 2019). In these cases, analysing eQTLs in a relevant tissue or context is essential to evaluate the functional impact of a specific SNP.

The Genotype-Tissue Expression (GTEx) project was founded to overcome such restrictions with the aim of measuring the functional impact of genetic polymorphisms on human transcriptomes by enabling the study of eQTLs across diverse tissues of the human body (GTEx consortium 2015). For that, RNA was isolated from postmortem samples derived of multiple tissues from donors enrolled in the study. In total, the GTEx v6p analysis freeze contains data from 44 tissues from 449 donors, including 31 solid organ tissues, whole blood, two cell lines derived from blood and skin samples (lymphoblastoid cell lines and fibroblast cultures), and 10 brain regions. Sampled donors were mostly from European ancestry (83.7%) while the next largest group was African Americans (15.1%). For all individuals, DNA was genotyped from blood samples using arrays and quantification of gene expression in the different samples was performed through massively-parallel sequencing of RNA. Therefore, the GTEx project provides a comprehensive landscape of gene regulation across a wide variety of tissues (GTEx consortium 2017).

The GTEx project has taught us several lessons. First, pervasive *cis*-

eQTL effects affecting the majority of human genes were found. In total, considering all GTEx tissues, more than 150,000 *cis*-eQTLs associated with almost 20,000 genes were discovered (GTEx consortium 2017). These numbers represent around 50% and 86% of all known autosomal long intergenic non-coding RNAs (lincRNAs) and protein-coding genes, respectively. Same conclusion has been obtained by other similar studies, such as V̄osa et al. (2018) analysis of 31,684 individuals, in which they detected that 88% of the genes expressed had a *cis*-eQTL in blood. This suggests that lowly expressed genes without associated eQTL so far could also have eQTLs in other contexts that remain to be identified. Second, overall, *cis*-acting eQTLs displayed two differentiated patterns with regard to tissue sharing, acting in either most of the 44 tissues or just in a specific and small subset of tissues. Third, in general, *trans*-eQTLs exhibited greater tissue specificity. In fact, other studies with fewer tissues already reported similar estimates of minimal sharing of *trans*-eQTLs (Grundberg et al. 2012; Buil et al. 2016). However, despite of the extensive tissue-specificity, some examples of *trans*-eQTLs shared across multiple tissues were observed. For instance, rs7683255 was associated with *NUDT13* in *trans* with a consistent effect direction across the majority of GTEx tissues. Additionally, rs60413914 was linked to *RMDN3* across the subset of brain-related tissues, but this *trans*-eQTL shows little or no effect in other tissues. Fourth, significant variants acting as eQTLs were enriched in promoters and enhancers. Also, eQTL activity was likely to be shared between a pair of tissues if this polymorphism was placed at a similar chromatin state in both contexts. Remarkably, secondary *cis*-eQTLs were enriched for chromosomal contact with target promoters according to Hi-C data. This result means that, in spite of being located further away from the transcriptional start site in the DNA linear sequence compared to primary eQTLs, both secondary and primary eQTLs are located physically close to regulated genes in the 3D space. On the other hand, functional characterization of *trans*-eQTL mechanisms revealed that *trans*-acting variants are enriched in enhancers, but not in promoter regions, consistent with tissue-specificity of enhancer activity. As we will see in the next sections, several studies support that *trans*-eQTLs are likely affecting the regulatory machinery (Degner et al 2012; Moen et al. 2013;

Waszak et al. 2015; Delaneau et al. 2019). Fifth, a strong evidence for regulation of genes in *trans* through expression changes of other genes in *cis* was also observed.

GTEX data has been also useful to boost further research in diverse areas. For instance, sex biases in the expression of chromosome X genes has allowed to establish a systematic catalogue of chromosome X inactivation across human tissues, pinpointing examples of variability in the degree of chromosome X inactivation escape that may contribute to phenotypic diversity (Tukiainen et al. 2017). Moreover, identification of individuals with extreme measures of gene expression has uncovered the presence of nearby conserved rare variants, demonstrating the contribution of rare genetic variation in large gene-expression changes and likely in disease risk as well (Li et al. 2017). Another study has reported the landscape of the dynamic patterns and regulation processes of adenosine-to-inosine RNA editing, illustrating editing trends across diverse tissues (Tan et al. 2017). Among other relevant studies, there are also the development of methods for transcriptome data analysis (Mohammadi et al. 2017), estimating the causal variants and tissues for both expression and complex traits (Brown et al. 2017; Ongen et al. 2017), the analysis of the impact of structural variation on gene expression (Chiang et al. 2017) and the characterization of regulation co-expression networks (Saha et al. 2017). Therefore, GTEX data has provided a critical and necessary resource for the scientific community that have aided in the interpretation of GWAS findings, making possible to identify regulatory variants acting in a tissue-specific manner and link them to human disease phenotypes.

In contrast to eQTL analysis addressing gene-expression profiles at steady-state, such as in the commented LCLs (Lappalainen et al. 2013) or post-mortem tissues (The GTEX Consortium 2017), recent studies have reported genetic variants associated to transcription variation in human cells challenged to immune antigens or infectious agents (Nedelec et al. 2016; Quach et al. 2016). Infectious diseases are among the most powerful selective driving forces in evolution. Indeed, immune-related genes and variants, which are enrolled in a vital combat against microorgan-

isms for host survival, have been reported to exhibit elevated adaptive evolution (Barreiro and Quintana-Murci 2010). Thus, being part of a complex phenotype such as our primary interface with the external world, immune defense mechanisms have been continuously shaped as humans faced contrasting pathogenic environments, leading to strong selective patterns across different populations. Either a disproportionate or irrelevant immune response can result in an inappropriate autoimmune or inflammatory disease or a mortal infection (Barreiro and Quintana-Murci 2010; Brinkworth and Barreiro 2014). Since there is a strong evidence that regulatory variants are involved in human traits and disease, it is consistent that many polymorphisms also play a significant role in immune response through changes in transcript and protein levels differences. In this regard, variants with effects specific of certain infections stimuli have been identified (Barreiro et al. 2012; Lee et al. 2014; Nedelec et al. 2016; Quach et al. 2016). Additionally, splicing QTLs affecting differential transcript isoform usage in the human antiviral response have also been found (Ye et al. 2018).

The study of pathogenesis of type 2 diabetes (T2D) illustrates the decisive value of targeting specific cell types for complex biomedical conditions of interest, instead of relying on eQTL data generated from easily accessible tissues or other tissues that can be considered as a substitute. First, thanks to RNA-seq data from subcutaneous adipose biopsies from 766 female twins it was possible to confirm that SNP rs4731702, which is within a T2D-associated haplotype, was an eQTL of the *KLF14* gene (Small et al. 2011). Despite *KLF14* is expressed in several tissues, *cis* and *trans* associations discovered were completely specific of adipose tissue, with no detectable signals in skin, whole blood or lymphoblastoid cell lines from the same subjects. Additionally, rs4731702-C allele was also associated with increased methylation levels around 3 kb upstream of *KLF14*. Again, this association was only present in subcutaneous adipose tissue and not in whole blood or skin. The *KLF14* regulatory variants also modulate, in *trans*, the expression of 385 genes, which consisted in a large adipose network linked to glucose uptake, lipogenesis and cell size. In fact, mice with adipose tissue-specific knockout of *Klf14* displayed insulin-resistant

phenotypes (Small et al. 2011). Second, previous studies have demonstrated that rs7903146, which is a major GWAS locus for T2D, influences chromatin accessibility and enhancer activity in islets, but they failed to characterize the influence on gene expression. A targeted analysis of 420 human islet preparations from the InsPIRE consortium (Viñuela et al. 2019) provided a detailed perspective of gene expression regulation in this tissue and the mechanisms underlying T2D predisposition. The authors of this study examined to which degree islet eQTLs overlapped with eQTLs reported in 44 tissues from the GTEx project. In this regard, 5% of genes discovered to be regulated in islets had no significant eQTLs in any of the 44 tissues, while none of the GTEx tissues could replicate more than 73% of them, indicating that no alternative tissue is completely useful to understand the genetic effects in islets. Remarkably, whole pancreas, which has been often employed as surrogate for its islet fraction, performed as a limited proxy for islets, and did not show any particular advantage. However, rs7903146 was found to be an eQTL for *TCF7L2* specific to islets (Viñuela et al. 2019), and no additional regulated genes could be found in other tissues. Thus, tissue-specific regulation plays a very important role. These findings indicate that, although many eQTLs are shared across cell types, tissue-specific regulation plays a very important role and must be taken into account to estimate the relative contribution of each tissue to a given trait and infer where genetic causality arises.

1.2.2 Gene activity is regulated by interaction with epigenetic states

Nearly all cells in multicellular organisms have the same genetic information. However, there are differences in the sets of genes that are expressed or inactivate, which determine heterogeneous cell types specialized to carry out distinct functions. This differential expression arises during embryonic development, in which pluripotent cells become differentiated cell lines. Thus, cellular functional diversity does not involve a change in the DNA sequence, but rather constitute what is called 'epigenetics'. Literally, epigenetics means "over the genetics", which states that these

changes are other than those encoded in the genetic sequence itself and may last through cell divisions. Examples of molecular events that mediate epigenetic phenomena include covalent histone modifications, cytosine methylation of DNA or chromatin structure. Although epigenetic regulation is essential for cell differentiation, the influence of environmental factors can also trigger epigenetic changes. Moreover, it is known that epigenetic modifications show inter-individual variation and can be directly affected by underlying genetic polymorphisms in the DNA sequence.

As we have mentioned before, we do not know yet how to satisfactorily interpret the effect of eQTLs located in non-coding regions of the genome, but many have been reported to modulate expression levels through regulatory changes that can be mediated by epigenetics. Indeed, genetic variants can alter chromatin structure and the recruitment of epigenetic regulatory enzymes, thereby affecting gene expression. Importantly, the identification of epigenetic changes associated to complex phenotypic traits may potentially be reversible and result in therapeutic targets. In this section, we review some of the key insights and novel ideas developed over last years about the epigenetic processes underlying eQTL function.

DNA methylation

Probably the most comprehensively studied epigenetic modification of DNA is cytosine methylation (Jaenisch and Bird 2003; Jones 2012; Tirado-Magallanes et al. 2017). DNA methylation is defined as the enzymatic process by which a methyl group is added through a covalent bond to the C-5 position of a DNA cytosine. DNA methylation is widely present at CpG-dinucleotides in mammalian genomes, although non-CpG methylation events also occur, especially in embryonic stem cells (Ramsahoye et al. 2000). CpG-dinucleotides are often found clustered in the so-called CpG islands, which are frequently found at regulatory regions. This epigenetic mark has typically been involved in repressing gene transcription. For instance, methylated promoters are associated with a maintained inactive gene state, such as in imprinted genes or in chromosome X inactivation

(Messerschmidt et al. 2014). Besides, aberrant DNA hypermethylation in promoters is associated with genome instability and oncogenesis, and methylation inhibitors have been used as a therapy for specific tumors such as myelodysplastic syndrome (Sato et al. 2017). However, evidences over the past few years confirm that the relationship between DNA methylation and gene expression is more complex and effects may vary (Wagner et al. 2014). Paradoxically, methylation in gene bodies seems to stimulate transcription (Hellman et al. 2007; Yang et al. 2014) and intragenic DNA methylation has also been associated with alternative promoter usage (Maunakea et al. 2010) and splicing (Shukla et al. 2011; Lev et al. 2015). Additionally, DNA methylation has been implicated in recruitment of transcription factors (TFs), silencing of transposable elements, chromosome stability and nucleosome positioning, cellular proliferation, differentiation, pluripotency, embryonic development, aging, drug response and disease (Jaenisch and Bird 2003; Jones 2012; Tirado-Magallanes et al. 2017).

5-methylcytosine (5mC) levels are known to vary in the population. Moen and colleagues (2013) analysed DNA methylation in 133 LCLs derived from individuals of European and African ancestry, finding that a substantial proportion of differences in cytosine modifications could be explained by local genetic variation (methylation quantitative trait loci (mQTL)). Since then, different large-scale studies with hundreds or even thousands of samples and multiple tissues have tried to uncover the genetic architecture underlying DNA methylation both in *cis* and *trans* (Gibbs et al. 2010; Gamazon et al. 2013; Hannon et al. 2016b; Gaunt et al. 2016; Bonder et al. 2017; McRae et al. 2018). As happened with gene expression, differential methylation patterns have been associated to distinct human diseases, such as diabetes (Davegardh et al. 2018), body-mass index (Wahl et al. 2017) or neurological disorders like autism (Hannon et al. 2018), schizophrenia (Hannon et al. 2016a) or Alzheimer's disease (Lunnon et al. 2014). There are different scenarios in which genetic variants can impact DNA methylation. For example, SNPs within epigenetic regulatory enzymes, such as DNA methyltransferases, may alter their function, affecting genome-wide epigenetic patterns. Also, SNPs in TFs or located at DNA

motifs may impact the binding and recruitment of proteins implicated in the methylation of a particular locus (Wang et al. 2019a; Wang et al. 2019b). An interesting example of the interplay among genetic variation, DNA methylation and gene expression leading to disease susceptibility is the effect of the intronic SNP rs3774937, which is associated with ulcerative colitis. The minor allele of rs3774937 is associated with higher levels of *NFKB1* expression in *cis* and differential DNA methylation at hundreds of distal CpG sites, many of which regulate further expression changes (Bonder et al. 2017).

Several technologies have been developed to study DNA methylation patterns on a genome-wide scale. One of these methods consists in immunoprecipitation of denatured methylated DNA fragments using an antibody to 5mC, which are subsequently sequenced (Weber et al. 2005). Another interesting approach is the use of the Illumina Infinium HumanMethylation450 BeadChip (450K array) for high-throughput profiling of 5mC levels across the human genome by analyzing bisulfite-treated DNA (Bibikova et al. 2011). This high density microarray can assay over 480,000 cytosine positions, the majority CpG dinucleotides, providing a comprehensive coverage across CpG islands, promoters and gene body regions. In fact, many alternative approaches take advantage of bisulfite conversion as central procedure to map 5mC. Treatment of DNA with bisulfite converts cytosines (C) to uracils (U), which are subsequently amplified as thymines (T), while 5mCs are resistant to conversion, allowing to detect the C to T difference by hybridization with different probes or sequencing. In this regard, the highest coverage at single base-pair resolution to identify 5mC is achieved by shotgun sequencing of bisulfite-converted DNA (Harris et al. 2010). Contrary to arrays, sequencing-based methods can interrogate 5mC positions in repetitive sequences and perform allele-specific methylation, although at a higher cost. The recent discovery of high levels of 5-hydroxymethylcytosine in Purkinje neurons and embryonic stem cells, which may also play a role in gene expression, introduces a problem in the interpretation since bisulfite sequencing cannot distinguish this conformation from 5mC, although novel detection methods are now available (Yu et al. 2012).

Histone modifications

Histones are small alkaline proteins made up of a central globular domain and a flexible charged amino terminal end, also called histone “tail”, that protrudes from the nucleosome. These tails are target of post-translationally chemical modifications, being the most well-known acetylation, phosphorylation and methylation, which have been involved in chromatin compaction and gene expression (Jenuwein and Allis 2001; Bannister et al. 2002; Kouzarides 2007; Lawrence et al. 2016). Two mechanisms have been described by which histone modifications can trigger their effects. First, some of these marks would directly shape the inter-nucleosomal interactions and perturb the overall physical state of the chromatin (Lu et al. 2008; Lawrence et al. 2016). For example, lysine 16 of histone H4 acetylation (H4K16ac) reduces chromatin compaction and gene transcription increases (Akhtar and Becker 2000; Shogren-Knaak et al. 2006). Second, histone modifications are supposed to act by the recruitment of remodeling enzymes and TFs (Kouzarides 2007). In fact, the “histone code” hypothesis postulates that diverse combinations of modification marks on the histone tails are required to provide binding sites for proteins and obtain a particular effect (Jenuwein and Allis 2001). In this case, histone modifications can cooperate in order to recruit factors more efficiently, but also adjacent modifications can disrupt the binding of a protein to a particular mark (Bannister and Kouzarides 2011). Currently, three well-studied histone modifications (H3K27ac, H3K4me1 and H3K4me3) are known to report both promoter and enhancer activity. H3K4me1 and H3K27ac are found at active enhancers, whereas H3K4me3 predominantly locates at active promoters (Heintzman et al. 2007; Rada-Iglesias et al. 2010). Other known modifications include H3K27me3 and H3K9me3 at repressed regions or H3K36me3 at gene body of active genes (Zhou et al. 2011).

Histone modifications are generally profiled by chromatin immunoprecipitation followed by sequencing (ChIP-seq) (O’Geen et al. 2011), which is based on the selection of protein-DNA complexes by specific antibodies. Briefly, proteins such as histones or transcription factors are covalently

cross-linked to the DNA region in which they are located. Next, chromatin is fragmented, proteins of interest are immunoprecipitated with the attached DNA, and DNA is sequenced to assess which regions are most frequently bound to the proteins captured by the antibody. Therefore, ChIP-seq technology is a powerful tool that can characterize DNA-protein interactions *in vivo* and identify genome-wide patterns of histone marks.

Thousands of genetic variants have been found to influence histone tail modifications (Kasowski et al. 2013, Kilpinen et al. 2013, McVicker et al. 2013). As we have commented above, functional variants in TF binding sites can alter the recruitment of modifying enzymes, such as histone methyltransferases, and result in changes in the local chromatin composition (Grubert et al. 2015; Waszak et al. 2015). An interesting study from Delaneau and colleagues (2019) showed that chromatin activity levels (i.e. enrichment of epigenetic marks) of H3K27ac, H3K4me1 and H3K4me3 histone modifications are structured in thousands of coordinated regions across LCLs and primary fibroblast lines. That is, inter-individual correlation between adjacent chromatin peaks confirmed the existence of a widespread coordination of chromatin activity in modules, termed as *cis*-regulatory domains (CRDs). CRDs, as chromatin peaks, are tightly regulated by nearby genetic variants and their activity can be modulated (Grubert et al. 2015; Waszak et al. 2015; Delaneau et al. 2019). Moreover, a small proportion of genetic variants can also affect CRD structure, disrupting this coordinated regulatory activity among histone peaks (Delaneau et al. 2019). Interestingly, gene expression is often strongly associated with chromatin activity and variants acting as QTLs of chromatin peaks and CRD (cQTLs and CRD-QTLs, respectively) are frequently eQTLs (Waszak et al. 2015; Delaneau et al. 2019). Besides, a high concordance has been observed between physical contacts derived from Hi-C and chromatin activity. This reflects that local three-dimensional conformation of the DNA can be translated into functional links of coordinated activity between promoters and enhancers (Grubert et al. 2015; Delaneau et al. 2019). In fact, CRDs delimit regulatory units within TADs. Finally, as shown in other studies, cQTLs and CRD-QTLs significantly overlap GWAS-associated variants, indicating that these loci may affect gene ex-

pression and phenotypes through histone modification changes.

Higher-order chromatin structure and accessibility

By using DNase I sequencing, Degner et al (2012) showed that most eQTLs also modify chromatin accessibility. This technique can identify regions sensitive to cleave by the DNase I enzyme (DNase I-hypersensitive sites) and measure chromatin accessibility. Thousands of variants correlate with DNase-seq read depth, called DNase I sensitivity QTLs or dsQTLs, and are strongly enriched within TF binding sites. A substantial fraction of dsQTLs were linked to differences of gene expression, which would seemingly reflect that the strengthening or weakening of TF binding is a major mechanism through which genetic variation influences gene transcription, as suspected before. Increased chromatin accessibility was usually associated with higher expression levels, thus suggesting that the majority of TFs that are bound to hypersensitivity sites act as enhancers, which in turn change nucleosome occupancy and hence measured DNase I cut estimates.

Chromosome conformation capture methods have been useful to reveal the spatial organization of the genome. In particular, the high-throughput version Hi-C has showed the partitioning of chromosomes into numerous condensed structures termed topologically associated domains or TADs. These chromatin domains are characterised by high intradomain contact frequency and insulation from adjacent TADs (Vietri et al. 2015). TAD boundaries are largely conserved across cell types (Schmitt et al. 2016) and species (Dixon et al. 2012), and are enriched of CTCF binding sites. TAD size ranges between 40 kb to 3 Mb with a median size of 185 kb (Rao et al. 2014), though these estimates heavily depend on the resolution of the Hi-C data employed. It has been proposed that TADs play a role in bringing enhancers into spatial proximity with target genes (Jin et al. 2013). As we have seen above, genetic variants, including inversions, can alter TAD boundaries and thereby the three-dimensional structure of the genome with consequences on gene expression and congenital diseases (Lupianez et

al. 2015; Flavahan et al. 2016; Franke et al. 2016; Hnisz et al. 2016; Kraft et al. 2019). Moreover, gene regulation by chromatin domain organization has been also studied in cell differentiation (Fraser et al. 2015; Narendra et al. 2016) and cancer (Dixon et al. 2018). While these studies highlight the functional relevance of these domains, a recent work has suggested a more moderated role (Ghavi-Helm et al. 2019). In this case, the authors induce major genomic rearrangements in *D. melanogaster* chromosomes, such as duplications, inversions or deletions, and found that the majority of genes do not have changes in expression, indicating that the relationship between chromatin topology and gene expression is much more complex than previously thought.

1.2.3 Genetic regulation of protein levels

The central dogma of molecular biology describes that information in DNA flows into proteins by a two-step process: DNA is transcribed to RNA, which is translated into protein. Since proteins are likely to be more directly implicated in the generation of the whole-organism phenotypes, alterations to protein levels can have consequences in human health and its study could provide insight into mechanisms behind disease origin and development. As we have already commented in previous sections, there is a vast body of literature dedicated to human genetic influence on gene expression. Consequently, multiple protein quantitative trait loci (pQTLs) analyses have emerged as the third element of this jigsaw puzzle in order to reveal the genetic architecture behind variation in protein abundance. However, although there has been a significant progress (Table 1.1), so far efforts dedicated to identify proteome QTLs lag behind when compared to published studies on eQTL identification due to the technical difficulties of protein quantification compared to DNA and RNA.

Although changes in mRNA levels of protein-coding genes are expected to be reflected in differential protein expression levels, it has been observed that variation in mRNA expression is not a perfect surrogate for the corresponding proteins since many processes can be involved in post-

transcriptional regulation. Thus, correlation between protein and mRNA levels has been described as weak or modest (de Sousa et al. 2009; Schwan-hausser et al. 2011), which highlights the importance of deciphering the association of genomic architecture with protein levels in humans directly. In fact, variants associated to gene expression tend to have reduced effect sizes on protein levels, indicating that their impact is often attenuated (Battle et al. 2015). Most of these studies have focused on different methods - aptamer-, immunoassay- and mass-spectrometry-based proteomics- that allow to quantify protein levels using blood plasma samples, which can be easily collected. Proteomic approaches based on high-resolution mass spectrometry detection of digested peptides separated by chromatography has been successfully applied to quantify relative protein expression measurements (Johansson et al. 2013; Wu et al. 2013; Liu et al. 2015; Battle et al. 2015). Nonetheless, this approach has been criticised due to the limited resolution by the peptide spectra and detection sensitivity, as well as other issues such as unreliably monitoring of post-translational modifications and isoform level differences, reproducibility and cost. Other methods include two-dimensional difference gel electrophoresis (2D DIGE) technology (Garge et al. 2010) or distinct immunoassays such as ELISA (Melzer et al. 2008; Kim et al. 2013; Enroth et al. 2014). In particular, many recent studies have used aptamer-based technology – slow off-rate modified aptamer (SOMAmers) – to measure protein levels (Lourdusamy et al. 2012; Suhre et al. 2017; Sun et al. 2018). Aptamers are synthetic single-stranded oligonucleotides, which can take diverse but well-defined shapes by folding into convoluted molecular structures that can bind to a wide array of target proteins, peptides or even small molecules in a manner that conceptually resembles the function of antibodies. Thus, SOMAmers act as recognition elements with high specificity and affinity, and are chemically stable (Gold et al. 2010).

In any case, quantifying protein abundance at large-scale is still a notable methodological challenge, and many of these population-based proteomic studies are limited with regard to the number of proteins assayed or the number of individuals. The most comprehensive study so far carried out by Sun and colleagues (2018) assessed how genetic variation influences

plasma protein abundance in humans by measuring 2,994 plasma proteins in 3,301 healthy donors of European descent from the INTERVAL study. In this analysis, more than 1,900 significant associations between 1,478 proteins and 764 genomic regions were found, with 89% of these pQTLs being previously unreported. Two thirds of these pQTLs had local associations and the rest were associations in trans. As happened in gene expression analyses, proteins can also have two or more independent pQTL signals. Although plasma protein levels may not reflect abundance within cells or tissues, 44% of cis pQTLs were reported by previous studies as eQTLs for the same gene in at least one tissue or cell line, showing greater overlap with blood, liver and LCLs, as expected. These findings highlight the fact that plasma protein levels are driven to some extent by effects acting at mRNA regulation, but not exclusively, since pQTLs not overlapping eQTLs may reflect other regulatory processes such as altered protein clearance, secretion, degradation or binding. Importantly, 88 pQTLs overlapped with disease susceptibility loci, providing not only new understanding of the genetic control of protein regulation, but also of the molecular impact of disease-associated variants.

Table 1.1 – Published pQTL studies in human blood derived samples.

Reference	Study sample	Proteins quantified (tissues analyzed)	Approach	Number of proteins with pQTLs	Proteomic technique
Melzer et al. (2008)	1,200 fasting European individuals from the population based InCHIANTI study	42 proteins (serum and plasma)	Genome-wide and cis-only (up to 300 kb away from the gene) association with 496,032 autosomal SNPs (MAF >1%)	9 proteins: 8 pQTLs in cis and 1 in trans	Distinct immunoassays
Garge et al. (2010)	24 individuals from HapMap CEU population	544 proteins (LCLs)	Genome-wide and cis-only (up to 200 kb away from the gene) association with 1.7 M SNPs from the HapMap Project	15 proteins: 24 pQTLs in cis and 2 in trans	2D difference gel electrophoresis
Lourdusamy et al. (2012)	96 elderly healthy Europeans cohort (mean age 72.1 years)	778 proteins (plasma)	cis-only (up to 300 kb away from the gene) association with 776,864 genetic variants (MAF >5%)	60 proteins: 60 pQTLs in cis	Aptamer-based proteomic technology
Johansson et al. (2013)	1,029 individuals from two population-based cohorts (KA06 and KA09)	163 proteins (plasma)	cis-only (up to 100 kb away from the gene) association with 7.83 M SNPs in the discovery cohort (KA06) and 8.78 M in the replication cohort (KA09)	5 proteins: 5 pQTLs in cis	Mass-spectrometry
Wu et al. (2013)	95 ethnically-diverse individuals genotyped in the HapMap Consortium	4021 proteins (LCLs)	cis-only (up to 20 kb away from the gene) association with HapMap phase III genotypes (MAF >10%)	77 proteins: 77 pQTLs in cis	Mass-spectrometry

Table 1.1 continued from previous page

Reference	Study sample	Proteins quantified (tissues analyzed)	Approach	Number of proteins with pQTLs	Proteomic technique
Kim et al. (2013)	521 Caucasian participants in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort and 59 participants in the Indiana Memory and Aging Study (IMAS) cohort	132 analytes (plasma)	cis-only (up to 100 kb away from the gene) association with 1,992 SNPs belonging to 137 genes (MAF >5 %)	28 proteins: 28 pQTLs in cis	Immunoassay
Enroth et al. (2014)	970 individuals from a longitudinal cross-sectional population-based study in Sweden (KA06 and KA09 cohorts)	77 biomarkers for cancer and inflammation (plasma)	Genome-wide association with 4,840,842 SNPs and indels	14protein s: 16 pQTL s in cis and 2 in trans	Immunoassay
Liu et al. (2015)	113 female fasting MZ and DZ twins from the Twins UK Adult Twin Registry	303 proteins from 1,904 peptides (plasma)	cis-only (up to 1 kb away from the gene) association with 758 SNPs	13 proteins: 13 pQTLs in cis	Mass-spectrometry
Battle et al. (2015)	62 individuals from HapMap YRI population	4,381 proteins (LCLs)	cis-only (up to 20 kb away from the gene) association with 15.8 million variants from the HapMap Project (MAF >10%)	278 proteins: 278 pQTLs in cis	Mass-spectrometry

Table 1.1 continued from previous page

Reference	Study sample	Proteins quantified (tissues analyzed)	Approach	Number of proteins with pQTLs	Proteomic technique
Suhre et al. (2017)	1,000 individuals of the population-based KORA study and 338 participants of the QMDiab study	1,124 proteins (plasma)	Genome-wide association with 509,946 common autosomal SNPs	284 proteins: 391 pQTLs in cis and 55 in trans	Multiplexed, aptamer-based, affinity proteomics platform (SOMAscan)
Sun et al. (2018)	3,301 healthy blood donors of European descent from the INTERVAL study	2,994 proteins (plasma)	Genome-wide association with 10.6 million imputed autosomal SNPs	1,478 proteins: 374 pQTLs in cis, 925 in trans, and 179 both in cis and in trans	SOMAmers

1.2.4 Current methods for QTL analysis

Earliest eQTL mapping studies in humans used dozens of samples, but current analyses can involve hundreds or even thousands of individuals with expression values (Lappalainen et al. 2013; The GTEx Consortium 2017; Võsa et al. 2018). As large-scale expression data were generated, it became obvious that novel statistical methods were required to take full advantage of sample size power. In this regard, high-throughput and efficient techniques to ensure computational tractability, while preserving acceptable accuracy of the results obtained, have been developed. We have to take into account that the goal of QTL mapping studies is to determine which DNA variants are really responsible for the phenotypic variation. However, a second round of experimentation would be needed to confirm the detected effects.

Identifying the variables that explain how data is structured is a crucial process in any molecular QTL analysis. For instance, population stratification refers to systematic population-specific differences among sample individuals. Thus, human genetic diversity, originated from migration across the world and genetic drift, can lead to false positive associations due to variation in allele frequencies. Principal Component Analysis (PCA) on genotype data has been established as the most widely used method to identify and quantify population structure (He et al. 2011). The top principal components can explain differences in the genetic data and reflect genetic variation due to ancestry. Thus, individuals with similar origin would be characterized by close PC values (Novembre et al. 2008; Raj et al. 2014). Moreover, stratification may be present even in a population apparently homogenous. For example, PCA analysis was applied to genetic data derived from Finland samples (Sabatti et al. 2009), identifying regional structure that corresponded very well to geographic territories in this country. It has been proposed that stratification can be controlled by genotyping few dozens of unlinked genetic markers. To eliminate redundant SNPs in high LD, selection of autosomal variants is done by LD pruning and trimming for minor allele frequency (Delaneau et al. 2017).

Batch effects and other confounding factors, such as date of sequencing, library preparation, environmental influences, gender or unknown factors, can also affect expression data and are known to reduce the power to find eQTLs (Plagnol et al. 2008). The software suite PEER (probabilistic estimation of expression residuals) (Stegle et al. 2012) is a widely used tool that implements statistical models to infer hidden determinants of variability on population-scale expression data and to improve the sensitivity and interpretation of genetic associations. Briefly, PEER is based on Bayesian factor analysis approaches that infer broad variance components in expression measurements; i.e. learned variables are assumed to have broad influence and to affect large fractions of all genes, explaining large variance components. PEER takes as input transcription profiles from a set of individuals and outputs the hidden determinants that explain much of the expression variability. Therefore, these methods allow us to account for such global confounders of variability in expression data and include them in our models, both boosting the power to detect associations and reducing spurious false-positive eQTLs.

Despite the goal of QTL mapping studies is to determine which DNA variants are really responsible for the phenotypic variation, a second round of experimentation would be needed to confirm the detected effects. Finally, future studies will be focus on (i) increasing the sample size, (ii) expanding the number and diversity of available tissues and cell types, and (iii) performing eQTL mapping on single cells. First, larger sample sizes are expected and several datasets will be combined into large-scale meta-analyses that will permit us to find more eQTLs with smaller effects (Võsa et al. 2018). In fact, there are already methods that allow to perform meta-analysis over distinct tissues simultaneously (Li et al. 2017b). Second, exploring a complete array of cells and conditions will be important to obtain a comprehensive catalogue of novel effects. For example, a recent study has identified retina specific eQTLs (Ratnapriya et al. 2019). Finally, the next natural step would be to study individual cells. In this sense, some studies have detected eQTLs that are only present in single cells, and are lost when the expression is averaged over multiple cells (van der Wijst et al. 2018; Igor et al. 2019). In summary, eQTL studies will

continue to contribute to our understanding of regulatory variants and yield substantial biological insights into many diseases.

1.3 Objectives

The present doctoral thesis is focused on the functional impact of polymorphic inversions in the human genome. Specifically, we seek to investigate how human inversion polymorphisms may have affected gene expression and epigenetic changes, and identify potential candidates to have consequences in human traits and disease. Thus, this work addresses one of the basic questions in biology and human molecular genetics in particular: which are the genetic basis of phenotypic characteristics and how genetic variation is related to the molecular mechanisms responsible for such phenotypes. For that, we have carried out an integrative and global bioinformatic analysis of their possible functional effects at different levels by making use of the unique knowledge about human inversions accumulated during the last years through the new methods for inversion genotyping developed within the InvFEST Project and the great amount of molecular data available in humans. The specific objectives are illustrated in Figure 1.8 and are described in more detail below:

1. **Analysis of the gene-expression changes associated to inversions.** Inversions can affect transcription by several mechanisms, including the disruption of gene sequences and the repositioning of functional elements. Therefore, this study aims to perform an in-depth association study of gene expression data from LCLs and inversions genotyped by different techniques (MLPA, ddPCR and iPCR). Given the importance of testing the proper cell type or tissue to infer hidden eQTL effects, these analysis will be extended to gene expression data in multiple tissues from the GTEx project. Finally, we will determine if inversions can create fusion transcripts with novel added sequences derived from promoters belonging to disrupted genes.

2. **Analysis of the epigenetic changes associated to inversions.**
Given the importance of regulatory elements of gene expression, the second objective of this thesis is to characterize histone modification and DNA methylation patterns associated to inversion orientation. Moreover, the correlation between gene expression and the enrichment of epigenetic marks could explain the mechanisms by which some inversions exert their effects.
3. **Identify potential inversions associated with disease susceptibility and other phenotypic traits.** The information of the effects of inversions on particular genes could be useful to determine their association with diseases. In addition, thanks to the generated genotypes we can study the role of inversions in phenotypic variation. Therefore, one of our goals is to fill the gap in the knowledge of the role of human inversions in disease susceptibility and other relevant traits.
4. **Perform a functional characterization of particular inversion candidates.** Besides the previous global analysis, we will carry out a more detailed analysis of some candidates of interest. We will take advantage of the multiple levels of information available and combine genomic functional information, gene expression in relevant cell types and conditions -like infection-, epigenetics and protein levels to establish clear relationships between these inversions and phenotypes.

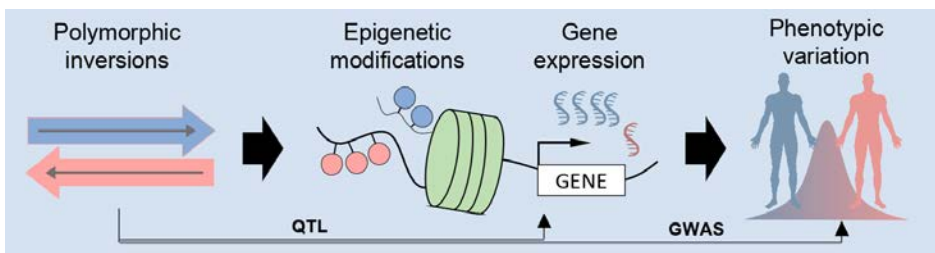


Figure 1.8 – Outline of our integrative analysis of omics data exploring the contribution of polymorphic inversions to human traits.

Chapter 2

Materials and Methods

In this thesis, I used the most complete and accurate set of inversion genotypes available. This data set contains information from 111 common inversions experimentally genotyped by targeted methods in a large number of individuals from diverse human populations. For the sake of clarity, I divided this chapter in three parts that summarise the main methods employed for each section in the Results chapter. Nonetheless, the first two sections of the Results chapter are articles that include part of my work, and the methods can be consulted directly on the corresponding article (Giner-Delgado et al. 2019 and Puig et al. 2019).

2.1 Methods for “Functional and phenotypic effects of a human inversion regulating *RHOH* isoforms and RhoH protein levels”

RHOH gene structure model

For improved clarity on *RHOH* structure, we used a collapsed gene model in which overlapping exon intervals were merged. For that, we retrieved GENCODE release 26 (Harrow et al. 2012). basic annotation and *RHOH*

overlapping exons from distinct transcripts were merged into meta-exons, and posterior expression quantification was calculated for the whole meta-exon. Exons were numbered according to their order in transcription direction from 5' to 3'.

Expression analysis in LCLs

RNA-Seq data from LCL samples from the Geuvadis project (EMBL-EBI ArrayExpress experiment E-GEUV-1) (Lappalainen et al. 2013) were mapped with STAR v2.4.2a (Dobin et al. 2013) to the human reference genome GRCh38.p10 using GENCODE v26 annotations (Harrow et al. 2012). Major chromosomes (chr. 1-22, chr. X, chr. Y and mitochondrial DNA) together with un-placed and un-localized scaffolds were included. HsInv0102 inversion genotypes for 173 individuals in common with the Geuvadis project were extracted from Giner-Delgado et al. (2019). Given that the error rate in 1000GP genotypes of this inversion for the 434 samples shared in both datasets is low (2.53%), we extended our analysis to the whole set of 445 individuals with expression data by using the extra 1000GP Ph3 HsInv0102 genotypes (labeled as esv3600303) (Sudmant et al. 2015). To distinguish HsInv0102 and rs7699141 effect on *RHOH* exon E8, we defined HsInv0102 *Std2* allele as *Std2* when the alternative allele of rs7699141 is detected, and *Std1* otherwise. Other neighboring genomic variants from the 1000GP Ph3 (The 1000 Genomes Project Consortium 2015) (Sudmant et al. 2015) were included in a joint analysis to estimate the real contribution to observed exon expression changes.

Expression levels were estimated as reads per kilobase per million mapped reads (RPKM). We only kept for downstream analysis protein-coding genes that had expression values higher than 0.1 RPKM in at least one third of samples. Gene expression was normalized by quantile transformation across all samples and each gene expression values were adjusted to a standard normal distribution by rank-based inverse normal transformation. Differential exon usage was analyzed with DEXSeq v1.20.2 (Anders et al. 2012) and expression measures obtained for *RHOH* meta-exons

based on the collapsed gene model were transformed to match normal distributions. Gene and exon expression values were adjusted by gender, population membership and the laboratories in which RNA-Seq was performed to control by confounder factors and batch effects, and associations were performed by linear regressions implemented in R `lm()` function (R Core Team 2016). For the protein-coding gene analysis, P values obtained for each gene by regressing with HsInv0102 or rs7699141 genotypes were corrected by false discovery rate (FDR) and significance was established at 5%.

RhoH protein levels in LCLs

We retrieved normalized peptide expression data measured through mass spectrometry using stable isotope labeling with amino acids in cell culture for 62 HapMap Yoruba LCLs (Battle et al. 2015). Although RhoH was originally excluded in this study due to the low number of peptide measures and individuals quantified, we tried to recover RhoH protein abundance levels in the following way. First, from the seven different peptides detected for the RhoH protein, we removed those peptides supported by only three or fewer observations. Second, we evaluated the agreement between measurements of distinct RhoH peptides. Expression of [SNLPCTPVLVVATQTDQR] and [GVQQVFEC AVR] fragments was highly correlated ($P = 5.58 \times 10^{-4}$), as happens among fragments for similar proteins with single-exon coding sequences such as ATP6V1E2 (P range: $6.11 \times 10^{-4} - 9.67 \times 10^{-5}$), REPIN1 (P range: $9.89 \times 10^{-4} - 1.08 \times 10^{-4}$), RAP2B (P range: $1.73 \times 10^{-7} - 4.65 \times 10^{-13}$) or F8A3 (P range: $9.79 \times 10^{-3} - 6.69 \times 10^{-19}$). Conversely, [TSELLVR] fragment quantification was not consistent with the other RhoH peptides ($P = 0.36$ and $P = 0.89$, respectively) and it showed significantly higher expression than the longer fragments (Wilcoxon test, $P = 1.79 \times 10^{-4}$). Moreover, a protein blast (Madden, 2002) of [TSELLVR] peptide against UniProtKB/Swiss-Prot (Boutet et al. 2016) generated several alignments with other protein peptides differing in just one amino acid, which together with the fact that it is the smallest fragment, suggests a potential low quality quantification

of this peptide. Therefore, we excluded [TSLLRV] and used [SNLPCT-PVLVVATQTDQR] and [GVQQVFEC AVR] to estimate RhoH protein expression levels. The median $\log_2(\text{sample}/\text{standard})$ ratio of both peptide measures was used when available in the same individual (including duplicate peptide measurements or replicates) and it was normalized to a standard normal for protein-level quantification. Finally, we recovered 42 individuals with inversion genotypes and protein level profile.

Validation of expression changes in genes in trans and network analysis We employed a permutation approach to confirm the reliability of significant expression changes found in distal genes in LCLs. First, we repeated 100 times the eQTL analysis by exchanging HsInv0102 or rs7699141 genotypes relative to expression and covariate phenotypes to break real genotype-phenotype correlations, finding that only 3% of permutations had significant differentially expressed genes after FDR correction. Second, we computed a permuted-based P value for each gene by the number of P values from these null distributions of no-association smaller than each ranked nominal P value and divided by its rank and the number of permutations.

For protein analysis, we correlated each protein normalized expression estimate against RhoH levels and we ran GOrilla analysis tool (Eden et al. 2009) to identify enriched gene ontology terms in a ranked list of the proteins the most correlated with RhoH.

HsInv0102 imputation through common genotyping arrays

To determine whether HsInv0102 or rs7699141 are associated to specific traits or diseases, we first searched unsuccessfully for SNPs in high LD ($r^2 > 0.8$) with these variants in any population included in the NHGRI Catalog of published GWAS (MacArthur et al. 2017). Thus, to determine whether these variants have been overlooked in GWAS, we checked if 76 commercial commonly-used genotyping arrays available through the LDLink web portal (Machiela and Chanock 2015) included SNPs in high LD ($r^2 > 0.8$) with HsInv0102 or rs7699141 in EUR, AFR and EAS populations separately. Second, we chose 27 of these arrays to check

if HsInv0102 can be imputed accurately through genotyped SNPs. For each array, we merged our HsInv0102 genotypes together with 1000GP Ph3 variants filtered by those SNPs included in the array for the 434 individuals in common, and used this as reference panel in the imputation. Imputation was performed with IMPUTE v2.3.2 (Howie et al. 2009) adapted to unphased reference genotypes with an effective population size of 20,000. We found similar performance considering different buffer regions of 250 kb, 500 kb and 1 Mb. We checked the imputation accuracy by a leave-one-out strategy: i.e. artificially removing one sample genotyped for the inversion from the reference panel and then imputed it. Genotypes were called with the highest posterior probability. To evaluate imputation accuracy, we estimated correlation coefficients r^2 between true inversion genotypes and imputed genotypes.

HsInv0102 association with blood cancer

To identify *Std* and *Inv* HsInv0102 alleles among blood cancer patient samples from WGS ICGC PCAWG datasets, we prepared fasta sequences of 150 bases surrounding the two inversion breakpoints in both conformations. Reads from bam files (CLLE-ES: EGAD00001001466; CMDI-UK: EGAD00001002664; LAML-KR: EGAD00001002119; MALY-DE: EGAD00001002123) aligned to the RHOH gene region were retrieved and mapped to our custom library with a modified version of BreakSeq software as described in a previous publication (Lucas-Lledó et al. 2014). Reads from both tumor and normal pair were joined, and those mapping uniquely and overlapping at least 20 bases to each side of the breakpoint were considered as evidence of *Std* or *Inv* alleles. We repeated this process with different overlapping lengths (5, 7 and 10 bases at each breakpoint side) to ensure the reliability of inversion genotypes. Alleles with fewer than 10 hits in total were excluded to avoid spurious callings. Only reads with a MAPQ ≥ 15 were analyzed.

To match ethnically control groups, for CLLE-ES project in Spain, we used three Iberian groups: IBS population from 1000GP (HsInv0102 im-

puted as rs7676043 variant), elder individuals from Spain genotyped by PCR (CNTPK), and individuals from the GCAT cohort from Catalonia genotyped by BreakSeq as commented above. We performed fisher tests to compare allele frequencies among control groups and no significant differences were detected. For CMDI-UK project in United Kingdom, MALY-DE in Germany and LAML-KR in South Korea, we employed, respectively, GBR, CEU and CHB populations from 1000GP. To increase control sample size, we also included additional CEU and CHB individuals not in 1000GP genotyped in Giner-Delgado et al. (2019). We excluded sample DO52739 from CMDI-UK, because it has Asian ancestry according to Campbell et al. (2017). Allele frequency differences between cases and controls were tested with fisher test, whereas genetic models implemented in SNPAssoc (González et al. 2007) were employed for genotype comparison. To infer HsInv0102 genotypes, we chose rs7676043 in 1000GP European control populations, since the SNP tags better the inversion than the esv3600303 genotypes provided by the 1000GP, whereas esv3600303 was used to infer the inversion orientation in CHB. Those genotypes in 1000GP samples that we knew that were wrong were corrected.

Finally, to check the possible contribution of each variant to blood cancer susceptibility, we carried out a variant calling of SNP genotypes in *RHOH* region in samples with WGS data with GATK Best Practices somatic variant calling protocol (Van der Auwera et al. 2013), using 1000GP individuals as controls. We confirmed reliability in variant calling by extracting WGS reads for rs7676043 and rs7699141 and obtaining the same genotypes by processing with SAMtools (Li et al. 2009). Allele frequencies were compared between cases and controls for rs7699141 and other biallelic variants located at *RHOH* gene as we did with HsInv0102. For rs7699141, we also imputed this SNP in samples not included in 1000GP Ph3.

2.2 Methods for “A human inversion influences antiviral response through different regulatory effects on interferon response genes”

HsInv0124 experimental genotypes

We retrieved inversion status obtained previously by the InvFEST project (Giner-Delgado et al 2019) from 550 individuals with African (AFR) (100 YRI and 90 LWK populations), European (EUR) (90 CEU and 90 TSI), South-Asian (SAS) (90 GIH) and East Asian (EAS) (90 CHB and 90 JPT) ancestry. A subset of 431 individuals included in 1000GP Phase 3 (Ph3) (The 1000 Genomes Project Consortium 2015) with available data were used for analysing LD patterns and imputation accuracy. On the other hand, genomic DNAs of 120 samples from the Genomes For Life (GCAT) cohort (Obón-Santacana et al. 2018), which recruits population of the north-east region of Spain (Catalonia), were used for experimental genotyping of the inversion by iPCR using previously described assays (Aguado et al. 2014; Giner-Delgado et al 2019).

Imputation of HsInv0124 genotypes

We measured LD (r^2) between HsInv0124 and surrounding biallelic 1000GP variants (+/-150 kb from inversion outer coordinates) using PLINK v1.90 (Purcell et al. 2007). This analysis allowed us to detect that HsInv0124 is only tagged by SNPs in East Asian populations. Thus, we decided to infer inversion status by imputation. For that, we merged available HsInv0124 genotypes with 1000GP Ph3 variants for the 431 individuals in common, and used this as reference panel. Imputation was performed with IMPUTE v2.3.2 (Howie et al. 2009) with an effective population size of 20,000 and adapted to unphased reference genotypes, since HsInv0124 is a moderately recurrent inversion. We found similar performance considering different buffer regions of 0, 25, 50, 100, 150, 200, 250, 300 and 500 kb. We checked the imputation precision by a leave-one-out

strategy: i.e. artificially removing one genotyped sample for the inversion from the reference panel and then imputed it. Genotypes were called with the highest posterior probability. To evaluate imputation accuracy, we estimated correlation coefficients r^2 between true inversion genotypes and imputed genotypes, confirming high rates of concordance.

We imputed HsInv0124 in all populations from 1000GP and using the commented reference panel and a buffer region of 150 kb at each side of the inversion. Moreover, inversion orientation was inferred in 120 GCAT samples and 200 individuals with expression data under different immune responses (100 european and 100 african individuals) provided by Quach et al. 2016. Imputed genotypes from GCAT individuals were compared with those obtained experimentally, whereas only samples with expression data from European origin were used in downstream analysis since inversion genotypes for some Africans differed when imputing several times.

RNA-Seq and *de novo* transcriptome

Total RNA was isolated from cell culture of LCLs of 5 *O1/O1* and 5 *O2/O2* individuals after stimulation with interferon. TruSeq libraries were prepared and 2 x 101 bp paired-end stranded sequencing was performed by the company Beckman Coulter using recommended protocols.

To generate the inverted orientation in human genome assembly GRCh38/hg38, we reverse-complemented *in silico* HsInv0124 sequence plus breakpoints. LCL raw RNA-Seq reads from the 10 LCL samples exposed to IFN were aligned against both reference (hg38-124O1) and modified (hg38-124O2) genome using STAR v2.4.2a in 2-pass mode to improve accuracy of spliced alignment (Dobin et al. 2013). Only the primary assembly was considered (i.e. chromosomes and scaffolds). We also ran the mapping with GENCODE version 29 annotations, although *IFITM1* coordinates were adjusted to the inverted region and *IFITM2-201* (ENST00000399815.2) transcript was removed for hg38-124O2 genome. We collected all the novel discovered splice junctions in the first round detected by mapping reads from the 10 individuals with each *O2/O2* and *O1/O1*

genotype and they were subsequently employed together with GENCODE annotations to align reads for the second mapping for each sample. We extracted uniquely aligned reads from HsInv0124 locus (chr11:305000-333000, hg38) and reconstructed a *de novo* transcriptome with StringTie per each sample (Pertea et al. 2015). Next, we combined all reconstructed isoforms across samples per inversion orientation with StringTie merge mode with a minimum input transcript length of 300 bp. Transcripts reconstructed in the hg38-124 O2 genome were mapped on hg38-124O1 genome and only transcripts supported by the correct stranded alignments were kept. To adjust external transcript limits and obtain a simplified and consistent annotation, we compared reconstructed transcripts with GENCODE annotation and followed several criteria: (i) for those isoforms among datasets overlapping the first or the last exon and sharing the same splice site/s, the exterior border of these exons were adjusted to the longest evidence; (ii) overlapping exons sharing the same donor splice site were enlarged to the extended border; (iii) the largest isoform was selected for overlapping mono-exonic transcripts; (iv) the first exon from *AC136475.1* and *AC136475.2* transcripts were adjusted according to the rules above; and (v) redundant transcripts were removed.

To demonstrate the presence of *IFITM2-IFITM1* fusion transcript described in GENCODE v29 annotation, we checked both reads mapping at splice junctions and read mates that support this isoform. However, no uniquely mapped reads crossing the junction between *IFITM2* and *IFITM1*, or read mates mapping each in one gene were found.

Differential expression analysis in IFN-stimulated LCL

IFN-stimulated LCL RNA-Seq reads were aligned against the human reference genome GRCh38.p10 using STAR v2.4.2a (Dobin et al. 2013) and gene-expression levels were estimated as counts based on a slightly-modified version of GENCODE annotations. We excluded genes located at *IFITM* locus from GENCODE version 29 and included our *de novo* annotation to guide the alignments. We summed all reads from the 10 samples

for each gene and those with 20 or fewer reads in total were excluded as low expressed. We also filtered genes that included one expression value in a sample that diverged more than 2.5 standard deviations from the median expression. To detect differentially expressed genes between *O1/O1* and *O2/O2* samples, we applied DESeq2 (Love et al. 2014) and selected genes with a false discovery rate (FDR) < 0.1 .

LCLs and CD14⁺ monocytes gene expression analysis

We performed pseudoalignment to estimate transcript abundance from RNA-Seq data of 445 LCLs in common with 1000GP from the Geuvadis consortium (EMBL-EBI ArrayExpress experiment E-GEUV-1) using Kallisto v0.46.0 (Bray et al. 2016). We generated a reference transcriptome index derived from a modified version of GENCODE v29 with our annotation. Additionally, we quantified RNA-Seq collected from unstimulated and stimulated CD14⁺ monocytes yielding a final dataset of 493 expression experiments from 100 individuals of European origin (100 non-stimulated, 96 LPS, 100 Pam3CSK4, 98 R848 and 99 IAV) (Quach et al. 2016). Gene-level abundance in each condition was calculated as the sum of estimated counts for all transcripts of a gene with tximport package (Soneson et al. 2015). Estimated counts were transformed to RPKMs.

eQTL analysis

We normalized RPKM values by applying rank-based inverse normal transformation to each gene and transcript. For LCLs, expression values were adjusted by gender, the first three principal components from genotype data, laboratory of RNA sequencing and a set of components calculated for normalized RPKMs to capture technical confounding factors. We tested technical variation with PEER software (Stegle et al. 2012), taking up to 50 expression-derived factors in intervals of 5. The number of components used was chosen to both maximize the number of

eQTL associations and consistency among different sets of technical covariates; i.e. the group of genes in which the inversion was the lead variant that appeared more frequently by adjusting expression levels and, in case of tie, we selected the set with more differentially expressed genes. In LCLs and monocytes, we used 30 and 40 factors for genes and transcripts, respectively. eQTL analysis was done including HsInv0124 and neighboring 1000GP variants ($MAF > 0.05$) to estimate the contribution of each polymorphism to expression variation and identify sentinel eQTLs. Linear regressions were implemented in QTLtools v1.1 (Delaneau et al. 2017) for each gene or transcript expression and the variants located within 1.5 Mb from the transcription start site. All P values from lead variants were corrected by bonferroni and significance was established at an adjusted $P < 0.05$. Only cases in which HsInv0124 was the lead variant were reported.

Chromatin peaks and CRD

We retrieved chromatin activity -enrichment of epigenetic marks- from three well-studied histone modification marks (H3K27ac, H3K4me1 and H3K4me3) that are known to capture enhancer and promote activity across 145 LCLs from 1000GP European individuals (Delaneau et al. 2019). Direct linear regressions were performed with QTLtools v1.1 (Delaneau et al. 2017).

2.3 Methods for “Comprehensive analysis of the influence of human inversions on gene expression, epigenetic changes and phenotypic variation.”

Validation and genotyping of inversions

To extend our prior catalogue of human inversions (Martínez-Fundichely et al. 2014; Aguado et al. 2014; Vicente-Salvador et al. 2017; Giner-

Delgado et al. 2019; Puig et al. 2019), we first tried to validate and define the exact breakpoints of as many as possible of inversions that have been already predicted more or less precisely using different methods (Chaisson et al. 2015; Sudmant et al. 2015; Hehir-Kwa et al. 2016; Huddleston et al. 2016). Only inversions with a predicted size of more than 50 bp and in which the inverted region does not correspond entirely to a repetitive element were considered. Also, for those inversions with population frequency data (Sudmant et al. 2015), we prioritized those with a global frequency above 5%. For that, we checked the presence of the inversions in downloaded the genomic sequences of 53 new human genome assemblies and two updated versions from 35 different individuals that have been submitted to the NCBI Assemblies database (ncbi.nlm.nih.gov/assembly). hg19/hg38 sequences were retrieved from inversion regions plus 1 kb of flanking sequence at each side using the coordinates reported in the corresponding publication. Next, we systematically assessed each inversion by comparing the inversion sequence in the reference human genome with the downloaded sequence of all the genome assemblies using BLAST. If the top hit is a contiguous BLAST hit extending over the inversions length plus 1,800 bp out of the 1 kb added side, the assembly is marked as “Reference”. Otherwise, it is determined if there is an inversion or other type of structural variant by manual inspection of pairwise Blast alignments. Real inversions discovered in the assemblies were further validated by PCRs of both breakpoints and genotyped in multiple individuals. In addition, a few other inversions mediated by inverted repeats (IRs) predicted in these studies were validated and genotyped by inverse PCR (iPCR). Finally, all other validated inversions by PCR based assays for which there is no population genotyping data yet (Martínez-Fundichely et al. 2014) were also genotyped in multiple individuals.

Imputation of inversion genotypes

For those inversions that do not have variants in perfect LD ($r^2 = 1$) across all the genotyped individuals, we assessed inversion calling *in silico* by imputation with IMPUTE2. For that, we merged the available exper-

imental genotypes for the inversions in our datasets and all the variants from the 1000GP Phase 3 at 500 kb at each side of the inversion (including SNPs, indels and structural variants) for the European and African individuals in common, and used these as reference panels. Imputation was performed with IMPUTE v2.3.2 adapted to unphased reference genotypes, due to the difficulty of phasing correctly recurrent NAHR-mediated inversions. We adapted the imputation protocol for chr. X inversions by coding hemizygote males as homozygous females. Inversion genotypes were inferred by the highest posterior probability and were classified accordingly. To estimate imputation accuracy, we followed a leave-one-out strategy by masking the known genotype of one individual and subsequently impute it with the rest of the panel. Imputation accuracy was measured as the LD between imputed and experimental genotypes and those inversions with an r^2 higher than 0.8 were selected for *in silico* genotyping in downstream analysis. For association studies, the European reference panel was used for imputing inversions on the rest of 1000GP individuals from European ancestry and Caucasian GTEx individuals, whereas African panel was employed for the 1000GP samples with this origin.

Impact of inversions on gene expression and epigenetics

We mapped *cis* INV-eQTLs by testing associations between inversion genotypes and gene-expression measures in 45 tissues and two cell lines from the GTEx project: brain (cerebellum, cortex, caudate, anterior cingulate cortex, hippocampus, hypothalamus, nucleus accumbens, putamen, amygdala, spinal cord -cervical c-1-, substantia nigra), artery (tibial, aorta, coronary), adrenal gland, adipose (subcutaneous, visceral -omentum-), colon (transverse, sigmoid), esophagus (mucosa, muscularis, gastroesophageal junction), heart (left ventricle, atrial appendage), skeletal muscle, spleen, vagina, uterus, prostate, pituitary, ovary, minor salivary gland, kidney (cortex), tibial nerve, pancreas, skin (sun exposed -lower leg-, not sun exposed -suprapubic-), testis, thyroid, whole blood, breast (mammary tissue), liver, small intestine (terminal ileum), lung, stomach, cultured fibroblasts and EBV-transformed lymphocytes. Gene-level quan-

tifications (transcripts per million) were retrieved from the GTEx project webpage (gtexportal.org/home/). To estimate the relative contribution of each variant to gene expression changes, we performed a joint eQTL analysis by including all common variants detected in GTEx together with imputed inversions. We used a window of 1 Mb at either side of the TSS that included at least one autosomal or chr. X inversion. Since the variation of RNA-seq levels can be due to technical or biological causes, we applied a set of covariates to correct gene expression by potential confounders, while retaining biological variation: gender, five genotyping principal components and a variable number of technical covariates. Principal Component Analysis (PCA) was done on trimmed genotype data (one variant with MAF > 0.05 every 50 kb) to obtain the principal components that reflect the population membership and stratification. We applied the same approach to expression levels to capture technical confounding factors. The number of components used as experimental covariates were determined on the basis of the number of samples, adding 5 components every 50 samples. Each gene expression was transformed to a standard normal distribution for linear regression against variant genotypes. These procedures are implemented in QTLtools software (Delaneau et al. 2017). We carried out a multiple-testing correction for each tissue on all inversion-expression tests using the Benjamini-Hochberg method at 5% FDR and further filtering by inversions acting as sentinel markers or in high LD with the top lead eQTLs ($r^2 > 0.8$), ensuring the reliability of INV-eQTL findings.

Gene expression levels from LCLs of the Geuvadis project (Lappalainen et al. 2013) were also analysed in 358 European samples from 1000GP. In this case, the analysis was done with imputed inversion genotypes for all the individuals. RNA-seq reads (EMBL-EBI ArrayExpress experiment E-GEUV-1) were aligned against the human reference genome GRCh38.p10 (excluding patches and alternative haplotypes) with STAR v2.4.2a (Dobin et al. 2013) and gene expression levels were quantified as RPKM based on GENCODE version 26 annotations (Harrow et al. 2012). RPKM values were normalized as described above for GTEx samples. In addition to gene expression, we also assessed other molecular phenotypes from LCL

samples included in the 1000GP by direct linear regressions. In particular, we downloaded normalized levels from DNase-seq data from 59 YRI LCLs that measured chromatin accessibility (Degner et al. 2012) and chromatin activity -enrichment of epigenetic marks- from three well-studied histone modification marks (H3K27ac, H3K4me1 and H3K4me3) that are known to capture enhancer and promote activity across 145 LCLs from European individuals (Delaneau et al. 2019). 100 bp windows of DNase-seq levels were lifted over from hg18 to hg19 and, due to their large number, association was centered in 5 kb on each site including an inversion. Finally, we tested cytosine modification profiles obtained by HumanMethylation450 BeadChip in 103 LCLs derived from European and African ancestry (Moen et al. 2013). Beta values were converted to M values through a logistic transformation using the R package “lumi” (Du et al. 2010) and they were then quantile normalized. Methylation levels were adjusted as described before for gene expression by using 10 technical components. As with eQTL analysis, we retained lead associations or in high LD with the top QTLs ($r^2 > 0.8$) after multiple-testing correction at 5% FDR. All analyses were performed by using the appropriate inversion set.

Inversions and phenotypes

To check if polymorphic inversions are associated with specific traits or diseases, we took advantage of the NHGRI Catalog of published GWAS (<http://www.ebi.ac.uk/gwas/>) [release 2019-07-30, v1.0], which stores a curated collection of the most significant SNPs from each independent locus highly associated ($P < 10^{-5}$) to a particular phenotype. First, we investigated if there was enrichment in the number of trait-associated signals in the inversion and flanking regions (± 20 kb) compared to what should be expected by chance. To do so, we lifted over GWAS Catalog coordinates to hg19 and crossed signals with 1000GP variants, obtaining a final number of 142,997 associations (95.4% from the initial 149,855 associations). GWAS SNPs in high LD ($r^2 \geq 0.8$) and associated exactly to the same phenotype were grouped together to obtain a non-redundant set of GWAS signals. We carried out a permutation strategy to generate 100

random genomic regions as a null model for each inversion and tested observed versus expected GWAS hits according to the random distribution. Chr. Y was excluded of this analysis. Also, permuted regions could not overlap genome gaps in order to avoid biasing the results. We extended this analysis to explore which individual inversions were more significantly enriched in GWAS signals. Therefore, we repeated the same procedure, but using a one-tailed permutation test for each inversion to deal with inversions including zero GWAS signals.

We also crossed variants that have been reported in GWAS with those SNPs in high LD with the inversion ($r^2 \geq 0.8$). Since each GWAS study is focused on populations with different origin, the LD employed to evaluate the association between inversions and GWAS signals was based on individuals with the corresponding ancestry or the closest one available (e.g. TSI for Sardinian, JPT for Japanese, CHB for Han Chinese or Singapore Chinese, and GIH for South Asian, Indian or Bangladeshi). In the case of individuals from the same population group (European, African, East Asian or Asian), we used the LD for the whole continent (e.g. European for Ashkenazi, Framingham, British, Caucasian or Hutterite, and East Asian for Korean), whereas the global LD was selected if populations from different continents were studied.

Detailed characterization of HsInv0014 and HsInv0030

Strong expression changes were associated to inversions HsInv0014 and HsInv0030, both of which affect directly gene sequences. To study in detail possible gene rearrangements caused by these inversions, we first modified the human reference genome sequence by reversing *in silico* the sequence between inversion breakpoints (chr16:75205642-75223319 for HsInv0030 and chr17:18622233-18823774 for HsInv0014; hg38 assembly). We used STAR 2-pass to map the RNA-Seq reads extracted from BAM files stored by GTEx project against the genome with the inverted conformation and only uniquely mapped reads were selected. Tissues assayed were pancreas for *CTRB1* and *CTRB2* (HsInv0030), and adipose subcutaneous

and testis for *AKR1C1* and *AKR1C2* (HsInv0014), where the genes affected by the inversions are expressed at higher levels. The number of individuals selected was dependent on the availability of aligned data. For HsInv0014, the same process was repeated for Geuvadis LCL reads, where the genes are also expressed. Homozygous samples were aligned and RNA-seq profiles were computed separately for each inversion orientation. Transcript structures were reconstructed with Cufflinks default parameters by merging all reads from each genotype.

Chapter 3

Results

In this chapter, I present my contribution to the functional analysis of polymorphic inversions in the human genome. As already commented in the Materials and Methods, the first two sections are a published article and a preprint under review that include part of my work:

Carla Giner-Delgado*, Sergi Villatoro*, **Jon Lerga-Jaso***, Magdalena Gayà-Vidal, Meritxell Oliva, David Castellano, David Izquierdo, Isaac Noguera, Bárbara Bitatello, Iñigo Olalde, Alejandra Delprat, Antoine Blancher, Carles Lalueza, Tonu Esko, Paul O'Reilly, Aida Andrés, Luca Ferretti, Lorena Pantano, Marta Puig, Mario Cáceres (2019). “Evolutionary and functional impact of polymorphic inversions in the human genome”. *Nature Communications.*, 10:4222. <https://doi.org/10.1038/s41467-019-12173-x>

* Equal contribution

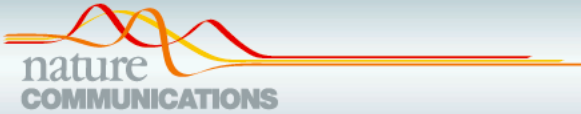
Marta Puig, **Jon Lerga-Jaso**, Carla Giner-Delgado, Sarai Pacheco, David Izquierdo, Alejandra Delprat, Magdalena Gayà-Vidal, Jack F. Regan, George Karlin-Neumann, Mario Cáceres (2019). “Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR.” *bioRxiv*. <https://doi.org/10.1101.766915>

In the next three sections, I describe a global analysis of inversion effects

at different levels and the functional characterization of two particular interesting inversions.

3.1 Evolutionary and functional impact of common polymorphic inversions in the human genome

In this article, we have carried out an exhaustive characterization of the evolutionary and functional impact of 45 common polymorphic inversions by genotyping them experimentally in a large sample of 550 individuals from 7 HapMap human populations. In order to make such large-scale genotyping possible, a new experimental high-throughput assay based on probe hybridization was developed and optimized. Specifically, I was responsible of two distinct analysis. First, I investigated the determinants of inversion frequency in human populations. Second, I explored the potential functional consequences of polymorphic inversions. In this regard, I focused on the effect of inversions on genes and gene expression, and their association with phenotypic traits.















ARTICLE

<https://doi.org/10.1038/s41467-019-12173-x>

OPEN

Evolutionary and functional impact of common polymorphic inversions in the human genome

Carla Giner-Delgado ^{1,2,13}, Sergi Villatoro ^{1,13}, Jon Lerga-Jaso^{1,13}, Magdalena Gayà-Vidal ^{1,3}, Meritxell Oliva¹, David Castellano ¹, Lorena Pantano ¹, Bárbara D. Bitarello ⁴, David Izquierdo ¹, Isaac Noguera¹, Iñigo Olalde⁵, Alejandra Delprat¹, Antoine Blancher^{6,7}, Carles Lalueza-Fox ⁵, Tõnu Esko⁸, Paul F. O'Reilly ⁹, Aida M. Andrés^{4,10}, Luca Ferretti ¹¹, Marta Puig ¹ & Mario Cáceres ^{1,12}

Inversions are one type of structural variants linked to phenotypic differences and adaptation in multiple organisms. However, there is still very little information about polymorphic inversions in the human genome due to the difficulty of their detection. Here, we develop a new high throughput genotyping method based on probe hybridization and amplification, and we perform a complete study of 45 common human inversions of 0.1–415 kb. Most inversions promoted by homologous recombination occur recurrently in humans and great apes and they are not tagged by SNPs. Furthermore, there is an enrichment of inversions showing signatures of positive or balancing selection, diverse functional effects, such as gene disruption and gene expression changes, or association with phenotypic traits. Therefore, our results indicate that the genome is more dynamic than previously thought and that human inversions have important functional and evolutionary consequences, making possible to determine for the first time their contribution to complex traits.

¹Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain ²Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain ³CIBIC/InBIO Research Center in Biodiversity and Genetic Resources, Universidade do Porto, Vairão, Distrito do Porto 4485 661, Portugal ⁴Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Saxony 04103, Germany ⁵Institute of Evolutionary Biology, CSIC Universitat Pompeu Fabra, Barcelona 08003, Spain ⁶Laboratoire d'immunologie, CHU de Toulouse, IFB Hôpital Purpan, Toulouse 31059, France ⁷Centre de Physiopathologie Toulouse Purpan (CPTP), Université de Toulouse, Centre National de la Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (Inserm), Université Paul Sabatier (UPS), Toulouse 31024, France ⁸Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu 51010, Estonia ⁹Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London SE5 8AF, UK ¹⁰UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK ¹¹Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, UK ¹²ICREA, Barcelona 08010, Spain ¹³These authors contributed equally: Carla Giner Delgado, Sergi Villatoro, Jon Lerga Jaso Correspondence and requests for materials should be addressed to MC (email: mcaceres@icrea.cat)

In the last decade a great effort has been devoted to characterizing all the variation in the human genome^{1,5}, which opens the door to determining the genetic basis of phenotypic traits and disease susceptibility. Nevertheless, despite the initial expectations, a significant fraction of the genetic risk for common and complex diseases is still unexplained^{6,7}. Furthermore, not all variants have been studied at the same level of detail. In particular, inversions are a type of structural variant that changes the orientation of a genomic segment, usually without gain or loss of DNA, and they often have highly-identical inverted repeats (IRs) at their breakpoints. These characteristics make inversion detection very challenging, even with next-generation sequencing methods, and they have been largely overlooked in humans^{8,9}.

Genome-wide inversion discovery has been typically based on genome sequence comparison^{10,11} or paired-end mapping (PEM)^{4,12}, although recent studies have exploited newer techniques that could be especially useful for inversion detection, such as long-read sequencing^{13,15}, Strand-seq¹⁶, BioNano optical maps¹⁷, or a combination of them¹⁸. In most cases around 100–200 inversions have been predicted, with a maximum of 786 predictions in the 1000 Genomes Project (1000GP)^{4,19}. However, these methods are not suitable for high-throughput genotyping, and with few exceptions^{4,16,20}, just a reduced number of individuals (1–15) have been analyzed. Moreover, the presence of repetitive sequences at the breakpoints influences the inversions that can be detected by each technique and results in high error rates for inversion validation compared to other variants^{4,18,20,21}.

Apart from the intensely-studied 17q21.31 and 8p23.1 inversions^{22,23}, genotyping efforts have been restricted to a small number of inversions and samples. For example, four other large inversions have been genotyped by FISH²⁴ and five smaller inversions by PCR²⁵ in 27 and 42 individuals of four populations, respectively. In addition, PCR and inverse PCR (iPCR) have been used for targeted studies of 34 inversions in 70–90 Europeans^{21,26} and a more worldwide characterization of three inversions^{25,27,28}. Also, although inversion genotypes might be predicted based on SNP data, these methods can only detect inversions above a certain size or associated with specific SNP combinations and the error rate can be high^{21,23,25,29}. Therefore, it is not yet clear how many polymorphic inversions really exist in humans and very little is known about their global frequency and distribution¹⁹.

Actually, inversions have been a model in evolutionary biology for almost 90 years^{30,31} and there are numerous examples of their phenotypic consequences and adaptive significance in diverse organisms, from plants to birds³². One of their main effects is related to recombination, since single crossovers within the inverted region in heterozygotes generate unbalanced gametes and, at the same time, the resulting inhibition of recombination could protect favorable allele combinations^{30,31}. In addition, inversion breakpoints can directly alter the expression patterns of adjacent genes^{9,33}.

From the little information available, it is clear that inversions can have important consequences in humans⁹. Inversions are associated with haemophilia A³⁴, increased risk of neurodegenerative diseases^{35,37}, autoimmune diseases^{23,29} or mental disorders³⁸. They could also predispose to other genomic rearrangements with negative phenotypic consequences in the offspring⁹. Moreover, there is evidence that the 17q21.31 inversion increases the fertility of the carriers and has been positively selected in Europeans²². Finally, inversions have been shown to affect gene expression^{23,28,29,39}. However, most of these effects are associated with just the two best-known inversions. Attempts to associate inversions with gene-expression and phenotypic variation in large datasets have been limited to those with simple breakpoints, and only a couple of additional candidates have been identified so far^{4,40,41}. Thus, specific genotyping studies of a

diverse range of inversions in multiple individuals are necessary to determine their functional and evolutionary impact.

Here, we have developed a new high-throughput genotyping method and we have characterized in detail 45 common polymorphic inversions. By combining accurate inversion genotypes in 551 individuals of different populations and the available genomic information, we show that a large fraction of inversions are not linked to other variants and have occurred recurrently. In addition, several of them have signatures of selection and/or functional effects, emphasizing the role of inversions in the human genome.

Results

High-throughput genotyping of inversions. We focused on a representative set of 45 paracentric inversions from the InvFEST database¹⁹, which comprised most of those experimentally validated by PCR-based techniques when the project started¹⁹ and corresponds approximately to half of the estimated number of real variants with >5% frequency in human populations^{4,14} (Supplementary Fig. 1, Supplementary Data 1). These inversions were originally detected in the comparison of the hg18 and HuRef genome assemblies¹⁰ or a fosmid PEM survey in nine individuals¹², and between 1 and 36 of them have been identified in different recent studies (Supplementary Data 1). The main limitation for inversion genotyping was due to breakpoint IRs, that had to be of less than 25–30 kb and with target sites for certain restriction enzymes at both sides but not within the IRs²⁶, which excluded previously-known large inversions mediated by big repeat blocks¹⁹. Overall, the studied inversions are located throughout the genome (37 in the autosomes, 7 in chr. X and 1 in chr. Y), with sizes ranging from 83 bp to 415 kb. Also, 24 (53%) have been generated by non-allelic homologous recombination (NAHR) between >90% identical IRs (from 654 bp to 24.2 kb). The rest (47%) were probably generated by non-homologous mechanisms (NH), including 18 with small deletions or insertions in the derived allele that may have been created in a single complex FoSTeS/MMBIR event^{4,21} (Supplementary Fig. 1, Supplementary Data 1).

Of those, 41 inversions were genotyped simultaneously using high-throughput assays derived from the multiplex ligation-dependent probe amplification (MLPA) technique⁴². For inversions with simple breakpoint sequences (17), we carried out directly custom MLPA assays with minor modifications. However, for inversions with repetitive sequences at the breakpoints (24), which are difficult to detect by most techniques, we developed a new method combining the principles of iPCR²⁶ and MLPA⁴² named iMLPA. In both cases, two pairs of oligonucleotide probes were used to interrogate the two alternative orientations for each inversion, orientation 1 (*O1*) and orientation 2 (*O2*) (Fig. 1). Four additional inversions not initially included in the MLPA-like assays were tested independently by PCR or iPCR (Supplementary Data 1). The 45 inversions were genotyped in 551 individuals from seven populations studied in HapMap and 1000GP^{1,3} with African (AFR) (YRI, LWK), European (EUR) (CEU, TSI), South-Asian (SAS) (GIH) or East-Asian (EAS) (CHB, JPT) ancestry, here referred as population groups (Supplementary Table 1).

MLPA and iMLPA inversion genotypes were carefully validated through several analyses and quality controls (Fig. 1): (1) comparison with 3377 available genotypes¹⁹ (see Supplementary Data 1 for data source); (2) identification of potential iPCR or iMLPA problems caused by restriction site polymorphisms; and (3) association between inversions and other variants (see below). As part of the validation, we repeated by PCR or iPCR 2160 extra genotypes with discrepancies or possible errors plus

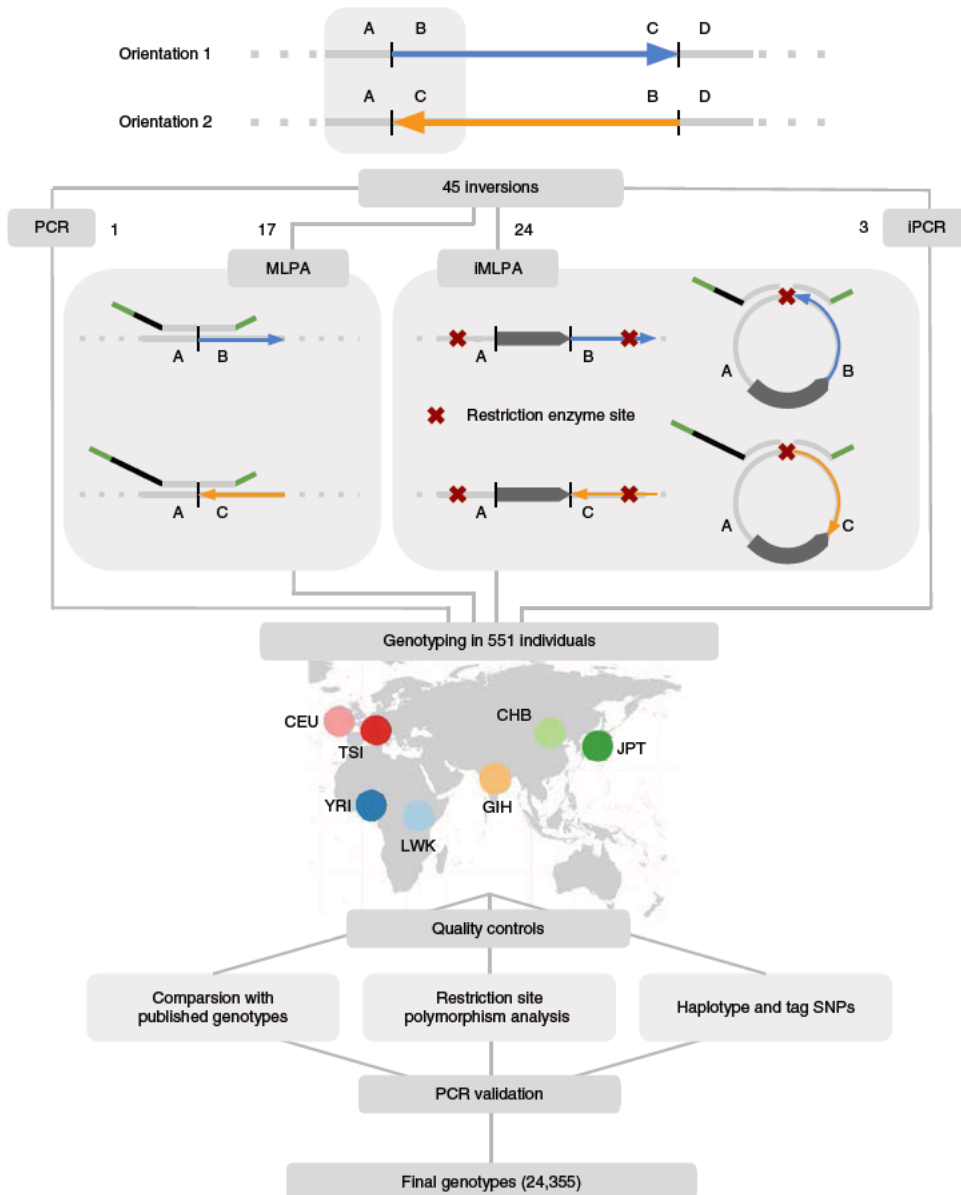


Fig. 1 Schematic representation of inversion genotyping strategy. High throughput genotyping of the 45 inversions in 551 individuals from different populations was done by MLPA (17), iMLPA (24), regular PCR (1) and iPCR (3). To avoid confusion about the ancestral status, the two inversion orientations have been named as 1 (AB and CD breakpoints) and 2 (AC and BD breakpoints), using the hg18 genome assembly as reference (Supplementary Data 1). In MLPA and iMLPA, two pairs of oligonucleotide probes (represented in top of the genome sequence) that are able to hybridize contiguously to the target region through a specific sequence complementary to the genome (light grey) were used to interrogate the two alternative orientations of each inversion. These probes, which include a stuffer sequence of variable length (black), are ligated together in a subsequent step and the resulting products are amplified for all the analyzed inversions at the same time with common primers (in green). IRs or other repetitive sequences at the breakpoints are represented as a dark pointed rectangle

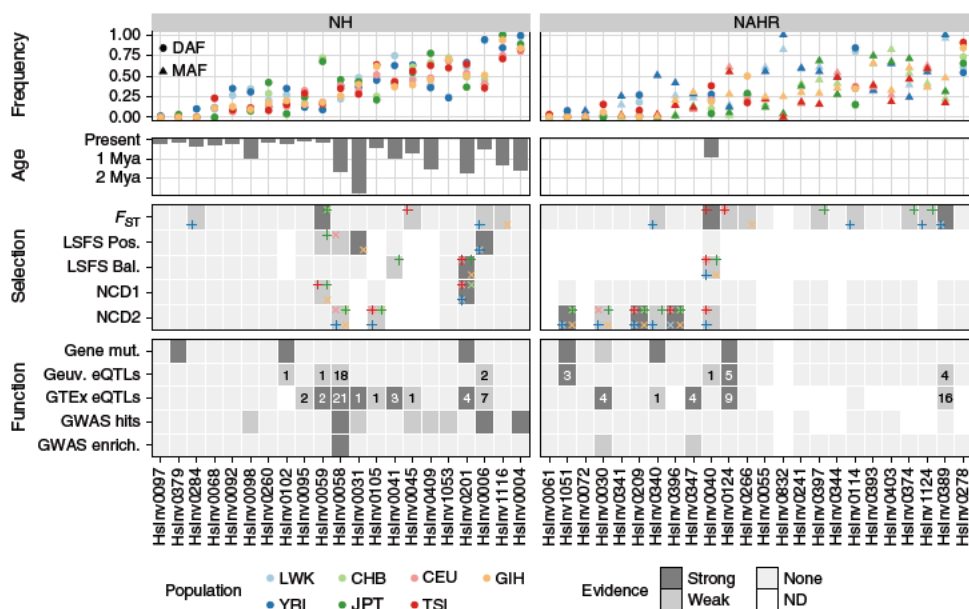


Fig. 2 Evolutionary and functional information for human polymorphic inversions. Frequency: inversion frequencies in the 480 unrelated individuals from seven populations, showing either the derived allele frequency (DAF) if the ancestral orientation is known or the minor allele frequency (MAF) according to the global frequency of the inversion otherwise (which enables the MAF to be higher than 0.5 in specific populations). Age: average age for 22 inversions in which it can be calculated from the divergence between orientations using three different substitution rate estimates. Selection: summary of inversion selection signatures from F_{ST} , LSFS (positive or balancing selection) and NCD1 and NCD2 tests. Populations where the signal was detected are indicated by different colors in the corners of each cell, with alternating vertical and diagonal crosses to avoid visual overlap. Criteria for classification of strong and weak selection evidence are explained in Supplementary Data 7. Function: functional effects of inversions summarizing direct gene mutations, which include gene or transcript disruption (strong) or exchange of genic sequences (weak) (Supplementary Table 3), eQTLs in the GEUVADIS or GTEx datasets (showing the number of affected genes and labeled as strong if inversion is lead eQTL for at least one gene) (Supplementary Data 8 and Supplementary Data 9), and association with GWAS hits (strong if for one of the associations $P < 1 \times 10^{-6}$) (Supplementary Table 4) or GWAS signal enrichment (strong if enrichment empirical test $P < 0.01$ in both GWAS databases). Source data are provided as a Source Data file

others randomly selected, including the whole set of 551 individuals for the three inversions that had the highest error rate (HsInv0045, HsInv0055 and HsInv0340) (Supplementary Data 1). This showed that the new inversion genotyping technique is very robust, with missing data (0.7%) and genotype errors for MLPA (0.1%) and iMLPA (0.9%) accumulating mainly in specific problematic inversions or DNA samples (Supplementary Fig. 2A). When compared to the global inversion data from the 1000GP⁴, just 14 of our inversions (31%) were detected, with nine of them having 2.5–71.5% incorrect genotypes and extremely low genotype agreement in the only two inversions mediated by large IRs in common (Supplementary Fig. 2B). Therefore, we have generated the largest and most accurate dataset of different types of inversions in humans (Supplementary Data 2).

As expected, the 45 inversions show correct genetic transmission in the 30 CEU and 30 YRI father-mother-child trios, and allele frequencies do not deviate from Hardy-Weinberg equilibrium in any population ($P > 0.01$). Minor allele frequencies (MAF) range globally from 0.5% to 49.7%, with 41 inversions spread through several population groups and only the four with the lowest frequencies being present in a single population group (HsInv0097, HsInv0284 and HsInv1051 in Africa; HsInv0379 in East Asia) (Fig. 2; Supplementary Table 1). On average, an

African and non-African individual carry the O2 allele, respectively, for 28 and 24 inversions.

Nucleotide variation and haplotype distribution. Thanks to the accurate genotypes, we were able to explore the linkage disequilibrium (LD) between inversions and neighboring variants (SNPs and small indels) from HapMap and 1000GP^{1,3}. While most NH inversions (20/21) have variants in complete LD ($r^2 = 1$) either inside or up to 100 kb from the breakpoint, among the 24 NAHR inversions only HsInv0040 and HsInv1051 have at least one such variant (Fig. 3a; Supplementary Table 2). Maximum r^2 values between the remaining 22 NAHR inversions and 1000GP variants range from 0.14 to 0.91 (Fig. 3a).

Also, we checked the presence of shared SNPs (not including indels) in both orientations. Consistent with recombination inhibition in heterozygotes, most NH inversions do not have shared SNPs within the inverted region in accessible 1000GP positions or HapMap data, with the exception of a few individual SNPs that might be genotype errors or gene conversion events (Fig. 3a; Supplementary Table 2). Outside of the inversion, the average proportion of shared 1000GP SNPs increases progressively after the last fixed variant, up to ~20% (Fig. 3b). This allowed us to define an extended area on each

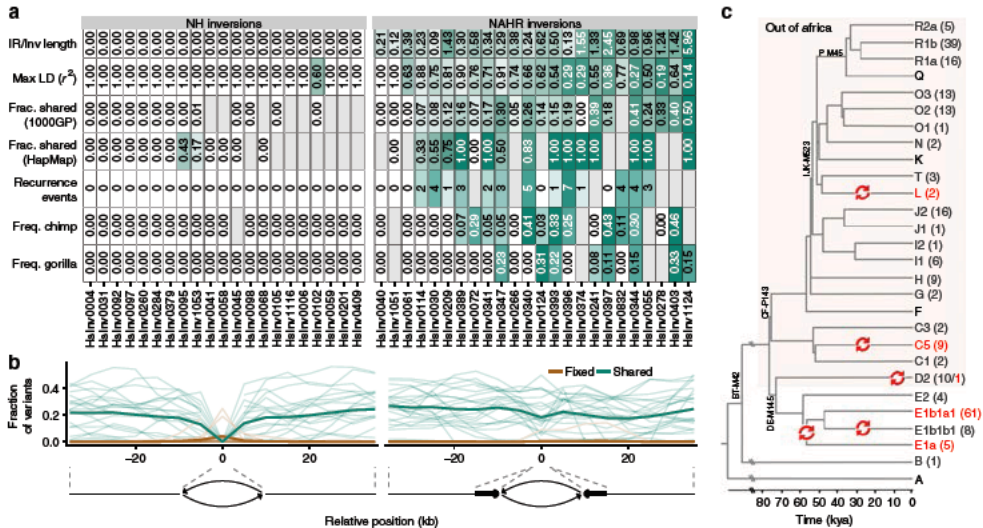


Fig. 3 Evidence of unique or recurrent origin of NH and NAHR generated inversions. **a** Summary of the ratio between breakpoint IR and inversion length, maximum LD with neighboring variants in 1000GP and HapMap, fraction of shared SNPs inside the inverted region from the total number of SNPs analyzed in 1000GP or HapMap (which includes less SNPs and results in higher shared fractions), estimated number of recurrence events from haplotype analysis, and DAF or MAF for the inversions in non human primates. Gray cells indicate values which could not be calculated. NAHR inversions show clear differences in LD with nearby variants, proportion of shared polymorphisms between orientations, recurrence events, frequency in ape species and other measures, with values that could be associated with higher levels of recurrence represented in stronger shades of green. **b** Distribution of fixed (brown) and shared (green) variants between the two orientations estimated from inversion genotypes and 1000GP data. The shared/fixed fraction with respect to the total number of polymorphic variants was computed for the whole inverted region plus 10 kb overlapping windows with a 5 kb step size in the flanking regions, and the horizontal axis represents the distance of the window central position to the inversion breakpoints (indicated by dashed lines). Thin lines represent individual inversions and thick ones represent the average for each inversion class. **c** Overview of the human chr. Y phylogenetic tree showing five different possible inversion events in the Hslnv0832 region (red arrows). Divergence dates (bottom scale) and tree topology are based on Poznik et al.⁷⁷. For each chr. Y haplogroup, the number of males genotyped for Hslnv0832 with each orientation is indicated in parenthesis (O1 in dark grey and O2 in red). Only the main branches and those including genotyped individuals are shown, with some characteristic mutations indicated in the tree. For haplogroup E, one of the two possible scenarios is represented, with the alternative being two inversion events in E1a and E1b1a1 haplogroups. Source data are provided as a Source Data file

side of the inversion of no or little recombination between orientations, which ranges from 0 to more than 20 kb (Supplementary Table 2). In contrast, 20 of the 24 NAHR inversions have a considerable number of shared SNPs scattered throughout the inverted and flanking regions (Fig. 3a, b; Supplementary Table 2).

Next, we visualized the haplotype diversity and distribution across orientations and populations using haplotype networks and a new representation integrating a hierarchical clustering of haplotypes and the differences between them (Supplementary Fig. 3). This analysis was focused on 1000GP data, although consistent results were obtained with HapMap SNPs. After taking into account possible phasing errors, two clearly differentiated patterns were observed again. In the 20 NH inversions with sequence variation information, the haplotypes of one of the orientations tend to cluster together, supporting a unique origin of the inversion (Supplementary Fig. 3). In NAHR inversions, this is true only for Hslnv0040, Hslnv0061 and Hslnv1051, with the other 21 having O1 and O2 haplotypes mixed throughout the network and hierarchical cluster, including in many cases identical haplotypes with both orientations (Supplementary Fig. 3). Such pattern is consistent with a multiple origin of these inversions and explains the absence of fixed SNPs and the high number of shared SNPs between orientations.

Based on the results of the different analyses, we inferred the minimum number of recurrent inversion events and their distribution in human populations (Fig. 3a; Supplementary Data 3). However, this relies on having differentiated haplotype clusters in which the existence of the alternative orientation cannot be easily explained by other factors (such as gene conversion or genotype/phasing errors of a few variants). Another problem is that recombination generates mixed sequences between haplotype groups and makes it difficult to accurately quantify recurrence. Thus, a nice example is chr. Y inversion Hslnv0832, in which there is no recombination. We used available haplogroup information of 232 males from the known chr. Y genealogical tree (Supplementary Data 4) to identify five independent inversion events in the last ~60,000 years (Fig. 3c). This results in an inversion rate of 5.35×10^{-5} per generation (see Methods), which is ~1000 times higher than that of single bases. For the inversions in which it is possible to quantify recurrence, we estimated a total of 40 additional inversion and re-inversion events (ranging for each inversion from 0 to 7 with an average of 2.2) (Supplementary Data 3). Of those, 12 are distributed globally and could precede the out-of-Africa migration, 17 are restricted to African individuals, and 11 probably appeared more recently in non-African populations. The fact that many of the recurrence events are shared by several individuals indicates that they are not

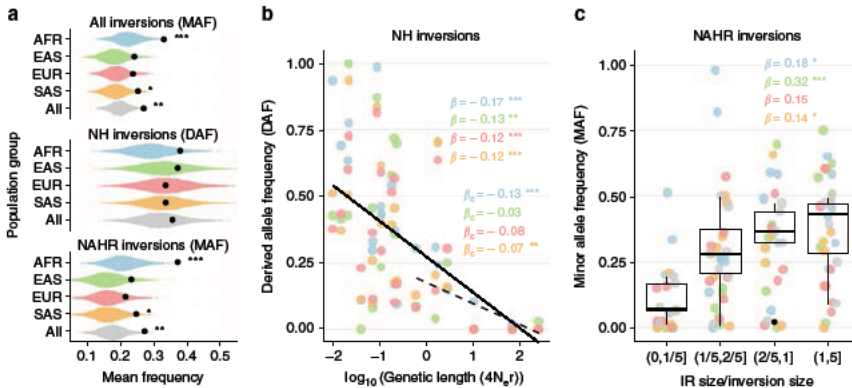


Fig. 4 Determinants of inversion frequency in human populations. **a** Observed mean inversion frequency per population group and mechanism of generation (black dots) compared with that expected from the detection method simulations using a null distribution of 10,000 SNP samples. Graphs represent DAF only if the ancestral orientation is known for all the inversions included and MAF otherwise. **b** Logarithmic robust regression between DAF of NH inversions and genetic length (measured in $4N_e r$ units) in the different population groups, showing a significant negative trend for all inversions (solid line; β regression coefficient) and for inversions larger than 2 kb to correct in part the detection bias against low frequency small inversions (dashed line; β_c regression coefficient). **c** Boxplots showing a positive relationship between the frequency of the minor allele in all populations together and IR/inversion length ratio for NAHR inversions (centre, median; bounds of box, first and third quartiles; and whiskers, smallest/largest value within 1.5 interquartile range). β indicates the robust regression coefficient in every population group. In **b** and **c** dots indicate the frequency of each inversion with population groups represented with the same colors as in **a**. Two tailed t test (**a**) and robust regression t test (**b** and **c**) P values: * $P < 0.05$; ** $P < 0.01$; and *** $P < 0.001$. Source data are provided as a Source Data file

artifacts from lymphoblastoid cell line (LCL) culture. In addition, as part of the validation of inversion genotypes, all the recurrence events were confirmed by checking at least one of the supporting individuals by PCR/iPCR. We have therefore extended considerably the previous recurrence analysis of some of the inversions in just the CEU population^{21,26}.

Ancestral orientation and inversion age. The published data on the ancestral orientation of 32 of the inversion regions^{21,26,28} was complemented and expanded by experimentally testing 42 inversions for which the human or modified assays generated reliable results in a panel of 23 chimpanzees (40 inversions) and seven gorillas (41 inversions) (Supplementary Data 5). Inversion orientation was also assessed in available genome assemblies of both species plus orangutan and rhesus macaque (Supplementary Data 6). In total, we could infer the ancestral allele for 29 inversions, with 15 showing the ancestral and 14 the derived allele in the human reference genome. For the 21 NH inversions, orientation was consistent in all the non-human samples and genomes analyzed, and with existing deletions and insertions occurring in the derived allele (Supplementary Data 1 and Supplementary Data 6). In contrast, 14 of the 22 NAHR inversions experimentally genotyped were polymorphic in at least one of the apes (six in chimpanzees, two in gorillas and six in both species) and several had opposite orientation in different primate assemblies (Supplementary Data 6). This agrees with previous analyses in a much smaller set of samples^{21,26}, but we found five additional polymorphic inversions in each species. In fact, the lower number of polymorphic inversions in gorilla indicates that more inversion regions might be identified as polymorphic in non-human primates by analyzing more individuals. Given the species divergence times, the most likely scenario is that shared inversions have appeared independently in chimpanzee and gorilla lineages, providing additional support for inversion recurrence (Fig. 3a).

Moreover, we checked the presence of the breakpoint sequences of 19 NH inversions without IRs in available Neanderthal, Denisovan and two ancient modern human genomes (Supplementary Data 6). Five inversions (HsInv0004, HsInv0006, HsInv0201, HsInv0409, and HsInv1116) showed the derived orientation in the Neanderthal or Denisovan genomes (including two with the derived orientation in both). These inversions are distributed through African and non-African populations, which suggests that they are not the result of introgression and they appeared before the divergence of the most recent *Homo* groups, around 550,000–750,000 years ago (ya)⁴³.

Finally, we dated more precisely 22 unique inversions from the sequence divergence between orientations (Fig. 2; Supplementary Data 6). Six inversions were estimated to have appeared more than 1 Mya, including four with the derived orientation in Neanderthal or Denisovan, plus HsInv0031 and HsInv0058. For HsInv0006, the estimated age (407,795–495,470 ya) is slightly more recent than the Neanderthal-Denisova and modern human divergence, but it is very close to its lower bound. Ages of the rest of inversions are consistent with their geographical distribution, with inversions restricted to either African (HsInv0097 and HsInv0284) or non-African populations (HsInv0379) having a relatively recent origin. Only in two cosmopolitan inversions age estimates are lower than population split times (HsInv1053 with a negative age and HsInv0095 with 22,582–41,258 ya), probably due to an underestimation of the divergence between orientations caused by the limited sequence information available.

Analysis of inversion frequencies. To evaluate whether there are selective pressures acting globally on inversions, we compared the inversion frequency spectrum with that of SNPs sampled according to neutral expectations (see Methods). Although the ascertainment bias associated with inversion detection predicts an enrichment of high-frequency inversions (~0.20 expected MAF), the observed frequencies tend to be higher than expected

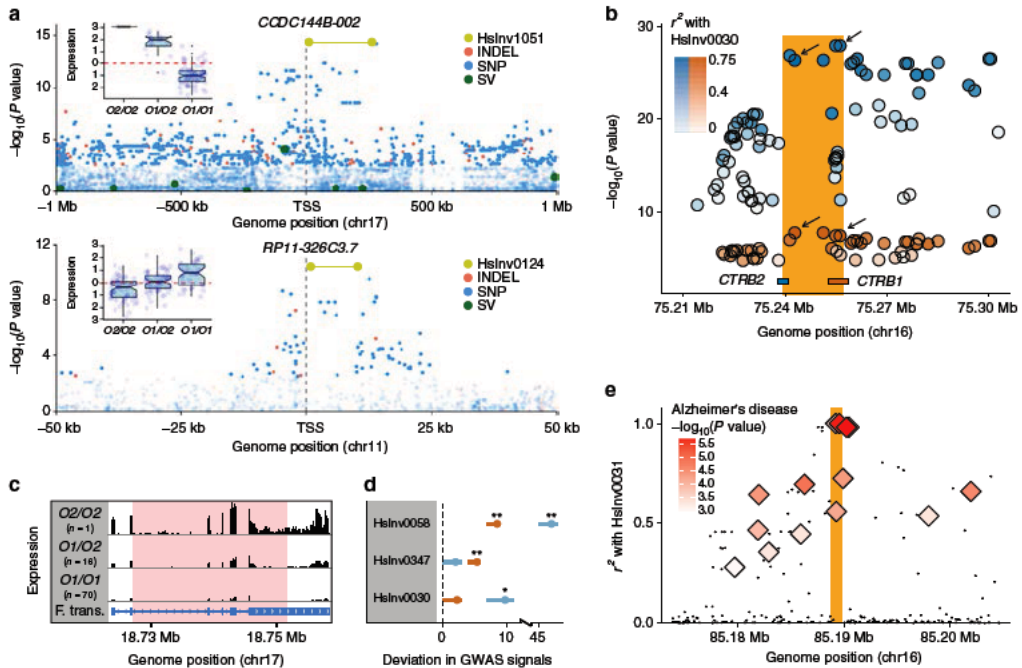


Fig. 5 Examples of inversion functional effects. **a** Manhattan plots of logarithm transformed linear regression t test P values for *cis*eQTL associations in LCLs of transcript *CCDC144B-002* and gene *RP11-326C3.7*, showing inversions Hslnv1051 (top) and Hslnv0124 (bottom) as lead variants, together with boxplots of rank normalized expression and inversion genotype (centre, median; bounds of box, first and third quartiles; whiskers, smallest/largest value within 1.5 interquartile range; and notches, 95% confidence interval of the median). **b** Manhattan plot of logarithm transformed GTEx P values of pancreas eQTLs for *CTRB1* (red) and *CTRB2* (blue) mapping around the Hslnv0030 region (orange bar). Top lead eQTLs for each gene were the variants in highest LD with the inversion (rs9928842, $r^2 = 0.75$; rs8057145, $r^2 = 0.73$; black arrows). **c** Schematic diagram of average RNA-seq expression profile in Hslnv0151 genotypes of GEUVADIS LCL reads mapped to the inverted allele, showing the generation of a fusion transcript (F. trans.) with new 3' sequences that loses the last two thirds of the gene exons. Inversion breakpoint is indicated in pink. **d** GWAS Catalog (orange) and GWASdb (blue) signals enrichment within individual inversions and flanking regions (± 20 kb). Dot represents the mean and error bars the 0.95 confidence interval of the difference between the observed number of GWAS signals and a null distribution from 1000 random genomic regions for each inversion. One tailed empirical test P value: ** $P < 0.01$; * $P < 0.05$. **e** LD (r^2) between Hslnv0031 (orange bar) and 1000GP variants (black points) or Alzheimer's disease GWAS signals (diamonds)⁷⁸. Source data are provided as a Source Data file

in all population groups (Fig. 4a). When inversions are separated by the generation mechanism, the increase in frequency is significant only in NAHR inversions (Fig. 4a). We also investigated how diverse genomic variables affect autosomal and chr. X inversion frequencies. The most significant predictor is inversion genetic length, which is negatively correlated with DAF of NH inversions and explains 23–55% of the frequency variance in the different population groups (Fig. 4b), followed by physical length (19–54% variance explained). For NAHR inversions, only 5–14% of MAF variance is accounted by their genetic length, whereas the ratio between IR and inversion length is positively correlated with inversion frequency in all population groups and explains 13–45% of MAF variance (Fig. 4c). This suggests that the higher frequency of NAHR inversions might be related to their repeated generation by recurrence.

Selection on human inversions. To investigate signals of natural selection acting on specific inversions, we first measured inversion frequency differences between populations using the fixation index

(F_{ST}), which can identify positive selection leading to a rapid increase in allele frequency in some areas. The global F_{ST} value was 0.11 for autosomal inversions, 0.21 for chr. X inversions, and 0.73 for the one in chr. Y, with the largest frequency variation between continents. Three inversions were within the top 1% of the F_{ST} distributions derived from SNPs with the same frequency (Fig. 2; Supplementary Data 7): Hslnv0040, with frequency differences between European populations, and Hslnv0389 and Hslnv0059, with high frequency in Africa or East Asia, respectively. Eleven more inversions fell within the top 5% of the empirical F_{ST} distribution, and were considered to have weak evidence of selection (Fig. 2; Supplementary Data 7).

Second, we applied a novel test based on the frequency spectrum of linked sites (LSFS), which is well suited to detect deviations from neutrality in low-recombination regions, such as inversions. We used optimized tests to identify positive and long-term balancing selection maintaining a polymorphism in several populations, and significance was assessed empirically using the null LSFS distribution from autosomes. Only 18 unique autosomal inversions with nearby perfect tag variants

that could be reliably phased were analyzed, including most NH inversions and HsInv0040 (Fig. 2; Supplementary Data 7). The strongest signals (empirical test $P < 0.01$) were in HsInv0201 for balancing selection and HsInv0006 and HsInv0031 for positive selection. In addition, four other inversions showed weaker evidence of balancing or positive selection. Consistent with the F_{ST} results, in HsInv0006 and HsInv0059 positive selection was detected in those populations with increased DAF (Fig. 2; Supplementary Data 7).

As independent confirmation of balancing selection signatures, we also used the recently developed non-central deviation statistics, NCD1 and NCD2, which detect site frequency spectrum shifts towards an equilibrium frequency and an excess of polymorphic sites⁴⁴. However, the results of these tests summarize the data of all the SNPs in a region and are not necessarily linked to the inversion, as before. Focusing on signals detected in at least three populations, we found respectively four and six inversions with strong and weak signatures of balancing selection for NCD1 or NCD2 (Fig. 2; Supplementary Data 7). Many of these candidates could not be analyzed with the LSFS method because of the lack of tag SNPs or correspond to low-frequency inversions, such as HsInv1051 and HsInv0209, that are unlikely the targets of selection, but consistent results were found for HsInv0201.

Effect of inversions on genes and gene expression. As previously described, some of the analyzed inversions can have important effects on genes^{4,9,21,25,28}. Although half of our inversions (21/45) are located in intergenic regions, three of them invert genes, eight are located within introns, seven might exchange gene sequences overlapping the IRs at the breakpoints, and six affect genes more directly through the inversion or deletion of an internal exon (HsInv0102, HsInv0201) and the disruption of the whole gene (HsInv0340, HsInv0379, HsInv1051) or an alternative transcript isoform (HsInv0124) (Fig. 2 and Supplementary Table 3).

We measured the effect of the 42 autosomal and chr. X inversions with $MAF > 0.01$ on expression of nearby genes (<1 Mb away) by linear regression between inversion genotypes and LCL transcriptome data from the GEUVADIS consortium⁴⁵. To increase statistical power and reliability, the analysis was replicated in two datasets: (1) 173 CEU, TSI and YRI individuals with inversion genotypes; and (2) the complete GEUVADIS set of 445 European (358) and African (87) individuals in which the genotypes of 33 inversions could be imputed accurately (Supplementary Fig. 4A). Considering the largest sample size for each inversion, we uncovered eight inversions significantly associated with LCL expression of 27 genes and 44 transcripts (Supplementary Fig. 4B; Supplementary Data 8), with highly concordant results for those analyzed in both datasets (7/7 genes and 11/12 transcript effects were replicated) (Supplementary Fig. 4C). As negative control, no associations were observed by permuting inversion genotypes relative to expression levels (Supplementary Fig. 4D). Moreover, significant expression effects were robust when applying different analysis approaches (see Supplementary Fig. 4E-F and Supplementary Methods), and inversions acting as expression quantitative trait loci (eQTLs) located significantly closer to the transcription start site (TSS) of the differentially expressed genes (<100 kb) (Supplementary Fig. 4D).

Next, we examined inversion expression effects in other tissues through variants already reported as eQTLs in the GTEx project⁴⁶. We found 62 genes with eQTLs in different tissues in high LD ($r^2 \geq 0.8$) with 11 of the 26 analyzed inversions, including seven not detected in LCL data (Supplementary Fig. 5;

Supplementary Data 9). By searching for eQTL signals in moderate LD ($r^2 \geq 0.6$) with some of the recurrent inversions, we found additional potential expression differences associated with HsInv0124 and in the genes affected by HsInv0030, which exchanges the first exon and promoter of chymotrypsinogen precursor genes *CTRB1* and *CTRB2* expressed only in pancreas^{21,25}, and HsInv0340, which disrupts the long non-coding gene *LINC00395* expressed in testis (Supplementary Fig. 5; Supplementary Data 9). In total, 17/27 of inversion-gene associations in LCLs were also identified in the smaller GTEx sample (Supplementary Fig. 5).

To assess if inversions were the main cause of the observed expression changes, we performed a joint eQTL analysis in LCLs including our inversions together with SNPs, indels and structural variants from the 1000GP^{3,4}. Two inversions, HsInv0124 and HsInv1051, were the most likely causal variant for two genes and three transcripts in LCLs (Fig. 5a). Six other inversions show the highest LD ($r^2 \geq 0.9$) with variants reported as first or second lead eQTL in a given tissue by the GTEx project (Supplementary Fig. 6). Similarly, for recurrent inversions HsInv0124 and HsInv0030, eQTL significance in GTEx data increases with LD with the inversions, supporting their causal role (Fig. 5b). In general, some of the strongest effects are related to inversions affecting exonic sequences, although the consequences can be complex and need to be investigated in detail. For example, HsInv1051 breaks the *CCDC144B* gene and the apparent upregulation of specific isoforms (Fig. 5a) is actually due to the creation of a fusion transcript with new sequences at 3' (Fig. 5c; Supplementary Fig. 7). HsInv0124 is the lead variant for the antisense RNAs *RP11-326C3.7* and *RP11-326C3.11*, which overlap respectively the *IFITM2* and *IFITM3* genes located at the breakpoints, and it has opposite effects in the two pairs of overlapping transcripts (Supplementary Data 8; Supplementary Data 9). Also, HsInv0102 removes the *RHOH* isoforms with the alternative non-coding exon that gets inverted, but its effect is masked by a more frequent lead eQTL (SNP rs7699141) acting in the same direction. On the other hand, HsInv0058 is associated with chr. 6 MHC haplotypes APD, COX, DBB, QBL and SSTO, which extend ~4 Mb and harbor important functional differences⁴⁷, suggesting that other variants in these haplotypes are responsible for the observed effects.

Inversions and phenotypic traits. The role of inversions in phenotypic variation was investigated using available genome-wide association studies (GWAS) data. We found a 1.26- and 1.95-fold increase in GWAS Catalog⁴⁸ and GWASdb⁴⁹ variants in the inversion and flanking regions (± 20 kb). The top inversion driving this result was the MHC-inversion HsInv0058, but HsInv0030 and HsInv0347 showed similar enrichment of GWAS hits in both datasets (Fig. 5d; Supplementary Fig. 8A-B). GWAS signals close to the latter two inversions are consistent with their effect on genes, involving type 1 and 2 diabetes, pancreatic cancer, insulin secretion, and cholesterol and triglyceride levels for HsInv0030, and glaucoma and optic disc and nerve characteristics for HsInv0347, which is associated to the expression of *c14orf39* (*SIX6OS1*) and *SIX6*, related to eye development.

We also explored whether inversions were in strong LD ($r^2 \geq 0.8$) with known GWAS hits in the population where the association was reported. That is the case of HsInv0004, which is in complete LD in Europeans with a nearly genome-wide significant GWAS SNP related to asthma susceptibility in children and another one associated with body mass index in asthmatic children (Supplementary Table 4). Moreover, HsInv0006 is linked to schizophrenia in Ashkenazi Jews and glaucoma in Europeans (Supplementary Table 4). Remarkably,

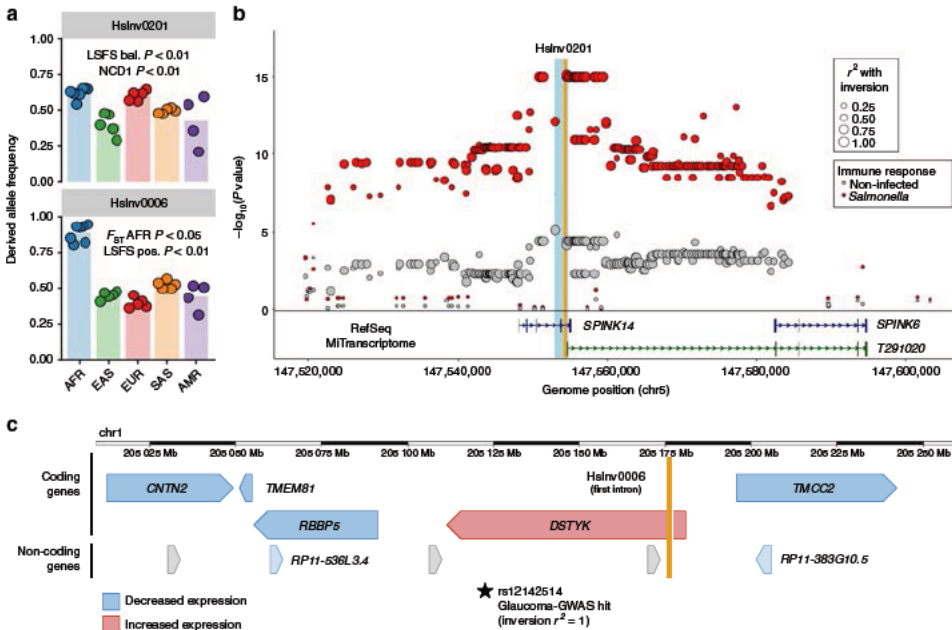


Fig. 6 Integrative evolutionary and functional analysis of inversion candidates. **a** Frequency of Hslnv0201 and Hslnv0006 across worldwide human populations from 1000GP Ph3 (colored dots) and in each population group (colored bars) estimated from global inversion tag SNPs (rs200056603 and rs79619752, respectively), showing a summary of test results for balancing (bal.) and positive (pos.) selection. **b** Manhattan plot of logarithm transformed linear regression t test P values for *cis* eQTL associations with expression variation of gene *SPINK6* in infected and non infected primary macrophages⁵², showing SNPs in complete LD with Hslnv0201 (orange bar) as lead eQTLs in *Salmonella* infection. Variants are represented as circles with varying size depending on the LD (r^2) with the inversion. The location of the genes in the region is shown below, including a new *SPINK6* isoform discovered by MiTranscriptome assembly⁷⁹, whose first exon is removed by a deletion associated to the inversion breakpoints (light blue bar). **c** Diagram of Hslnv0006 (orange bar) genomic region showing the effect of the inverted allele on the expression of neighboring genes in different tissues according to the GTEx data and the inversion tag SNP in Europeans associated to increased risk of Glaucoma (Supplementary Table 4). Arrowheads indicate the direction of transcription of the genes. Source data are provided as a Source Data file

several of these inversions affect gene expression as well (Fig. 2). A good example is Hslnv0031, which is associated with lower levels of *FAM92B* in cerebellum and is in almost perfect LD in Europeans ($r^2 = 0.98$) with SNP rs2937145 associated with Alzheimer's disease risk (Fig. 5c; Supplementary Table 4). Nevertheless, for many inversions the low LD with SNPs means that any effects would be missed in typical array-based GWAS (Supplementary Fig. 8C).

Discussion

This work represents the most exhaustive and accurate study of human inversions so far, including a significant fraction of common inversions¹⁹, and it is focused on inversions mediated by IRs, which generally escape detection. New genome-wide techniques are providing a more complete inversion catalogue^{13, 18}, but they tend to be laborious and expensive, and population data in well-characterized individuals are still scarce. In addition, in many cases the high genotype error rate has precluded identifying tag variants that could be used to infer inversion effects^{4, 40}. Thus, the new genotyping method developed here and the reliable information from multiple individuals of different populations generated are crucial to fill the void in the knowledge of this type of variation.

Despite the effort to include as many inversions as possible, only those validated and with well-characterized breakpoints could be analyzed¹⁹. In addition, important limitations exist for the study of those mediated by IRs. Although it is not clear how many of them are real, in the InvFEST database there are ~100 inversion predictions with 1–25 kb inverted segmental duplications (SDs) at the breakpoints, which could be potentially interrogated using our method, and ~250 with larger SDs, which are currently out of reach of the methodology¹⁹. Therefore, further improvements and complementary strategies, such as the possibility of making directed cuts in specific positions or estimating the distance between regions separated by larger repeats, are necessary to expand the range of analyzed inversions. Nevertheless, the present work already offers a good picture of the contribution of common inversions to genetic diversity, adaptation, and phenotypic traits in humans.

In particular, we found clearly contrasting patterns for inversions generated by homologous and non-homologous mechanisms, supporting a high-degree of recurrence of all inversions mediated by highly-identical IRs except three (two of which have very low frequency). Recurrent inversions are characterized by low LD with other variants, a large amount of shared SNPs and shared haplotypes between O1 and O2 chromosomes (Fig. 3). Also, recurrence is not limited to humans but extends to other

great ape species. Similar results had been found in previous small-scale studies^{21,24,26}, but thanks to the analysis of many more individuals and populations, multiple controls (such as genotype confirmation by PCR of both breakpoints and different sources of nucleotide variation data) and a detailed analysis of haplotype relationships, we have obtained a better estimate of the independent inversion events. Specifically, we found nine more recurrence events in the five inversions originally predicted as recurrent in humans and that seven inversions considered to be unique or lacking information in CEU are recurrent as well^{21,26}. This suggests that, like other repeats, IRs are rearrangement hotspots and the genome is more dynamic than previously thought. However, recurrence has been detected only through the sequences associated with the inversion, and we need more direct ways to quantify inversion generation rates precisely.

As for other types of recurrent changes^{50,51}, the lack of association between SNPs and many inversions and poor coverage by common arrays (Supplementary Fig. 8C) means that their phenotypic effects have been largely missed in GWAS. For example, half of the NAHR inversions cannot be accurately predicted with typical imputation algorithms (Supplementary Fig. 4A). We have found 23 and 22 inversions with different selection and functional signatures, respectively (Fig. 2). More importantly, although not all analyses could be applied to every inversion, there is a significant enrichment of inversions with both effects on genes or gene expression and selection signals directly linked to them (Fisher's exact test $P = 0.0320$) (see Supplementary Methods). This combination of the two independent types of evidence strongly indicates that inversions can have important consequences in humans.

One particularly interesting example is HsInv0201, an old inversion (>1.5 Mya) with intermediate frequency around the globe and clear signals of balancing selection (Fig. 6a), which deletes an exon of *SPINK14* and is lead eQTL for two nearby genes (Supplementary Data 9). Moreover, the inversion haplotype is the main responsible for the lower *SPINK6* expression during immune response to *Salmonella* infection^{52,53} (Fig. 6b). In fact, HsInv0201 eliminates the promoter and first exon of a putative novel *SPINK6* transcript (Fig. 6b) and it is in high LD ($r^2 = 0.971$ in EUR) with SNPs accounting for plasma levels of *SPINK6* protein⁵⁴. Together with the role of several of the affected genes in lung and extracellular mucosae, this suggests that the inversion could be related to immune response. In HsInv0006, its particular distribution pattern and the selection test results point to positive selection of the derived allele in Africa (Fig. 6a). Furthermore, the inversion is located within *DSTYK* first intron²¹ and is associated with expression changes in the proximal genes, including *DSTYK* upregulation in different tissues (Fig. 6c; Supplementary Fig. 5). *DSTYK* deletion causes pigmentation problems and elevated cell death after ultraviolet irradiation⁵⁵. Thus, positive selection on these traits could explain the inversion increase in Africa. Incidentally, the inverted orientation has been linked to higher risk of glaucoma in Europeans (Supplementary Table 4) and glaucoma is more common and severe in individuals from African ancestry⁵⁶. Other interesting candidates include HsInv0031, HsInv0059, HsInv0124, HsInv0340, and HsInv0389 (Fig. 2).

Inversions differ from other genetic variants because of their expected negative consequences in fertility resulting from the generation of unbalanced gametes by recombination^{30,31}, which is exemplified by the reduction in frequency with genetic length (Fig. 4b) and the small number of inversions described compared to CNVs¹⁹. According to this, there could be a maximum length for an inversion to behave neutrally in terms of its fertility effects. Above that size, some type of compensatory selection, perhaps related to advantageous regulatory changes on nearby genes, would be necessary for the inversions to reach a certain

frequency. Therefore, this may explain the observed enrichment of inversions with functional and selection signals. Future in-depth studies of the identified candidates and other inversions will help uncover their real role in human evolution and the unexplained part of the genetic basis of complex traits.

Methods

Human and ape DNA samples. We used 550 human samples included in the last phase of the HapMap Project and many of them in 1000GP phase 1 (Ph1) and 3 (Ph3)¹⁻³, which belong to seven populations of four main population groups (AFR, EUR, SAS, and EAS), plus individual NA15510 of unknown origin (see Supplementary Table 1 and Supplementary Data 2 for details). Most individuals were unrelated, but 70 are either children of mother father child trios (30 in YRI and 30 in CEU) or cryptic first and second degree relatives (9 in LWK and 1 in GIH)^{2,3}. Genomic DNAs of 70 CEU and 10 YRI samples and NA15510 were extracted from LCLs commercialized by the Coriell Cell Repository (Camden, NJ, USA), while the rest of DNAs were acquired from Coriell^{21,26,28}. Chimpanzee and gorilla DNAs include six already used to genotype most of the inversions from frontal cortex tissue samples of the Banc de Teixits Animals de Catalunya (N457/03, Z01/03 and Z02/03) or LCLs from Barcelona Zoo individuals (PTR1211, PTR1213, and PTR1215)^{21,26}, plus 19 chimpanzee and five gorilla DNAs extracted from a previously existing collection of primate LCLs of one of the authors (Supplementary Data 5). Ape samples comprise four mother father child trios and one father son pair in chimpanzees and one father son pair in gorillas. As for humans, cells were grown in 75 ml flasks to nearly confluency and high molecular weight genomic DNA was obtained using a standard phenol chloroform extraction²⁶. All procedures that involved the use of human and non human primate samples were approved by the Animal and Human Experimentation Ethics Committee (CEEAH) of the Universitat Autònoma de Barcelona.

Experimental genotyping of inversions. Initial genotypes for 41 inversions were obtained by newly developed assays based on the MLPA technique (Supplementary Data 1), which has been widely used for genome copy number analysis and consists on the multiplex amplification of fragments of different sizes with common primers that are fluorescently labeled, and their detection by capillary electrophoresis⁴². Most NH inversions (17) were genotyped simultaneously from 100–150 ng of DNA with a slightly modified MLPA protocol using two pairs of probes that bind specifically at the breakpoint sequences of each orientation (AB and CD or AC and BD), with one of the probes that could be common to both pairs (Fig. 1). For 24 inversions with IRs or other repetitive sequences at the breakpoints, we developed a new iMLPA method that uses a combination of iPCR and MLPA. iMLPA requires some extra processing of the DNA with a restriction enzyme that cuts at each side of the breakpoint repetitive sequences, followed by self circularization of all the digested DNA molecules together by ligation in diluted conditions with T4 DNA ligase and DNA purification. Then, MLPA was carried out as usual with a pair of probes that bind specifically at the self ligation site of the circular molecules from each orientation (see Supplementary Methods for MLPA and iMLPA details). Supplementary Data 10 lists the sequences and concentrations of the probes used for MLPA (68) and iMLPA (87).

PCRs and iPCRs were carried out to genotype seven inversions in the 551 human samples (including four not analyzed in the MLPA/iMLPA assays) and to validate many of the MLPA/iMLPA genotypes (Fig. 1; Supplementary Data 1). Multiplex PCRs and iPCRs of each inversion were done with primers flanking either the breakpoint (PCR) or the self ligation site of circularized molecules (iPCR) from the two orientations^{21,26} (Supplementary Data 11). For the restriction site polymorphism analysis, we downloaded dbSNP (version 142) SNPs and indels around the inversion region (± 50 kb) and determined all possible restriction site gains or losses affecting the iPCR/iMLPA experiments. To ensure that the genotypes were completely accurate, in most potential discrepancies both breakpoints of each orientation were tested.

For inversion genotyping in chimpanzees and gorillas, first we ran the same MLPA and iMLPA assays as described above, and those inversions that did not work were genotyped by PCR or iPCR. In some cases, new chimpanzee or gorilla specific primers and restriction enzymes for iPCR were used to overcome human assay problems²⁶ (Supplementary Data 5; Supplementary Data 11). However, this was not always possible due mainly to deletions or genome gaps, and a few inversions in one or both species could not be tested. All polymorphic inversions in chimpanzees or gorillas were validated by PCR or iPCR of at least one breakpoint to make sure that there were no errors in the iMLPA results.

Analysis of nucleotide variants associated with inversions. We measured pairwise LD (r^2) between inversions and overlapping and neighboring biallelic variants up to 200 kb at each side of the inversion from 1000GP Ph3 (including SNPs and indels for 434 unrelated individuals) and HapMap release 27 (including fewer SNPs but all the 480 unrelated individuals) using either plink v1.90⁵⁷ or Haploview v4.1⁵⁸. Variants located within the breakpoint interval and associated deletions or IRs were excluded for this and subsequent analysis to avoid possible genotyping errors. Supplementary Data 12 lists the inversion tag variants with $r^2 \geq$

0.8 from 1000GP and HapMap data considering all the individuals together, as well as each population and population group. This analysis allowed us to detect a few inversion genotypes that did not match those expected from the tag variants and most of them were confirmed to be genotyping errors by independent PCR or iPCR validation (Supplementary Fig. 2). SNPs and indels around inversion regions were further classified directly from the genotype data according to its distribution across orientations in fixed ($r^2 = 1$), shared (unambiguously polymorphic in both orientations) and polymorphic in *O1* or *O2* chromosomes using in house perl and bash scripts^{21,26}. To minimize possible genotyping errors, only the most reliable variants according to the 1000GP strict accessibility mask were included. Non recombining flanking regions were defined from the 1000GP data (which has more resolution than HapMap) as the longest sequence outside the breakpoints up to a maximum distance of 20 kb that: (1) does not contain reliable shared variants compatible with a crossing over event between orientations; and (2) includes most of the fixed variants (Supplementary Table 2).

Phasing and visualization of haplotypes. Each orientation haplotypes were determined following two complementary strategies to minimize errors and obtain more robust results. First, after testing several commonly used phasing programs, we selected PHASE 2.1⁵⁹ because it avoids switch errors in inversion heterozygotes by fixing the phase of the two orientation alleles added at the breakpoint positions^{21,26}. Phasing was done independently for the YRI, LWK, EUR, SAS, and EAS populations or population groups, using the available trio information when possible and five iterations ($-x5$) for HapMap data and two iterations ($-x2$) with the hybrid model ($-MQ$) for 1000GP data. Only variants within the inverted region plus 20 kb of flanking sequence for 1000GP Ph1 data (which was the only one available at that time) or 200 kb for HapMap data were selected. Second, we took advantage of the 1000GP Ph3 phased haplotypes³ to impute directly the inversion orientation based on the presence of perfect tag variants ($r^2 = 1$). For inversions without perfect tag variants, only homozygotes or hemizygotes for each orientation, in which the inversion status of the haplotypes can be assigned unambiguously, were analyzed.

The relationships between the different haplotypes of each dataset were visualized by combining several data sources and representation methods. Phased 1000GP Ph1 and HapMap haplotypes were used to build Median Joining (MJ) networks⁶⁰ with the NETWORK v.4.6.1.3 software (www.fluxus-engineering.com). In addition, we devised our own representation of the haplotype relationships and the distribution of nucleotide changes along the sequence, named integrated haplotype plot (iHPlot), which combines a hierarchical clustering, distance matrix and visual alignment of the alleles in each polymorphic position, plus the haplotype orientation and the populations in which it is present (see Supplementary Methods for details).

Inversion origin was estimated from the information of the different phasing and haplotype visualization strategies, which overall showed consistent results (see Supplementary Methods). For inversions with perfect tag variants, the analysis was based mainly on 1000GP Ph3 haplotypes (including when possible the flanking non recombining region), which allowed a better discrimination of haplotypes, filtering of accessible SNPs according to the pilot accessibility mask and had less phasing errors due to the use of sequences from more individuals³. For inversions without tag variants, we relied mainly in the phased 1000GP Ph1 haplotypes iHPlots, since all the genotyped individuals in common could be analyzed. Inversion recurrence events were conservatively estimated by identifying clusters of haplotypes with both orientations that differ significantly from all others with the same orientation as the potential recurrence event, after eliminating possible phasing errors in inversion heterozygotes (see Supplementary Methods). Around 125 individuals supporting each of the independent recurrence events were validated by PCR or iPCR (in this case testing mostly both breakpoints) to discard possible inversion genotyping errors (Supplementary Data 3). Also, we validated the genotype of many more individuals with unexpected orientation haplotype combinations and of other inversions with a high proportion of shared SNPs for which recurrence events could not be clearly identified. HsInv0832 inversion rate was estimated from the publicly available information of 232 of the 282 genotyped males (Supplementary Data 4) by calculating the total number of mutations and generations along the genealogical tree of the analyzed Y chromosomes using the approach of Repping et al.⁶¹ and Hallast et al.⁶² (see Supplementary Methods).

Ancestral orientation and inversion age estimate. Besides the experimental analysis in chimpanzees and gorillas, the ancestral state of inversions was complemented by bioinformatic and manual inspection of four of the best non human primate genome assemblies (see Supplementary Methods). Due to their fragmented status, the orientation of several available ancient hominin genomes (Supplementary Data 6) was determined by identifying the reads that span the *O1* or *O2* breakpoints of 19 NH inversions without IRs using a library with four 100 bp sequences centered at the two breakpoints of both orientations (Supplementary Data 13) and a slightly modified version of the BreakSeq pipeline^{21,27,28}.

Age of unique inversions was obtained with the usual divergence based approach^{63,64}, using the pairwise differences between orientations from all SNPs (excluding indels) in the available 1000GP Ph3 haplotypes and the largest of the two average pairwise differences within *O1* or *O2* sequences. To have more information in short inversions, we considered the inverted region and any extra

non recombining flanking region (up to 20 kb), excluding breakpoint intervals and associated IRs or indels to avoid sequence errors. Confidence intervals of age estimates were calculated by bootstrap sampling the same number of total individuals with replacement 1000 times and using both a constant substitution rate and two local substitution rates from the divergence with chimpanzee and gorilla (see Supplementary Methods).

Inversion frequency analyses. To control the effect of the study design ascertainment bias in the observed frequency of inversions, we simulated the detection and genotyping process in biallelic SNPs from 1000GP Ph3. The process was simulated in two steps: (1) selection of those SNPs for which the alternative allele is present in 1000GP individuals matching the demographic and gender composition of the detection panel (nine individuals for 38 autosomal or chr. X inversions detected from the fosmid PEM data¹² and one individual for six inversions detected exclusively from the genome assembly comparison¹⁰); and (2) for each of the PEM inversions, generate a random sample of 10,000 SNPs from the total pool of polymorphic SNPs in the PEM panel according to the simulated detection probability calculated from the SNP frequency and the inversion characteristics (see Supplementary Methods). Mean and median frequencies of inversions and SNPs in the 434 1000GP Ph3 individuals were compared by sampling 10,000 sets of SNPs without replacement, with one matched SNP per inversion at a time, and empirical *P* values were estimated as twice the fraction of samples with values more extreme or equal than the observed.

Different genomic variables were tested to determine their effect on inversion frequency: (1) physical length of the inverted region; (2) inversion genetic length; (3) number of genes within the inversion or breakpoint regions; (4) distance to the closest coding gene; (5) number of mammalian constrained sites inside the inversion;⁶⁵ (6) direct functional effect of inversions on genes (Supplementary Table 3); and (7) size of breakpoint IRs. Inversion genetic length was estimated as the cumulative $4N_e r$ for all the genotyped chromosomes using the 1000GP Ph3 SNP data and the LDhat v2.2 rhomap function⁶⁶. Due to high correlation between several variables, to keep only those with significant regression coefficients and reduce the effect of potential outliers in reduced samples, we built robust regression models of the autosomal and chr. X inversion frequency in each population group by stepwise forward selection of predictors with the lmer function from robustbase R package⁶⁷.

Inversion selection tests. Frequency differences between populations were calculated with *vctools* (v0.1.15) using the F_{ST} statistic⁶⁸. F_{ST} values were compared with empirical null distributions in the same individuals obtained from biallelic SNPs accessible according to the strict mask, with a defined ancestral allele, and that have similar frequency and chromosome type (autosome or chr. X) as the inversion (see Supplementary Methods).

LSFS tests are a newly developed family of neutrality tests especially appropriate for inversions, which are a direct application of nearly optimal linear tests for neutrality⁶⁹. The summary statistic was the frequency spectrum of variants closely linked to the inversion, including their linkage pattern (nested or disjoint) with the inverted allele⁷⁰, and we tested strong positive or balancing selection coefficients. LSFS was calculated from biallelic 1000GP Ph3 SNPs in the genotyped individuals, after removing those with a GERP score⁶⁵ higher than 2 and within 0.5 Mb of any of the inversions in our dataset. Only 18 autosomal inversions unambiguously phased into the 1000GP Ph3 haplotypes with perfect tag SNPs within 20 kb from the breakpoints were analyzed using 3 kb non overlapping windows localized within the inverted or non recombining flanking regions (skipping the breakpoints, IRs and indels to avoid genotype errors). Inversion windows were compared against the empirical LSFS computed around all autosomal SNPs and tests were conditioned on the inversion frequency in the different populations. Each population and window was tested separately and population *P* values of the same windows were combined via Edgington's method⁷¹, whereas the results across different windows of an inversion were combined using a conservative and an approximate approach (see Supplementary Methods).

NCD1 and NCD2 statistics⁴⁴ to test long term balancing selection acting on autosomal and chr. X inversion regions were computed for three target frequencies (0.3, 0.4, and 0.5) in overlapping windows of 2 kb (with 1 kb step), defined with the same criteria as in the LSFS test, using 1000GP Ph3 SNPs (accessible according to the pilot mask) from all individuals of the seven studied populations. Only windows with a minimum of eight informative sites (either human polymorphisms or fixed differences with chimpanzee in NCD2 only) and at least 16.7% of positions covered by hg19 panTro4 alignments were considered. Finally, an empirical *P* value was assigned for each inversion region and population by comparing the tests results with a null genome wide distribution obtained by sampling regions of the same size as the inversion (see Supplementary Methods).

Gene-expression analysis in LCLs. We analyzed 42 inversions (excluding two with $MAF < 1\%$ and HsInv0832 in chr. Y) in 173 experimentally genotyped individuals (42 CEU, 84 TSI, and 47 YRI) with GEUVADIS⁴³ and 1000GP Ph3 data. Besides, we imputed 33 inversions in the complete set of 445 GEUVADIS individuals in common with 1000GP Ph3 (89 CEU, 91 TSI, 86 GBR, 92 FIN, and 87 YRI) using the already identified perfect tag SNPs ($r^2 = 1$) (19 inversions) or

IMPUTE v2.3.2⁷² (14 inversions with >90% average imputation accuracy) (Supplementary Fig. 4A). LCL raw RNA Seq reads (ArrayExpress experiment E GEUV 1) were aligned against the GRCh38.p10 human genome with STAR v2.4.2a⁷³ using GENCODE version 26 annotations⁷⁴. Gene expression levels were estimated as reads per kilobase per million mapped reads (RPKM) and transcript expression levels were quantified with RSEM v1.2.31⁷⁵. *cis* eQTL analysis was done through linear regressions implemented in QTLtools v1.1⁷⁶, considering 850 genes and 3318 transcripts expressed in at least 20% of the samples and with TSS within 1 Mb from inversions. First, we carried out a targeted study to test only the association with the genotypes of each inversion. Second, we performed a joint eQTL analysis with inversions and neighboring 1000GP variants to estimate their contribution to observed gene expression changes and identify lead eQTLs. Expression values were adjusted by gender, the first three principal components from 1000GP data (corresponding to continent, population and population structure) and a set of expression principal components to capture technical confounding factors (for genes and transcripts, respectively, 10 and 20 in the experimental dataset and 35 and 50 in the imputed dataset) (see Supplementary Methods). Next, they were transformed to match normal distributions $N(0,1)$ to avoid false positive associations due to outliers and significance was established at 5% false discovery rate (FDR).

Inversions as eQTLs in other tissues and conditions. Gene expression effects in other tissues for 26 inversions were examined by looking whether their highest associated SNPs across all populations ($r^2 \geq 0.8$) have been identified as *cis* eQTLs in different human tissues by the GTEx project (GTEx Analysis Release v7)⁴⁶. Also, we extended the analysis to seven recurrent inversions with SNPs in moderate LD ($r^2 \geq 0.6$). To determine the potential causal variants, we checked if those SNPs in highest LD with the inversions were reported as being the first or second lead eQTL in a specific tissue. The same strategy was applied to link inversions to immune eQTLs associated to the transcriptional response to *Listeria* and *Salmonella* of primary macrophages from African American ($n = 76$) and European American ($n = 99$) individuals⁵² and of macrophages differentiated from 123 induced pluripotent stem cell (iPSC) lines of European origin to *Salmonella* plus interferon γ stimulation⁵³.

Association of inversions with GWAS SNPs. Association of inversions with specific traits or diseases was based on the NHGRI Catalog of published GWAS (release 2017 07 17)⁴⁸ and the GWASdb (release 2015 08 19)⁴⁹ databases. To remove redundant entries, the strongest signal per locus (± 100 kb genomic region) was selected. Only inversions in high LD ($r^2 \geq 0.8$) with a GWAS signal in the studied population or the closest one available were considered (Supplementary Table 4). To investigate if the number of GWAS signals in the inversion and flanking regions (± 20 kb) was higher than expected by chance, we crossed GWAS Catalog and GWASdb signals with 1000GP variants and grouped together those in high LD ($r^2 \geq 0.8$) and associated to the same phenotype. Enrichment P values were calculated by comparison with a null distribution from 1000 random genomic regions as a background model for each inversion, controlling by inversion size and SNP frequencies (average SNP frequency ± 0.01 and Chi square test $P > 0.05$ for the number of SNPs with MAF < 0.2 and ≥ 0.2 compared to the inversion region) and excluding gaps and chr. Y. To test the enrichment in specific inversions, we repeated the same analysis computing a one tailed empirical P value for each inversion.

Statistical information. Details of the statistical tests are described in the corresponding sections. In many cases P values were derived from empirical genome wide null distributions from at least 1000 random samples and two tailed tests were always used, except when specifically mentioned.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data described in this article are available in the Supplementary Information and in the InvFEST database (<http://invfestdb.uab.cat/>). In addition, inversion genotypes have been deposited in the dbVar database (<https://www.ncbi.nlm.nih.gov/dbvar/>) under accession number nstd169 [<https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd169/>]. The source data underlying Figs. 2, 3a c, 4a c, 5a, c, e and 6a b and Supplementary Figs. 1, 2a b, 4a, f, 5 and 8a c are provided as a Source Data file. All other relevant data are available upon request.

Code availability

Computer code is available upon request.

Received: 21 December 2018 Accepted: 27 August 2019

Published online: 17 September 2019

References

- The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
- The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1092 human genomes. *Nature* **491**, 56–65 (2012).
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- Sudmant, P. H. et al. An integrated map of structural variation in 2504 human genomes. *Nature* **526**, 75–81 (2015).
- Walter, K. et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
- Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
- Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
- Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
- Puig, M., Casillas, S., Villatoro, S. & Cáceres, M. Human inversions and their functional consequences. *Brief. Funct. Genom.* **14**, 369–379 (2015).
- Levy, S. et al. The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
- Catacchio, C. R. et al. Inversion variants in human and primate genomes. *Genome Res.* **28**, 910–920 (2018).
- Kidd, J. M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
- Huddleston, J. et al. Discovery and genotyping of structural variation from long read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
- Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675 (2019).
- Shao, H. et al. nPInv: accurate detection and genotyping of inversions using long read sub alignment. *BMC Bioinforma.* **19**, 261 (2018).
- Sanders, A. D. et al. Characterizing polymorphic inversions in human genomes by single cell sequencing. *Genome Res.* **26**, 1575–1587 (2016).
- Li, L. et al. OMSV enables accurate and comprehensive identification of large structural variations from nanochannel based single molecule optical maps. *Genome Biol.* **18**, 230 (2017).
- Chaisson, M. J. P. et al. Multi platform discovery of haplotype resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
- Martínez Fundichely, A. et al. InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Res.* **42**, D1027–D1032 (2014).
- Hehir Kwa, J. Y. et al. A high quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7**, 12989 (2016).
- Vicente Salvador, D. et al. Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. *Hum. Mol. Genet.* **26**, 567–581 (2017).
- Stefansson, H. et al. A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
- Salm, M. P. A. et al. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Res.* **22**, 1144–1153 (2012).
- Antonacci, F. et al. Characterization of six human disease associated inversion polymorphisms. *Hum. Mol. Genet.* **18**, 2555–2566 (2009).
- Pang, A. W. C., Migita, O., Macdonald, J. R., Feuk, L. & Scherer, S. W. Mechanisms of formation of structural variation in a fully sequenced human genome. *Hum. Mutat.* **34**, 345–354 (2013).
- Aguado, C. et al. Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. *PLoS Genet.* **10**, e1004208 (2014).
- Lucas Lledó, J. I., Vicente Salvador, D., Aguado, C. & Cáceres, M. Population genetic analysis of bi allelic structural variants from low coverage sequence data with an expectation maximization algorithm. *BMC Bioinforma.* **15**, 163 (2014).
- Puig, M. et al. Functional impact and evolution of a novel human polymorphic inversion that disrupts a gene and creates a fusion transcript. *PLoS Genet.* **11**, e1005495 (2015).
- González, J. R. et al. A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. *Am. J. Hum. Genet.* **94**, 361–372 (2014).
- Hoffmann, A. A. & Rieseberg, L. H. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* **39**, 21–42 (2008).
- Kirkpatrick, M. How and why chromosome inversions evolve. *PLoS Biol.* **8**, e1000501 (2010).
- Wellenreuther, M. & Bernatchez, L. Eco evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018).
- Imsland, F. et al. The *Rose comb* mutation in chickens constitutes a structural rearrangement causing both altered comb morphology and defective sperm motility. *PLoS Genet.* **8**, e1002775 (2012).

34. Lakich, D., Kazazian, H. H., Antonarakis, S. E. & Gitschier, J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat. Genet.* **5**, 236–241 (1993).
35. Myers, A. J. et al. The H1c haplotype at the *MAPT* locus is associated with Alzheimer's disease. *Hum. Mol. Genet.* **14**, 2399–2404 (2005).
36. Zabetian, C. P. et al. Association analysis of *MAPT* H1 haplotype and subhaplotypes in Parkinson's disease. *Ann. Neurol.* **62**, 137–144 (2007).
37. Webb, A. et al. Role of the tau gene region chromosome inversion in progressive supranuclear palsy, corticobasal degeneration, and related disorders. *Arch. Neurol.* **65**, 1473–1478 (2008).
38. Okbay, A. et al. Genetic variants associated with subjective well being, depressive symptoms and neuroticism identified through genome wide analyses. *Nat. Genet.* **48**, 624–633 (2016).
39. de Jong, S. et al. Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue specific manner. *BMC Genom.* **13**, 458 (2012).
40. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
41. Kehr, B. et al. Diversity in non repetitive human sequences not found in the reference genome. *Nat. Genet.* **49**, 588–593 (2017).
42. Schouten, J. P. et al. Relative quantification of 40 nucleic acid sequences by multiplex ligation dependent probe amplification. *Nucleic Acids Res.* **30**, e57 (2002).
43. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
44. Bitarello, B. D. et al. Signatures of long term balancing selection in human genomes. *Genome Biol. Evol.* **10**, 939–955 (2018).
45. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
46. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
47. Horton, R. et al. Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60**, 1–18 (2008).
48. MacArthur, J. et al. The new NHGRI EBI Catalog of published genome wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
49. Li, M. J. et al. GWASdbv2: an update database for human genetic variants identified by genome wide association studies. *Nucleic Acids Res.* **44**, D869–D876 (2016).
50. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
51. Gymrek, M. et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
52. Nédelec, Y. et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell* **167**, 657–669 (2016).
53. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
54. Sun, B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
55. Lee, J. Y. W. et al. Large intragenic deletion in *DSTYK* underlies autosomal recessive complicated spastic paraparesis, SPG23. *Am. J. Hum. Genet.* **100**, 364–370 (2017).
56. Tielsch, J. M. et al. Racial variations in the prevalence of primary open angle glaucoma. The Baltimore Eye Survey. *JAMA* **266**, 369–374 (1991).
57. Purcell, S. et al. PLINK: a tool set for whole genome association and population based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
58. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
59. Stephens, M. & Donnelly, P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**, 1162–1169 (2003).
60. Bandelt, H. J., Forster, P. & Röhl, A. Median joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).
61. Repping, S. et al. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat. Genet.* **38**, 463–467 (2006).
62. Hallast, P., Balaresque, P., Bowden, G. R., Ballereau, S. & Jobling, M. A. Recombination dynamics of a human Y chromosomal palindrome: rapid GC biased gene conversion, multi kilobase conversion tracts, and rare inversions. *PLoS Genet.* **9**, e1003666 (2013).
63. Hasson, E. & Eanes, W. F. Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. *Genetics* **144**, 1565–1575 (1996).
64. Corbett Detig, R. B. & Hartl, D. L. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1003056 (2012).
65. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
66. Auton, A. & McVean, G. Recombination rate estimation in the presence of hotspots. *Genome Res.* **17**, 1219–1227 (2007).
67. Maechler, M. et al. robustbase: Basic Robust Statistics R package version 0.93 2. (2018).
68. Weir, B. S. & Cockerham, C. C. Estimating F statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
69. Ferretti, L., Perez Enciso, M. & Ramos Onsins, S. Optimal neutrality tests based on the frequency spectrum. *Genetics* **186**, 353–365 (2010).
70. Ferretti, L. et al. The neutral frequency spectrum of linked sites. *Theor. Popul. Biol.* **123**, 70–79 (2018).
71. Edgington, E. S. An additive method for combining probability values from independent experiments. *J. Psychol.* **80**, 351–363 (1972).
72. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
73. Dobin, A. et al. STAR: ultrafast universal RNA seq aligner. *Bioinformatics* **29**, 15–21 (2013).
74. Harrow, J. et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
75. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
76. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
77. Poznik, G. D. et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y chromosome sequences. *Nat. Genet.* **48**, 593–599 (2016).
78. Pérez Palma, E. et al. Overrepresentation of glutamate signaling in Alzheimer's disease: network based pathway enrichment using meta analysis of genome wide association studies. *PLoS One* **9**, e95413 (2014).
79. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).

Acknowledgements

We thank Cristina Aguado, Olga Dolgova, Teresa Soos, Esteban Urrea, David Vicente Salvador, and Roser Zaurin for help with inversion genotyping, Antonio Barbadilla, Ruth Gómez, José Ignacio Lucas Lledó, Sebastián Ramos Onsins, and Alfredo Ruiz for help with the evolutionary analysis, Mariona Bellet, Robert Castelo, Diego Garrido, and Roderic Guigó for help with the gene expression analysis, Sònia Casillas and Alexander Martínez Fundichely for help with the inversion selection, and Xavier Estivill, Tomás Marqués Bonet, Aurora Ruiz Herrera, the Coriell Institute for Medical Research, the Barcelona Zoo and the Banc de Teixits Animals de Catalunya (BTAC) for providing the human and non human primate samples used in this study. This work was supported by research grants ERC Starting Grant 243212 (INVEST) from the European Research Council under the European Union Seventh Research Framework Programme (FP7), BFU2013 42649 P and BFU2016 77244 R funded by the Agencia Estatal de Investigación (AEI, Spain) and the European Regional Development Fund (FEDER, EU), and 2014 SGR 1346 and 2017 SGR 1379 from the Generalitat de Catalunya (Spain) to M.C., a PIF PhD fellowship from the Universitat Autònoma de Barcelona (Spain) to C.G.D., a La Caixa Doctoral fellowship to J.L.J., and a FPI PhD fellowship from the Ministerio de Economía y Competitividad (Spain) to M.O. and I.N. M.G.V. was supported in part by POCI 01 0145 FEDER 006821 funded through the Operational Programme for Competitiveness Factors (COMPETE, EU) and UID/BIA/50027/2013 from the Foundation for Science and Technology (FCT, Portugal).

Author contributions

M.C. conceived the inversion genotyping strategy, devised the study and oversaw all the steps; S.V., M.P., and M.C. designed the genotyping assays; S.V., D.L., A.D., and M.P. carried out experiments; C.G.D., M.G.V., I.O., C.L.F., M.P., and M.C. analyzed evolutionary data; C.G.D., J.L.J., D.C., B.B., I.N., P.O., A.A., and L.F. performed selection tests; J.L.J., M.O., L.P., T.E., and M.C. analyzed functional effects; A.B. provided samples; M.C., C.G.D., J.L.J., M.G.V., L.F., and M.P. wrote the paper and all the authors contributed comments to the final version of the manuscript.

Additional information

Supplementary Information accompanies this paper at <https://doi.org/10.1038/s41467-019-12173-x>.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Peer review information *Nature Communications* thanks Megan Dennis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

Supplementary Information

Evolutionary and functional impact of common polymorphic inversions in the human genome

Carla Giner-Delgado^{1,2 †}, Sergi Villatoro^{1 †}, Jon Lerga-Jaso^{1 †}, Magdalena Gayà-Vidal^{1,3}, Meritxell Oliva¹, David Castellano¹, Lorena Pantano¹, Bárbara D. Bitarello⁴, David Izquierdo¹, Isaac Noguera¹, Iñigo Olalde⁵, Alejandra Delprat¹, Antoine Blancher^{6,7}, Carles Lalueza-Fox⁵, Tõnu Esko⁸, Paul F. O'Reilly⁹, Aida M. Andrés^{4,10}, Luca Ferretti¹¹, Marta Puig¹, Mario Cáceres^{1,12 *}

¹ Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, 08193, Spain.

² Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, 08193, Spain.

³ CIBIO/InBIO Research Center in Biodiversity and Genetic Resources, Universidade do Porto, Vairão, Distrito do Porto, 4485-661, Portugal.

⁴ Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Saxony, 04103, Germany.

⁵ Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona, 08003, Spain.

⁶ Laboratoire d'immunologie, CHU de Toulouse, IFB Hôpital Purpan, 31059, Toulouse, France.

⁷ Centre de Physiopathologie Toulouse-Purpan (CPTP), Université de Toulouse, Centre National de la Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (Inserm), Université Paul Sabatier (UPS), 31024, Toulouse, France.

⁸ Estonian Genome Center, University of Tartu, Tartu, 51010, Estonia.

⁹ Social, Genetic, and Developmental Psychiatry, King's College London, London, SE5 8AF, United Kingdom.

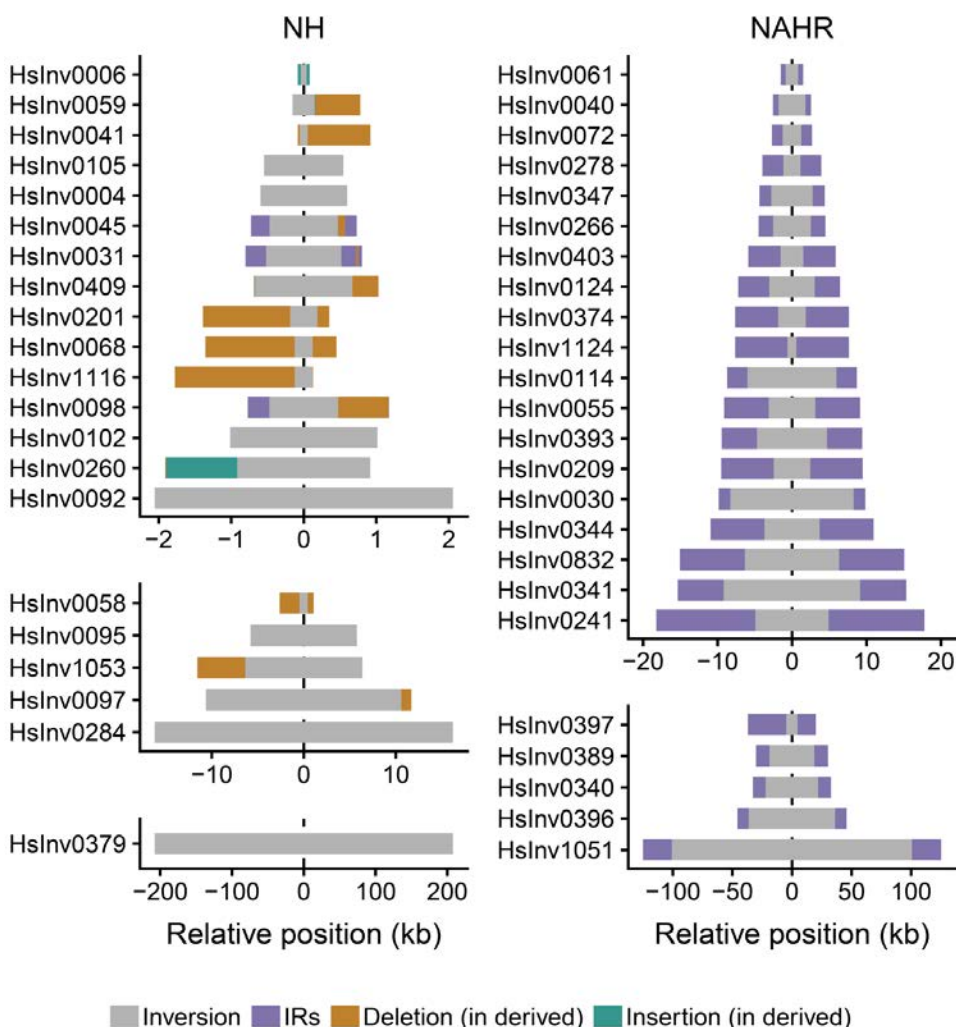
¹⁰ UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, United Kingdom.

¹¹ Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, OX3 7LF, United Kingdom.

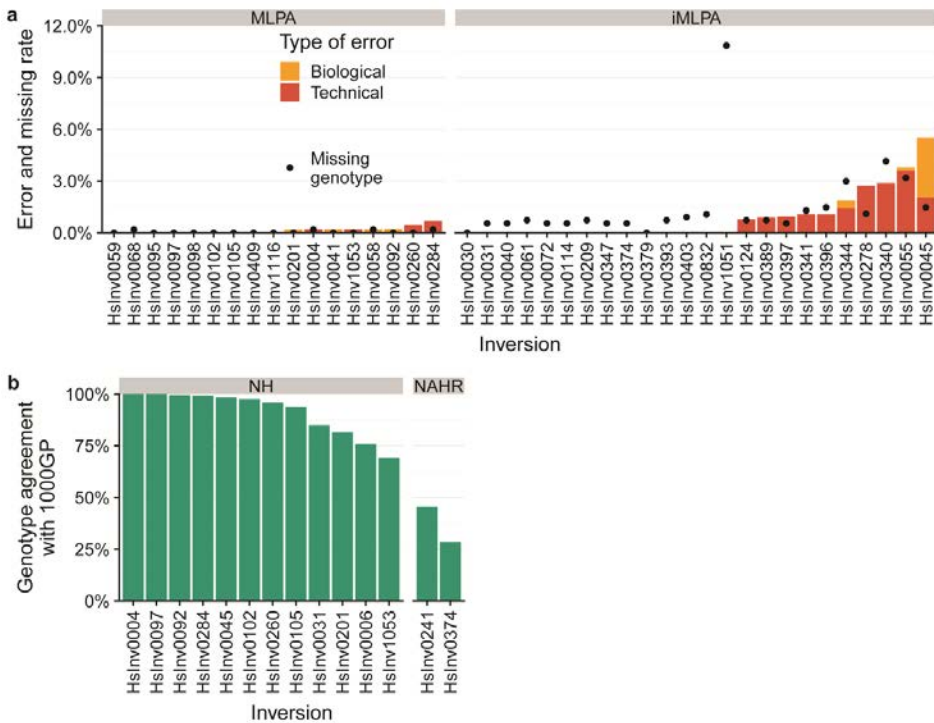
¹² ICREA, Barcelona, 08010, Spain.

† These authors contributed equally to this work.

* Corresponding author.

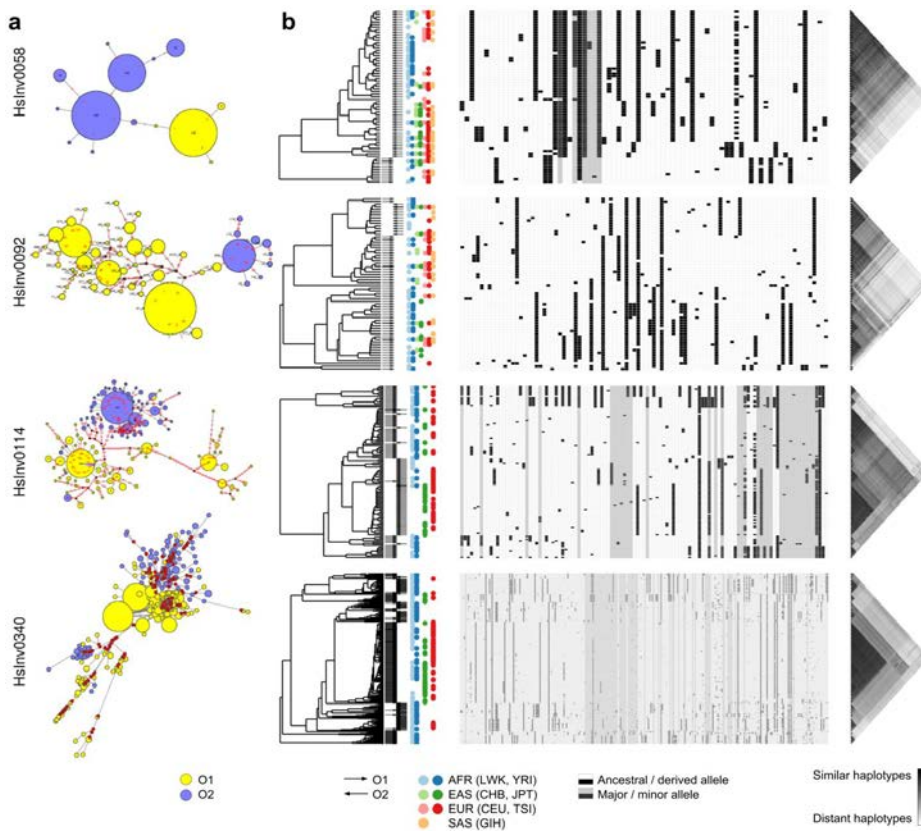


Supplementary Fig. 1. Size and breakpoint complexity of the 45 studied inversions. The graphs illustrate the main characteristics of inversions created by non-homologous mechanisms (NH) or non-allelic homologous recombination (NAHR), with the inverted region represented as a gray bar and flanking inverted repeats (IRs) or other structural changes in different colors. In NH inversions, deletions are sequences present in the original orientation that are eliminated in the derived orientation, and insertions are sequences gained. Three of these inversions (HsInv0031, HsInv0045 and HsInv0098) have also short low-identity IRs (249-297 bp, 83.2-86.2% identity) in the ancestral orientation that are partially deleted in the derived orientation. Source data are provided as a Source Data file.

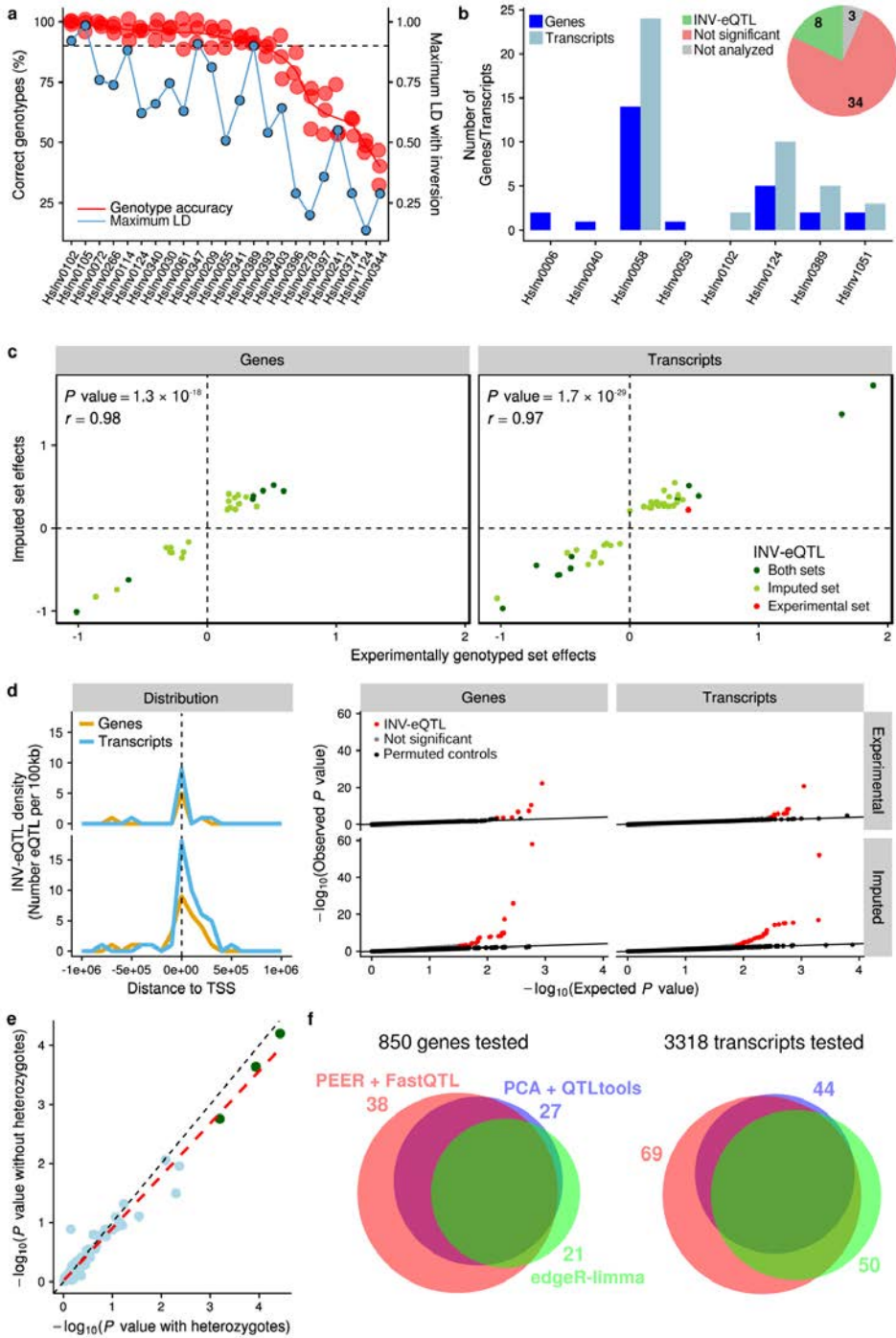


Supplementary Fig. 2. Inversion genotype accuracy by PCR-based validation and published data.

a. Genotyping performance of MLPA and iMLPA assays. Inversion genotypes from MLPA/iMLPA were compared with those obtained from PCR or iPCR, plus those imputed from perfect tag SNPs in 1000GP Ph3 data. Genotyping success rate was 99.96% for MLPA and 98.56% for iMLPA. The lower success in iMLPA was due to a lower self-ligation efficiency of large restriction DNA fragments compared to shorter ones (as in the case of HsInv1051), which reduces the amount of specific probe target region and results in smaller amplification peaks, and to problems in specific samples (with one third of missing genotypes accumulating in just three samples). Biological errors correspond to known problems due to restriction site polymorphisms in a few specific inversions or DNA contamination, while technical errors do not have a clear cause and appear to be mainly due to problems in MLPA probe amplification in certain inversions. **b.** Genotype agreement between the 14 inversions in common with the 1000GP structural variant release¹ according to the InvFEST database² for the 434 samples shared in both datasets. Of the genotypes that differ between studies, 99.1% are due to 1000GP incorrectly assigning the reference genome orientation to one of the alleles, whereas according to our experiments it should be the alternative, which leads to underestimating the frequency of the inversion. Also, with a few exceptions, 1000GP error rates tend to be much higher in inversions flanked by indels or inverted repeats than in those with clean breakpoints. Source data are provided as a Source Data file.

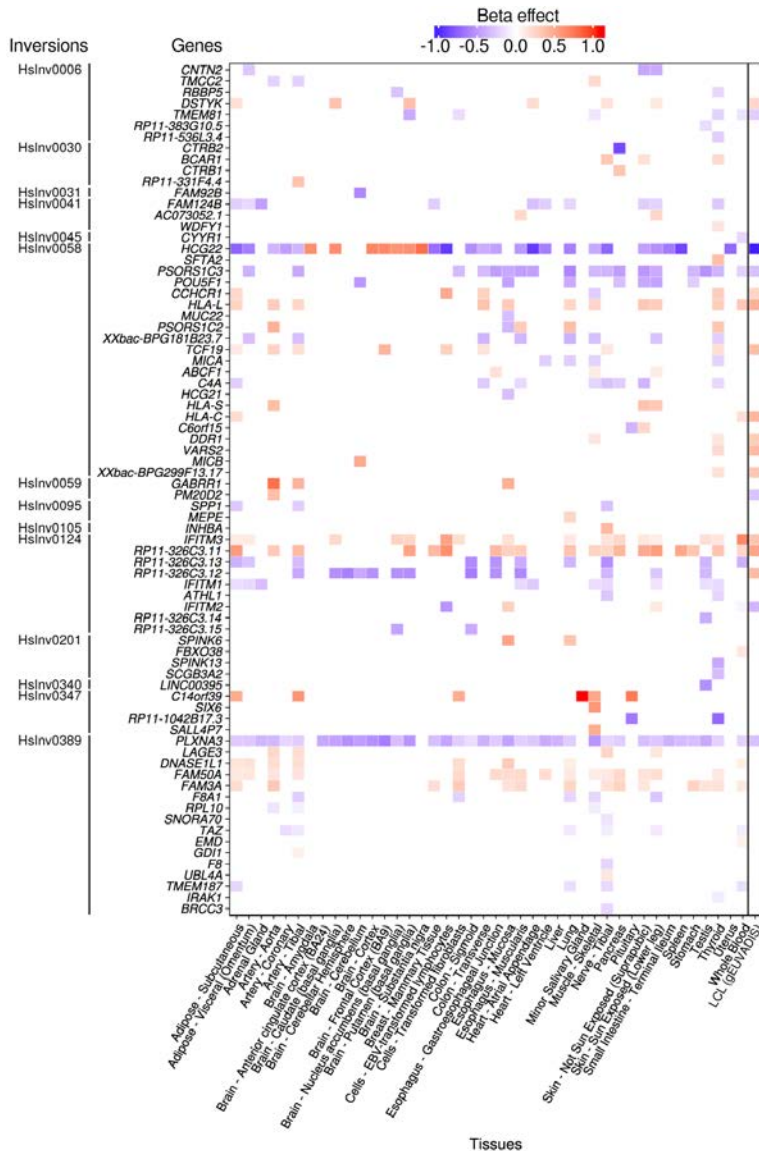


Supplementary Fig. 3. Summary of haplotype relationships for different inversions. **a.** Representative median-joining networks from 1000GP Ph1 haplotypes obtained with PHASE 2.1³. Each circle represents a haplotype, whose size is proportional to the number of chromosomes carrying that particular haplotype. Small red points are hypothetical haplotypes not found in the individuals analyzed, and the length of the branch connecting two haplotypes is proportional to the number of changes between them. **b.** Integrated haplotype plots (iHPlots) for the same four inversions. For unique inversions (HsInv0058 and HsInv0092), the haplotypes correspond to those from 1000GP Ph3 with the extended flanking region whenever possible, whereas for recurrent inversions (HsInv0114 and HsInv0340), the haplotypes are those obtained with PHASE 2.1³ from 1000GP Ph1 data including only the inverted region. O1 and O2 haplotypes of unique inversions can be clearly separated (*e.g.* HsInv0058), probably corresponding to old inversions that had time to diverge, or those with the derived orientation can be clustered together with haplotypes carrying the other orientation (*e.g.* HsInv0092), likely representing more recent or small inversions with few informative positions and little differentiation.

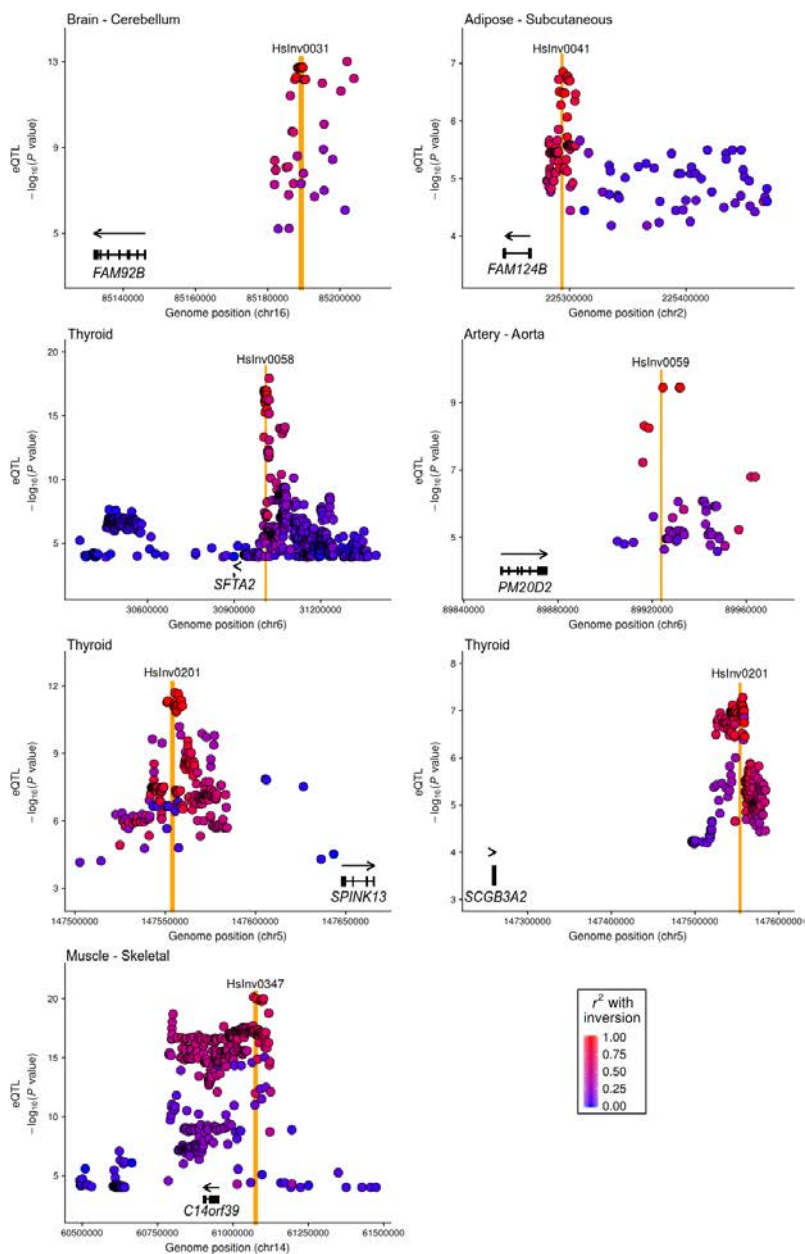


Chapter 3. Results

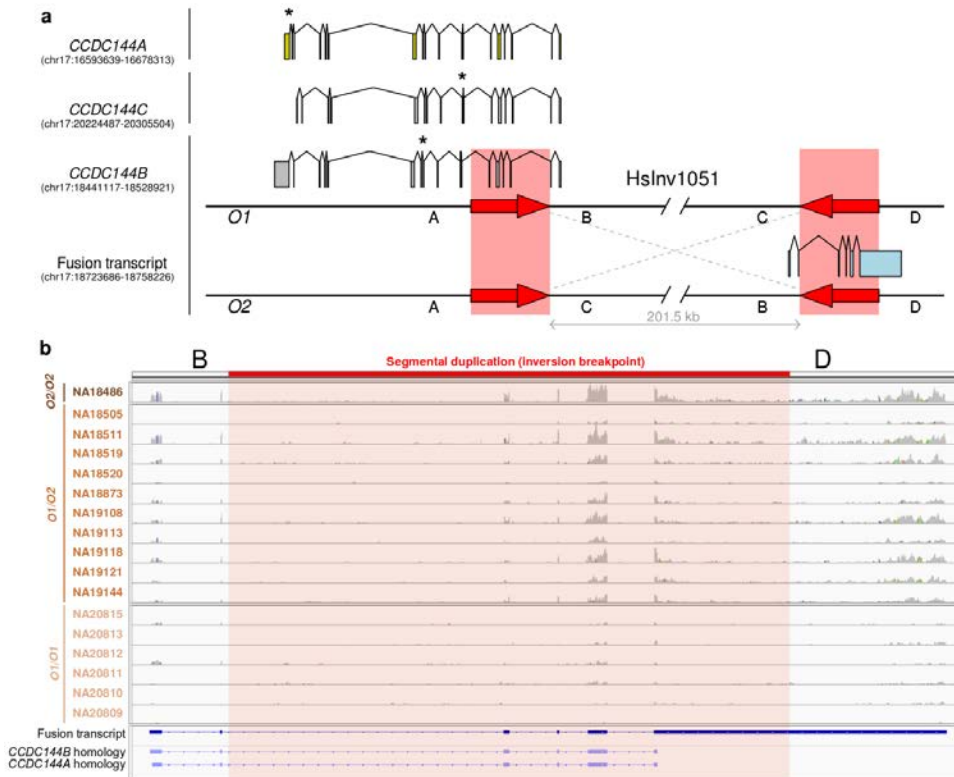
Supplementary Fig. 4. Summary of inversion gene-expression analysis in lymphoblastoid cell lines (LCL). **a.** Genotype imputation accuracy of 23 autosomal and chr. X inversions without perfect tag SNPs ($r^2 = 1$) based on all 1000GP Ph3 variants (including both SNPs and structural variants). Imputation was performed with IMPUTE v2.3.2⁴ adapted to unphased reference genotypes, due to the difficulty of phasing correctly recurrent inversions, using a region of 1.5 Mb at each side of the inversion and an effective population size of 20,000 (15,000 in chr. X). Genotypes were called with a posterior probability higher than 0.7, and were classified as missing otherwise. Imputation accuracy was checked by removing three random sets of 30 individuals from European (CEU, TSI) and YRI populations from the reference panel of 173 GEUVADIS individuals with known inversion genotypes, which were subsequently imputed under the same criteria. The red line represents the mean percentage of right calls in the three test samples (red dots) and 14 inversions with >90% imputation accuracy (dashed line) were used for gene-expression analysis in other GEUVADIS individuals. Maximum LD between inversions and surrounding genomic variants (blue line) was lower for inversions with worse imputation accuracy. In Hslnv0102, which does not have SNPs in high LD, its imputation is based on the 1000GP genotypes for the inversion itself. **b.** Pie chart and graph summarizing the effects in *cis* of the 45 inversions on LCL expression variation and the number of genes or transcripts affected. Results represented correspond to those from the experimentally-genotyped set (173 individuals) for the 9 inversions that could not be imputed and the extended imputed set (445 individuals) for the 33 imputed inversions. **c.** Comparison of effect sizes on genes and transcripts from inversion *cis*-eQTLs (INV-eQTLs) identified in the experimentally-genotyped and imputed sets in LCLs, showing concordant results (dark green, replicated in both sets; light green, specific of the imputed set; and red, specific of the experimentally-genotyped set). **d.** *Cis*-eQTL analysis of inversions in LCL expression data. Left: Distribution of INV-eQTLs with respect to the transcription start site (TSS) of the affected genes and transcripts. Inversions tend to locate closer (<100 kb) to genes or transcripts affected compared to all association tests performed both for the experimental (top, Fisher test $P = 0.013$ and $P = 0.0005$, respectively) and imputed data sets (bottom, Fisher test $P = 0.018$ and $P = 3 \times 10^{-6}$, respectively). Right: Quantile-quantile plot of associations between inversions and gene or transcript expression for the experimentally-genotyped and the imputed sets: red dots, significant INV-eQTLs (FDR < 0.05); grey dots, not significant associations; and black dots, negative controls obtained by permuting sample labels from the inversion genotype matrix relative to covariates and expression levels, which follow the expected P value distribution assuming no-association. **e.** Correlation of gene eQTL analysis P values for inversions located in chr. X with and without heterozygous females (to eliminate the effect of the random inactivation of one copy of this chromosome). Significant associations (FDR < 0.05) in both analyses are indicated as green dots, and the similarity between the observed and perfect 1:1 correlation (red and black dashed lines, respectively), with slightly lower eQTL P values when including all samples, suggests that the consequences of silencing the chr. X with or without the inversion get averaged across all cells. **f.** Results of inversion effects in gene and transcript expression when using different approaches: “PCA+QTLtools”, which corresponds to the pipeline used in this work⁵ (blue); “PEER+FastQTL”, which corresponds to the pipeline used in the GTEx Project⁶ (red); and “edgeR-limma”^{7,8} (green). Numbers indicate the significant inversion-gene or inversion-transcript pairs with each analysis method. Venn diagram was done with BioVenn⁹. Findings using the different pipelines were highly coincident, although a larger number of significant genes/transcripts were estimated by the GTEx pipeline, indicating that our chosen method based in PCA and QTLtools is more conservative. Source data are provided as a Source Data file.



Supplementary Fig. 5. Summary of inversion effects on GTEx gene-expression data. Inversion effects were estimated through variants in high LD ($r^2 \geq 0.8$), or moderate LD ($r^2 \geq 0.6$) for recurrent inversions, reported as eQTLs in GTEx Analysis Release v7 (Supplementary Data 9). The direction and strength of the beta effect of the eQTL is indicated in different color, with blue and red representing respectively lower and higher expression associated to the O2 orientation of the inversion. Inversion eQTLs also identified in the LCL analysis from the GEUVADIS data are represented in the last column. Source data are provided as a Source Data file.

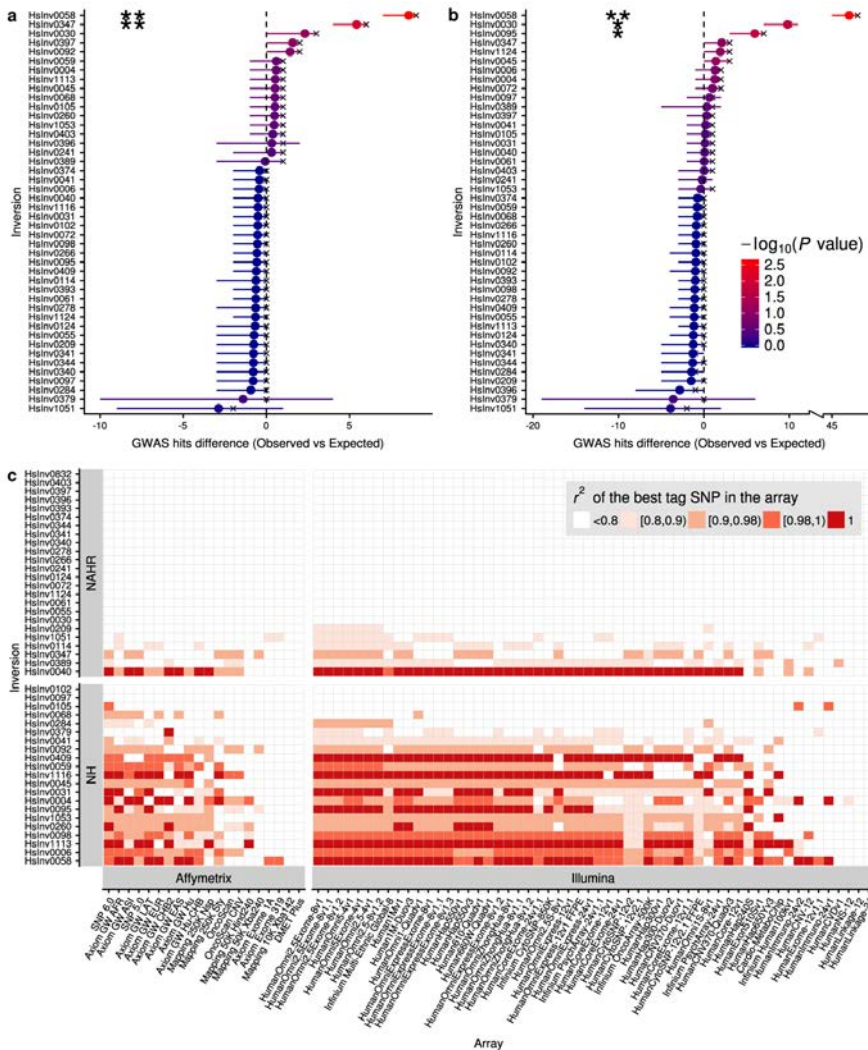


Supplementary Fig. 6. Examples of potential expression effects of six inversions in different tissues. Manhattan plots of logarithm-transformed linear regression t-test P values for *cis*-eQTLs associations from the GTEx project in which an inversion shows the highest LD ($r^2 \geq 0.9$) with the two first lead markers in the corresponding tissue. The orange bar pinpoints the inversion position and its LD to each variant is represented in different colors. The affected genes are shown in black and arrowheads indicate the direction of transcription.



Supplementary Fig. 7. Representation of the fusion transcript created by Hslnv1051. **a.** Diagram of *CCDC144B* gene disruption by inversion Hslnv1051 and the novel fusion transcript created by including additional 3' sequences from region D (light blue), with the segmental duplications at the inversion breakpoints represented as red arrows. *CCDC144B* is part of a family with two other members, *CCDC144A* and *CCDC144C*, that have ~99% identity and very similar exon-intron structure (shown on top). Nevertheless, whereas *CCDC144A* encodes a 1,427-amino acid protein, *CCDC144B* and *CCDC144C* have different frameshift mutations that reduce their coding capacity to 725 and 646 amino acids, respectively (with stop codons shown by asterisks). *CCDC144B* premature stop codon is not included in the fusion transcript from the inverted allele. **b.** RNA-Seq profiles from GEUVADIS LCL reads mapped to the inversion BD breakpoint, which was created *in silico* the sequence between the Hslnv1051 breakpoints in the human reference genome (hg19). Reads were remapped to this construct using STAR 2-pass¹⁰ to improve the accuracy of alignments, revealing a novel fusion transcript expressed only in O1/O2 heterozygotes and at higher levels in O2/O2 homozygotes. The chimeric transcript structure is shown below, after its precise reconstruction with Cufflinks default parameters¹¹ by merging all reads from these samples around the breakpoint region. In addition, its homology with the first six exons of *CCDC144B* and *CCDC144A* is also shown. RNA-seq profiles were visualized on Integrative Genomics Viewer¹².

Chapter 3. Results



Supplementary Fig. 8. Potential phenotypic effects of inversions from GWAS data. **a-b.** Enrichment of GWAS signals around 44 autosomal and chr. X inversions (inverted region ± 20 kb) in the GWAS Catalog (**a**) and GWASdb (**b**) databases. Error bars show the 0-0.95 confidence interval of the difference in the observed number of GWAS hits compared with a background model from 1,000 random genomic regions for each inversion, together with the mean (filled circle) and the median (cross) of the differences. The color indicates the one-tailed empirical test P value of the enrichment according to the scale shown. Hslnv0058 showed significant enrichment of GWAS hits in both datasets, whereas Hslnv0030 and Hslnv0347 showed similar trends in both datasets and significant differences from the expected number in at least one. *, $P < 0.05$; **, $P < 0.01$. **c.** Coverage of SNPs associated with inversions in 76 commonly-used genotyping arrays by checking the presence of inversion structural global tag SNPs ($r^2 \geq 0.8$) in the arrays through the LDLink web portal¹³. LD with the inversion of the best global tag SNP in each array is indicated in different colors, showing that for the great majority of NAHR inversions

Chapter 3. Results

and several of the NH inversions there are not tag SNPs or they are not present in the array (represented as white squares). The best performing arrays assessed, HumanOmni5-4v1 and HumanOmni5Exome-4v1 (Illumina), could detect up to 23 inversions (51%), with only 7 being represented by perfect global tag SNPs ($r^2 = 1$), and 16 by variants with lower LD. Source data are provided as a Source Data file.

Chapter 3. Results

Supplementary Table 1. Frequencies of the 45 inversions in seven human populations. The derived allele frequency (DAF) is shown whenever the ancestral orientation is known, and the frequency of the minor allele considering the seven populations together (MAF) is indicated otherwise. The total number of genotyped individuals, as well as those unrelated (Unrel) and included in either the 1000 Genome Project Phase 1 (Ph1) or Phase 3 (Ph3), are also indicated for each population and population group at the bottom. Inversion frequency was estimated from the 480 unrelated individuals of the seven known populations, although for some analyses only the 434 individuals in common with 1000GP Ph3 were used. Deviation from Hardy-Weinberg equilibrium was calculated with Plink –hardy option¹⁴ and for all populations and inversions an exact test $P > 0.01$ was obtained. Populations are: Luhya in Webuye, Kenya (LWK); Yoruba in Ibadan, Nigeria (YRI); Utah residents (CEPH) with Northern and Western European ancestry (CEU); Toscani in Italia (TSI); Gujarati Indians in Houston, Texas, USA (GIH); Han Chinese in Beijing, China (CHB); and Japanese in Tokyo, Japan (JPT). Population groups are: African ancestry (AFR); European ancestry (EUR); South-Asian ancestry (SAS); and East-Asian ancestry (EAS).

Inversion	Allele	Population or population group											Total
		LWK	YRI	AFR	CEU	TSI	EUR	GIH	SAS	CHB	JPT	EAS	
Derived allele frequency (DAF)													
Hslnv0004	O1	0.981	0.993	0.987	0.800	0.828	0.817	0.837	0.837	0.856	0.889	0.872	0.884
Hslnv0006	O1	0.931	0.943	0.937	0.408	0.356	0.377	0.511	0.511	0.456	0.400	0.428	0.587
Hslnv0030	O1	0.012	0.021	0.017	0.158	0.156	0.157	0.062	0.062	0	0	0	0.066
Hslnv0031	O1	0.481	0.371	0.430	0.317	0.281	0.295	0.399	0.399	0.433	0.432	0.433	0.383
Hslnv0040	O1	0.228	0.279	0.252	0.208	0.382	0.312	0.191	0.191	0.100	0.044	0.072	0.225
Hslnv0041	O2	0.747	0.629	0.692	0.442	0.428	0.433	0.371	0.371	0.411	0.422	0.417	0.500
Hslnv0045	O2	0.463	0.636	0.543	0.450	0.561	0.517	0.393	0.393	0.611	0.544	0.578	0.514
Hslnv0058	O1	0.222	0.384	0.297	0.383	0.350	0.363	0.258	0.258	0.456	0.456	0.456	0.340
Hslnv0059	O1	0.093	0.093	0.093	0.183	0.161	0.170	0.180	0.180	0.722	0.678	0.700	0.247
Hslnv0061	O1	0	0	0	0.025	0.034	0.030	0.006	0.006	0	0.022	0.011	0.013
Hslnv0068	O1	0.099	0.116	0.107	0.225	0.233	0.230	0.112	0.112	0	0	0	0.126
Hslnv0092	O2	0.265	0.350	0.305	0.067	0.089	0.080	0.129	0.129	0.078	0.089	0.083	0.160
Hslnv0095	O1	0.173	0.121	0.149	0.325	0.289	0.303	0.157	0.157	0.256	0.244	0.250	0.218
Hslnv0097	O2	0.019	0.014	0.017	0	0	0	0	0	0	0	0	0.005
Hslnv0098	O2	0.346	0.307	0.328	0.117	0.111	0.113	0.096	0.096	0.078	0.078	0.078	0.171
Hslnv0102	O2	0.272	0.350	0.308	0.133	0.156	0.147	0.202	0.202	0.033	0.044	0.039	0.188
Hslnv0105	O1	0.469	0.450	0.460	0.517	0.639	0.590	0.618	0.618	0.256	0.211	0.233	0.488
Hslnv0114	O1	0.800	0.843	0.820	0.375	0.354	0.362	0.348	0.348	0.144	0.156	0.150	0.463
Hslnv0201	O2	0.611	0.664	0.636	0.533	0.644	0.600	0.506	0.506	0.489	0.367	0.428	0.561
Hslnv0209	O2	0.184	0.271	0.225	0.017	0.084	0.057	0	0	0.022	0.011	0.017	0.091
Hslnv0260	O2	0.204	0.129	0.169	0.125	0.083	0.100	0.174	0.174	0.289	0.422	0.356	0.183
Hslnv0266	O2	0.269	0.271	0.270	0.208	0.178	0.190	0.500	0.500	0.250	0.289	0.270	0.288
Hslnv0278	O1	0.608	0.543	0.577	0.898	0.909	0.905	0.843	0.843	0.733	0.656	0.694	0.751
Hslnv0284	O2	0.105	0.101	0.103	0	0	0	0	0	0	0	0	0.032
Hslnv0379	O2	0	0	0	0	0	0	0	0	0.022	0.033	0.028	0.005
Hslnv0409	O1	0.385	0.359	0.373	0.478	0.630	0.569	0.455	0.455	0.662	0.776	0.719	0.515
Hslnv1051	O2	0.029	0.080	0.055	0	0	0	0	0	0	0	0	0.018
Hslnv1053	O2	0.247	0.236	0.242	0.692	0.600	0.637	0.702	0.702	0.722	0.689	0.706	0.538
Hslnv1116	O2	0.722	0.843	0.778	0.750	0.711	0.727	0.938	0.938	1	1	1	0.833

(Continued in next page)

Chapter 3. Results

Inversion	Allele	Population or population group											Total
		LWK	YRI	AFR	CEU	TSI	EUR	GIH	SAS	CHB	JPT	EAS	
Minor allele frequency (MAF)													
Hslnv0055	O1	0.605	0.557	0.583	0.217	0.225	0.221	0.247	0.247	0.156	0.122	0.139	0.325
Hslnv0072	O1	0.050	0.078	0.063	0.011	0.008	0.009	0.007	0.007	0	0	0	0.024
Hslnv0124	O1	0.146	0.121	0.134	0.608	0.551	0.574	0.281	0.281	0.022	0.056	0.039	0.281
Hslnv0241	O2	0.582	0.597	0.589	0.167	0.184	0.177	0.301	0.301	0.405	0.378	0.391	0.370
Hslnv0340	O2	0.513	0.507	0.510	0.008	0.034	0.024	0.011	0.011	0	0	0	0.167
Hslnv0341	O2	0.152	0.257	0.201	0.025	0.022	0.024	0.017	0.017	0	0.023	0.011	0.076
Hslnv0344	O2	0.487	0.493	0.490	0.542	0.483	0.507	0.354	0.354	0.411	0.278	0.344	0.442
Hslnv0347	O2	0.234	0.286	0.258	0.092	0.118	0.107	0.303	0.303	0.133	0.122	0.128	0.195
Hslnv0374	O2	0.392	0.243	0.322	0.467	0.461	0.463	0.601	0.601	0.533	0.711	0.622	0.475
Hslnv0389	O2	0.958	1	0.978	0.178	0.173	0.175	0.478	0.478	0.221	0.313	0.267	0.499
Hslnv0393	O2	0.317	0.330	0.323	0.367	0.391	0.381	0.657	0.657	0.647	0.746	0.696	0.474
Hslnv0396	O2	0.263	0.417	0.335	0.159	0.144	0.150	0.209	0.209	0.029	0.015	0.023	0.195
Hslnv0397	O2	0.608	0.553	0.583	0.156	0.158	0.157	0.291	0.291	0.456	0.687	0.570	0.393
Hslnv0403	O2	0.650	0.650	0.650	0.267	0.189	0.221	0.328	0.328	0.824	0.672	0.748	0.475
Hslnv0832	O2	0.821	1	0.908	0	0	0	0.250	0.250	0	0.043	0.022	0.339
Hslnv1124	O2	0.375	0.551	0.456	0.625	0.589	0.603	0.365	0.365	0.409	0.600	0.506	0.495
Total indiv.	-	90	100	190	90	90	180	90	90	45	45	90	550
Unrel. indiv.	-	81	70	151	60	90	150	89	89	45	45	90	480
Ph1 indiv.	-	87	48	135	35	90	125	0	0	41	39	80	340
Ph3 indiv.	-	75	58	133	45	89	134	82	82	40	45	85	434

Chapter 3. Results

Supplementary Table 2. Summary of the total number of analyzed, shared and fixed variants in human inversions from 1000 Genomes Project (1000GP) Phase3 and HapMap genotype data. For 1000GP data only accessible variants according to the strict criteria were used. As expected, most NH inversions have no shared variants between orientations within the inverted region. The only exceptions were HsInv1053, with a single shared SNP near the second breakpoint in 1000GP and two shared SNPs in HapMap, and HsInv0095, with shared variants only in HapMap. These variants tend to be grouped in certain positions and are likely the result of gene conversion, SNP genotyping errors or even independent mutations. In addition, there is considerable variation in the number of fixed variants between inversions, which is probably related to the recombination events outside the inverted region. Non-recombining flanking region was estimated according to the distribution of fixed and shared variants up to a maximum 20 kb from the breakpoints. NA, Not applicable.

Inversion	Inside				Inside + 200 kb		Non-recombining flanking region (kb)	
	Analyzed variants		Shared variants		Fixed variants		Upstream	Downstream
	1000GP	HapMap	1000GP	HapMap	1000GP	HapMap		
Inversions generated by non-homologous mechanisms (NH)								
HsInv0004	21	2	0	0	18	2	1.9	11.7
HsInv0006	0	0	NA	NA	4	0	3.1	0.1
HsInv0031	8	4	0	0	9	2	0.5	0
HsInv0041	0	1	NA	0	2	0	0	0.2
HsInv0045	4	1	0	0	1	0	0	1.3
HsInv0058	0	1	NA	0	8	2	0.9	2.6
HsInv0059	0	0	NA	NA	1	0	4.3	4.7
HsInv0068	0	1	NA	0	1	1	3	7.7
HsInv0092	56	1	0	0	1	0	2.8	3.2
HsInv0095	101	7	0	3	4	1	3.9	2.1
HsInv0097	247	18	0	0	17	0	20	20
HsInv0098	8	0	0	NA	8	0	10.9	0.4
HsInv0102	14	0	0	NA	0	0	0	0.6
HsInv0105	1	0	0	NA	0	1	0.5	20
HsInv0201	0	0	NA	NA	16	3	3.3	4.7
HsInv0260	12	1	0	0	3	0	19.8	1.8
HsInv0284	418	13	0	0	3	0	1.6	18.3
HsInv0379	3426	155	0	0	3	0	18.1	15.6
HsInv0409	0	0	NA	NA	1	1	0	0.3
HsInv1053	185	12	1	2	2	0	0.2	0.3
HsInv1116	0	0	NA	NA	22	3	7.7	2

(Continued in next page)

Chapter 3. Results

Inversion	Inside				Inside + 200 kb		Non-recombining flanking region (kb)	
	Analyzed variants		Shared variants		Fixed variants		Upstream	Downstream
	1000GP	HapMap	1000GP	HapMap	1000GP	HapMap		
Inversions generated by non-allelic homologous recombination (NAHR)								
Hslnv0030	225	11	17	6	0	0	NA	NA
Hslnv0040	33	0	0	NA	43	0	9.1	20
Hslnv0055	21	2	5	2	0	0	NA	NA
Hslnv0061	10	0	0	NA	0	0	NA	NA
Hslnv0072	14	1	1	0	0	0	NA	NA
Hslnv0114	141	12	10	4	0	0	NA	NA
Hslnv0124	58	0	8	NA	0	0	NA	NA
Hslnv0209	71	4	9	3	0	0	NA	NA
Hslnv0241	46	3	18	3	0	0	NA	NA
Hslnv0266	40	0	2	NA	0	0	NA	NA
Hslnv0278	18	0	6	NA	0	0	NA	NA
Hslnv0340	170	6	45	5	0	0	NA	NA
Hslnv0341	196	11	33	11	0	0	NA	NA
Hslnv0344	37	2	15	2	0	0	NA	NA
Hslnv0347	44	2	13	1	0	0	NA	NA
Hslnv0374	1	1	0	1	0	0	NA	NA
Hslnv0389	247	7	39	7	0	0	NA	NA
Hslnv0393	73	1	11	1	0	0	NA	NA
Hslnv0396	399	14	75	14	0	0	NA	NA
Hslnv0397	98	0	18	NA	0	0	NA	NA
Hslnv0403	10	0	4	NA	0	0	NA	NA
Hslnv0832	0	0	NA	NA	0	0	NA	NA
Hslnv1051	67	30	0	0	1	0	NA	NA
Hslnv1124	6	1	3	1	0	0	NA	NA

Chapter 3. Results

Supplementary Table 3. Summary of inversion mutational effects on gene sequences. Gene annotations are based on GENCODE Version 26 Comprehensive Gene Annotation Set, including gene isoforms with a Transcript Support Level of at least 3, single-exon genes not labelled as "problem", and pseudogenes. Effects of inversions and associated indels at the breakpoints were classified conservatively in six different categories: (1) gene disruption, if there is at least one transcript that encompasses the complete area of one breakpoint; (2) exchange of genic sequences, if two genes of the same family overlap each inversion breakpoint and extend outside of them; (3) inversion of a gene/exon, if the entire gene/exon is situated within the inverted region; (4) inversion of part of an intron, if the inversion and breakpoints are contained inside an intron; (5) overlap of breakpoints with genes within IRs, if there are genes completely embedded within IRs at the inversion breakpoints; and (6) intergenic, if none of the above conditions are fulfilled. For genes overlapping inversion breakpoints within IRs, there could be a potential disruption or exchange of gene sequences, although it is difficult to determine its precise effect due to the high identity of the IRs.

Inversion	Effect	Protein-coding genes	Long non-coding RNAs	Pseudogenes	Other
Hslnv0006	Inversion of part of intron	<i>DSTYK</i>			
Hslnv0030	Exchange of genic sequences	<i>CTRB1, CTRB2</i>			
Hslnv0055	Inversion of part of intron			<i>AC016561.1</i>	
Hslnv0059	Inversion of part of intron	<i>GABRR1</i>			
Hslnv0061	Inversion of part of intron		<i>RP1-60019.1</i>		
Hslnv0098	Inversion of part of intron	<i>ULK4</i>			
Hslnv0102	Inversion or deletion of an exon	<i>RHOH</i> isoform			
Hslnv0105	Inversion of part of intron	<i>SUGCT</i>			
Hslnv0124	Gene disruption	<i>IFITM2</i> isoform			
	Inversion of whole gene	<i>IFITM1</i>			
	Breakpoints overlap genes within IRs		<i>RP11-326C3.7, RP11-326C3.11</i>		
Hslnv0201	Inversion or deletion of an exon	<i>SPINK14</i>			
Hslnv0209	Breakpoints overlap genes within IRs	<i>KRTAP5-10, KRTAP5-11, AP000867.1</i>		<i>AP000867.14, KRTAP5-14P</i>	
Hslnv0241	Breakpoints overlap genes within IRs	<i>AQP12A, AQP12B</i>			
	Inversion of whole gene	<i>AC011298.1</i>	<i>AC011298.2</i>		
Hslnv0278	Inversion of whole gene			<i>FOXO1B</i>	
Hslnv0340	Gene disruption		<i>LINC00395</i>		
Hslnv0344	Breakpoints overlap genes within IRs	<i>SNX6</i>	<i>RP11-671J11.7, antisense RNA RP11-671J11.4</i>		small nuclear RNAs <i>RNU1-27P</i> and <i>RNU1-28P</i>

Chapter 3. Results

Inversion	Effect	Protein-coding genes	Long non-coding RNAs	Pseudogenes	Other
Hslnv0347	Inversion of whole gene				Small nucleolar RNA U3
Hslnv0374	Inversion of part of intron			AC005562.1 (SMURF2P1-LRRC37BP1 readthrough transcribed pseudogene)	
	Inversion of whole gene			SH3GL1P2	
	Breakpoints overlap genes within IRs			RP11-271K11.6, LRRC37BP1	
Hslnv0379	Gene disruption	ZNF257		RP11-420K14.1	
	Inversion of whole gene	ZNF100, ZNF43, ZNF208	RP11-420K14.8, AC003973.4	MTDHP2, MTDHP3, MTDHP4, VN1R84P, RP11-420K14.6, BRI3BPP1, BNIP3P27, BNIP3P28	miRNAs AC092364.2 and AC092364.4
Hslnv0389	Inversion of whole gene	FLNA, EMD			
Hslnv0393	Breakpoints overlap genes within IRs	ARMCX6		ARMCX7P	
Hslnv0396	Breakpoints overlap genes within IRs	PABPC1L2A, PABPC1L2B	antisense RNAs PABPC1L2B-AS1 and RP11-493K23.4		
Hslnv0409	Inversion of part of intron	NLGN4X			
Hslnv1051	Gene disruption			CCDC144B	
	Breakpoints overlap genes within IRs	PRPSAP2		AC107982.4	small non-coding RNAs RN7SL639P and RN7SL627P
	Inversion of whole gene	TBC1D28, ZNF286B, TRIM16L, FBXW10, TVP23B	CTD-2145A24.3	RP11-815I9.3, AC026271.5, FOXO3B, UBE2SP2, RP11-815I9.5, TRIM16L	short non-coding RNA RP11-815I9.4
Hslnv1124	Breakpoints overlap genes within IRs		FAM225A, FAM225B		

Chapter 3. Results

Supplementary Table 4. Potential phenotypic effects of inversions from GWAS data. Inversion effects were estimated through linkage disequilibrium (LD) with GWAS signals reported in the GWAS Catalog (<http://www.ebi.ac.uk/gwas/>) or GWASdb (<http://ijwanglab.org/gwasdb>) databases with a P value of less than 1×10^{-7} . Because each study is focused on individuals with different ancestry, we included in the analysis only inversions in high LD ($r^2 \geq 0.8$) with the GWAS variants in the studied population, the closest one available (e.g. TSI for Sardinian, JPT for Japanese, CHB for Han Chinese or Singapore Chinese, and GIH for South Asian, Indian or Bangladeshi) or the same population group (e.g. EUR for Ashkenazi, Framingham, British, Caucasian or Hutterite, and EAS for Korean). Finally, if studied populations were from different continents or not specified, we used the LD in the global population (GLB). References to each of the GWAS studies are indicated with the same number as in the Supplementary References list or with the dbGAP accession number.

Inversion	Database	GWAS variant	Chr	Position (hg19)	Population of study	Inv. LD (r^2) (Population)	GWAS P value	Phenotypic trait and reference
Hslnv0006	GWAS Catalog	rs16937	chr1	205035455	Ashkenazi Jewish	0.811 (EUR)	5.00×10^{-7}	Schizophrenia ¹⁵
Hslnv0058	GWAS Catalog	rs2844685	chr6	31006855	European	1 (EUR)	3.00×10^{-7}	Drug-induced Stevens-Johnson syndrome or toxic epidermal necrolysis (SJS/TEN) ¹⁶
Hslnv0004	GWASdb	rs2488411	chr1	197058799	British	0.872 (EUR)	3.83×10^{-7}	Height ¹⁷
Hslnv0004	GWASdb	rs1775456	chr1	197733055	European	1 (EUR)	3.00×10^{-7}	Asthma ¹⁸
Hslnv0004	GWASdb	rs1924518	chr1	197738327	European	1 (EUR)	2.90×10^{-7}	Body mass index (asthmatics) ¹⁹
Hslnv0006	GWASdb	rs12142514	chr1	205122529	European	1 (EUR)	2.68×10^{-5}	Glaucoma (primary open-angle) ²⁰
Hslnv0006	GWASdb	rs10900468	chr1	205163057	Not specified	0.881 (GLB)	5.30×10^{-5}	Blood pressure (dbGAP pha003046)
Hslnv0006	GWASdb	rs10900468	chr1	205163057	Not specified	0.881 (GLB)	9.14×10^{-5}	Blood pressure (dbGAP pha003048)
Hslnv0031	GWASdb	rs2937145	chr16	85190230	European	0.981 (EUR)	2.02×10^{-6}	Alzheimer's disease ²¹
Hslnv0045	GWASdb	rs485446	chr21	28022267	Caucasian	0.971 (EUR)	5.79×10^{-7}	Response to TNF antagonist treatment ²²
Hslnv0045	GWASdb	rs366384	chr21	28024225	European	0.986 (EUR)	6.50×10^{-6}	Urinary metabolites ²³
Hslnv0058	GWASdb	rs2844685	chr6	31006855	European	1 (EUR)	3.00×10^{-7}	Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS-TEN) ¹⁶
Hslnv0058	GWASdb	rs2517538	chr6	31013541	Korean	1 (EAS)	2.60×10^{-5}	Height ²
Hslnv0058	GWASdb	rs2517538	chr6	31013541	Hutterite	0.964 (EUR)	2.07×10^{-21}	Lymphocyte counts ²⁵
Hslnv0063	GWASdb	rs10269258	chr7	70440091	European	1 (EUR)	1.60×10^{-5}	Urinary metabolites ²³
Hslnv0098	GWASdb	rs10510717	chr3	41332490	Framingham	0.882 (EUR)	5.00×10^{-6}	Volumetric brain MRI ²⁶
Hslnv0098	GWASdb	rs1487589	chr3	41388428	Not specified	0.804 (GLB)	9.44×10^{-5}	Coronary artery disease ²⁷
Hslnv0098	GWASdb	rs9311269	chr3	41374821	European	0.922 (EUR)	2.62×10^{-5}	Statin-induced myopathy ²⁸
Hslnv0409	GWASdb	rs5916341	chrX	6135980	Not specified	1 (GLB)	2.47×10^{-7}	Amyotrophic lateral sclerosis ²⁹

Supplementary Methods

Inversion genotyping by MLPA and iMLPA

The multiplex ligation-dependent probe amplification (MLPA) technique enables the specific detection of a region of interest by using a pair of oligonucleotide probes (left and right probes) that hybridize contiguously to the target genome sequence in order to be ligated together in a subsequent step. The probes include a variable stuffer sequence and the sequence of the forward or reverse common primers, which are used for the simultaneous amplification of fragments of different sizes formed by the ligation of the left and right probes, and their detection by capillary electrophoresis³⁰. In order to genotype at the same time multiple inversions with IRs or other repetitive sequences at the breakpoints, we developed a new method based on inverse PCR and MLPA that we termed inverse MLPA (iMLPA). iMLPA differs from normal MLPA by the addition of several extra initial steps that are necessary to obtain an orientation-specific unique target sequence for these inversions before probe hybridization.

The iMLPA protocol optimization was carried out by comparison with the known genotypes from the panel of nine individuals in which the inversions had been previously validated² (Supplementary Data 1). A detailed description of iMLPA steps can be found in the patent application EP13382296.5³¹. Briefly, 400 ng of genomic DNA of each sample were first digested overnight at 37°C in six separated 20- μ l reactions with 5 U of the appropriate restriction enzyme (*EcoRI*, *HindIII*, *SacI* or *BamHI* from Roche, and *NsiI* or *BglII* from New England Biolabs), followed by restriction enzyme inactivation for 15 min at 65°C (with the exception of *BglII* that was inactivated for 20 min at 80°C). Next, DNA self-ligation was performed overnight for 16-18 hours at 16°C by mixing together the six restriction enzyme digestions with 1x ligase buffer and 400 U of T4 DNA Ligase (New England Biolabs) in a total volume of 640 μ l (resulting in an optimal concentration of 0.625 ng/ μ l of the DNA fragments generated by each restriction enzyme). Then, the ligase was inactivated and the DNA was fragmented by heating at 95°C for 5 min, purified with the ZR-96 DNA Clean & ConcentratorTM-5 kit (Zymo Research) and resuspended in 7.5 μ l of water.

The last step was the regular MLPA assay using the SALSA MLPA kit (MRC-Holland), according to the manufacturer instructions with minor modifications. In particular, the ligated DNA was denatured at 98°C for 1.5 min, and probe hybridization was carried out adding 1.5 μ l of our iMLPA probe mix (Supplementary Data 10) plus 1.5 μ l of SALSA MLPA Buffer (MRC-Holland) and incubating for 1.5 min at 95°C and 16 hours at 60°C. Ligation of adjacent probes was then performed for 25 min at 54°C by adding 25 μ l water, 1 μ l SALSA Ligase 65, 3 μ l Ligase Buffer A and 3 μ l Ligase Buffer B (MRC-Holland). After ligase inactivation (5 min at 95°C), PCR amplification of ligated probes was performed separately in three groups of 8-9 inversions (Supplementary Data 10) using a common reverse primer and one of three forward primers labeled with a different fluorochrome (FAM, VIC or NED) (Supplementary Data 11). Each PCR was done in 25 μ l with 5-6 μ l of the iMLPA hybridization-ligation reaction, 2 μ l SALSA PCR buffer (MRC-Holland), 0.25 mM each dNTP, 0.2 μ M each primer, 1 μ l PCR buffer without MgCl₂ (Roche), and 2.5 U of Taq DNA polymerase (Roche). PCR conditions were 95°C for 15 sec, 47 cycles of 95°C for 30 sec, 60°C for 30 sec and 72°C for 60 sec, and final extension at 72°C for 25 min. Finally, 0.67 μ l of the three PCRs of each sample were mixed together, analyzed by capillary electrophoresis using an ABI PRISM 3130 Genetic Analyzer (Applied Biosystems), and the peaks were visually inspected using the GeneScan version 3.7 software (Applied Biosystems). For the regular MLPA, the process was identical with the exception that it started directly at the denaturation step of 100-150 ng of genomic DNA in 5 μ l for 5 min at 98°C and that the ligated probes

Chapter 3. Results

were amplified in only two multiplex PCRs with 8-9 inversions each (Supplementary Data 10). In both cases all the successive reactions were carried out in a 96-well plate format to maximize speed and throughput and, with the exception of those used for optimization of the technique, only one MLPA or iMLPA reaction was done for every sample.

Visualization of inversion haplotypes and quantification of recurrence events

Reticulated networks are able to accommodate past recombination events, but each sequence is reduced to a node or edge, making it difficult to understand at the same time haplotype relationships and the spatial distribution of nucleotide changes along the sequence. Therefore, apart from building Median-Joining networks³², we devised our own way to represent the similarities between haplotypes, named integrated haplotype plot (iHPlot), which are similar to the Haplostrips plots that have been recently developed independently³³. Specifically, distances between simplified haplotypes after removing singleton positions were computed as the number of pairwise differences and were clustered with the UPGMA average method implemented in R³⁴ base function `hclust`. The corresponding dendrogram was then created using `ggdendro` R package³⁵ and all the information was integrated with a custom R script using `ggplot2` and `cowplot` packages³⁶. iHPlots were applied to the phased 1000GP Ph1 haplotypes of the inverted region and the imputed 1000GP Ph3 haplotypes based on inversion tag SNPs or on homozygotes for each orientation. For 1000GP Ph3 data, we used only accessible SNPs (excluding indels) according to the pilot accessibility mask that includes more SNPs than the strict mask³⁷. In addition, besides the inverted region, whenever possible, we extended the analysis to the non-recombining region flanking the breakpoints (excluding associated indels and IRs) to increase the resolution of haplotype discrimination.

To determine more reliably the evolutionary history of each inversion, we combined the information from the different strategies for phasing and visualization of the inverted region haplotypes: 1) Median-joining networks of 1000GP Ph1 phased data; 2) iHPlots of 1000GP Ph1 phased data; and 3) iHPlots of 1000GP Ph3 published haplotypes (including the flanking non-recombining region if available). Moreover, HapMap phased data was used to confirm 1000GP results, although in many cases there was information from just a few SNPs. All inversions could be analyzed by at least some of the method combinations, except HsInv0041, which did not have enough variants and was excluded. Results of inversions with perfect tag variants ($r^2 = 1$) were determined mainly from the extended 1000GP Ph3 haplotypes, but consistent conclusions were obtained in the different analyses. The only exceptions were a few phasing errors by PHASE 2.1³ in 1000GP Ph1 data in several inversions and a likely imputation error in HsInv0409 O2/O2 individual NA20530 in 1000GP Ph3 (in which one of the haplotypes is typical of O1 chromosomes, whereas in 1000GP Ph1 both haplotypes belong clearly to the O2 group).

On the other hand, the estimation of recurrence events for inversions without tag variants relied mainly in the analysis of the iHPlots from phased 1000GP Ph1 haplotypes, since they contain all genotyped individuals in common, although there could be more phasing errors in inversion heterozygotes. First, we defined the putative original inversion event based on the ancestral allele information, the haplotype diversity within each orientation, and the frequency and geographical distribution of haplotypes, tending to favor as the first event those occurring in Africa. Next, we conservatively identified additional inversion or re-inversion events in differentiated clusters of haplotypes with both orientations. In order to consider that there has been inversion recurrence, these clusters have to differ from all other ones, and especially from those with the same orientation as the

potential recurrence event, by three or more sequence changes along most of the inverted region (and spanning at least 2 kb). Therefore, the presence of these nucleotide differences cannot be explained easily by other mechanisms, such as gene conversion or sequence errors. Direction of recurrence events was defined based on the relationship between the clusters and the frequency of the haplotypes with each orientation (Supplementary Data 3). Possible phasing errors in inversion heterozygotes were checked manually by determining if switching the orientation of both haplotypes still supports unequivocally the existence of recurrence. The same analysis was also repeated with 1000GP Ph3 iHPlots, in which just the orientation from haplotypes of *O1* and *O2* homozygotes is assigned, and only those clear recurrence events not invalidated with the new data were considered. It is important to take into account that since recurrence detection relies on differentiated haplotype clusters, it is not possible to distinguish more than one event within a cluster and there is a bias to predict more potential recurrence in larger inversions with more variants. For example, in six of the smallest NAHR inversions, *O1* and *O2* haplotypes are too similar to identify individual recurrence events (Supplementary Data 3). In two others (HsInV0124, HsInV0397), most *O1* and *O2* haplotypes belong to the same big cluster with just few differences between them and no clear recurrence can be identified. As a consequence, these results have to be interpreted with caution.

In the case of HsInV0832, we gathered publicly available information of the chr. Y haplogroups of 232 of the 282 genotyped males from different sources³⁸⁻⁴³, as listed in Supplementary Data 4. Most of these studies determined also the evolutionary relationship between the chr. Y haplogroups, which were largely consistent and are shown in Fig. 3 in a simplified genealogical tree using the branch lengths of Poznik *et al.*⁴³. This allowed us to identify with confidence five independent inversion events in the HsInV0832 region, assuming the most parsimonious scenario. HsInV0832 inversion rate was estimated dividing the number of inversion events (n) by the number of generations (g) encompassed in the phylogeny that relates the 217 Y-chromosomes for which sequence data was available⁴³. To estimate g , we used the data from the B-T branch split to the leaves from Poznik *et al.*⁴³, including a B-T branch split time of 105.8 kya, a total number of mutations in all branches involved in the phylogeny that relates those 217 males of 17,332, and an average number of mutations of all branches of 784.57, plus a generation time of 25 years as in Repping *et al.*⁴⁴. This results in 93,489 generations and an inversion rate of 5.35×10^{-5} per generation. In addition, to have another estimate of the inversion rate, we also used the approach of Hallast *et al.*⁴⁵, which was based on Repping *et al.*⁴⁴ that resequenced 80 kb in 47 Y chromosomes covering most major branches of the phylogenetic tree to obtain the nucleotide divergence in an unbiased manner. According to their data, we estimated a lower and upper bound of g of 127,467-336,533 generations by calculating the maximum (631) and minimum (239) total number of mutations spanning all the informative branches and an average number of mutations to the root in the different branches of 8.85, assuming a divergence time of 118 kya and a generation time of 25 years^{44,45}. This yields an inversion rate of $1.48-3.92 \times 10^{-5}$, which is quite similar to the previous one.

Bioinformatic analysis of inversion orientation in non-human primate genomes

The bioinformatic analysis of inversion orientation in the available genome assemblies of chimpanzee (panTro5), gorilla (gorGor5), orangutan (ponAbe2) and rhesus macaque (macRhe8) was done using an automated bash script based on the command-line blat tool (v35x1)⁴⁶. For each inversion, three separate hg18 sequences were extracted using twoBitToFa UCSC utility: the inverted region (or alternatively two separate internal 10-kb sequences adjacent to each breakpoint when the inverted region is longer than 20 kb) and the two 10-kb segments flanking each breakpoint outside the

Chapter 3. Results

inversion. We excluded the breakpoint intervals and their associated IRs and indels to avoid ambiguous mappings. Then, each sequence was aligned with *blat* to the genomes of interest, which were downloaded from the UCSC Genome Browser website in 2bit format. The longest hit was kept as the likely homologous region in the target assembly and orientation was defined as *O1* if all best hits mapped in the same strand, and as *O2* if the internal best hit(s) mapped in the opposite strand than those from the external sequences. As quality control, all best hits needed to be in the same scaffold or chromosome and the total span in the target assembly had to be 0.5-2-times that in hg18. In addition, in those cases in which the orientation could not be reliably defined or was inconsistent across species or with published data^{47,48}, results from the automated analysis were revised by aligning the sequences spanning the entire region from each assembly with the Gepard dotplot application⁴⁹ and Blast2seq⁵⁰, using default parameters.

Inversion age estimate

Inversion age was estimated from the net number of differences accumulated between sequences in opposite orientations. This number was obtained by subtracting from the mean pairwise nucleotide differences between *O1* and *O2* chromosomes, the expected average pairwise differences in the original population (before the generation of the inversion), which was approximated by the largest value of the average pairwise differences within sequences with the same orientation (either *O1* or *O2*). To ensure the maximum reliability of the divergence estimates, we considered all SNPs available in the extended 1000GP Ph3 haplotypes and sequence orientation was determined by the presence of tag variants or by using only *O1* and *O2* homozygous individuals. For two low-frequency inversions, HsInv0061 and HsInv1051, divergence could not be estimated because there were no tag variants in the analyzed region and all inversion carriers are heterozygous. A first age estimate was obtained by using a constant substitution rate of 10^{-9} changes per base-pair per year⁵¹. Moreover, in order to control for local differences in substitution rates, we obtained two additional local estimates from the divergence between human and chimpanzee or gorilla genomes, considering, respectively, a split time of 6 and 8 million years in each case (Supplementary Data 6). Pairwise LASTZ alignments⁵² of human hg19 assembly with chimpanzee (assembly CSAC 2.1.4/panTro4) and gorilla (assembly gorGor3.1) genomes were retrieved from ENSEMBL GRCh37 portal⁵³, using the Compara Perl API⁵⁴. We then used Kimura's two-parameter substitution model⁵⁵ to calculate the divergence between human and outgroup assemblies in the same inversion region analyzed above, after removing alignment gaps and non-syntenic alignment blocks. Alignments shorter than 1 kb were discarded, including the missing chimpanzee alignment for inversion HsInv0045 due to a deletion of the whole region and both outgroup alignments for inversion HsInv0041.

Simulation of inversion detection ascertainment bias

Given the heterogeneous origin of the inversions included in the study (Supplementary Data 1), to take into account the effect of ascertainment bias associated to inversion detection, we simulated two different processes using a bash and R³⁴ pipeline: one for the 38 autosomal or chr. X inversions detected from the fosmid paired-end mapping (PEM) data of nine individuals⁵⁶, and another for the six inversions detected exclusively by comparison of two genome assemblies⁵⁷. First, we built panels from 1000GP individuals with matching demographic and gender composition to the detection samples. For the PEM study, corresponding to eight females and one male with African (4 YRI), East Asian (1 CHB and 1 JPT) and European ancestry (2 CEU and presumably European individual NA15510)^{56,58}, we were able to use all the original individuals except for NA19240 and NA15510, which were replaced with NA18502 and NA12717. For the genome comparison study, we used a

randomly selected European male (NA12872) and selected all SNPs that contained the alternative allele. Variants were filtered from 1000GP vcf files (<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502>) with bcftools v1.7 view command⁵⁹. Additional filters were applied to the SNPs to simplify comparisons (keeping only SNPs with assigned ID and ancestral allele in 1000GP vcf files), to use only putatively neutral variants (conservation GERP score⁶⁰ below 2 in functionally annotated 1000GP vcf files ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/supporting/functional_annotation/unfiltered/), and to ensure high SNP quality (accessible according to the 1000GP Ph3 strict accessibility mask). However, the effect of these extra filters on the final frequency distribution was negligible, affecting just <1.5% of the detectable SNPs, which have similar average frequency to the rest of SNPs.

Second, we simulated the detection process of the inversions with the methods employed. This step was only simulated for the PEM data, since the limitations of inversion detection by assembly comparison are likely independent of variant frequency (and instead probably just related to repeat content and complexity of the genomic region). PEM detection, on the other hand, is affected by the sizes of the inversion and the IRs at the breakpoints, both of which limit the number of PEMs supporting it. To that end, we modeled the detection of an inversion that is present in the PEM panel as a function of these two characteristics and the number of chromosomes with the alternative orientation. Specifically, the probability of having two discordant PEMs in the whole panel (the minimum number necessary to detect an inversion) was calculated by a Poisson distribution with a lambda parameter equal to the expected number of discordant PEMs ($E(\text{disc})$). Following Equation 1 in Lucas-Lledó *et al.*⁶¹, $E(\text{disc})$ for the two breakpoints of a given inversion (*inv*) and IR (*ir*) size, considering the average PEM insert length (*ins*) and read length (*read*), was estimated as:

$$E(\text{disc}) = 2 \frac{\min(\text{inv} - \text{read}, \text{ins} - 2 \text{read} - \text{ir})}{g} n f$$

where g is the sequenced haploid genome size (approximated to 3 Gb), n is the total number of fosmid sequences^{56,62}, and f is the fraction of chromosomes carrying the mutation in the nine individuals analyzed. For each PEM inversion, a custom R script was used to generate a matching random sample of 10,000 SNPs. These SNPs were selected according to the detection probability of the SNPs based on their frequency in the PEM panel and the inversion characteristics, including the chromosome type (autosomes or chr. X).

Frequency differences between populations (F_{ST})

To calculate F_{ST} ⁶³, we created vcf files containing the inversion genotypes for the 434 individuals common to 1000GP Ph3 and used the `--weir-fst-pop` option from vcfTools (v0.1.15). F_{ST} values were obtained for each pair of populations within the same population group, each pair of population groups, and globally. Genome-wide F_{ST} null distributions were obtained from 1000GP Ph3 biallelic SNPs polymorphic in the same 434 individuals that are accessible according to the strict criteria and have a defined ancestral allele. To control F_{ST} dependence on chromosome type and allele frequency, empirical P values for each inversion and comparison were estimated as the fraction of the SNP distribution (including always a minimum of 10,000 SNPs) from the same chromosome type (autosome or chr. X) and global MAF bin (from 0 to 0.5 in 0.05 increases) as the inversion with equal or larger F_{ST} (Supplementary Data 7). Reduced levels of population differentiation are sometimes interpreted as evidence of balancing selection. However, power to detect the extreme low F_{ST} values

Chapter 3. Results

was very low. Global population differentiation for all inversions together was measured by a hierarchic analysis of molecular variance (AMOVA) according to geographic criteria using Arlequin v3.5⁶⁴. Resulting variation was mainly due to the difference between the three continental (CT) groups for both autosomal inversions ($F_{ST} = 0.13$, $P < 0.0001$; $F_{CT} = 0.11$, $P < 0.0001$; $F_{SC} = 0.03$, $P = 0.02$) and chr. X inversions ($F_{ST} = 0.24$, $P < 0.0001$; $F_{CT} = 0.20$, $P < 0.0001$; $F_{SC} = 0.05$, $P < 0.0001$).

Linked site frequency spectrum (LSFS) selection tests

For LSFS tests we used a simplified version of the tests, i.e. weights were chosen computing the covariance in the approximation of unlinked sites, and we assumed strong selection coefficients in two scenarios: (1) classical selective sweep (positive selection); and (2) long-term balancing selection. The frequency spectrum of variants closely linked to the inversion, including their linkage pattern (nested or disjoint) with the inverted allele⁶⁵, was calculated in relatively-small non-overlapping windows of 3 kb in order to reduce the effects of recombination within each window on the empirical null spectrum. The windows tested were localized either within the inversion or the non-recombining flanking regions and skipped the breakpoint interval and IRs to avoid errors from associated indels or incorrect short-read mappings. The autosome-wide empirical spectrum was computed on windows of the same size (3 kb) around all autosomal SNPs. The LSFS was calculated from biallelic 1000GP Ph3 SNPs in the 434 samples with inversion genotypes. We removed from the analysis all SNPs with a GERP score⁶⁰ higher than 2 to reduce the effect of linked selection, as well as those SNPs within 0.5 Mb of any of the inversions in our dataset, since their dynamics could be heavily influenced by the inversion itself. Tests were conditioned on the inversion frequency in the different populations. For each test distribution conditioned on minor allele counts of at least 6, a local cubic smoothing was finally applied to the frequency dependence of the distribution, considering derived allele counts between +5 and -5 with respect to that of the inversion. In addition, to control for the complex demographic history of human populations, we used the empirical autosome-wide first and second moments of the empirical linked frequency spectrum of SNPs in each population as a substitute for the null spectrum.

Edgington's method⁶⁶ was used to combine the P values of the same windows of each population. Combining the results across different windows of an inversion is complicated by the correlation of their P values, since in the absence of recombination they share the same evolutionary history. We dealt with this in two ways. The first approach (conservative) was to assume an arbitrary dependence between windows, and compute the False Discovery Rate (FDR) correcting for multiple correlated testing via Benjamini-Hochberg-Yakutieli⁶⁷ for each inversion separately and for all inversions together (in the latter case, HsInV0379 was removed from the analysis due to its size and unbalanced contribution to the statistical noise) (Supplementary Data 7). The second approach (approximate) is to approximate the joint distribution across correlated windows as a multidimensional Gaussian distribution by: (1) applying a Gaussian transformation to the P values; (2) computing the empirical correlation across all pairs of windows of the same inversion; (3) computing the average Gaussian score for each inversion; (4) building an equicorrelated matrix of the same size as the number of windows in the inversion, with elements equal to 1 on the diagonal and to the empirical correlation off the diagonal; and (5) comparing the average Gaussian score with the average score extracted from a multidimensional Gaussian distribution with covariances distributed as the equicorrelated matrix. This approach was applied both to each population separately and to the combined P values from all populations (Supplementary Data 7).

Non-central deviation (NCD) selection tests

NCD statistics were adapted to test long-term balancing selection acting on autosomal and chr. X inversion regions. NCD1 detects site frequency spectrum shifts towards an equilibrium frequency as expected under balancing selection, whereas NCD2 incorporates also information on polymorphism density and is most powerful to detect long-term balancing selection⁶⁸. NCD1 and NCD2 were computed genome-wide as previously described⁶⁸ using overlapping windows of 2 kb (with 1 kb step), which fit well the size of the smaller inversions, and three target frequencies (0.3, 0.4 and 0.5). Human polymorphism data was obtained from 1000GP Ph3 SNPs from all individuals of the seven studied populations accessible according to the pilot accessibility mask, and human-chimpanzee differences were obtained from the hg19-panTro4 alignments available at the UCSC Genome Browser⁴⁶. Windows of the 44 inversions were defined with the same criteria as in the LSFS test, including the inverted and flanking non-recombining region, while avoiding breakpoint, IR and indel intervals. Nine inversions did not have any window passing the filtering criteria and were not analyzed (HsInv0031, HsInv0041, HsInv0045, HsInv0055, HsInv0061, HsInv0072, HsInv0344, HsInv0409, and HsInv1124).

A raw empirical P value was assigned to each inversion window corresponding to their quantile in the null genome-wide distribution of the statistic in that population computed with the target frequency most similar to the inversion global MAF⁶⁸, and the lowest P value of all the windows for each inversion and population was selected. To correct for the fact that some inversions have more than one window, we then sampled 1,000 sets of regions of equal size and from the same chromosome as each of the inversions, selected the lowest P value of all the windows of each region, and obtained the empirical distribution of minimum P values equivalent to that of the inversion. Finally, size-corrected P values for each inversion and population were estimated from the quantile in the corresponding minimum- P -value distribution (Supplementary Data 7). Since balancing selection signals are expected to be shared across multiple populations⁶⁸, we chose as candidates those inversions with three or more populations with size-corrected P values < 0.01 (strong candidates) or P values < 0.05 (weak candidates). The main limitation of these tests is that, by reducing recombination, inversions may affect the expected empirical distribution. For example, inversions increase variance in the SFS or the age of alleles. Nevertheless, the reduced recombination means stronger effects of background selection, which results in lower levels of diversity and younger alleles, which are the opposite to the signatures detected by the NCD statistics. An additional limitation is that the signatures of balancing selection could be due to any SNP within the windows, rather than the inversion itself. However, the functional effects of the inversion are expected to be much stronger than those of a single nucleotide change.

Validation of lymphoblastoid cell lines (LCLs) gene-expression analysis results

We employed different strategies to confirm the reliability of the results of the gene expression analysis from LCLs, which are summarized in Supplementary Fig. 4C-F. In particular, we compared our results with those of two additional commonly-used eQTL mapping methods: the one described by the GTEx Project⁶ and edgeR-limma^{7,8}. In the GTEx analysis, RPKM values were quantile normalized across all samples and gene/transcript expression levels were subsequently adjusted by rank-based inverse normal transformation per each gene and transcript. In this case, technical confounding variation was accounted with the PEER software⁶⁹. The number of technical covariates was chosen to optimize eQTL identification by maximizing consistent eQTL calls and minimizing differences between GTEx and QTLtools pipelines, but avoiding overfitting the model. We tested up to

Chapter 3. Results

the top 60 expression-derived PEER factors and 60 principal components of the PCA, taken in groups of 5 in decreasing order of the variance explained, and determined the optimal number according to the results overlap. Linear regressions were then done with FastQTL v2.0⁷⁰, including the selected PEER factors (for gene and transcript analysis, respectively, 5 and 20 in the experimental dataset and 35 and 55 in the imputed dataset), gender, and the three population principal components as covariates. In the edgeR-limma workflow, raw read counts were corrected by library size in counts per million. Genes and transcripts that passed the expression-level cutoff (0.1 counts per million in at least two samples) were normalized with trimmed mean of M-values (TMM)⁷¹ and transformed with voom⁷. Next, limma fit an additive linear model to contrast differentially expressed genes across genotypes, including gender, population and sequencing laboratory as covariates. Other potential batch effects were uncovered with the SVA package (1 and 2 for experimental and imputed sets, respectively)⁷². All *P* values were corrected by Storey & Tibshirani FDR⁷³.

As an independent replication of these results, we also examined the available gene-expression data from blood samples of ~2,000 Estonian individuals obtained by hybridization with Illumina HumanHT-12 v3.0 Gene Expression BeadChip arrays. In this case, we checked directly the effects of 1,541 SNPs that were in high LD ($r^2 \geq 0.8$) with 33 inversions either globally (27) or just in Europeans (6). These SNPs were already imputed in Estonian samples based on 1000GP Ph1 variants. In total, six potential inversion-eQTL effects were identified in this study in blood (FDR < 5%): Hslnv0006 and *DSTYK*; Hslnv0058 and *HLA-E* and *HLA-C*; Hslnv0095 and *SPP1*; Hslnv0201 and *FBXO38*; and Hslnv0209 and *FOLR3*. Of those, five were also found in the GTEx or GEUVADIS data, which represents a good degree of consistency considering the different expression quantification platforms and analysis methods used.

Integrative analysis of functional and selection evidence

Overlap of functional and selection signals for the 44 autosomal and chr. X inversions analyzed was calculated by a Fisher's exact test of independence. To reduce possible spurious signals, we focused on selection signatures calculated on the inversion itself (excluding NCD1 and NCD2 test results) and all functional effects except those from GWAS data, which in most cases are related to diseases and could have detrimental consequences during evolution. Criteria for classification of strong and weak selection and functional evidence are explained in Supplementary Data 7 or Fig. 2. The association was replicated considering only the strongest functional effects and selection signals for the 44 inversions (Fisher's exact test $P = 0.0130$) or just the 21 inversions with perfect tag SNPs that were included in most analyses, which comprise all NH inversions, except Hslnv0102, plus Hslnv0040 (Fisher's exact test $P = 0.0300$).

Supplementary References

1. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
2. Martínez-Fundichely, A. *et al.* InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Res.* **42**, D1027-32 (2014).
3. Stephens, M. & Donnelly, P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**, 1162–9 (2003).
4. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* **5**, e1000529 (2009).
5. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
6. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–13 (2017).
7. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
8. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–40 (2010).
9. Hulsen, T., de Vlieg, J. & Alkema, W. BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9**, 488 (2008).
10. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
11. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325–9 (2011).
12. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–6 (2011).
13. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–7 (2015).
14. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–75 (2007).
15. Goes, F. S. *et al.* Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **168**, 649–59 (2015).
16. Génin, E. *et al.* Genome-wide association study of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in Europe. *Orphanet J. Rare Dis.* **6**, 52 (2011).
17. Strachan, D. P. *et al.* Lifecourse influences on health among British adults: effects of region of residence in childhood and adulthood. *Int. J. Epidemiol.* **36**, 522–31 (2007).
18. Ferreira, M. A. R. *et al.* Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378**, 1006–14 (2011).
19. Melén, E. *et al.* Genome-wide association study of body mass index in 23 000 individuals with and without asthma. *Clin. Exp. Allergy* **43**, 463–74 (2013).
20. Gibson, J. *et al.* Genome-wide association study of primary open angle glaucoma risk and quantitative traits. *Mol. Vis.* **18**, 1083–92 (2012).
21. Pérez-Palma, E. *et al.* Overrepresentation of glutamate signaling in Alzheimer’s disease: network-based pathway enrichment using meta-analysis of genome-wide association studies. *PLoS One* **9**, e95413 (2014).
22. Plant, D. *et al.* Genome-wide association study of genetic predictors of anti-tumor necrosis factor treatment efficacy in rheumatoid arthritis identifies associations with polymorphisms at seven loci. *Arthritis Rheum.* **63**, 645–53 (2011).
23. Suhre, K. *et al.* A genome-wide association study of metabolic traits in human urine. *Nat. Genet.* **43**, 565–9 (2011).
24. Cho, Y. S. *et al.* A large-scale genome-wide association study of Asian populations uncovers

Chapter 3. Results

- genetic factors influencing eight quantitative traits. *Nat. Genet.* **41**, 527–34 (2009).
25. Cusanovich, D. A. *et al.* The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. *Hum. Mol. Genet.* **21**, 2111–23 (2012).
 26. Seshadri, S. *et al.* Genetic correlates of brain aging on MRI and cognitive test measures: a genome-wide association and linkage analysis in the Framingham Study. *BMC Med. Genet.* **8 Suppl 1**, S15 (2007).
 27. Samani, N. J. *et al.* Genomewide Association Analysis of Coronary Artery Disease. *N. Engl. J. Med.* **357**, 443–53 (2007).
 28. Isackson, P. J. *et al.* Association of common variants in the human eyes shut ortholog (EYS) with statin-induced myopathy: evidence for additional functions of EYS. *Muscle Nerve* **44**, 531–8 (2011).
 29. Schymick, J. C. *et al.* Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet. Neurol.* **6**, 322–8 (2007).
 30. Schouten, J. P. *et al.* Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* **30**, e57 (2002).
 31. Cáceres, M., Villatoro, S. & Aguado, C. Inverse Multiplex Ligation-dependent Probe Amplification (iMLPA), an in vitro method of genotyping multiple inversions. (2015).
 32. Bandelt, H. J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).
 33. Marnetto, D. & Huerta-Sánchez, E. Haplotrips: revealing population structure through haplotype visualization. *Methods Ecol. Evol.* **8**, 1389–92 (2017).
 34. R Core Team. R: A Language and Environment for Statistical Computing. (2017).
 35. de Vries, A. & Ripley, B. D. gg dendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'. (2016).
 36. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).
 37. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
 38. Wei, W. *et al.* A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res.* **23**, 388–95 (2013).
 39. Magoon, G. R. *et al.* Generation of high-resolution a priori Y-chromosome phylogenies using 'next-generation' sequencing data. *bioRxiv* 802 (2013). doi:10.1101/000802
 40. Wang, C.-C. & Li, H. Discovery of Phylogenetic Relevant Y-chromosome Variants in 1000 Genomes Project Data. *arXiv* 1310.6590 (2013).
 41. Van Geystelen, A., Decorte, R. & Larmuseau, M. H. D. Updating the Y-chromosomal phylogenetic tree for forensic applications based on whole genome SNPs. *Forensic Sci. Int. Genet.* **7**, 573–80 (2013).
 42. Hallast, P. *et al.* The Y-Chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. *Mol. Biol. Evol.* **32**, 661–73 (2015).
 43. Poznik, G. D. *et al.* Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.* **48**, 593–9 (2016).
 44. Repping, S. *et al.* High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat. Genet.* **38**, 463–7 (2006).
 45. Hallast, P., Balaesque, P., Bowden, G. R., Ballereau, S. & Jobling, M. A. Recombination dynamics of a human Y-chromosomal palindrome: Rapid GC-biased gene conversion, multi-kilobase conversion tracts, and rare inversions. *PLoS Genet.* **9**, e1003666 (2013).
 46. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
 47. Aguado, C. *et al.* Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. *PLoS Genet.* **10**, e1004208 (2014).
 48. Vicente-Salvador, D. *et al.* Detailed analysis of inversions predicted between two human

- genomes: errors, real polymorphisms, and their origin and population distribution. *Hum. Mol. Genet.* **26**, 567–81 (2017).
49. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–8 (2007).
 50. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10 (1990).
 51. Moorjani, P., Gao, Z. & Przeworski, M. Human germline mutation and the erratic evolutionary clock. *PLoS Biol.* **14**, e2000744 (2016).
 52. Harris, R. S. Improved pairwise alignment of genomic DNA. (The Pennsylvania State University, 2007).
 53. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res.* **44**, D710–6 (2016).
 54. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database* **2016**, bav096 (2016).
 55. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–20 (1980).
 56. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
 57. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
 58. Korb, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–6 (2007).
 59. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–93 (2011).
 60. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
 61. Lucas-Lledó, J. I. & Cáceres, M. On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. *PLoS One* **8**, e61292 (2013).
 62. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–32 (2005).
 63. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution (N. Y.)* **38**, 1358–70 (1984).
 64. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–7 (2010).
 65. Ferretti, L. *et al.* The neutral frequency spectrum of linked sites. *Theor. Popul. Biol.* **123**, 70–9 (2018).
 66. Edgington, E. S. An additive method for combining probability values from independent experiments. *J. Psychol.* **80**, 351–63 (1972).
 67. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–88 (2001).
 68. Bitarello, B. D. *et al.* Signatures of long-term balancing selection in human genomes. *Genome Biol. Evol.* **10**, 939–55 (2018).
 69. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–7 (2012).
 70. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–85 (2016).
 71. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
 72. Leek, J. *et al.* sva: Surrogate Variable Analysis. R package version 3.28.0 (2018).
 73. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100**, 9440–5 (2003).

Chapter 3. Results

74. Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7**, 12989 (2016).
75. Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–85 (2017).
76. Sanders, A. D. *et al.* Characterizing polymorphic inversions in human genomes by single cell sequencing. *Genome Res.* **26**, 1575–87 (2016).
77. Li, L. *et al.* OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. *Genome Biol.* **18**, 230 (2017).
78. Audano, P. A. *et al.* Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–75 (2019).
79. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
80. Martínez-Fundichely, A. *et al.* Accurate characterization of inversions in the human genome from paired-end mapping data with the GRIAL algorithm. *In prep.*
81. Lucas-Lledó, J. I., Vicente-Salvador, D., Aguado, C. & Cáceres, M. Population genetic analysis of bi-allelic structural variants from low-coverage sequence data with an expectation-maximization algorithm. *BMC Bioinformatics* **15**, 163 (2014).
82. Puig, M. *et al.* Functional impact and evolution of a novel human polymorphic inversion that disrupts a gene and creates a fusion transcript. *PLoS Genet.* **11**, e1005495 (2015).
83. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–9 (2014).
84. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–6 (2012).
85. Keller, A. *et al.* New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* **3**, 698 (2012).
86. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225–8 (2014).
87. Pantano, L., Armengol, L., Villatoro, S. & Estivill, X. ProSeek: A web server for MLPA probe design. *BMC Genomics* **9**, 573 (2008).

3.2 Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR

In this article, we have carried out a complete characterization of human inversions flanked by long inverted repeats. For that, a new method based on droplet digital PCR (ddPCR) technology was developed to genotype these inversions in a large number of samples for the first time. Specifically, I designed a statistical clustering method for genotype calling from ddPCR linkage ratio measurements. I also estimated the factors influencing inversion recurrence levels. Finally, I checked the potential functional impact of 16 inversions with experimental genotyping information across 95 African, European and East-Asian individuals. Specifically, I tested effects on gene expression in multiple tissues and associated inversions with GWAS signals.

Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR

Marta Puig^{1*}, Jon Lerga-Jaso¹, Carla Giner-Delgado¹, Sarai Pacheco¹, David Izquierdo¹, Alejandra Delprat¹, Magdalena Gayà-Vidal², Jack F. Regan³, George Karlin-Neumann³, Mario Cáceres^{1,4*}

¹ Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.

² CIBIO/InBIO Research Center in Biodiversity and Genetic Resources, Universidade do Porto, Vairão, Distrito do Porto, Portugal.

³ Digital Biology Center, Bio-Rad Laboratories, Pleasanton, CA, United States.

⁴ ICREA, Barcelona, Spain.

* Corresponding authors:

Mario Cáceres: mcaceres@icrea.cat

Marta Puig: marta.puig@uab.cat

Running title: Characterization of human inversions by ddPCR

Keywords: polymorphic inversions, structural variants, ddPCR technology, recurrent mutation, functional effects

ABSTRACT

Despite the interest in characterizing all genomic variation, the presence of large repeats at the breakpoints of many structural variants hinders their analysis. This is especially problematic in the case of inversions, since they are balanced changes without gain or loss of DNA. Here we tested novel linkage-based droplet digital PCR (ddPCR) assays on 20 inversions ranging from 3.1 to 742 kb and flanked by long inverted repeats (IRs) of up to 134 kb. Among these, we validated 13 inversions predicted by different genome-wide techniques. In addition, we have generated new experimental human population information across 95 African, European and East-Asian individuals for 16 of them, including four already known inversions for which there were no high-throughput methods to determine directly the orientation, like the well-characterized 17q21 inversion. Through comparison with previous data, independent replicates and both inversion breakpoints, we have demonstrated that the technique is highly accurate and reproducible. Most of the studied inversions are frequent and widespread across continents, showing a negative correlation with genetic length. Moreover, all except two show clear signs of being recurrent, and the new data allowed us to define more clearly the factors affecting recurrence levels and estimate the inversion rate across the genome. Finally, thanks to the generated genotypes, we have been able to check inversion functional effects in multiple tissues, validating gene expression differences reported before for two inversions and finding new candidate associations. Our work therefore provides a tool to screen these and other complex genomic variants quickly in a large number of samples for the first time, highlighting the importance of direct genotyping to assess their potential consequences and clinical implications.

Chapter 3. Results

INTRODUCTION

During the last years a substantial amount of information has accumulated about different types of genomic changes, ranging from single nucleotide polymorphisms (SNPs) to more complex structural variants (SVs) (The 1000 Genomes Project Consortium 2015; Sudmant et al. 2015; Handsaker et al. 2015; Audano et al. 2019). However, inversions remain as one of the most difficult classes of variation to identify and characterize. Polymorphic inversions have been studied for many years and are known to have adaptive value and be associated with phenotypic effects in many species, including latitudinal clines in *Drosophila*, mimicry in butterflies, reproductive isolation in plants, adaptation to freshwater in stickleback fishes or different behaviors in sparrows, among many other examples (Umina et al. 2005; Joron et al. 2011; Lowry and Willis 2010; Jones et al. 2012; Thomas et al. 2008). In humans, hundreds of inversions have been predicted (Martínez-Fundichely et al. 2014; Puig et al. 2015a). However, for many of them it is not possible to detect reliably both orientations due to their balanced nature and the complexity of their breakpoints, and only a few have been analyzed in detail (Stefansson et al. 2005; Salm et al. 2012; González et al. 2014; Puig et al. 2015b; Giner-Delgado et al. 2019).

Approximately half of human inversions have IRs at their breakpoints (Martínez-Fundichely et al. 2014; Puig et al. 2015a), which can be up to hundreds of kilobases long. Therefore, short reads from next generation sequencing technologies (~100-150 bp) (Sudmant et al. 2015; Hehir-Kwa et al. 2016; Collins et al. 2019), and even longer reads from single-molecule sequencing technologies (~10 kb on average) (Huddleston et al. 2017; Shao et al. 2018; Audano et al. 2019) or paired-end mapping (PEM) data from large fragments (~40-kb fosmid clones) (Kidd et al. 2008), are often not able to jump across the IRs at inversion breakpoints, precluding the detection of these inversions. New methods like single-cell sequencing of one of the two DNA strands by Strand-seq (Sanders et al. 2016; Chaisson et al. 2019) or generation of genome maps based on linearized DNA molecules labeled at particular sequences by optical mapping (Li et al. 2017; Levy-Sakin et al. 2019) have demonstrated their ability to detect inversions despite the presence of long IRs. Nevertheless, these techniques are not suitable for the analysis of large numbers of individuals.

The recent genotyping of 45 inversions in multiple individuals by a combination of inverse PCR (iPCR) and multiplex ligation-dependent probe amplification (MLPA) (Giner-Delgado et al. 2019) has increased extraordinarily the amount of data available on human inversions, although there are still limitations in the size of the IRs that can be spanned (up to ~25-30 kb in this case). At the other end, low-throughput techniques like fluorescence in situ hybridization (FISH) have been used to study inversions with large IRs, but FISH is time consuming and can be applied only to inverted segments in the Mb scale where probes can be detected as separate signals (Antonacci et al. 2009; Salm et al. 2012). This leaves a set of potential inversions with IRs too large for iPCR-based techniques and inverted regions too small for FISH analysis that cannot be validated or genotyped in a large sample to determine their functional effects and evolutionary history.

Moreover, the great majority of inversions mediated by IRs have been shown to occur recurrently several times both within the human lineage and in non-human primates (Aguado et al. 2014; Vicente-Salvador et al. 2017; Giner-Delgado et al. 2019; Antonacci et al. 2009). Unlike inversions without IRs, which always have a single origin, these recurrent inversions tend to show low linkage disequilibrium (LD) to nearby nucleotide variants (Giner-Delgado et al. 2019), and as a consequence, they have been probably missed in existing genome-wide association studies (GWAS) of different phenotypes. In fact, several inversions have been associated to potential important effects on gene expression, phenotypic traits and disease susceptibility (de Jong et al. 2012; Myers et al. 2005; Zabetian et al. 2007; Webb et al. 2008; Salm et al. 2012; González et al. 2014; Puig et al. 2015b; Giner-Delgado et al. 2019). Inversion recurrence therefore highlights the need to expand the set of analyzed inversions by developing new tools able to genotype them directly through the order of the sequences around the breakpoints, rather than relying on linkage to other variants easier to genotype or computational methods based on SNP combinations (Ma and Amos 2012; Cáceres et al. 2012; Cáceres and González 2015).

In this context, droplet digital PCR (ddPCR) provides the opportunity to fill the void in inversion characterization by detecting linkage between amplicons located at both sides of the breakpoint repeats, making it possible to jump long genomic distances. The ddPCRTM technology has already been used to detect copy number variation (Boettger et al. 2012), viral load (Strain et al. 2013), low-frequency transcripts (Hindson et al. 2013), rare mutations or cell free DNA (Olmedillas-López et al. 2017; Camunas-Soler et al. 2018) among other applications. Recently, it was shown that this technique could also be used to phase variants separated by up to 200 Kb (Regan et al. 2015), which has already been applied to fusion transcript detection (Hoff et al. 2016) or to phase deletions into haplotypes (Boettger et al. 2016). Here, we have developed new ddPCR assays to genotype quickly and reliably human polymorphic inversions flanked by large IRs, and thanks to the genotype data we demonstrate that most of these inversions are recurrent and that inversion alleles are associated to gene expression changes.

RESULTS

Inversion genotyping

To test the ddPCR application for inversion genotyping, we analyzed a representative sample of 20 inversions mediated by IRs of different sizes from the InvFEST database (Martínez-Fundichely et al. 2014). Due to the high error rate of inversion detection in repetitive sequences (Vicente-Salvador et al. 2017; Chaisson et al. 2019; Giner-Delgado et al. 2019), we selected well-supported inversions, excluding predictions on very complex regions full of segmental duplications (SDs) or with genome assembly gaps. In particular, 14 were initially predicted from fosmid PEM data in nine individuals (Kidd et al. 2008), although five had additional supporting evidence (Supplemental Table S1). The rest were validated inversions for which there was no available experimental genotyping method to study them in multiple individuals (the well-known 17q21 inversion or HsInv0573, HsInv0390, HsInv0290 and HsInv0786) (Stefansson et al. 2005; Beck et al. 2015; Feuk et al. 2005; Martin et al. 2004) and two control inversions already genotyped by

Chapter 3. Results

iPCR-based assays (HsInv0241 and HsInv0389) (Aguado et al. 2014; Giner-Delgado et al. 2019) (Supplemental Table S1). Minimal inversion sizes range from 3.1 to 741.7 kb and IRs at breakpoints between 11.3 and 134 kb (Fig. 1A, Table 1), large enough to hinder inversion detection by conventional genome sequencing methods (see below).

ddPCR technology allows us to quantify how close two independent sequences are within a DNA molecule based on their simultaneous amplification in a higher number of droplet partitions than expected by chance (Regan et al. 2015). Thus, for each inversion we designed three amplicons in the unique sequence outside the IRs (A or D) and at both ends of the inverted segment (B and C) (Fig. 1B). Then, we determined the percentage of DNA molecules supporting orientation 1 (*O1*) and orientation 2 (*O2*) from the linkage values between the amplicons outside and inside the inverted region, which correspond to the percentage of molecules containing both products (either A and B or C and D in the case of *O1* linkage, and A and C or B and D for *O2* linkage). Inversions were finally genotyped in each sample by the ratio between *O1* linkage and the total linkage for both orientations (*O1+O2* linkage), with ideal values for the three expected genotypes of 1 (*O1/O1*), 0.5 (*O1/O2*) and 0 (*O2/O2*). A different strategy was used for inversion 17q21, where all breakpoints contain variable repeat blocks >200 kb (Boettger et al. 2012; Steinberg et al. 2012), except for the AB breakpoint junction with a 132-kb repeat. Therefore, we measured only the *O1* linkage (AB) and compared it to a reference linkage (Ref) between two products located at the same relative distance in both orientations, resulting in *O1/Ref* linkage ratios of 1 (*O1/O1*), 0.5 (*O1/O2*) and 0 (*O2/O2*). In addition, to simplify the process and reduce costs, we developed triplex assays where the three amplicons from a given inversion are amplified simultaneously using different amounts of two probes labeled with FAM, which allows us to genotype an individual in a single reaction (Fig. 1B; Supplemental Table S2). These assays were tested and optimized in a small sample of 7-15 individuals, including those in which the inversions were predicted (see Methods).

Table 1. Inversion features and frequencies. Inversion sizes correspond to the distance between both inverted repeats (IR). The O1 allele corresponds to the orientation in the hg18 genome assembly, except for HsInv1111 in which it is the orientation in hg38 that represents the first complete sequence of the region. Inversion frequencies for the three analyzed populations together (global) and each population independently are those of the indicated global minor allele AFR, Africans; EAS, East-Asians; EUR, Europeans

Inversion	Previous status	Genomic position (hg38)	Inversion size (bp)	IR size (bp)	Minor allele	MAF			
						Global	YRI	EAS	CEU
HsInv0233	Predicted	chr1:108221621-108472664	73,074	89032 / 88938	-	-	-	-	-
HsInv0228	Predicted	chr1:149817486-149876379	6,775	26033 / 26086	O1	0.436	0.469	0.219	0.633
HsInv0012	Predicted	chr1:248459787-248588406	9,529	59480 / 59611	-	-	-	-	-
HsInv0241	Validated	chr2:240672507-240701802	3,177	13283 / 12836	O2	0.420	0.625	0.403	0.226
HsInv1057	Predicted	chr6:167165783-167392651	160,959	31131 / 34779	O2	0.044	0.129	0.000	0.000
HsInv0290	Validated	chr7:5893638-6832887	741,736	99169 / 98345	O2	0.089	0.125	0.078	0.065
HsInv1111	Predicted	chr8:2232511-2486876	35,033	110841 / 108492	O2	0.500	0.688	0.403	0.403
HsInv0370 ¹	Predicted	chr16:20411068-20607534	20,967	90311 / 85189	O2	0.142	0.297	0.047	0.081
HsInv0786	Validated	chr16:28337952-28777130	171,288	133941 / 133950	O2	0.253	0.233	0.250	0.274
HsInv0573	Validated	chr17:45495836-46707124	589,240	131964 / 490085	O2	0.063	0.000	0.000	0.204
HsInv0382	Predicted	chr20:25752457-26103777	165,605	92696 / 93020	O2	0.247	0.204	0.083	0.450
HsInv1126	Predicted	chrX:51667325-51725664	7,400	25639 / 25301	O2	0.356	0.333	0.419	0.318
HsInv0605	Predicted	chrX:52037053-52213786	89,288	42265 / 45181	O2	0.348	0.438	0.255	0.348
HsInv0395	Predicted	chrX:55453530-55519672	12,922	26612 / 26609	O2	0.355	0.521	0.319	0.217
HsInv0390	Validated	chrX:103918062-104113598	59,994	71373 / 64170	O2	0.489	0.292	0.638	0.543
HsInv0822	Predicted	chrX:149652865-149750398	41,016	28263 / 28255	O2	0.421	0.426	0.532	0.304
HsInv0389	Validated	chrX:154335936-154396222	37,621	11311 / 11355	O2	0.489	1.000	0.255	0.196
HsInv0830	Predicted	chrX:154555685-154648782	21,769	35643 / 35686	O2	0.309	0.646	0.128	0.136
HsInv0608	Predicted	chrX:155336693-155504550	67,255	50035 / 50568	O2	0.234	0.396	0.085	0.217

Chapter 3. Results

HsInv0416 ²	Predicted	chrY:21005927-21063547	32,293	15776 / 15773	<i>O1</i>	0.347	0.000	0.353	0.688
-------------------------------	-----------	------------------------	--------	---------------	-----------	-------	-------	-------	-------

¹ There is a third allele in which the inverted region is deleted with 0.031 frequency in YRI and 0.011 globally

² Due to an error in the human reference genome where IR2 is assembled in the opposite orientation (Vicente-Salvador et al 2017), HsInv0416 IR and inversion sizes are those obtained by correcting the genome with the sequence of fosmid clone ABC8-724240H6 (AC226836.2) that contains the *O2* second breakpoint

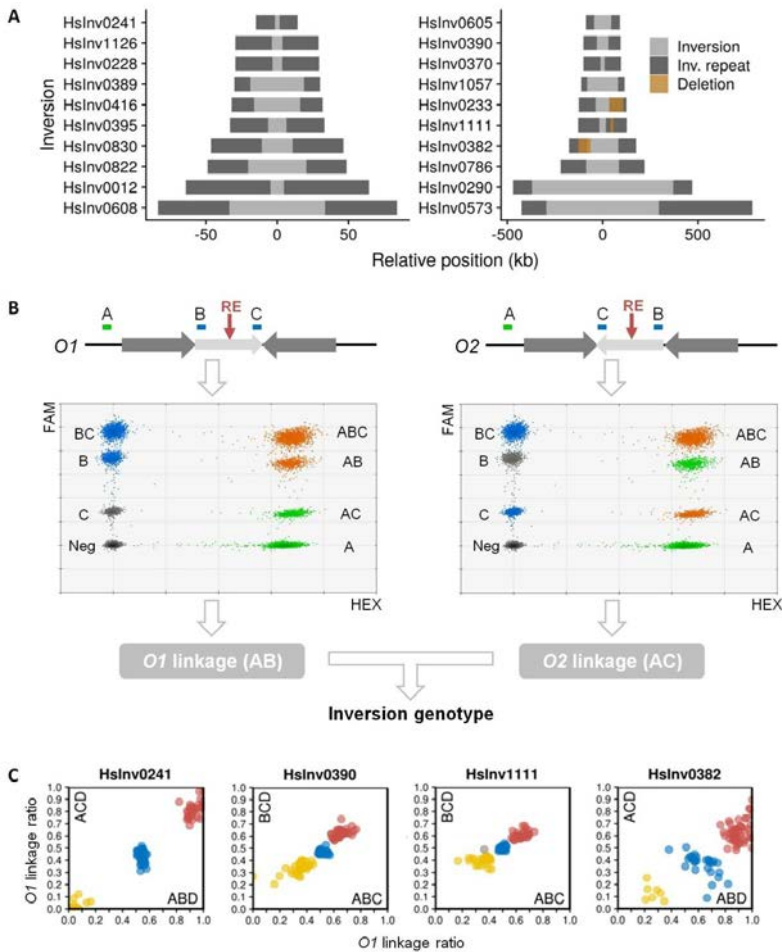


Figure 1. Inversion characterization by ddPCR genotyping. (A) Main features of the studied polymorphic inversions, with the inverted region shown in grey and the inverted repeats as dark boxes. Brown boxes denote deletions affecting inversion breakpoints. (B) Strategy used to genotype inversions. Dark grey arrows represent IRs at inversion breakpoints flanking the inverted region (light grey arrow), with ddPCR amplicons A, B and C shown on top. *O1* linkage (AB, left) and *O2* linkage (AC, right) were calculated separately from the triplex ddPCR results, ignoring respectively amplicons C or B. Colors of droplet clusters indicate if they are used to estimate the molecules containing only amplicon A (green), only the B or C internal amplicon (blue), two amplicons at the same time, either A and B or A and C (orange), or they are considered as negative (grey). (C) Plots of *O1* linkage ratios for both breakpoints of four inversions in 95 analyzed individuals. Colors indicate genotype groups (*O1/O1*, red; *O1/O2*, blue; *O2/O2*, yellow) and the grey dot is a sample with altered amplicon ratios in HsInv1111 ABC breakpoint. HsInv0241 clusters show the expected linkage ratios since DNA was digested to separate both breakpoints, whereas for the

Chapter 3. Results

other inversions DNA digestion was not possible and homozygote and heterozygote clusters are closer. The effect of a large deletion within breakpoint AB in HsInv0382 can be seen as higher *O1* linkage ratio values in the horizontal axis.

The main problems in distinguishing ddPCR genotypes were related to inversion and IR size, variation in IR length and altered amplicon ratios. First, in small inversions linkage can be detected between A and C in *O1* chromosomes, or A and B in *O2* chromosomes, resulting in linkage ratios for homozygotes that depart from the 0/1 expected values. To solve this, whenever possible we separated both breakpoints by DNA digestion with a restriction enzyme that cuts the inverted sequence but not the IRs (Fig. 1B, Supplemental Table S2). In those inversions without a suitable restriction enzyme, linkage ratios for homozygotes were closer to those of heterozygotes, but we could still distinguish the three genotypes reliably (Fig. 1C). The exception was HsInv0012, in which the inverted region is so small compared to the IRs (Fig. 1A, Table 1) that amplicons B and C were equally linked to D in all samples and it was not analyzed further. Second, increasing distance between the ddPCR amplicons due to the IRs at the breakpoints reduces the linkage detected because there are less molecules long enough to contain both of them. In the inversions with the largest IRs, true linkage values become closer to those indicating no linkage, and the intrinsic variation of the technique can affect the final genotype. Thus, DNA length and quality limits the ability to resolve these inversion genotypes. Third, IR size variation caused by large polymorphic indels can change the distance between amplicons in different chromosomes of the population. These indels can also be moved to a different breakpoint by the inversion, altering more than one linkage value. Three of the studied inversions had polymorphic deletions within one of the IRs, but they only affected significantly the linkage in HsInv0382 (62-kb deletion within a 92.7-kb IR) and HsInv0233 (74.5-kb deletion within an 89-kb IR) (Supplemental Table S2). In HsInv0382 *O1/O2* individuals carrying the AB deletion, *O1* linkage is higher than *O2*, but three genotype groups can still be clearly separated (Fig. 1C). In HsInv0233, the large deletion in breakpoint CD combined with the relatively small inversion size leaves amplicons B and D at a similar distance in deleted *O1* chromosomes and in the inverted orientation, and it was not possible to assign genotypes with confidence (Supplemental Fig. S1). Finally, although all amplicons were carefully designed in unique single-copy sequences, we detected a few individuals that carry deletions or duplications that affect consistently one or more amplicons, which in the case of copy number increases makes it very difficult to interpret linkage values (Supplemental Table S3).

Next, we genotyped 19 inversions (excluding HsInv0012) in 95 individuals included in different genomic projects (The 1000 Genomes Project Consortium 2015; Lappalainen et al. 2013) with African (32 YRI), East Asian (EAS) (16 CHB and 16 JPT), and European (31 CEU) ancestry (Supplemental Table S4). All experiments were performed in duplicate, except for HsInv0389 that had been previously genotyped (Giner-Delgado et al. 2019), and five inversions for which the two breakpoints were analyzed independently (HsInv0241, HsInv0233, HsInv0382, HsInv0390 and HsInv1111). In HsInv0233, we identified different

samples showing distinctive *O1/O1* and *O2/O2* genotypes with no signs of the large deletion and we consider this inversion validated, although we cannot genotype reliably all individuals. In general, genotypes for the rest of inversions were very clear and consistent between replicates. However, samples with problems (low linkage values in inversions with the largest IRs, intermediate linkage ratios that cannot be easily interpreted into genotypes, or altered amplicon ratios) were repeated several times. Final genotypes were called using a statistical clustering method that groups the *O1* linkage ratios of all the analyzed samples taking into account the information of every replicate (see Methods for details), which was especially important for inversions analyzed using undigested DNA (Fig. 1C). In total, we obtained 1635 genotypes for the 18 inversions with complete data (98.3%), and only 29 genotypes were not determined because of low linkage (13), inconclusive clustering results (12) or altered amplicon ratios due to CNVs (4) (Supplemental Table S3).

To assess the accuracy of the ddPCR results, we compared them with the published genotypes for all the samples for HsInv0241 and HsInv0389 (Giner-Delgado et al. 2019), plus a few from other inversions obtained by FISH or Southern blot (Supplemental Table S5). In addition, we genotyped by haplotype-fusion PCR (HF-PCR) (Turner et al. 2006) inversion HsInv0395 (90 CEU individuals) and HsInv0605 (8 individuals) with medium-sized IRs (26.6-45.2 kb) (Supplemental Fig. S2). In HF-PCR a fusion product is created from amplicons at both sides of a breakpoint, but it is difficult to set up and not very robust. Out of the 244 compared genotypes from 6 different inversions, there were only four discordant genotypes (1.6%) (Supplemental Table S5), which correspond to African individuals for HsInv0241 that appear to be *O2/O2* by iPCR, but are heterozygotes by ddPCR assays of both breakpoints. To check how common this discrepancy was, we genotyped the whole 30 YRI trios from the HapMap project (including 68 extra samples) (Supplemental Table S4) and only one child was discordant as well. Also, in three out of four candidates to have the same problem identified by SNP analysis from the LWK population, ddPCR genotypes were indeed *O1/O2* but *O2/O2* in iPCR (Supplemental Table S4). This suggests that there are some unknown variants affecting HsInv0241 *O1* chromosome detection by iPCR (Aguado et al. 2014; Giner-Delgado et al. 2019). Similarly, the 88 ddPCR 17q21 inversion genotypes agree with those inferred from the commonly-used tag SNPs (Steinberg et al. 2012). In contrast, in HsInv0786, 9 of the 81 genotypes (11.1%) computationally inferred from SNP genotype data (González et al. 2014) differ from ddPCR results (including eight heterozygotes misclassified as *O1/O1* and one *O1/O1* homozygote as heterozygote) (Supplemental Table S5), indicating errors in the imputation.

When ddPCR data were compared to recent inversion predictions in multiple human genomes with different techniques, quite contrasting results were obtained (Supplemental Table S1). Despite most of the inversions being relatively common, only HsInv0241, with some of the smallest IRs, was detected in one of the short-read analyses (Sudmant et al. 2015; Hehir-Kwa et al. 2016; Collins et al. 2019) and two more (HsInv0370 and HsInv1126) using long reads (Audano et al. 2019). On the other hand, two studies relying in a multiplatform strategy based mainly on Strand-seq (Chaisson et al. 2019) or Bionano optical maps (Levy-

Chapter 3. Results

Sakin et al. 2019) identified 14 inversions each, with 9 detected by both (Supplemental Table S1). Although in most cases genotypes were not provided, based on the presence or not of the inverted allele of the identified inversions, short reads (Sudmant et al. 2015) and long reads (Audano et al. 2019) showed the lowest performance (with respectively 66% and 83% error rate). The Bionano analysis missed the inverted alleles in 23% of individuals, and the multiplatform strategy genotyped correctly 10 of 11 comparable inversions in the only individual in common (9% error rate). This emphasizes the amount of inversion information still missing from the studied genomes and the need of specific methods for the analysis of IR-mediated inversions.

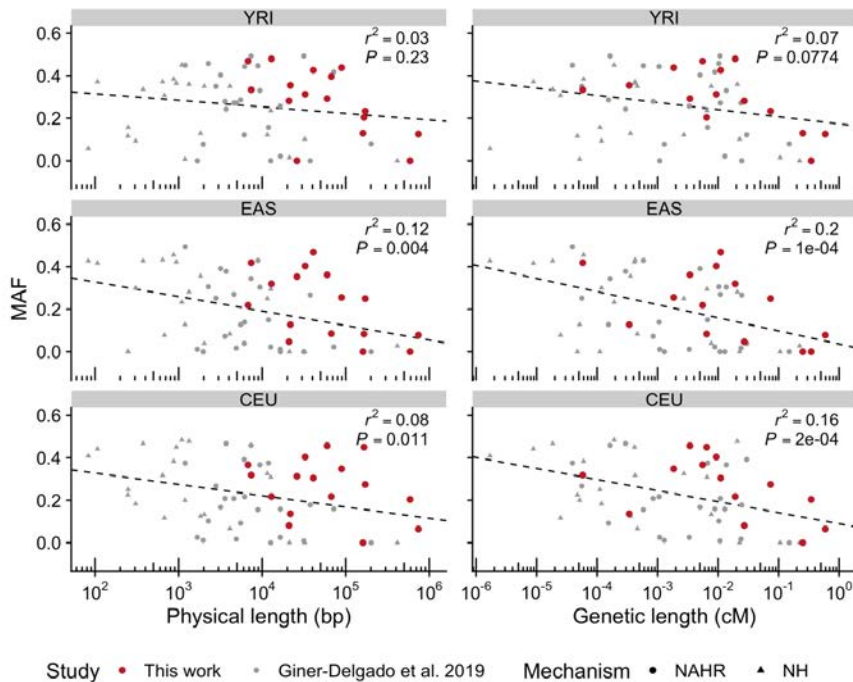


Figure. 2 Correlation between inversion size and frequency. Negative correlations between the logarithm of the minimal physical or genetic length and minor allele frequency (MAF) of the inversion in three populations with African (YRI), East-Asian (EAS) and European (CEU) ancestry are observed for the 16 ddPCR-analyzed inversions (in red) and 45 polymorphic inversions described in Giner-Delgado et al. (2019) (in grey), including inversions generated by non-homologous mechanisms (NH) or non-allelic homologous recombination (NAHR).

With regard to the distribution and frequency of these inversions (Table 1), all are in Hardy-Weinberg equilibrium in the analyzed populations and most of them are widespread in the three continents with global

minor allele frequencies (MAF) >0.1 . The only exceptions are HsInv1057, found exclusively in Africans, the 17q21 inversion, detected only in Europeans where the frequency of the H2 allele is known to be higher (Stefansson et al. 2005; Steinberg et al. 2012; Alves et al. 2015), and HsInv0290 with low frequency in the three populations (0.065-0.125). Several inversions also have variable frequencies between continents, with chr. X inversions HsInv0830 and HsInv0389, and HsInv0416 in chr. Y showing the highest F_{ST} values, although only those of HsInv0389 are within the top 5% of the expected distribution (Supplemental Table S6) (Giner-Delgado et al. 2019). As observed before (Giner-Delgado et al. 2019), we found a negative correlation between inversion size and MAFs in each population either for the inversions analyzed here only or all inversions together (Fig. 2). These correlations tend to be higher with the inversion genetic length than the physical length (Fig. 2), which suggests that in large inversions negative selection against the generation of unbalanced gametes by recombination in heterozygotes (Hoffmann and Rieseberg 2008; Kirkpatrick 2010) might be an important factor in determining their frequency.

Nucleotide variation analysis

We used the accurate genotypes of 15 inversions to check the linkage disequilibrium (LD) with 1000 Genome Project (1000GP) nucleotide variants (92 individuals in common), excluding HsInv0416 in chr. Y without 1000GP data and HsInv0241 and HsInv0389 already analyzed in a larger number of individuals (Giner-Delgado et al. 2019). Only population-specific inversions 17q21 and HsInv1057 have tag SNPs in complete LD ($r^2 = 1$), while for the rest of inversions the maximum LD (r^2) values range between 0.21 and 0.79 and just five of them have tag SNPs ($r^2 > 0.8$) in at least one of the populations (CEU, EAS or YRI) (Fig. 3A, Supplemental Table S7). We also classified the SNPs within the inverted region as shared between orientations, private to one of them or fixed (i.e. in complete LD) (Fig. 3A, Supplemental Table S7). The 17q21 and HsInv1057 inversions show no reliable shared variants, consistent with the inhibition of recombination across the entire inversion length. In contrast, shared variants represent an important fraction (6-48%) of the variation within the entire length of the remaining inversions (Supplemental Fig. S3), a pattern suggestive of several inversion events on different haplotypes.

Next, we estimated the independent inversion events by phasing inversion genotypes into 1000GP haplotypes of the inverted region (Fig. 4). While in the two inversions with perfect tag SNPs all $O2$ haplotypes cluster together, in the other inversions several different clusters of similar haplotypes containing both orientations can be identified, which are typical of recurrent inversion and re-inversion events (Giner-Delgado et al. 2019). We identified between 2 and 5 inversion events in the 14 new recurrent inversions analyzed here and a total of 33 additional inversion events, with 27 shared by more than one individual (Supplemental Table S8). This suggests that they are real evolutionary events rather than phasing errors or new variants generated in the lymphoblastoid cell lines (LCLs) used as DNA source. However, recurrence quantification might be complicated by inversion and SNP phasing errors and recombination between haplotypes. The exception is HsInv0416, in which the $O1$ and $O2$ distribution in the known phylogeny of chr. Y haplogroups (Poznik et al. 2016) clearly supports four independent inversion events (Supplemental

Chapter 3. Results

Table S8) and results in an inversion mutation rate of 1.29×10^{-4} inversions per generation, very similar to that previously described for another Chr. Y inversion (0.53×10^{-4} inversions per generation) (Giner-Delgado et al. 2019).

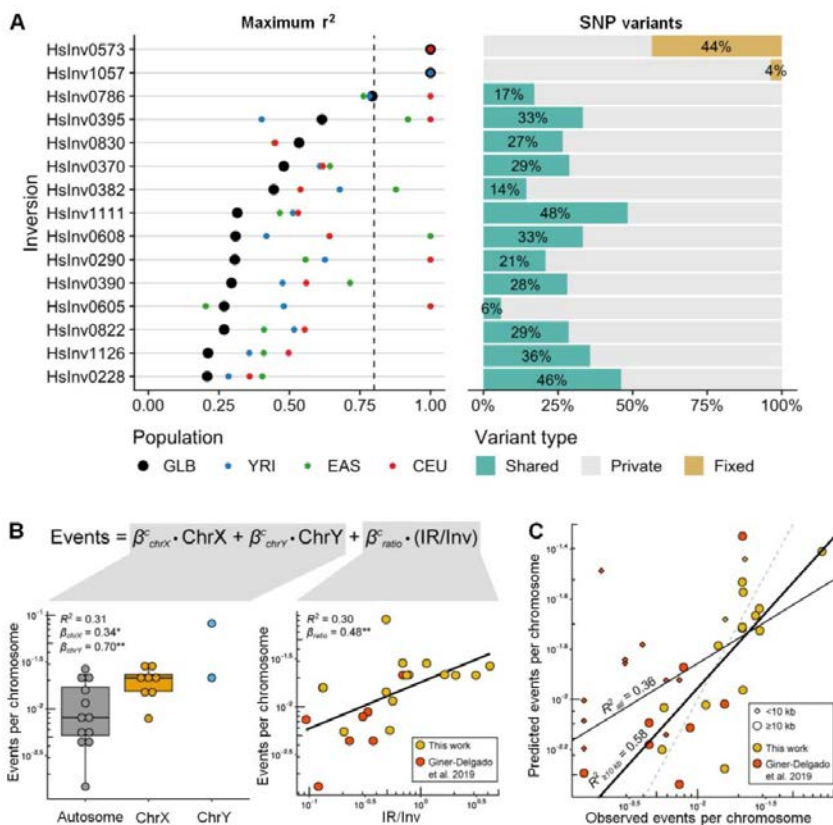


Figure 3. Nucleotide variation analysis of newly genotyped autosomal and chr. X inversions. (A) Left, maximum LD (r^2) between 1000GP variants located 1 Mb at each side and inversion alleles in all individuals together (black dots) and different populations (colored dots). Right, proportions of SNPs classified as fixed (yellow), shared between orientations (green) or private of one orientation (grey) within each inversion. (B) Correlation between the logarithm-transformed estimated number of independent inversion events per chromosome and chromosome type (left) and the IR/inverted region (Inv) size ratio (right) calculated with 22 inversions >10 kb using data from this work (yellow dots) and Giner-Delgado et al. (2019) (orange dots). (C) Adjustment of the observed recurrence events with the expected number calculated by applying the developed model. Number of events is underestimated in small inversions (<10 kb; diamonds), which results in a lower R^2 value for all inversions (all, thin black line) than for those greater than 10 kb (≥ 10 kb, thick black line). Dashed line represents the 1:1 equivalence.

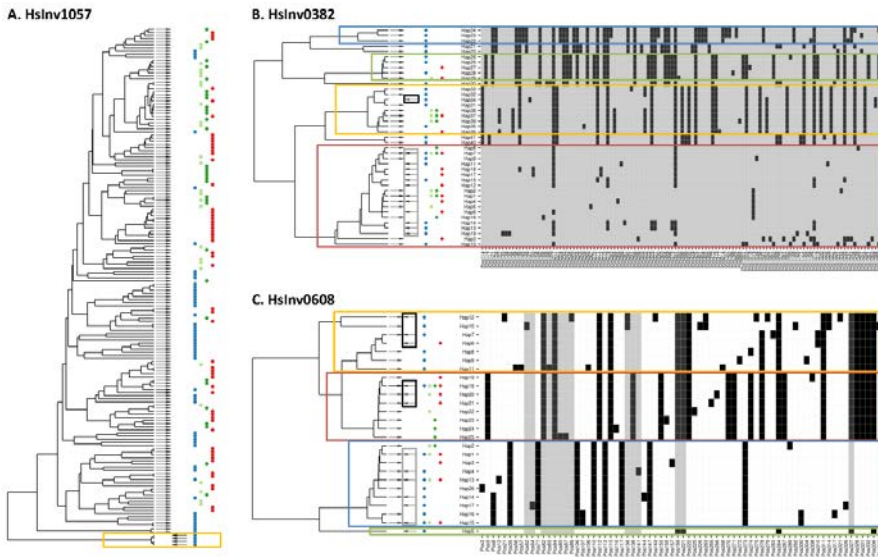


Figure 4. Estimation of the number of inversion events from inverted region haplotypes. Each inversion was analyzed using integrated haplotype plots (iHPlots) (Giner-Delgado et al. 2019), with the tree indicating the relationship between the different haplotypes, the rightwards (*O1*) and leftwards (*O2*) arrows the orientations observed for each haplotype, and dots the populations where each haplotype has been found (blue for YRI, light and dark green for CHB and JPT, and red for CEU). Inverted region haplotypes are represented by the variable positions (see Methods for variant selection) with different colors indicating the two alleles (white, ancestral; grey, hg19 reference; black, derived/alternative), and colored boxes showing the main differentiated haplotype clusters. For the unique inversion (A) only the first part of the plot is shown with a yellow box highlighting the single group of tightly-clustered inverted haplotypes. In the recurrent inversions (B, C), the orientation of the haplotypes of what we considered the original inversion event are included in a grey box and any additional inversion event in a black box (one in HsInv0382 and two in HsInv0608).

We also investigated the effect of different parameters on the observed recurrence levels by combining all the inversions mediated by NAHR analyzed here and in Giner-Delgado et al. (2019). Focusing only on the 22 inversions >10 kb, which have more resolution to detect recurrence events, the only significant variables were the ratio between IR and inversion sizes (IR/Inv ratio) and chromosome type (autosome, chr. X or chr. Y), with IR/Inv ratio and sex chromosomes being positively correlated with the number of inversion events (Fig. 3B). The resulting model fits very well the real data ($R^2 = 0.58$), although there is a clear underestimation of recurrence events in smaller inversions (<10 kb) because the lower number of SNPs reduces the ability to distinguish haplotype clusters (Fig. 3C).

Functional effects of inversions

The functional consequences of the genotyped inversions were first investigated through their association with available LCL gene expression data (Lappalainen et al. 2013) of 59 CEU and YRI samples in common. We identified two inversions associated to gene-expression changes in this small sample (Fig. 5A, Supplemental Table S9). In particular, inversion 17q21 is the lead variant ($r^2 \geq 0.8$ with top eQTLs) for an antisense RNA and three pseudogenes at the gene level, as well as for specific transcripts from protein-coding genes *KANSL1* and *LRR37A2* plus other pseudogenes and non-coding RNAs (Supplemental Table S9), which indicates a pervasive effect of the two inversion haplotypes on gene expression. To increase the power to detect associations, we extended the analysis to a larger sample of 387 Europeans for five inversions whose genotypes could be inferred through perfect tag SNPs ($r^2 = 1$) in the CEU population (Fig. 3A). We found that all of them were significantly associated to expression changes in 42 genes and 103 transcripts (Fig. 5A, Supplemental Table S9), including multiple additional genes for inversion 17q21. In addition, HsInv0786 appeared as potential lead eQTL for gene *NFATC2IP*, and *APOBR*, *IL27* and *EIF3C* transcripts.

Since more than 60% of genes inside or around our inversions (± 1 Mb) are not expressed in LCLs, we explored their effects in other tissues using eQTL information from the GTEx Project (GTEx Consortium 2017). We imputed inversion association P values by estimating LD patterns between inversion alleles and SNPs identified as eQTLs in GTEx (see Methods). We found 73 genes linked to 6 inversions in different tissues, although the majority of associations involve inversions 17q21 and HsInv0786, whose effects are easier to infer due to their high LD with neighboring eQTLs (including 37 genes for which the inversions were potential lead variants) (Fig. 5A-B, Supplemental Table S10, Supplemental Fig. S4). However, the low LD patterns with SNPs due to high recurrence levels of most analyzed inversions prevents inferring their contribution to gene expression variation.

We also checked whether inversions account for phenotypic variation by using the GWAS Catalog information (MacArthur et al. 2017). We found a 2.8-fold significant enrichment ($P = 0.026$) of GWAS Catalog signals within inversion regions compared to other genomic regions of similar size, with 17p21 ($P = 0.028$), HsInv0786 ($P = 0.009$) and HsInv0290 ($P = 0.157$) apparently driving this enrichment (Fig. 5C). When we mapped GWAS phenotypes to genes located within 150 kb of the analyzed inversions, several of them showed an enrichment of GWAS-reported genes for certain traits (Supplemental Table S11), such as brain-related traits (Parkinson's disease, neuroticism, intelligence and cognition) for 17p21 and immunological disorders (Crohn's disease, ulcerative colitis and type 1 diabetes) and body mass characteristics for HsInv0786, indicating their potential roles on these diseases. For seven inversions in high LD ($r^2 \geq 0.8$) with SNPs in at least one population, we also looked for GWAS hits associated to those SNPs in the corresponding population or continent (Supplemental Table S12). The 17p21 inversion has already been linked to many traits (Puig et al. 2015a) and a total of 64 potential associations with $P < 10^{-6}$ from 35

studies were found in this analysis (Fig. 5D), including lung function, neurological traits or disorders, ovarian cancer, blood profiles and diabetes, which illustrates the pleiotropic consequences of this inversion. HsInv0786, which was associated to an asthma and obesity phenotype (González et al. 2014), can also be associated to several pediatric autoimmune diseases (Li et al. 2015) and the presence of type 1 diabetes autoantibodies (Plagnol et al. 2011) among other traits, suggesting a role in the immune system that may be related to the observed expression changes in genes like *IL27* and *NFATC2IP* (Supplemental Tables S9-S10).

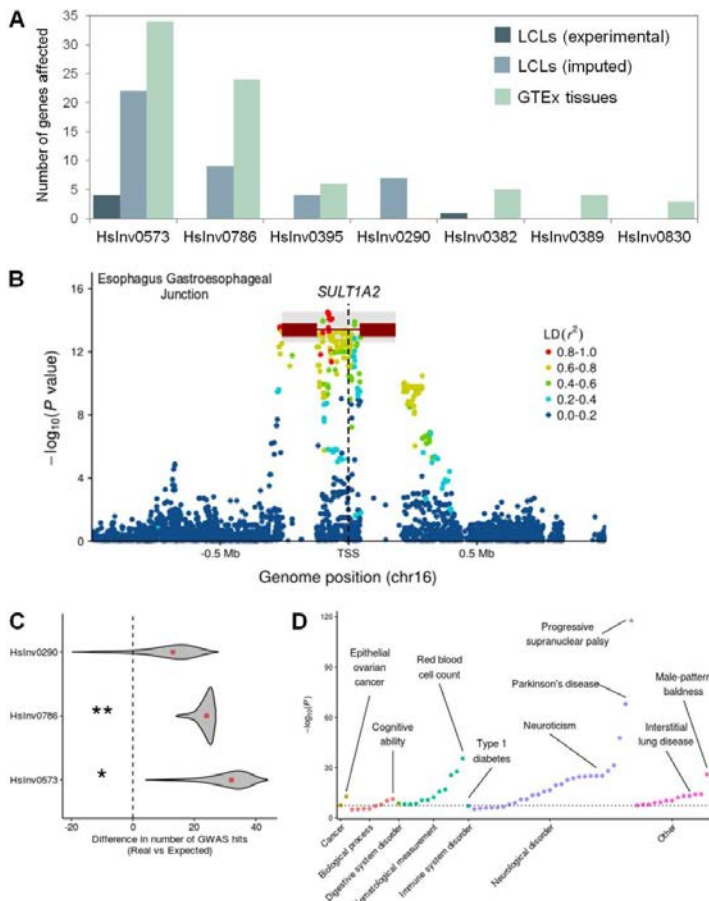


Figure 5. Inversion effects on gene expression and phenotypic traits. (A) Number of significant associations at the gene level in each of the differential expression analyses. (B) Manhattan plot for cis-eQTL associations reported in GTEx project of gene *SULT1A2*, showing inversion HsInv0786 (dark red line with rectangles representing the IRs) as potential lead variant. HsInv0786 eQTL P values and LD with neighboring variants were calculated by permuting samples with the same ancestry proportions as in GTEx samples (see Methods) and the P value imputation range is shown in grey. (C) Enrichment of GWAS

Chapter 3. Results

Catalog signals within individual inversions measured by the deviation in observed minus expected value. Density distribution represents the 95% one-sided confidence interval with the median indicated by a red dot (one-tailed permutation test: **, $P < 0.01$; *, $P < 0.05$). (D) Phenome-wide association study for inversion 17p21. Significant reported traits ($P < 10^{-5}$) were grouped by categories according to terms defined by GWAS Catalog ontologies. Dotted line indicates genome-wide significance threshold ($P = 5 \times 10^{-8}$).

DISCUSSION

The study of polymorphic inversions, and especially those flanked by large IRs, has lagged behind because of the lack of high-throughput techniques able to detect them. Here, we have taken advantage of ddPCR ability to measure linkage between sequences at relatively long distances to genotype inversions with IRs up to ~150 kb in a single reaction in an efficient, reliable and reproducible way. Using this method we have validated 13 inversion predictions and generated new experimentally-resolved genotypes for 16 inversions, most of which are analyzed here in detail for the first time. In comparison with the set of 24 inversions flanked by IRs recently analyzed (Giner-Delgado et al. 2019), on average our 16 inversions are longer (138.9 kb versus 20.7 kb) and have much bigger IRs (74.2 kb versus 6.9 kb), filling a void in the study of human variation. Almost all these inversions are missed by next-generation sequencing with short or long reads (Supplemental Table S1). In addition, although recent studies using other methods like Strand-seq or Bionano (Chaisson et al. 2019; Levy-Sakin et al. 2019) have been able to identify 95% of them (Supplemental Table S1), an acceptable genotyping accuracy (>90%) is only obtained by combining different techniques, which complicates the analysis of a large number of individuals. Therefore, the novel ddPCR application provides a very valuable and powerful resource for the targeted characterization of inversions and complex SVs, including accurate information on other associated variants in the region (like indels or CNVs) (Supplemental Table S3).

In fact, half of inversions mediated by inverted SDs in InvFEST (Martínez-Fundichely et al. 2014) have IRs between 25 and 150 kb long that could be analyzed using the ddPCR technology. The main limitations are restricted to extremely long IRs (>100 kb), small inversions (<5-10 kb) with relatively long IRs where breakpoints cannot be separated by restriction-enzyme digestion, and CNVs altering the distance between amplicons. We have overcome these problems by using good-quality high-molecular-weight DNA and a clustering tool that allows us to distinguish the three genotype groups independently of the magnitude of the linkage ratio differences. In the near future, the possibility to work with even longer DNA molecules and to interrogate several inversions simultaneously by multiplexing amplicons labeled with different fluorochromes will expand the range of studied inversions and reduce the costs, making easier to undertake more ambitious ddPCR genotyping projects in higher numbers of samples.

Consistent with previous results (Giner-Delgado et al. 2019), we have shown that the vast majority of inversions mediated by IRs are generated multiple times on different haplotypes by NAHR and cannot be

easily imputed from SNP data. For example, in HsInv0786, half of individuals incorrectly genotyped using SNP information (11.1%) (González et al. 2014), carry a recurrent chromosome with unexpected orientation according to its haplotype. Also, inversion recurrence or some other mechanism able to exchange variants between the two IRs had been already suggested for HsInv0390 (Beck et al. 2015), HsInv0830 (Aradhya et al. 2001) and HsInv0290 (Hayward et al. 2007) regions. The only exceptions are the two inversions with more restricted geographical distributions: HsInv1057 (found exclusively in Africans) and 17q21 (with a higher frequency in Europeans compared to all other populations) (Stefansson et al. 2005; Steinberg et al. 2012; Alves et al. 2015). This indicates that although the 17q21 and a few other unique inversions can be indirectly imputed by tag SNPs, the only way to genotype accurately recurrent inversions is to interrogate experimentally the sequences at both sides of the breakpoint with techniques like the one developed here.

Actually, we have found similar levels of recurrence for the previous smaller inversions (Giner-Delgado et al. 2019) and the longer inversions studied here, showing that it is a general characteristic of inversions flanked by IRs. Moreover, the higher resolution provided by the longer inversions has allowed us to estimate more precisely the number of recurrence events and determine the main factors affecting recurrence. Specifically, the chromosome type and IR/inversion size ratio together explain a very significant part (up to 58%) of the variance in recurrence levels between inversions. This fits well the expectations, since repeat length and distance has been found to affect the generation of recurrent pathological rearrangements (Liu et al. 2011), suggesting that the closer and longer the IRs are, the more likely they are to pair and recombine. In addition, the inversion causing hemophilia A is known to occur much more frequently in the male germ line, probably due to the increased NAHR within the single chr. X copy (Antonarakis et al. 1995). According to the estimated values for two chr. Y inversions, the model results in predicted NAHR mutation rates for the analyzed autosomal inversions of $0.9-4.4 \times 10^{-5}$ inversions/generation and of $1.9-7.4 \times 10^{-5}$ inversions/generation for chr. X inversions, which illustrate the importance of this phenomenon and its potential impact in the genome. However, other factors, like DNA 3D conformation or recombination motifs, might also affect recurrence.

On the other hand, the newly-generated genotypes have allowed us to carry out the first complete analysis of the potential functional consequences of the majority of these inversions. By taking advantage of different available datasets and analysis strategies, we were able to associate eight inversions with gene-expression changes across different tissues. In particular, the inversions with the largest effects were 17q21 and HsInv0786, which were the top eQTLs for many genes. In this case, our analyses confirmed most expression changes already reported for 17q21 in blood, cerebellum and cortex (de Jong et al. 2012) or for HsInv0786 in blood and LCLs (González et al. 2014) and estimated more precisely the real contribution of the inversions. Moreover, we identified expression differences in additional genes and tissues never examined before (Supplemental Table S10, Supplemental Fig. S4). We also found that the two same inversions are also enriched in GWAS signals, including certain phenotypic categories such as neurological traits or immunological disorders, and they are in LD with multiple variants associated to different phenotypes.

Chapter 3. Results

Thus, these results show that inversions could have important effects on both gene expression and clinically relevant disorders and uncover other interesting candidates for further study.

However, the inversion functional analysis has two main limitations. First, the small number of genotyped individuals with expression data (59) results in low statistical power and only large differences can be detected, especially for low-frequency inversions. This is exemplified by the additional expression associations identified for 17q21 and four other inversions when genotypes were imputed in additional European individuals (387). Second, for many inversions the lack of LD with neighboring variants that can be used as a proxy does not allow us to infer reliably their association with gene expression in other tissues or with phenotypic traits from non-genotyped individuals. Therefore, we are probably missing a significant fraction of inversion effects, especially those of modest sizes, emphasizing the need to genotype these inversions in a larger set of individuals.

Another important effect of inversions is the generation of aberrant chromosomes by recombination crossovers within the inverted region in heterozygotes (Hoffmann and Rieseberg 2008; Kirkpatrick 2010). By extending the analysis to a different set of inversions, here we have reinforced the idea that there is a negative correlation between the frequency and genetic length of inversions related to their negative impact in fertility (Giner-Delgado et al. 2019). In that sense, it is noteworthy that some of the longer inversions, such as 17q21 (589 kb) and Hsinv0786 (171 kb), are the best examples of inversions with functional consequences at different levels. This suggests that some of their effects on gene expression could compensate the potential negative costs associated to the longer inversions, and it has already been proposed that the 17p21 inversion has been positively selected in European populations through increased fertility in carrier females (Stefansson et al. 2005).

Finally, polymorphic inversions could also predispose to other pathological SVs in the region, due either to recombination problems in heterozygotes or changes in the orientation of repeats (Puig et al. 2015a). Recently, a complete catalogue of inversions in nine individuals has shown that a high proportion of them overlap the critical regions of microdeletion and microduplication syndromes (Chaisson et al. 2019). However, these inversions tend to be mediated by large and complex repeats and are difficult to characterize with simple methods. In our case, 7 of the 8 chr. X inversions analyzed are located in regions where additional SVs disrupting genes and resulting in disease have been described (Supplemental Table S13). Some of these diseases involve recurrent mutations mediated by repeats within the polymorphic inversions, like the inversion causing Hemophilia A and HsInv0608 (Antonarakis et al. 1995) or the deletion causing incontinentia pigmenti and HsInv0830 (Aradhya et al. 2001). In HsInv0389 and HsInv0390, DNA polymerase stalling during replication of the repeats has been associated to different duplication-inverted triplication-duplication (DUP-TRP/INV-DUP) rearrangements that affect dose-sensitive genes (*MECP2* and *PLP1*, respectively) (Carvalho et al. 2011; Beck et al. 2015). Similarly, different deletions with one breakpoint mapping nearby or within one of the HsInv1126 IRs and affecting the inversion region and the

~310 kb separating it from HsInv0605 have been associated to X-linked intellectual disability (Grau et al. 2017). Interestingly, within the 6-Mb chr. Xq28 region near the telomere, a total of six genomic disorders caused by the previous and several other rearrangements overlapping or in close proximity to polymorphic inversions HsInv0822, HsInv0389, HsInv0830 and HsInv0608 have been described (Bondeson et al. 1995; Small et al. 1997; Clapham et al. 2012; Aradhya et al. 2001; Fusco et al. 2012; Li et al. 2015), making this region a possible hotspot for genome reorganization (Fig. 6). The new ddPCR application thus offers now the opportunity to study easily these inversions in parents of patients and determine their role in the generation of pathological variants, contributing to having a more complete picture of the impact of inversions in the human genome.

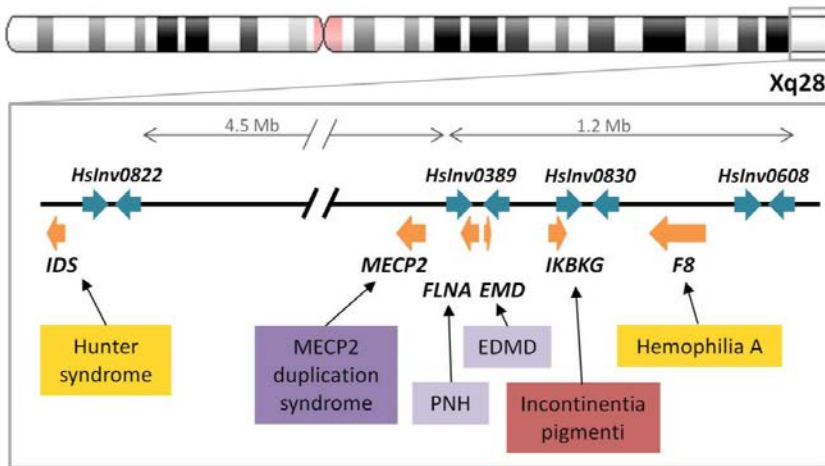


Figure 6. Polymorphic inversions in chromosome region Xq28 and diseases caused by structural rearrangements. Polymorphic inversion IRs are represented as blue arrows and genes as orange arrows. Colored boxes indicate diseases associated to different types of recurrent pathological rearrangements: inversions (yellow), deletion (red), and complex events resulting in deletion (light purple) or in a duplication-inverted triplication-duplication structure (DUP-TRP/INV-DUP) (dark purple). EDMD, Emyr-Dreifuss muscular dystrophy; PNH, periventricular nodular heterotopia.

METHODS

Human samples and DNA isolation

Genomic DNA of 95 unrelated human samples (Supplemental Table S4) was isolated from ~20-ml culture of an Epstein-Barr virus-transformed B-lymphoblastoid cell line of each individual (Coriell Cell Repository, Camden, NJ, USA) grown according to the recommended procedures. To obtain high-molecular-weight DNA, the cell pellet was resuspended in extraction buffer (10 mM Tris-HCl pH 8, 10 mM EDTA pH 8, 150 nM NaCl, 0.5% SDS) and incubated overnight with slow rotation at 37 °C with RNase cocktail (Invitrogen) and 100 µg/ml Proteinase K (Invitrogen). Four purification steps with one volume of TE-equilibrated phenol

Chapter 3. Results

pH 7.9 (twice), phenol:chloroform:isoamyl alcohol pH 7.9, and chloroform:isoamyl alcohol were performed by mixing by rotation until an emulsion was formed, and then centrifuging at 5,000 x g for 10-15 min to separate organic and aqueous phases. All steps that involved handling of the DNA sample were done pipetting gently with wide-bore tips. Finally, DNA was precipitated by adding 0.1 volumes of 3 M sodium acetate and 2 volumes of absolute ethanol, centrifuged, washed with 70% ethanol, and resuspended in 100-300 μ l of water. DNA was stored at 4 °C, which we observed that preserved DNA integrity over time better than freezing. Identity of all the isolated DNAs was confirmed by microsatellite analysis.

Droplet digital PCR (ddPCR)

Inversion genotyping by ddPCR assays was carried out by quantitative amplification of three different products simultaneously with six different primers and three fluorescent probes (Supplemental Table S14) in aqueous droplets within an oil phase (emulsion PCR). Since the QX200™ ddPCR system can detect only two colors, we labeled the probe in the amplicon common to both inversion alleles (A in an ABC experiment) with HEX and the other two (B and C) with FAM and used a lower concentration of one of the FAM probes to separate the clusters of droplets including the different amplicon combinations (Fig. 1B). All primers and probes were tested in duplex experiments before optimizing the triplex reactions. Final ddPCR reactions were prepared in a total volume of 22 μ l in a 96-well plate with 450 nM-2.5 μ M of each primer, 75-275 nM of each probe, 1x ddPCR Supermix for Probes (No dUTP) (Bio-Rad) and 50 ng of genomic DNA. DNA samples were always handled with wide-bore tips and the ddPCR mix was mixed by gently pipetting up and down 5-10 times to avoid breaking the DNA molecules. Droplet generation, thermal cycling and fluorescence reading were performed as explained before (McDermott et al. 2013) using the Bio-Rad QX200™ Droplet Generator, C1000™ or T100™ thermal cyclers and QX200™ Droplet Reader. QuantaSoft™ version 1.7.4 software (Bio-Rad) was used to analyze each sample twice, every time ignoring one of the FAM products and counting droplets with that product the same as negatives (Fig. 1B). The linkage between the two targeted amplicons (percentage of DNA molecules containing both amplicons) was obtained as the excess of double-positive droplets over what is statistically expected (Regan et al. 2015). From these values we could calculate the total linkage and the *O1* and *O2* linkage ratios that allow us to genotype inversions. For inversions with restriction enzyme targets inside the inverted region but not within the IRs at the breakpoints, DNA was digested prior to ddPCR quantification. This helps to: (1) Separate both breakpoints and avoid the detection of linkage between undesired products due to their proximity; and (2) Proper droplet formation, which can be hampered by extremely long DNA molecules. Digestion was performed in 10 μ l including 2.5 U of restriction enzyme, 1x of the corresponding buffer, and 250 ng of genomic DNA at 37 °C for 3 hours, and then 2 μ l (50 ng) were directly used as template in the ddPCR reaction. Since DNA molecule size reduces over time, ddPCR genotyping was performed by decreasing order of IR size, starting with the inversions with higher DNA quality requirements.

There were several reasons why a ddPCR result was not valid: (1) Total linkage >7.5% was required to be considered reliable, which was only an issue in those inversions with the largest IRs; (2) Droplet count

below 10,000 due to the presence of undigested high-molecular-weight DNA; (3) Intermediate linkage ratios between the expected values for homozygotes and heterozygotes (e.g. ~ 0.25 or ~ 0.75); and (4) Small deletions or duplications of one or more of the amplicons that result in ratios between them that were different than 1 in certain individuals (Supplemental Table S3). Except for the altered amplicon ratios, in most cases these problems were solved by repeating the ddPCR reactions using a fresh DNA dilution from stock or storing the diluted DNA at 4 °C for a few days.

Genotype clustering

Inversion genotypes were called by clustering all *O1* linkage ratios (Supplemental Fig. S5), except for those based on a total linkage $<7.5\%$, or $<15\%$ if only one measurement was available. Also, in HsInv0382, where a deletion increases the linkage in one of the breakpoints, samples genotyped only by one breakpoint were excluded. For each inversion, we calculated the euclidian distance between individuals (stats::dist R function) (R Core Team 2017) using two randomly-selected *O1* linkage ratios per sample scaled to normal scores (base::scale R function). Since in some cases there is a variable number of measurements, this process was repeated 200,000 times and a mean pairwise distance between individuals was obtained. We performed a hierarchical clustering analysis (ward.D implemented method) on this similarity matrix to determine group membership (stats::hclust, stats::cutree R functions) (R Core Team 2017). Clustering was run to find two or three clusters that were defined taking into account that heterozygotes should be centered around 0.5. To assess the uncertainty of sample classification, we clustered two thirds of our samples selected at random 10,000 times and their percentage of association to each genotype was measured. Individuals that appeared as outliers in the membership distribution and were included $>5\%$ of times in a different cluster were not genotyped. For chr. X inversions, we repeated the analysis only with males clustered into *O1* or *O2* and, if these genotypes were more robust, they were the ones used. Finally, we tried to recover samples without a clear genotype in an extra clustering step by selecting proportionally more often those *O1* linkage ratios based on a higher total linkage when calculating euclidean distances and repeating the bootstrapping to obtain the final genotype clusters.

Haplotype fusion PCR (HF-PCR)

HF-PCR was carried out in an oil and water emulsion to generate the fused amplification product followed by a regular PCR with nested primers (Supplemental Table S15). To separate both inversion breakpoints, genomic DNA was digested overnight at 37 °C in a volume of 20 μl including 5 U of SwaI for HsInv395 or Sall for HsInv605, 1x buffer and 250 ng DNA, the restriction enzyme was then heat inactivated at 65 °C for 15 min, and 25 ng of DNA were used as template. Emulsion PCR reactions were performed in 25 μl in 96-well plates for 40 cycles as previously described (Turner and Hurles 2009). The main differences were that we used SOLiD™ EZ™ Bead Emulsifier Oil Kit (Applied Biosystems) to form the emulsions, and that after amplification we carefully transferred the emulsion to a fresh plate, added 50 μl of 1x Phusion HF buffer (Thermo Scientific) to increase volume, centrifuged for 5 minutes at maximum speed and recovered the aqueous phase containing the amplification products. Next, we did a 30-cycle reamplification step with 1 μl

Chapter 3. Results

of a 1/10 dilution of the previous PCR, 1.5 U Taq DNA polymerase (Roche), and 200-400 nM of each of the three nested primers in a 25 μ l total volume (Turner et al. 2006). Finally, 10 μ l of the PCR reaction were loaded into a 3% agarose gel for visualization.

Analysis of inversion frequency

Frequency differences between populations were measured with Weir and Cockerham's F_{ST} estimator implemented in `vcftools` (v0.1.15) (Danecek et al. 2011), using the 92 samples common to the 1000GP and only females for chr. X inversions or paired male chromosomes for the chr. Y inversion. F_{ST} values of each inversion were compared with an empirical distribution from 10,000 genome-wide biallelic 1000GP SNPs polymorphic in at least two of the populations, matched by chromosome type (autosome or chr. X) and excluding those SNPs overlapping inversion regions ± 100 kb. Correlation between MAF and the logarithm of the physical and genetic lengths of inversions was measured with a linear model implemented in `robustbase::lmrob` R function (Maechler et al. 2018), including data from 45 inversions in a larger sample of the same populations (Giner-Delgado et al. 2019). Inversion physical length corresponds to the distance between IRs and genetic length was interpolated from Bh erer et al. (2017) (Bh erer et al. 2017) recombination map, using the female map for chr. X and the sex average map for autosomes. No genetic length was available for chr. Y inversions and HsInv0608 (chr. X), which falls outside the last marker in the map.

Linkage disequilibrium (LD) analysis

Pairwise LD between genotypes of inversions and neighboring biallelic single nucleotide changes and small indels from 1000GP Phase 3 (inversion region ± 1 Mb) was calculated using the r^2 statistic with PLINK v1.9 (Purcell et al. 2007) separately for each population group and the 92 samples common to both datasets. SNPs were further classified as shared, private or fixed, depending on whether they were unambiguously polymorphic in the two orientations, polymorphic only in one, or their alleles were in perfect LD with the inversion, respectively (Giner-Delgado et al. 2019). To minimize possible genotyping errors in 1000GP data, we based our analyses on reliable variants, defined as those located in accessible regions according to the 1000GP strict accessibility mask and that do not overlap known SDs (Bailey et al. 2002).

Recurrence analysis

To generate haplotypes of the inverted regions, we selected the same 1000GP reliable variants used in the previous analysis that were present in at least two genotyped chromosomes. When the number of variants in an inversion was below 50, those accessible according to the pilot criteria were also included to maximize information. Available 1000GP Phase 3 haplotypes were used as scaffolds and the phase of the inversion genotypes was inferred using MVNcall (Menelaou and Marchini 2013) by positioning inversions in the middle as a single-nucleotide variant. Haplotypes were clustered by similarity and their relationships were visualized using the iHPlots strategy (Giner-Delgado et al. 2019). The putative ancestral orientation and the original inversion event were defined based on the $O1$ and $O2$ haplotype diversity and the frequency and

geographical distribution of haplotypes (usually considering as ancestral those found in African samples) (Supplemental Table S8). Additional independent inversion or re-inversion events were identified conservatively as haplotype clusters with an unexpected orientation and clearly differentiated from other *O1* or *O2* haplotypes (≥ 3 differences spanning >3 kb that cannot be easily explained by a recombination event) (Fig. 4B-C). To avoid possible phasing errors, in inversion heterozygotes we tested whether switching the orientation of both haplotypes still supported recurrence or not. For HsInv0416 we used the known chr. Y haplogroup information to determine the number of inversion events (Poznik et al. 2016) and estimated the recurrence rate as in Giner-Delgado et al. (2019) (Giner-Delgado et al. 2019). Briefly, a number of 30,931.1 generations were calculated for all branches involved in the phylogeny that relates the 48 analyzed males (Poznik et al. 2016), including a C-T branch split time of 76,000 years, a total number of mutations of 5,591, and an average number of mutations of all branches of 549.5, plus a generation time of 25 years (Repping et al. 2006). Finally, we tested different variables to determine their effect on the number of recurrent events per chromosome: chromosome type (autosome, chr. X or chr. Y), inversion and IR length, IR/Inv size ratio, IR identity, and PRDM9 motifs/kb within IRs (Myers et al. 2008). The model was built by stepwise regression with forward selection using the `robustbase::lmrob` R function (Maechler et al. 2018) and logarithm-transformed values to remove outliers.

Gene expression analysis

Inversion effects on LCL gene expression were first analyzed in 30 CEU and 29 YRI experimentally-genotyped individuals from the Geuvadis project (Lappalainen et al. 2013). We excluded HsInv0416 in chr. Y due to the low statistical power to detect differences only in males. Inversions 17p21, HsInv0290, HsInv0395, HsInv0605 and HsInv0786 were also imputed in 328 additional Geuvadis European samples (59 CEU, 91 TSI, 86 GBR and 92 FIN) using a representative tag SNP in the CEU population (which except for HsInv0290 belongs to a larger set of SNPs in LD in the expanded population). HsInv1057, with tag SNPs in YRI, was imputed in 58 additional Geuvadis YRI individuals but no significant gene-expression associations were obtained. RNA-seq reads (EMBL-EBI ArrayExpress experiment E-GEUV-1) were aligned against the human reference genome GRCh38.p10 (excluding patches and alternative haplotypes) with STAR v2.4.2a (Dobin et al. 2013). We estimated gene expression levels as reads per kilobase per million mapped reads (RPKM) based on GENCODE version 26 annotations (Harrow et al. 2012) and quantified transcript expression with RSEM v1.2.31 (Li and Dewey 2011), filtering out non-expressed genes and transcripts with <0.1 RPKM in $>80\%$ of the samples. RPKM values were normalized by quantile transformation across all samples and expression of each gene/transcript was adjusted to a standard normal distribution by rank-based inverse normal transformation. Association with expression of 418 genes and 2,044 transcripts was calculated for all biallelic variants with MAF >0.5 (including the inversion) within 1 Mb at either side of the transcription start site using linear regressions implemented in FastQTL (Ongen et al. 2016). Since technical or biological confounders reduce the power to find associations, we adjusted expression values by the top three 1000GP genotyping principal components (corresponding to population structure), sequencing laboratory, gender, and an optimal number of PEER (probabilistic estimation of expression residuals)

Chapter 3. Results

components (Stegle et al. 2012) for eQTL finding (for genes and transcripts, respectively, 12 and 15 for the experimental and 25 and 30 for the imputed set). Significant INV-eQTLs correspond to a Q value false-discovery rate (FDR) < 0.05 (Storey & Tibshirani 2003).

Next, we estimated inversion gene-expression effects in other tissues through the LD between inversion alleles and eQTLs in GTEx V7 release (GTEx Consortium 2013, 2017) using FAPI v0.1 (Kwan et al. 2016). First, we randomly took three samples of 30 experimentally genotyped individuals following ethnic proportions of GTEx donors (25 individuals from CEU, 4 YRI and 1 EAS) per inversion-gene pair and tissue to calculate LD patterns between each inversion and neighbouring SNPs, which were subsequently used to impute the corresponding inversion association P values from GTEx eQTL P values. If any P value was lower than the genome-wide empirical threshold defined by GTEx for each gene and tissue (GTEx Consortium 2017), we generated 30 samples of 30 individuals to calculate statistical significance more accurately. The eQTL P value of the inversion was defined as the median of permuted P values and the association confidence interval as the 25th and 75th percentiles, since the small number of individuals for LD calculation can produce extreme P values. In addition, we filtered out those associations with estimated P value lower than GTEx significance threshold or with a confidence interval spanning more than two orders of magnitude. The conservative nature of this analysis is represented by HsInv0389, in which only four of the 16 genes previously associated to the inversion based directly on the LD with GTEx eQTLs from a larger number of samples were identified. For both Geuvadis and GTEx, the most significant associated variants for each gene/transcript were designated as lead eQTLs. Moreover, inversions in high LD with the top marker ($r^2 \geq 0.8$) were indicated as potential lead eQTLs. GTEx LD was estimated as the median LD of permutations as explained above. Effect sizes were calculated as a function of MAF and P value, whereas direction was determined through LD with eQTLs using PLINK v1.9 *--r2 in-phase* option (Purcell et al. 2007).

GWAS data analysis

The impact of inversions in relevant phenotypic traits was assessed with the GWAS Catalog curated collection of the most significant SNPs associated to a particular phenotype ($P < 10^{-5}$) (<http://www.ebi.ac.uk/gwas/>) [release 2018-06-25, v1.0] (MacArthur et al. 2017). First, to explore the enrichment of GWAS SNPs within inversions, we translated GWAS Catalog coordinates to hg19 using Ensembl REST API (Yates et al. 2016) and grouped together the signals associated to SNPs in high LD ($r^2 \geq 0.8$) in 1000GP data and corresponding exactly to the same phenotype, resulting in 67,035 non-redundant SNP-trait associations. Then, we created a background distribution of each inversion with 1,500 random genomic regions of the same size that the inverted segment to calculate the enrichment P values. We excluded from permutations chr. Y, gaps, and the major histocompatibility complex region (chr6:28,477,797-33,448,354), known to harbor a vast number of associations. In addition, we tested that the GWAS enrichment was not biased by the allele SNP frequencies by selecting 150 random regions per inversion with comparable patterns of common variants (number of 1000GP loci with global MAF > 0.05

Chapter 3. Results

per kb $\pm 20\%$) and without this criteria, which showed very similar results ($r^2 = 0.99$). To explore which inversions were driving the enrichment, we repeated the analysis for each inversion independently using a one-tailed permutation test (to account for inversions with zero GWAS signals). Also, we compared the proportion of genes related to particular clinically relevant traits or diseases as reported in the GWAS Catalog inside or around (± 150 kb) each inversion with respect to the rest of the genome (Supplemental Table S11). Only traits with at least four associated genes close to the inversion were considered and P values were calculated with a Fisher's exact test adjusted by Bonferroni correction. Finally, we crossed GWAS Catalog variants with those in high LD with our inversions ($r^2 \geq 0.8$) in the corresponding population or the closest one available in our data, while the global LD was used for GWAS with populations from different continents (Supplemental Table S12).

DATA ACCESS

All inversion genotypes are available at InvFEST (<http://invfestdb.uab.cat>) and have been deposited in NCBI's dbVar database of genomic structural variation (<https://www.ncbi.nlm.nih.gov/dbvar>) under accession number XXXXXX.

ACKNOWLEDGEMENTS

We would like to thank Salvador Bartolomé, Xavier Alba and James Thomas for technical support and advice, Alba Vilella, Marina Laplana and Sergi Villatoro for help with DNA isolations and microsatellite analysis, and Xavier Estivill and Marta Morell for the European and African lymphoblastoid cell lines. This work was supported by research grants BFU2013-42649-P and BFU2016-77244-R funded by the Agencia Estatal de Investigación (AEI, Spain) and the European Regional Development Fund (FEDER, EU), ERC Starting Grant 243212 (INVFEST) from the European Research Council under the European Union Seventh Research Framework Programme (FP7), and 2017-SGR-1379 from the Generalitat de Catalunya (Spain) to MC, and a La Caixa Doctoral fellowship to JLJ. MGV was supported by POCI-01-0145-FEDER-006821 funded through the Operational Programme for Competitiveness Factors (COMPETE, EU) and UID/BIA/50027/2013 from the Foundation for Science and Technology (FCT, Portugal). ddPCR reagents used in this study were provided by Bio-Rad Laboratories, Inc.

AUTHOR CONTRIBUTIONS

MP and MC designed the genotyping assays and oversaw all steps; MP, SP, DI and AD performed experiments; CGD, MGV, MP and MC analyzed evolutionary data; JLJ and MP analyzed functional effects; MP, SP, JFR, GKN and MC contributed to ddPCR assay design and optimization; MP, JLJ and MC wrote the paper.

DISCLOSURE DECLARATION

Chapter 3. Results

The authors declare the following competing financial interests: Bio-Rad Laboratories, Inc. markets and sells the QX200 Droplet Digital PCR System. JFR and GKN are or were employees of Bio-Rad Laboratories, Inc. at the time the study was performed.

REFERENCES

- Aguado C, Gayà-Vidal M, Villatoro S, Oliva M, Izquierdo D, Giner-Delgado C, Montalvo V, García-González J, Martínez-Fundichely A, Capilla L, et al. 2014. Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. *PLoS Genet* **10**: e1004208.
- Alves JM, Lima AC, Pais IA, Amir N, Celestino R, Piras G, Monne M, Comas D, Heutink P, Chikhi L, et al. 2015. Reassessing the evolutionary history of the 17q21 inversion polymorphism. *Genome Biol Evol* **7**: 3239–48.
- Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE. 2009. Characterization of six human disease-associated inversion polymorphisms. *Hum Mol Genet* **18**: 2555–66.
- Antonarakis SE, Rossiter JP, Young M, Horst J, de Moerloose P, Sommer SS, Ketterling RP, Kazazian HH, Négrier C, Vinciguerra C, et al. 1995. Factor VIII gene inversions in severe hemophilia A: results of an international consortium study. *Blood* **86**: 2206–12.
- Aradhya S, Bardaro T, Galgóczy P, Yamagata T, Esposito T, Patlan H, Ciccodicola A, Munnich A, Kenwrick S, Platzer M, et al. 2001. Multiple pathogenic and benign genomic rearrangements occur at a 35 kb duplication involving the NEMO and LAGE2 genes. *Hum Mol Genet* **10**: 2557–67.
- Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the major structural variant alleles of the human genome. *Cell* **176**: 663–75.
- Bailey JA, Gu Z, Clark RA, Reinert K, Samonte R V., Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–7.
- Beck CR, Carvalho CMB, Banser L, Gambin T, Stubbolo D, Yuan B, Sperle K, McCahan SM, Henneke M, Seeman P, et al. 2015. Complex genomic rearrangements at the PLP1 locus include triplication and quadruplication. *PLoS Genet* **11**: e1005050.
- Bhéret C, Campbell CL, Auton A. 2017. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat Commun* **8**: 14994.
- Boettger LM, Handsaker RE, Zody MC, McCarroll SA. 2012. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* **44**: 881–5.
- Boettger LM, Salem RM, Handsaker RE, Peloso GM, Kathiresan S, Hirschhorn JN, McCarroll SA. 2016. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat Genet* **48**: 359–66.
- Bondeson ML, Dahl N, Malmgren H, Kleijer WJ, Tønnesen T, Carlberg BM, Pettersson U. 1995. Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum Mol Genet* **4**: 615–21.
- Cáceres A, González JR. 2015. Following the footprints of polymorphic inversions on SNP data: from detection to association tests. *Nucleic Acids Res* **43**: e53.
- Cáceres A, Sindi SS, Raphael BJ, Cáceres M, González JR. 2012. Identification of polymorphic inversions

Chapter 3. Results

from genotypes. *BMC Bioinformatics* **13**: 28.

Camunas-Soler J, Lee H, Hudgins L, Hintz SR, Blumenfeld YJ, El-Sayed YY, Quake SR. 2018.

Noninvasive prenatal diagnosis of single-gene disorders by use of droplet digital PCR. *Clin Chem* **64**: 336–45.

Carvalho CMB, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, Fang P, McCall A, Pivnick EK, Hines-Dowell S, Seaver LH, et al. 2011. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat Genet* **43**: 1074–81.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784.

Clapham KR, Yu TW, Ganesh VS, Barry B, Chan Y, Mei D, Parrini E, Funalot B, Dupuis L, Nezarati MM, et al. 2012. FLNA genomic rearrangements cause periventricular nodular heterotopia. *Neurology* **78**: 269–78.

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Khera A V., Francioli LC, Gauthier LD, Wang H, Watts NA, et al. 2019. An open resource of structural variation for medical and population genetics. *bioRxiv* 578674.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–8.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Feuk L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, Scherer SW. 2005. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet* **1**: e56.

Fusco F, Paciolla M, Napolitano F, Pescatore A, D'Addario I, Bal E, Lioi MB, Smahi A, Miano MG, Ursini MV. 2012. Genomic architecture at the Incontinentia Pigmenti locus favours de novo pathological alleles through different mechanisms. *Hum Mol Genet* **21**: 1260–71.

Giner-Delgado C, Villatoro S, Lerga-Jaso J, Gaya-Vidal M, Oliva M, Castellano D, Pantano L, Bitarello B, Izquierdo D, Noguera I, et al. 2019. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat Commun*, In press.

González JR, Cáceres A, Esko T, Cuscó I, Puig M, Esnaola M, Reina J, Siroux V, Bouzigon E, Nadif R, et al. 2014. A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. *Am J Hum Genet* **94**: 361–72.

Grau C, Starkovich M, Azamian MS, Xia F, Cheung SW, Evans P, Henderson A, Lalani SR, Scott DA. 2017. Xp11.22 deletions encompassing CENPVL1, CENPVL2, MAGED1 and GSPT2 as a cause of syndromic X-linked intellectual disability. *PLoS One* **12**: e0175962.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204–13.

GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580–5.

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA. 2015. Large

- multiallelic copy number variations in humans. *Nat Genet* **47**: 296–303.
- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–74.
- Hayward BE, De Vos M, Valleley EMA, Charlton RS, Taylor GR, Sheridan E, Bonthron DT. 2007. Extensive gene conversion at the PMS2 DNA mismatch repair locus. *Hum Mutat* **28**: 424–30.
- Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, et al. 2016. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7**: 12989.
- Hindson CM, Chevillet JR, Briggs HA, Gallichotte EN, Ruf IK, Hindson BJ, Vessella RL, Tewari M. 2013. Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nat Methods* **10**: 1003–5.
- Hoff AM, Alagaratnam S, Zhao S, Bruun J, Andrews PW, Lothe RA, Skotheim RI. 2016. Identification of novel fusion genes in testicular germ cell tumors. *Cancer Res* **76**: 108–16.
- Hoffmann AA, Rieseberg LH. 2008. Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annu Rev Ecol Evol Syst* **39**: 21–42.
- Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al. 2017. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* **27**: 677–85.
- Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55–61.
- de Jong S, Chepelev I, Janson E, Strengman E, van den Berg LH, Veldink JH, Ophoff RA. 2012. Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. *BMC Genomics* **13**: 458.
- Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L, et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**: 203–6.
- Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56–64.
- Kirkpatrick M. 2010. How and why chromosome inversions evolve. *PLoS Biol* **8**: e1000501.
- Kwan JS, Li M-X, Deng J-E, Sham PC. 2016. FAPI: Fast and accurate P-value Imputation for genome-wide association study. *Eur J Hum Genet* **24**: 761–6.
- Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, González-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–11.
- Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, Young E, Lam ET, Hastie AR, Wong KHY, et al. 2019. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat Commun* **10**: 1025.

Chapter 3. Results

- Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.
- Li JN, Carrero IG, Dong JF, Yu FL. 2015. Complexity and diversity of F8 genetic variations in the 1000 genomes. *J Trombos Haemost* **13**: 2031–40.
- Li L, Leung AK-Y, Kwok T-P, Lai YYY, Pang IK, Chung GT-Y, Mak ACY, Poon A, Chu C, Li M, et al. 2017. OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. *Genome Biol* **18**: 230.
- Liu P, Lacaria M, Zhang F, Withers M, Hastings PJ, Lupski JR. 2011. Frequency of nonallelic homologous recombination is correlated with length of homology: Evidence that ectopic synapsis precedes ectopic crossing-over. *Am J Hum Genet* **89**: 580–8.
- Lowry DB, Willis JH. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol* **8**: e1000500.
- Ma J, Amos CI. 2012. Investigation of inversion polymorphisms in the human genome using principal components analysis. *PLoS One* **7**: e40224.
- MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**: D896–901.
- Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller, Manuel Conceicao ELT, di Palma MA. 2018. robustbase: Basic Robust Statistics R package version 0.93-2.
- Martin J, Han C, Gordon LA, Terry A, Prabhakar S, She X, Xie G, Hellsten U, Chan YM, Altherr M, et al. 2004. The sequence and analysis of duplication-rich human chromosome 16. *Nature* **432**: 988–94.
- Martínez-Fundichely A, Casillas S, Egea R, Ràmia M, Barbadilla A, Pantano L, Puig M, Cáceres M. 2014. InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Res* **42**: D1027–32.
- McDermott GP, Do D, Litterst CM, Maar D, Hindson CM, Steenblock ER, Legler TC, Jouvenot Y, Marrs SH, Bemis A, et al. 2013. Multiplexed target detection using DNA-binding dye chemistry in droplet digital PCR. *Anal Chem* **85**: 11619–27.
- Menelaou A, Marchini J. 2013. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**: 84–91.
- Myers AJ, Kaleem M, Marlowe L, Pittman AM, Lees AJ, Fung HC, Duckworth J, Leung D, Gibson A, Morris CM, et al. 2005. The H1c haplotype at the MAPT locus is associated with Alzheimer's disease. *Hum Mol Genet* **14**: 2399–404.
- Myers S, Freeman C, Auton A, Donnelly P, McVean G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* **40**: 1124–29.
- Olmedillas-López S, García-Arranz M, García-Olmo D. 2017. Current and emerging applications of droplet digital PCR in oncology. *Mol Diagn Ther* **21**: 493–510.
- Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. 2016. Fast and efficient QTL mapper for

- thousands of molecular phenotypes. *Bioinformatics* **32**: 1479–85.
- Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet* **48**: 593–9.
- Puig M, Casillas S, Villatoro S, Cáceres M. 2015a. Human inversions and their functional consequences. *Brief Funct Genomics* **14**: 369–79.
- Puig M, Castellano D, Pantano L, Giner-Delgado C, Izquierdo D, Gayà-Vidal M, Lucas-Lledó JI, Esko T, Terao C, Matsuda F, et al. 2015b. Functional impact and evolution of a novel human polymorphic inversion that disrupts a gene and creates a fusion transcript ed. J.M. Akey. *PLoS Genet* **11**: e1005495.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559–75.
- R Core Team. 2017. R: A Language and Environment for Statistical Computing. *R Found Stat Comput*.
- Regan JF, Kamitaki N, Legler T, Cooper S, Klitgord N, Karlin-Neumann G, Wong C, Hodges S, Koehler R, Tzonev S, et al. 2015. A rapid molecular approach for chromosomal phasing ed. D.E. Arking. *PLoS One* **10**: e0118270.
- Repping S, van Daalen SKM, Brown LG, Korver CM, Lange J, Marszalek JD, Pyntikova T, van der Veen F, Skaletsky H, Page DC, et al. 2006. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat Genet* **38**: 463–7.
- Salm MPA, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, Schadt EE, Cookson WO, Wierzbicki AS, Naoumova RP, et al. 2012. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Res* **22**: 1144–53.
- Sanders AD, Hills M, Porubský D, Guryev V, Falconer E, Lansdorp PM. 2016. Characterizing polymorphic inversions in human genomes by single cell sequencing. *Genome Res* **26**: 1575–87.
- Shao H, Ganesamoorthy D, Duarte T, Cao MD, Hoggart CJ, Coin LJM. 2018. npInv: accurate detection and genotyping of inversions using long read sub-alignment. *BMC Bioinformatics* **19**: 261.
- Small K, Iber J, Warren ST. 1997. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat Genet* **16**: 96–9.
- Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, et al. 2005. A common inversion under selection in Europeans. *Nat Genet* **37**: 129–37.
- Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**: 500–7.
- Steinberg KM, Antonacci F, Sudmant PH, Kidd JM, Campbell CD, Vives L, Malig M, Scheinfeldt L, Beggs W, Ibrahim M, et al. 2012. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat Genet* **44**: 872–80.
- Strain MC, Lada SM, Luong T, Rought SE, Gianella S, Terry VH, Spina CA, Woelk CH, Richman DD.

Chapter 3. Results

2013. Highly precise measurement of HIV DNA by droplet digital PCR. *PLoS One* **8**: e55943.
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.
- The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.
- Thomas JW, Cáceres M, Lowman JJ, Morehouse CB, Short ME, Baldwin EL, Maney DL, Martin CL. 2008. The chromosomal polymorphism linked to variation in social behavior in the white-throated sparrow (*Zonotrichia albicollis*) is a complex rearrangement and suppressor of recombination. *Genetics* **179**: 1455–68.
- Turner DJ, Hurles ME. 2009. High-throughput haplotype determination over long distances by haplotype fusion PCR and ligation haplotyping. *Nat Protoc* **4**: 1771–83.
- Turner DJ, Shendure J, Porreca G, Church G, Green P, Tyler-Smith C, Hurles ME. 2006. Assaying chromosomal inversions by single-molecule haplotyping. *Nat Methods* **3**: 439–45.
- Umina PA, Weeks AR, Kearney MR, McKechnie SW, Hoffmann AA. 2005. A rapid shift in a classic clinal pattern in *Drosophila* reflecting climate change. *Science* **308**: 691–3.
- Vicente-Salvador D, Puig M, Gayà-Vidal M, Pacheco S, Giner-Delgado C, Noguera I, Izquierdo D, Martínez-Fundichely A, Ruiz-Herrera A, Estivill X, et al. 2017. Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. *Hum Mol Genet* **26**: 567–81.
- Webb A, Miller B, Bonasera S, Boxer A, Karydas A, Wilhelmsen KC. 2008. Role of the tau gene region chromosome inversion in progressive supranuclear palsy, corticobasal degeneration, and related disorders. *Arch Neurol* **65**: 1473–8.
- Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. *Nucleic Acids Res* **44**: D710–6.
- Zabetian CP, Hutter CM, Factor SA, Nutt JG, Higgins DS, Griffith A, Roberts JW, Leis BC, Kay DM, Yearout D, et al. 2007. Association analysis of MAPT H1 haplotype and subhaplotypes in Parkinson's disease. *Ann Neurol* **62**: 137–44.

SUPPLEMENTAL FIGURES

Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR

Marta Puig¹, Jon Lerga-Jaso¹, Carla Giner-Delgado¹, Sarai Pacheco¹, David Izquierdo¹, Alejandra Delprat¹, Magdalena Gayà-Vidal², Jack F. Regan³, George Karlin-Neumann³, Mario Cáceres^{1,4}

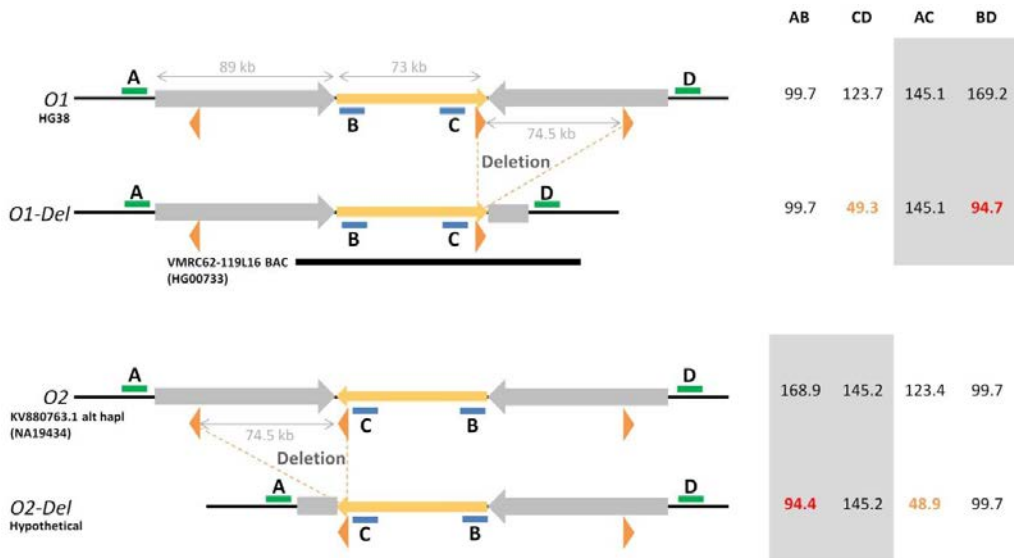
¹ Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.

² CIBIO/InBIO Research Center in Biodiversity and Genetic Resources, Universidade do Porto, Vairão, Distrito do Porto, Portugal.

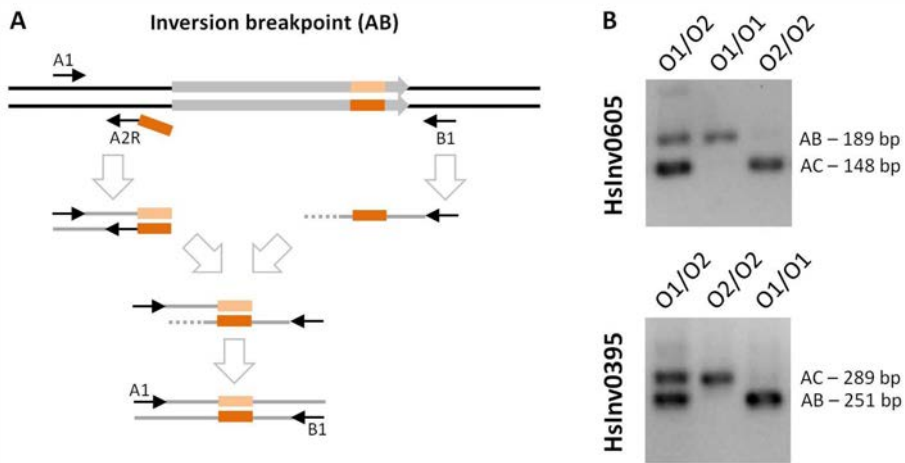
³ Digital Biology Center, Bio-Rad Laboratories, Pleasanton, CA, United States.

⁴ ICREA, Barcelona, Spain.

Chapter 3. Results

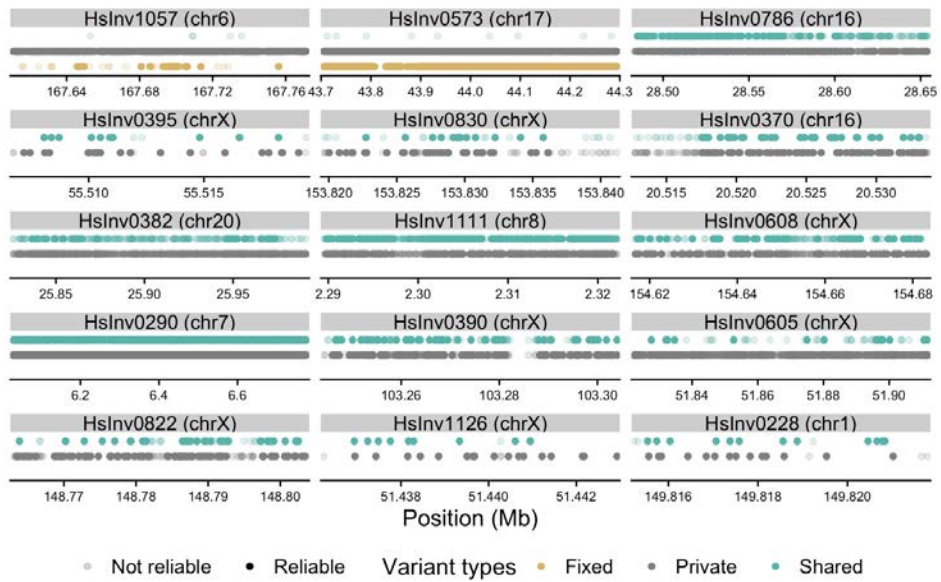


Supplemental Figure S1. ddPCR analysis of structural variation at the HsInv0233 inversion region. Grey arrows indicate inverted segmental duplications (SDs) and the yellow arrow the orientation of the inverted sequence (*O1* or *O2*). ddPCR amplicons are represented as green (outside the inversion) and blue (inside the inversion) bars labeled according to their position, and orange triangles correspond to the 11.6-kb repeats that mediate a 74.5-kb deletion affecting most of one of the SDs. The source of the sequences is indicated below the name of each conformation and in *O1-Del* a black bar below indicates the region included in the BAC clone that supports this structure. The distances in kb between each pair of amplicons are represented to the right of each structure, with combinations expected to have lower linkages shaded in grey and those affected by the deletion shown in orange and red. As a consequence of the deletion, the BD distance is similar in an *O1-Del* and *O2* chromosome (and the same happens with AB in the hypothetical *O2-Del* and *O1* conformation). In addition, when the distance between the two analyzed amplicons is different in the two chromosomes of an individual, they will contribute differently to the measured linkage, making it impossible to interpret without additional information and preventing accurate genotyping of the inversion orientation.

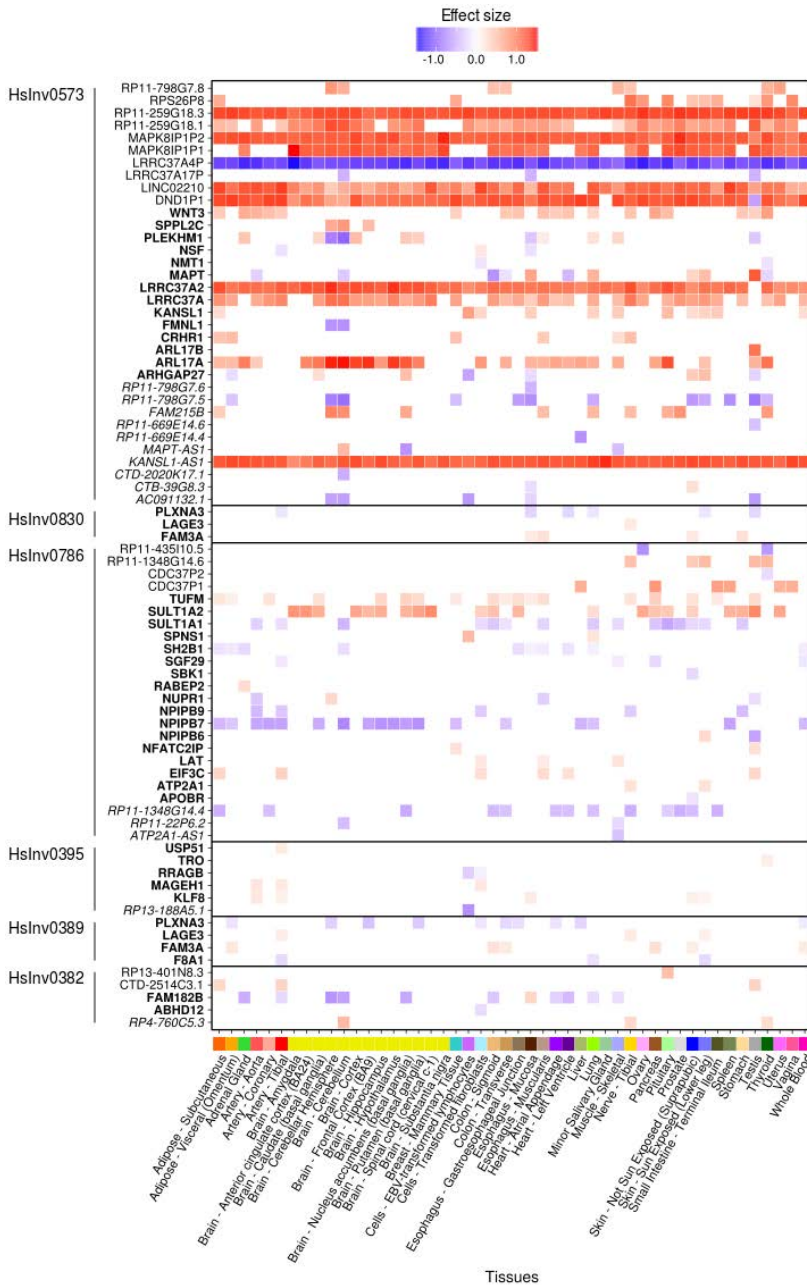


Supplemental Figure S2. Inversion genotyping by haplotype-fusion PCR. (A) Summary of haplotype-fusion PCR strategy to genotype the AB breakpoint of an inversion (Turner et al. 2006). A double-stranded amplicon located outside an inversion is amplified with two primers A1 and A2R (black arrows), one of which has a 5' extension (orange rectangle) with a sequence found within the inverted repeat (grey arrow) at the inversion breakpoint. In the same reaction, a single-stranded product is linearly amplified at the other side of the breakpoint with primer B1. This single-stranded product contains the sequence able to hybridize with amplicon A (orange rectangle) and an AB fusion product containing sequences from both sides of the breakpoint is amplified once primer A2R runs out. Since the PCR reaction takes place in an emulsion, only one DNA template molecule is expected to be found within a single droplet and the fusion product will indicate the presence in the sample of an AB junction. By adding a primer C1 able to amplify the other end of the inverted sequence, we can detect both AB and AC breakpoints and genotype the inversion. (B) Examples of inversion genotypes. After reamplification with nested primers, the three genotypes can be clearly distinguished for the two inversions analyzed in this work by visualizing the fusion products in an agarose gel. All primers used in this experiment are listed in Supplemental Table S15.

Chapter 3. Results

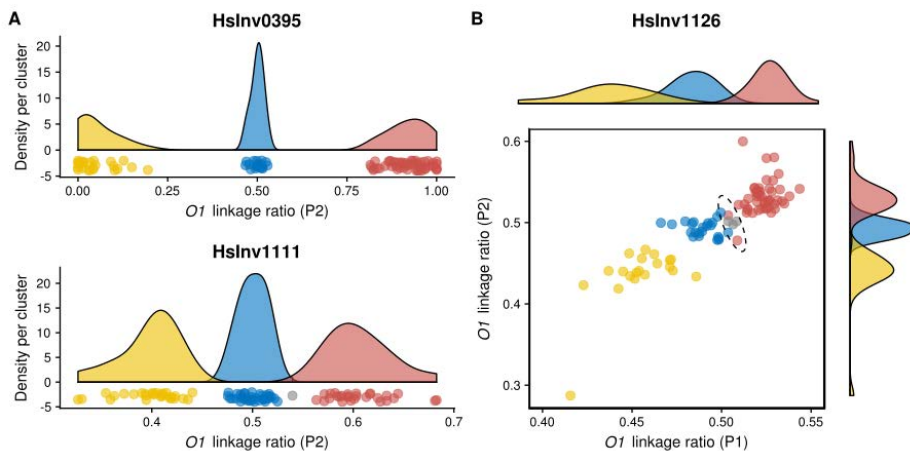


Supplemental Figure S3. Distribution of SNPs along the analyzed inverted regions. SNPs are classified as fixed between orientations (yellow), private to one orientation (grey) or shared between orientations (green), and color intensity indicates their reliability according to the 1000GP Phase 3 strict accessibility mask and their location outside segmental duplications. The presence of shared variants polymorphic in both orientations along the entire length of the inversion is difficult to explain by gene conversion events and it can be thus considered a sign of recurrence.



Supplemental Figure S4. Summary of gene expression changes associated to inversions across GTEx tissues. Inversion effects were estimated using FAPI (Kwan et al. 2016) through LD patterns with eQTLs in GTEx Analysis Release v7 (see Supplemental Table S10). The direction and strength of the beta effect is indicated in blue or red representing, respectively, a lower or higher expression associated to the O2 orientation. Gene names in bold correspond to protein coding genes, in italics to non-coding RNA genes, and the rest are classified as pseudogenes in GENCODE version 26.

Chapter 3. Results



Supplemental Figure S5. Inversion genotype calling by clustering of *O1* linkage ratios. Dots indicate the *O1* linkage ratio for each sample and colors mark genotype groups: *O1/O1*, red; *O1/O2*, blue; *O2/O2*, yellow. Grey dots are samples without a clear genotype. HsInv0395 and HsInv1111 (A) show separated genotype groups, while HsInv1126 (B) has overlapping clusters. P1 and P2 represent two randomly selected *O1* linkage ratio values for the different samples. Dots inside the ellipse correspond to samples included in both *O1/O1* and *O1/O2* clusters with similar probability and were not genotyped (with the exception of two males in red recovered in the male-specific clustering analysis for chr. X inversions).

3.3 Functional and phenotypic effects of a human inversion regulating *RHOH* isoforms and RhoH protein levels

A good example of an inversion with clear effects at the molecular level is HsInv0102, a 2 kb inversion that inverts an alternative 5'-UTR exon of the *RHOH* gene, which no longer can be part of the final transcript in the inverted orientation (Puig et al. 2015a; Giner-Delgado et al. 2019). The inversion was originally predicted by PEM (Korbel et al. 2007) and has been identified in many other studies since then (Sudmant et al. 2015; Hehir-Kwa et al. 2016; Audano et al. 2019; Chaisson et al. 2019). Recently, as part of a global study of multiple common human inversions, HsInv0102 was experimentally genotyped in 551 individuals of seven populations (Giner-Delgado et al. 2019), which showed relatively good consistency with the imputed genotypes from the 1000GP Phase 3 Structural Variation release (2.5% imputation error) (Sudmant et al. 2015). This work also identified differences in HsInv0102 worldwide distribution, with the frequency of the derived inverted orientation (*Inv*) ranging from 3.9% in East Asia to 30.3% in Africa, and confirmed the expected reduction of the expression of the alternatively-spliced isoform including the exon in the individuals with the inversion (Giner-Delgado et al. 2019). However, the possible functional consequences of this change were not studied in detail.

The gene *RHOH* encodes an atypical hematopoiesis-specific guanosine triphosphatase (GTPase) protein, RhoH, which contains different residues in two highly-conserved amino acid sites necessary for GTPase activity, remaining constitutively in a GTP-bound active conformation without cycling (Li et al. 2002; Lahousse et al. 2004; Gu et al. 2005). RhoH was found to block $I\kappa B$ degradation, suppressing the activation of NF- κB induced by TNF α and other Rho GTPases (Li et al. 2002), which is known to control the transcription of a wide range of genes involved in inflammation, cell proliferation and survival (Hodge and Ridley 2016; Phuyal and Farhan 2019). In fact, experimentally-induced overexpression

of RhoH in hematopoietic progenitor cells led to a significant reduction in cell growth and proliferation in response to cytokine stimulation, inhibited actin assembly and chemokine-induced cell migration, was associated with increased apoptosis, suppresses activation of Rac GTPases and contributed to defective engraftment *in vivo*, whereas down-regulation had the opposite effect (Gu et al. 2005). RhoH was also found to recruit to the plasma membrane Zap70 and LCK, key signaling molecules in the T cell receptor (TCR) complex, triggering TCR-mediated signaling pathways in T cell activation, and to contribute to T cell development (Gu et al. 2006; Chae et al. 2010). In addition, *RHOH* levels may influence the regulation of intracellular signal transduction in B-lymphocytes as well (Matsumoto et al. 2009). Finally, *RHOH* has also been found to regulate several other processes, including cell migration, microtubule dynamics, homing and chemotaxis (Wang et al. 2010; Troeger et al. 2012; Pan et al. 2018; Tajadura-Ortega et al. 2018), IL-3-mediated signalling (Gundogdu et al. 2010), leukotriene production in neutrophils (Daryadel et al. 2009), signal transduction in mast cell through Syk interaction (Oda et al. 2009) or it is needed to maintain the lymphocyte integrin LFA-1 in a nonadhesive state on resting T cells (Cherry et al. 2004). Consequently, expression, translation and degradation of RhoH is probably robustly controlled.

Numerous studies have shown how important RhoH physiological and immune role is, with *RHOH* mutations being involved in susceptibility to viral infections due to impaired T-cell function (Crequer et al. 2012), psoriasis (Tamehiro et al. 2019) and non-Hodgkin lymphoma (Dallery et al. 1995). Moreover, *RHOH* was described as an hypermutable gene in B-cell diffuse large-cell lymphoma (Pasqualucci et al. 2001), AIDS related non-Hodgkin lymphoma (Gaidano et al. 2003) and primary central nervous system lymphoma (Montesinos-Rongen et al. 2004). *RHOH* mutations have been found in a variety of other human cancers as well (Fueller and Kubatzky 2008). On the other hand, low RhoH levels were also detected in hairy cell leukaemia (HCL) (Galiegui-Zouitina et al. 2008) and acute myeloid leukaemia (AML), where they were a bad prognosis marker (Iwasaki et al. 2008). In this regard, RhoH has been hypothesized to act as a protector from Rac1 tumorigenic action, since RhoH overexpression

suppresses this activity and increases apoptosis (Li et al. 2002; Zhang et al. 2004). In contrast, *RHOH* mRNA levels are unusually increased in chronic lymphocytic leukemia (CLL), where RhoH seems to be required for cell survival through the interaction with supportive microenvironments (Sanchez-Aguilera et al. 2010; Troeger et al. 2012). Finally, in epithelial cancer cell lines such as prostate, it has been found that *RHOH* depletion reduces cell migration and cancer progression (Tajadura-Ortega et al. 2018). In any case, although all this evidence suggests a possible function of RhoH dysregulation in malignant transformation, the involvement of *RHOH* in cancer susceptibility has not been proven yet.

Interestingly, both recurrent chromosomal translocations and point mutations involved in cancer affected the same 5' region of the gene, highlighting its importance in *RHOH* regulation. Diverse alternative splicing patterns of *RHOH* 5' UTR exons have been reported among different hematopoietic cell lineages, together with distinct transcription start sites. This gives rise to high heterogeneity in transcript profiles, which suggest a complex post-transcriptional regulation (Lahousse et al. 2004).

3.3.1 HsInv0102 impact on *RHOH* gene structure and expression in LCLs

Up to 24 isoforms have been reported for the *RHOH* gene by the Comprehensive Gene Annotation Set from GENCODE release 26 (Harrow et al. 2012). Since some isoforms are partial or non-protein coding transcripts, we decided to use the 10 transcripts from GENCODE basic annotation as a representative subset for this gene. Overlapping exon intervals within *RHOH* were merged to obtain a standard gene model to work with (Figure 3.1), and exons were numbered following the direction of *RHOH* transcription with previously described exons (Lahousse et al. 2004) indicated in parenthesis: E1 (X1), E2 (1a), E3 (X2), E4 (1b), E8 (X3), E9 (X4) and the CDS-containing exon as E10 (2). According to this, inversion HsInv0102 inverts exon E8, which apparently originated by the existence of splicing sites within repeat elements. The 5' donor splice site is within an Alu

element, whereas E8 has two different 3' acceptor sites separated by 14 bp and provided by the L1 repeat element LIME3A.

We used RNA-Seq data from lymphoblastoid cell lines (LCLs) of 445 individuals with European (358) and African (87) ancestry (Lappalainen et al. 2013) in which we had previously imputed the inversion to estimate exon E8 and global *RHOH* expression levels. First, we estimated that the mean number of reads that span across exon-exon junctions using the E8 second acceptor splicing site was ~ 2.56 fold higher than those crossing the junction of the longer exon form (Figure 3.1). Second, we performed a joint splicing QTL (sQTL) analysis to assess the effect of HsInv0102 and other 1000GP variants ($MAF > 0.05$) on E8 inclusion (The 1000 Genomes Project Consortium 2015; Sudmant et al. 2015). As expected, the inverted allele is associated with the complete exclusion of this exon from the final mRNA. However, this effect is masked by the first lead sQTL SNP, rs7699141, since the *Inv* orientation appears only with rs7699141 C allele, which already reduces the expression of this exon around 93%. The SNP is located 5 bp away from the beginning of the intron, affecting a relatively conserved position within the donor sequence regulating splicing, and is linked to two small non-inverted haplotypes with high (*Std1*) and low (*Std2*) E8 expression. In fact, inversion HsInv0102 appears as a secondary independent lead sQTL of E8 after accounting by rs7699141 (Figure 3.1). Therefore, inversion HsInv0102 acts as a variant that potentiates the rs7699141 C allele effect on E8 exclusion.

To evaluate the potential functional impact of the alternative exon E8, we examined the presence of this exon in *RHOH* final mRNA. First, we checked which exons appear together with E8 by looking at junction reads and mate alignments of reads from LCLs. Main mRNA E8-containing transcripts include E4-E8-E9-E10, E2-E8-E9-E10 and E2-E4-E8-E9-E10, which were also supported by human expressed sequence tags (ESTs) (Kent et al. 2002). In these transcripts, an upstream open reading frame (uORF) is created between E8 and E9 exons, which is the longest *RHOH* uORF (108 nucleotides potentially encoding 35 amino acids) with at least 30 nucleotides found in *RHOH* (uORFs of 37-72 nucleotides in E2 and

E4) and the closest to the coding sequence. Second, we calculated the inclusion of E8 in these transcripts from all samples, although the non-uniformity of RNA-Seq read coverage prevents determining accurately the real incorporation of this exon. We calculated that E8 is present in $\sim 2.2\%$ of the isoforms assuming uniform coverage. Taking into account that the inversion of this exon completely removes it from *RHOH* mRNA, we observed that E8 inclusion increases to 6.2% in *Std1/Std1* samples, which is close to the 6.7% in cDNAs products generated in the Raji B-cell line by Lahousse et al (2004). These values were confirmed by linking the relative number of full reconstructed transcripts with E8 from Pacific Biosciences LCL transcriptome data from the LCL transcriptomes of a HapMap trio (Tilgner et al. 2014) to HsInv0102 and rs7699141 genotypes, resulting in E8 inclusion levels of 0% for NA12891 (*Inv/Inv*), 2.9% for NA12878 (*Std1/Inv*) and 4.9% for NA12892 (*Std1/Std1*). Therefore, exon E8 is present in a significant fraction of *RHOH* transcripts in *Std1/Std1* individuals.

Based on these results, we investigated the association of HsInv0102 and rs7699141 with *RHOH* expression, but no significant effects were found, indicating that E8 inclusion does not correlate with global transcript levels ($P = 0.77$) (Figure 3.1). In fact, we found that the lead eQTL, rs11723134 ($P = 8.97 \times 10^{-7}$), and other linked variants were located upstream of *RHOH* at the promoter region. By checking different available data sets, we found that the same SNP was also reported as eQTL for *RHOH* across populations using the same data (Wen et al. 2015), but eQTL mapping in blood from 31,684 individuals from the eQTLGen Consortium found as lead eQTL rs1397934, which is also located upstream of *RHOH* but is not linked to rs11723134 (Võsa et al. 2018) and no eQTLs were found in EBV-transformed lymphocytes, spleen or whole blood tissues according to the GTEx project (GTEx Analysis Release v7) (GTEx Consortium 2017). Therefore, the regulation of *RHOH* gene expression is still largely unknown.

Next, we took advantage of normalized peptide expression data from Battle et al. (2015) to estimate protein abundance in LCLs. In this study,

RhoH was originally excluded due to the low number of peptides and individuals measured for this protein. Nonetheless, we obtained accurate estimates of RhoH protein abundance in 42 YRI individuals from two peptides of the same protein that were highly correlated ($r = \text{XXX}$; $P = 5.58 \times 10^{-4}$) (see Methods). We found a decrease in RhoH protein levels when rs7699141 and HsInv0102 reference alleles are present (logarithmic regression, $P = 0.0023$) (Figure 3.1), suggesting that higher inclusion rates of this exon are linked to translation or protein degradation rather than transcription. In fact, rs7699141 was the lead variant acting as protein QTL (pQTL) and the combination of the SNP and inversion genotypes was the most significant signal in the pQTL analysis ($P = 0.0014$). We could not check if *RHOH* lead eQTL in LCLs, rs11723134, was also a pQTL, since this SNP is not polymorphic in YRI population. In order to verify that the observed associations were not an artifact due to the limited sample size, we compared transcript and protein expression levels in LCLs. We found a significant correlation despite the low number of common samples (32) between both datasets ($P = 0.0437$). However, whereas no linear or logarithmic correlation between inversion and rs7699141 genotypes and gene expression was found, correlation with protein levels remained significant in this sample subset ($P = 0.0221$). Finally, we used ribosome profiling data (Battle et al. 2015) to see if the E8-E9 uORF could account for the low RhoH translation rate, but no differences in ribosome occupancy data were found across HsInv0102 or rs7699141 genotypes, suggesting that RhoH levels are regulated through other mechanisms. Altogether, these results indicate that the E8 inclusion in the absence of the inversion or alternative rs7699141 allele increases to ~ 1 per 16 *RHOH* transcripts, and its presence correlates with a reduction of RhoH protein, uncovering rs7699141 and HsInv0102 as novel pQTLs not identified in previous studies (Battle et al. 2015; Suhre et al. 2017; Folkersen et al. 2017; Yao et al. 2018; Sun et al. 2018).

3.3.2 Additional effects on gene and protein expression

Since RhoH is a GTPase involved in multiple signaling pathways, we searched for other gene-expression changes in LCLs associated to HsInv0102 and rs7699141. For that, we focused on 13,460 expressed protein-coding genes and assumed linear additive effects as a function of the previous genotype combinations of these variants to mirror their impact on E8 levels. Only the gene *CLU*, which codes for a molecular chaperone responsible for apoptotic processes, was identified as differentially expressed using either linear (FDR-adjusted $P = 0.0046$) or logarithmic regression (FDR-adjusted $P = 0.0041$). We also correlated gene expression with HsInv0102 genotypes alone, since its breakpoint 1 (BP1) disrupts a core of transcription factor binding sites detected by ORegAnno (Lesurf

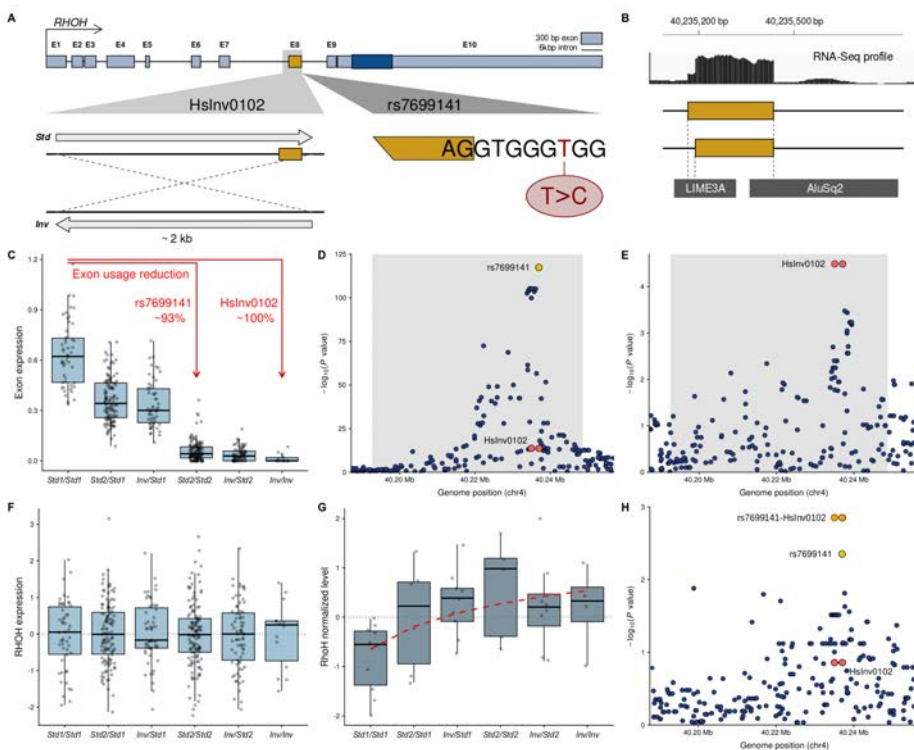


Figure 3.1 – HsInv0102 impact on *RHOH* expression in LCLs. Caption on the following page.

Figure 3.1 – HsInv0102 impact on *RHOH* expression in LCLs.

(A) *RHOH* gene structure, where overlapping exons were merged into meta-exons depicted as light blue boxes with the coding sequence in dark blue and an arrow showing the transcription direction. HsInv0102 (light gray shading) inverts exon E8, whereas SNP rs7699141 affects the splice donor site, with the exact position of the two alleles shown in red. Legend shows the different scale used to represent the size of exons and introns. (B) LCL RNA-Seq profile of E8, which is the result of splice sites provided by LINE and Alu repeat elements. The two forms of E8 depending on acceptor site usage are shown. (C) Expression levels of exon E8 for the different genotype combinations, showing that alternative alleles of HsInv0102 (*Inv*) and rs7699141 (*Std2*) reduce E8 inclusion. Manhattan plots for association of variants with exon E8 splicing in LCLs, showing SNP rs7699141 as a primary lead sQTL (D) and inversion HsInv0102 as secondary sQTL after adjusting by the effect of rs7699141 (E). Expression levels of *RHOH* mRNA (F) and RhoH protein (G) as a function of HsInv0102 and rs7699141. (H) Manhattan plot for association of variants to RhoH protein expression association in LCLs, showing SNP rs7699141 and the combination rs7699141-HsInv0102 as sentinel pQTLs. Grey bar in manhattan plots indicate *RHOH* localization and HsInv0102 is represented with two dots corresponding to both breakpoints.

et al. 2016) and ENCODE ChIP-seq data (Gerstein et al. 2012), and the complete elimination of E8 by the inversion could have additional expression effects. In this regard, we found significant differences for *CD44*, which encodes a glycoprotein involved in cell interaction and migration (FDR-adjusted $P = 0.0357$). Importantly, both *RHOH* and *CD44* have been related to NF- κ B pathway and cancer (refs). To assess the reliability of these effects in *trans*, as negative control we permuted genotypes relative to expression and covariate phenotypes to break real relationships. Only 3 out of 100 permutations in the previous two approaches have significant results after FDR correction. In fact, if we compute a permuted-based FDR from this null distribution of expected P values for each observed association, same genes appeared as significant ($P \leq 0.02$). Although

Clusterin protein expression could not be checked, we also found a significant correlation between RhoH and CD44 protein levels ($r = \text{XXX}$; $P = 0.0176$; $n = 42$), indicating a low probability of spurious associations in our analysis. Moreover, we found that this CD44 and RhoH correlation was lost when individuals with the lowest RhoH levels (*Std1/Std1* and *Std1/Std2*) were removed from the regression ($P = 0.419$, $n = 26$), and this was not an effect of the smaller sample size since 93.6% of associations were stronger when we took the same number of individuals at random.

We did not find any association with HsInv0102 after FDR correction when the same analysis was repeated for the expression of 4,381 proteins in 62 YRI individuals from Battle et al. (2015). Since effects in *trans* may be mediated by the actual RhoH protein levels, we followed a complementary approach as for CD44 by correlating abundance of each distinct protein and RhoH. Only QPRT showed a slightly significant correlation (FDR-adjusted $P = 0.028$), but up to 34 proteins appeared close to the significance level ($\text{FDR} < 0.10$). We used GOrilla analysis tool (Eden et al. 2009) to search for enriched functional relationships within the top ranked list of proteins correlated with RhoH. ATP-dependent chromatin remodeling (FDR-adjusted $P = 3.51 \times 10^{-6}$), positive regulation of I κ kinase/NF- κ B signalling (FDR-adjusted $P = 0.0182$) or regulation of I κ B kinase/NF- κ B signalling (FDR-adjusted $P = 0.0290$) appeared among the most significant biological processes, whereas the most significant enriched cellular component was SWI/SNF superfamily-type complex (FDR-adjusted $P = 8.36 \times 10^{-9}$). Thus, RhoH appears to be correlated with multiple effects on different proteins and pathways such as NF- κ B and SWI/SNF, which highlights the importance of this gene and the potential functional role of HsInv0102.

3.3.3 HsInv0102 association with blood cancer

Next, given the known effects of the genes in several signaling pathways and cancer, we investigated the possible implication of the inversion with blood cancer susceptibility. HsInv0102 has no tag SNPs or other variants

in linkage disequilibrium (LD) ($r^2 > 0.8$) across multiple human populations. Nonetheless, in Europeans there are nine SNPs highly linked to HsInv0102 ($r^2 > 0.9$) and an almost tag SNP (rs7676043, $r^2 = 0.97$), Neither any of them nor variants in LD with rs7699141 in any population ($r^2 > 0.8$) were associated to phenotypic traits according to the GWAS Catalog (MacArthur et al. 2017). To determine whether the absence of phenotypic associations is due to the array coverage of variants tagging HsInv0102 and rs7699141, we first checked which SNPs in LD ($r^2 > 0.8$) with these variants were included in 76 typical genotyping arrays available through the LDLink web portal (Machiela and Chanock 2015). The only SNPs that could report useful information about HsInv0102 in Europeans were rs73132503 ($r^2 = 0.94$), just interrogated by 11 Illumina HumanOmni arrays, and rs73132504 ($r^2 = 0.94$), included in 3 arrays designed for non-European populations (Figure 3.2). In contrast, rs7699141 has informative variants in 39 of the arrays for European and East-asian populations. Second, we used IMPUTE2 (Howie et al. 2009) to infer HsInv0102 genotypes in 1000GP individuals using only variants included in 30 different arrays. Briefly, we excluded one sample at a time from the 434 genotyped individuals in common between with 1000GP, which were used as reference panel, and imputed HsInv0102 status with the information of the rest of samples. The linkage coefficient r^2 between real and imputed genotypes was used to evaluate imputation accuracy. We found that HsInv0102 could be accurately imputed ($r^2 > 0.8$) in European and East-Asian individuals with 10 and 7 of the selected arrays, respectively. Individuals with African ancestry gave the worst estimates in imputation, with only 5 arrays with enough precision ($r^2 > 0.8$) (Figure 3.2). As expected, HsInv0102 imputation was better when the array coverage (i.e. the number of SNPs located +/- 1 Mb from the inversion) was higher (Figure 3.2). We calculated that ~600 SNPs in HsInv0102 region are needed to impute the inversion with $r^2 > 0.8$ accuracy, but only 37% of the commonly-used arrays analysed here reach this number, with Illumina HumanOmni arrays being the best performing platform. However, even in European cohorts, the imputation in many cases has some errors and that since the inversion has not been imputed in any of the previous studies, its potential phenotypic effects have been missed in current GWAS

data. Therefore, these results show the necessity of genotyping directly the inversion through PCR or whole-genome sequencing (WGS) to obtain reliable associations.

Next, to study the association of HsInv0102 with blood malignancies, we retrieved whole genome sequencing (WGS) data provided by the International Cancer Genome Consortium - PanCancer Analysis of Whole Genomes (ICGC PCAWG). First, we genotyped the inversion in 150 individuals with CLL (CLLE-ES project) (Puente et al. 2015) by searching directly for the breakpoint sequences with BreakSeq (Lam et al. 2010; Lucas-Lledo 2014). Also, HsInv0102 was genotyped using different strategies in three ethnically matched control groups: 107 Iberians (IBS population from 1000GP), 94 samples of elder Spanish population (CNTPK) and 786 individuals from the GCAT Genomes for life Cohort with Spanish origin (Obón-Santacana et al. 2018). No inversion frequency discrepancies were found among control groups ($P > 0.2$). Although we did not observe either significant differences ($P = 0.098$) in *Inv* allele frequency between CLL patients (11.7%) and controls (15.5%), we found a significant underrepresentation of heterozygotes in CLL patients with CLL relative to controls (*Std/Std + Inv/Inv* vs *Inv/Std* over-dominant genetic model: $P = 0.033$). We also found a non-significant but consistent lower frequency of rs7699141 C allele (associated to lower E8 levels) in CLL patients if we take into account the effect of both these variants on RhoH expression (67.7% vs 71.1%; $P = 0.24$). Taking into account these results and the effect of HsInv0102 and rs7699141 as pQTLs, we found a significant enrichment of genotype combinations with higher E8 inclusion levels (*Std1/Std1* and *Std1/Std2*) in CLL patients (40.3% vs 49.3%; $P = 0.040$), indicating that the potential inclusion of E8 may be detrimental, possibly through their effect on RhoH protein levels.

To confirm a possible association with other types of cancer and replicate this result, we analysed the other available ICGC PCAWG blood cancer datasets: 101 individuals with malignant lymphoma (MALY-DE), 31 individuals with chronic myeloid disorders (CMDI-UK) and 9 with acute myeloid leukemia (LAML-KR) (Figure 3.3). In addition, we collected 344

controls from matched CEU, GBR and CHB populations from 1000GP. We observed a consistent underrepresentation of the *Inv* allele in patients of each blood cancer type (Figure 3.3), which became significant when we analyzed all groups together (meta-analysis, $P = 0.046$), indicating a moderate protective effect of the inverted allele (OR = 0.75). Again, we observed a lower number of heterozygotes was found in patients than controls in the meta-analysis, with significant results in the over-dominant model ($Std/Std + Inv/Inv$ vs Inv/Std : $P = 0.019$), a dominant effect of the *Inv* allele as a protective locus (Std/Std vs $Inv/Std + Inv/Inv$: $P = 0.020$) or just an additive effect (log-additive genetic model: $P = 0.036$). We did not find the same effect for rs7699141 across these diseases ($P = 0.414$), since rs7699141 C allele was in higher frequency in CMDI patients than GBR individuals, although this could be a consequence of the small sample size. However, the general trend of a protective effect of low E8 inclusion was maintained (OR = 0.92). When the analysis was done by HsInv0102 and rs7699141 genotype combinations, we found again a marginal overrepresentation of $Std1/Std1$ and $Std2/Std1$ in patients (OR = 0.78), which are the genotypes with highest E8 inclusion and lower RhoH protein levels (dominant model, $P = 0.069$; additive model, $P = 0.076$).

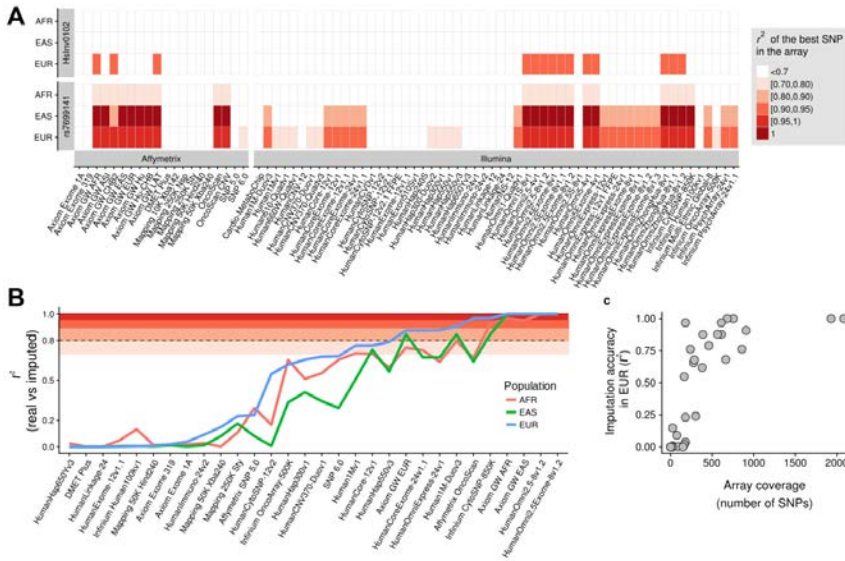


Figure 3.2 – Potential missed phenotypic associations of HsInv0102 by common genotyping arrays. (A) Coverage of SNP rs7676043 in high LD with HsInv0102 and SNP rs7699141 affecting E8 splicing in three population groups by commercial arrays from the LDLink web portal (Machiela and Chanock 2015). (B) Simulation of the imputation accuracy (r^2) of HsInv0102 using the SNP coverage of 30 different arrays in 1000GP samples from three population groups by IMPUTE2. (C) Correlation between HsInv0102 imputation accuracy (r^2) in Europeans and array coverage represented as the number of SNPs interrogated around 1 Mb from the inversion.

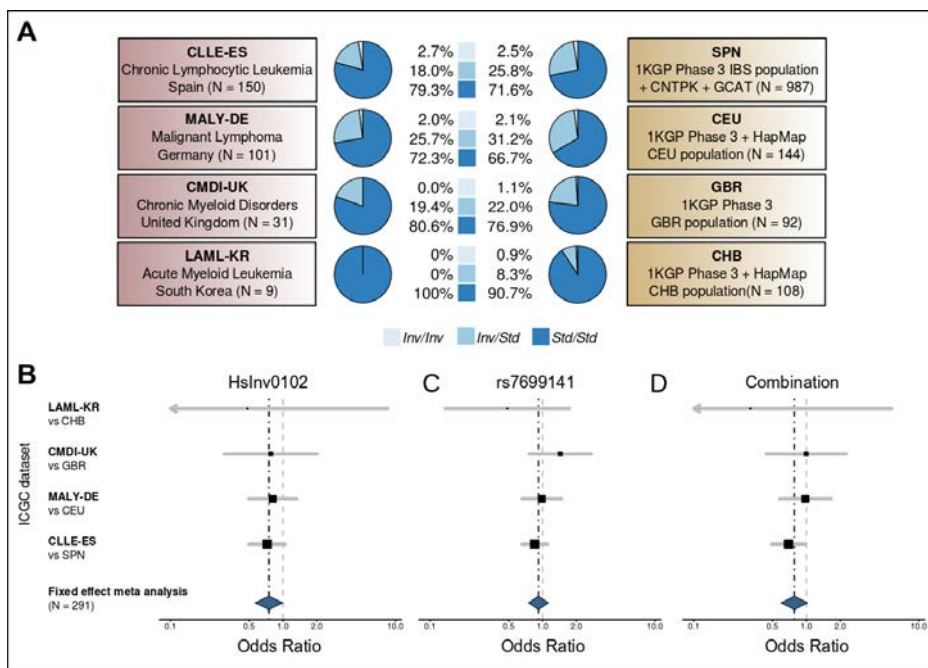


Figure 3.3 – Possible association of HsInv0102 with blood cancer susceptibility. (A) Cases and controls employed for association analysis showing the frequency of HsInv0102 genotypes. (B) Forest plots for the meta-analysis of the association between blood malignancies and genotypes for HsInv0102, rs7699141 and the combination of both variants.

3.4 A human inversion influences antiviral response through different regulatory effects on interferon response genes

Inversion HsInv0124 constitutes an interesting example of a human inversion with evidence of functional effects and possible positive selection. It is located on chromosome 11 within the interferon (IFN)-induced transmembrane protein (*IFITM*) genes locus, including the *IFITM1*, *IFITM2* and *IFITM3* genes (Aguado et al. 2014). These genes are relatively small and main isoforms encoding IFITM proteins consist of two coding exons interspersed by an intron. Paralogous genes *IFITM2* and *IFITM3* are embedded in the segmental duplications where inversion breakpoints are located, but apparently they are not exchanged by the inversion, whereas *IFITM1* is located within the inverted region and change its orientation with the inversion. HsInv0124 has been already reported to be lead eQTL of the non-coding RNAs that overlap with its breakpoints and it could be also associated to expression changes in *IFITM2* and *IFITM3* in LCLs and other tissues (Giner-Delgado et al. 2019), but the potential regulatory effects of the inversion have not been well characterized yet. Interestingly, this inversion shows clear frequency differences among populations, having intermediate frequencies in individuals of European ancestry and being almost fixed in Africans and especially East Asians (Figure 3.5). However, since it has long and highly-identical inverted repeats at the breakpoints (Aguado et al. 2014; Giner-Delgado et al. 2019), it has been missed by most studies describing structural variants in humans (The Genome of the Netherlands Consortium 2014; Sudmant et al. 2015; Chaisson et al. 2019).

IFITM family members have been established as potent antiviral restriction factors of multiple pathogenic infections (Bailey et al. 2014). Initially, *IFITM3* RNA interference knockdown was shown to promote influenza virus infection in cells, whereas its overexpression, together with that of *IFITM1* and *IFITM2*, suppressed not only influenza but also flaviviruses replication (Brass et al. 2009). The importance of *IFITM3* was sub-

sequently confirmed by Everitt et al. (2012), who discovered that mice lacking *Ifitm3* suffered a much severe influenza infection compared to the wild-type littermates. These findings were supported and validated by Bailey et al. (2012), who described that knockout heterozygotes exhibited an intermediate phenotype between knockout and wild-type, indicating a gene-dosage effect. Further analyses demonstrated that IFITM proteins have a pervasive inhibitory activity that can affect a comprehensive list of viruses (Schoggins et al. 2011; Perreira et al. 2013), including among many others Zika virus (Savidis et al. 2016; Monel et al. 2017), human immunodeficiency virus (HIV) (Lu et al. 2011), Ebola virus (Huang et al. 2011) and hepatitis C virus (Wilkins et al. 2013), or even bacterial infection by *Mycobacterium tuberculosis* (Ranjbar et al. 2015). Although the exact inhibitory mechanism remains to be elucidated, IFITMs seem to restrict viral membrane hemifusion by modifying membrane fluidity (Li et al. 2013) or by blocking the formation of fusion pores (Desai et al. 2014). Remarkably, IFITM1, IFITM2 and IFITM3 exhibit diverse specificities toward viruses and it has been proposed that their different subcellular localization may be important for blocking virus entry (Lu et al. 2011; Mudhasani et al. 2013). In this regard, while IFITM1 tends to be at the plasma membrane, IFITM2 and IFITM3 are mainly found in endosomal compartments and are more active against those viruses that enter inside late endocytic vesicles instead that through the cell surface (Feeley et al. 2011; Amini-Bavil-Olyaei et al. 2013; John et al. 2013; Desai et al. 2014; Savidis et al. 2016). For example, Foster et al (2016) revealed that HIV-1 infection susceptibility depends on which coreceptor is engaged, since IFITM2 and IFITM3 were more active against HIV-1 strains using CXCR4 as cofactor, whereas CCR5-tropic strains exhibited sensitivity to IFITM1. Moreover, CCR5-tropism was also sensitive to IFITM2 and IFITM3 when these proteins were relocalized to cell surface, while X4-tropic restriction disappeared, confirming that the sensitivity of IFITM proteins depends on the route of virus entry. The variety and diversification of IFITM genes may therefore result in a broad-spectrum viral restriction activity and is likely the result of positive selection (Zhang et al. 2012; Compton et al. 2016).

Given the relevance of the *IFITM* family, deciphering how genetic variation in these genes is linked to viral infectious diseases is of special interest. In this regard, an enrichment of the synonymous SNP rs12252 C allele in *IFITM3* was first found in 53 hospitalized adults during the 2009 H1N1 influenza pandemic and CC homozygotes were associated with lower levels of IFITM3 protein expression and with an increased susceptibility to infection and flu severity (Everitt et al. 2012). Although the underlying mechanism remains elusive, the rs12252 risk allele would supposedly alter an *IFITM3* splice acceptor site, leading to a 21-amino acid truncated protein (IFITM3-N21Δ). Higher rs12252-C allele frequency among patients with severe influenza infection was confirmed in two Chinese cohorts, where this allele is much more prevalent (Zhang et al. 2013; Wang et al. 2014). Besides, the same allele was over-represented in HIV-infected Chinese patients with rapid disease progression (Zhang et al. 2015). Despite two meta-analyses suggested a significant association between rs12252 T>C polymorphism and influenza risk in both Asian and Caucasian populations (Prabhu et al. 2018; Chen et al. 2018), recent studies show that CC-genotype carriers do not generate a truncated IFITM3 protein (Randolph et al. 2017; Makvandi-Nejad et al. 2018) and that the truncated form restricts virus entry and replication as well (Williams et al. 2014). These and other studies that did not find a clear association of rs12252 C allele with mild or severe influenza infection (Mills et al. 2014; López-Rodríguez et al. 2016; Randolph et al. 2017) have challenged the real impact of this allele on virus restriction. Novel analyses identified rs34481144 A allele, which is located at *IFITM3* 5' UTR, as a risk variant associated with severe influenza infection in three cohorts, but also correlated with a lower number of antiviral CD8⁺ T cells in patient airways during infection (Allen et al. 2017). This variant can decrease *IFITM3* mRNA levels by reducing IRF3 and increasing CTCF binding to the promoter region, and may also alter transcriptional landscape of neighboring genes due to possible CTCF effects on chromatin topology (Allen et al. 2017). However, additional variants might also be important for *IFITM* genes function in virus defense.

In this work, we make a complete functional characterization of inver-

sion HsInv0124. This inversion changes the promoters of two non-coding RNAs and moves the position of histone modification peaks affecting the regulation of *IFITM* genes. Finally, we show that these changes have consequences in viral restriction pathways and we investigated the inversion role on virus infection susceptibility.

3.4.1 HsInv0124 frequency and distribution

It has already been described that HsInv0124 is not in linkage disequilibrium (LD) ($r^2 > 0.8$) globally with other variants (Giner-Delgado et al. 2019), although there are some SNPs in high LD with the inversion in Europeans and East Asians (Figure 3.4). To obtain a more comprehensive view of its distribution across the world, we extended the existing experimentally-validated genotypes in 551 individuals from European (CEU, TSI), African (YRI, LWK), East-Asian (CHB, JPT) and South-Asian (GIH) ancestry by imputing HsInv0124 in all 1000GP populations. For that, we evaluated imputation accuracy with IMPUTE2 by excluding one sample at a time from the 431 individuals in common with 1000 Genomes project (1000GP) Phase 3 (The 1000 Genomes Project Consortium 2015) and inferring the inversion status employing the rest of individuals as reference panel. As shown in Figure 3.4, the inversion could be imputed accurately (Howie et al. 2009), especially in those populations with other variants in high LD and consistent inversion frequencies were obtained to those generated experimentally. We also used an alternative variant set from sequenced individuals from Spanish origin provided by the Genomes For Life (GCAT) project (Obón-Santacana et al. 2018) to verify inversion calling for other datasets. In all GCAT samples analysed (120/120) the same genotype was obtained by iPCR and in silico, demonstrating the reliability of the imputation.

Calling the inversion in the 26 populations from 1000GP allowed us to obtain more detailed geographical distribution patterns (Figure 3.5). As already described (Giner-Delgado et al. 2019), the *O1* orientation identical to that in the hg18-hg38 genome is the predominant form in Europeans,

whereas the alternative *O2* orientation is much more frequent in the rest of the world, being almost fixed in East Asia, where no *O1/O1* homozygotes were found in 504 individuals. Interestingly, the *O1* orientation is found at highest frequency in northern Europe regions (69% FIN and GBR) and it lowers to the south (60% CEU, 51% TSI, 57% IBS). It is also seen in America at moderate levels, which probably indicates European admixture, and in South-Asian populations. However, it is not clear if this distribution pattern has been created by some kind of positive selection.

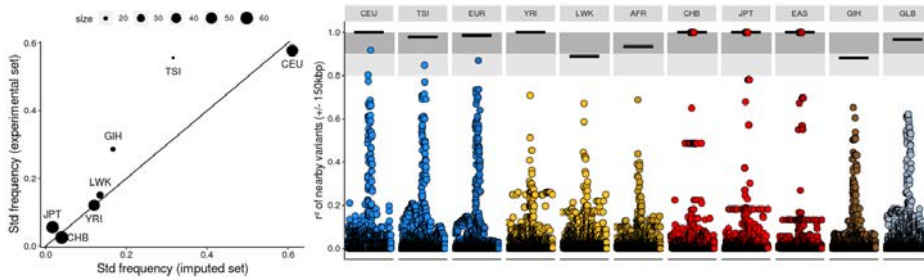


Figure 3.4 – HsInv0124 LD patterns and imputation accuracy. (Left) Position of nearby 1000GP variants (± 150 kb) and LD (r^2) with HsInv0124 across populations in which it has been experimentally genotyped. Imputation accuracy is estimated with r^2 between imputed and real HsInv0124 genotypes, indicated with a horizontal segment. (Right) Agreement between HsInv0124 *O1* allele frequency by experimental genotyping and imputation in all 1000GP Ph3 samples. The line represents the 1:1 relationship.

3.4.2 HsInv0124 impact on gene expression in LCLs

As already mentioned, HsInv0124 inverts protein coding gene *IFITM1*, while *IFITM2* and *IFITM3* are placed close to its breakpoints. Moreover, multiple lncRNAs are also expressed from this region, although the real structure of these transcripts remains elusive due to repetitive sequences at *IFITM* locus. In fact, isoforms described by the GENCODE project (Harrow et al. 2012) do not match with miTranscriptome (Iyer et al, 2015) or FANTOM (Hon et al. 2017) annotations, which described many gene

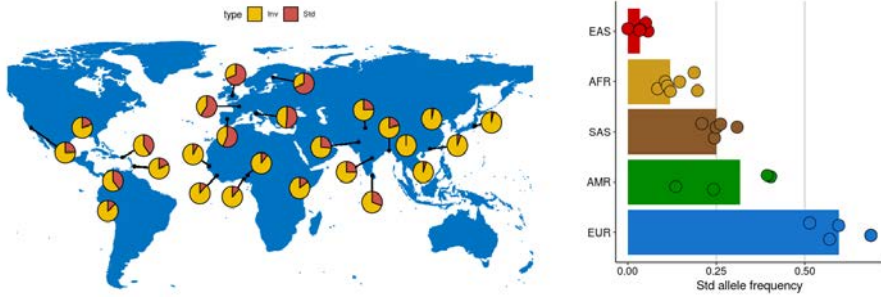


Figure 3.5 – HsInv0124 global distribution across 26 populations from 1000GP Ph3. (Left) Map of the worldwide HsInv0124 distribution, with the *O1* or *O2* orientation represented in red or yellow, respectively. (Right) Summary of *O1* allele frequency in different populations grouped by continents.

fusions. Thus, we took advantage of 101-bp, paired-end, strand-specific reads from RNA sequencing (RNA-seq) of 10 LCLs from five homozygotes for each inversion orientation, after IFN stimulation to ensure high expression from this locus (see below), to carry out a *de novo* transcriptome in each sample and then combine all of them together to define a reliable collection of transcribed isoforms. Our final assembled transcriptome contained 14 genes grouping 40 distinct transcripts (Figure 3.6). Half of the transcripts have structures that exactly match GENCODE annotated isoforms or include alternative splicing sites (8 and 12, respectively). Additionally, 7 (17.5%) were completely novel transcripts, all of which except one were antisense transcripts of *IFITM* genes. Finally, 13 GENCODE transcripts from this region could not be validated (32.5%), but they highly resembled the structures observed in our assembly and they were included in the final transcriptome in case they exist at low levels. This new annotation was well supported by uniquely mapped reads that were not confounded by repeated sequences. However, we excluded a potential *IFITM2-IFITM1* fusion transcript that would be broken by the inversion, since we did not find any evidence of its presence by aligning reads to the junction and by paired reads that map in *IFITM2* and *IFITM1*. This transcript is based in just one metastatic chondrosarcoma spliced EST (CA443766.1) that could align equally well in *IFITM1* with-

out generating a fusion transcript (the 31 bp that in theory are specific to *IFITM2* can map in *IFITM1* with just one change, corresponding to a very rare but existent polymorphism (rs763098917), so the whole sequence could map perfectly in *IFITM1*). Thanks to this annotation, we recovered an average of 180 more mapped reads from the 10 stimulated LCLs than using only GENCODE version, which, excluding *IFITM1*, *IFITM2* and *IFITM3* genes where a vast majority of reads map ($\sim 92\%$), it means on average a ~ 1.5 -fold higher number of mapped reads in the rest of genes. Therefore, this confirms the reliability of the new annotation and it was used for the downstream analysis.

We next tested for associations between HsInv0124 and expression of nearby genes or transcripts by linear regression between inversion genotypes (experimental and imputed) and LCL transcriptome data of 445 individuals from the Geuvadis consortium (Lappalainen et al. 2013). Expression was adjusted by the three first principal components of the genetic data to account for population membership and structure, gender, the laboratory of sequencing and a variable number of factors to capture technical confounding effects. *cis* eQTL mapping was carried out within 1.5 Mb of each transcript surrounding the inversion together with all neighboring 1000GP variants (MAF > 0.05). We found that HsInv0124 is the lead eQTL of *AC136475.2* and *AC136475.1* non-coding genes located at its breakpoints, as shown before (Giner-Delgado et al. 2019) (Figure 3.7). Interestingly, these are copies of the same gene located at the inverted segmental duplications, and the inversion effect may be caused by an exchange of the promoter region. In fact, in *O2/O2* individuals *AC136475.1* expression is up-regulated and *AC136475.2* down-regulated. Additionally, a novel discovered isoform (*TN2*), which is an *IFITM3* antisense transcript, also showed significant expression changes, indicating that HsInv0124 clearly affects several isoforms transcribed close to its breakpoints.

Since *IFITM* genes can modulate viral restriction response and are activated after infection, we searched for genome-wide expression changes associated to HsInv0124 by analyzing RNA-Seq data from the 10 LCL

samples stimulated with IFN (Figure 3.8). Since just five *O1/O1* and *O2/O2* individuals were compared, we controlled by low expressed genes and genes with high variation in expression that may bias the results. We employed DESeq2 (Love et al. 2014) to contrast gene expression between *O1/O1* and *O2/O2* samples. Consistent with Geuvadis analysis, a \log_2 fold change of -1.38 ($P = 7.82 \times 10^{-9}$) and 2.07 ($P = 1.55 \times 10^{-6}$) were found for *AC136475.2* and *AC136475.1* lncRNAs between *O2* and *O1* homozygotes. Within the HsInv0124 region, novel non-coding transcripts *TN1* and *TN3* also showed significant differences between inversion genotypes. After filtering genes with expression outliers, we detected 87 differentially-expressed genes across the genome (adjusted $P < 0.1$) with moderate fold changes (>95% of genes with \log_2 fold change below -3.24 and above 3.50), of which 24 were down-regulated and 62 up-regulated. Next, we used GOrilla enrichment analysis tool (Eden et al. 2009) to detect functional relationships among these differentially expressed genes, finding two enriched GO biological process categories (adjusted $P < 0.05$): type I interferon signaling pathway and defense response to virus. To confirm this enrichment, we permuted five times the list of genes uncovering no statistically significant gene ontology (GO) terms. Thus, we selected 6 genes involved in the interferon signaling pathway with different fold-change and significance values to validate their expression using real-time quantitative PCR. All these genes showed good validation. Remarkably, some of these genes did not show expression changes in non-stimulated LCLs, which could indicate that differences are enhanced after cell stimulation.

3.4.3 HsInv0124 and epigenetic changes

Since the activity of many genes is orchestrated by epigenetic regulation, we next investigated how HsInv0124 affects three well-studied histone modification marks (H3K4me1, H3K4me3 and H3K27ac) that are known to capture enhancer and promote activity, using available ChIP-seq data (Delaneau et al. 2019). Interindividual correlation between chromatin peaks showed that HsInv0124 is within a previously described *cis*-

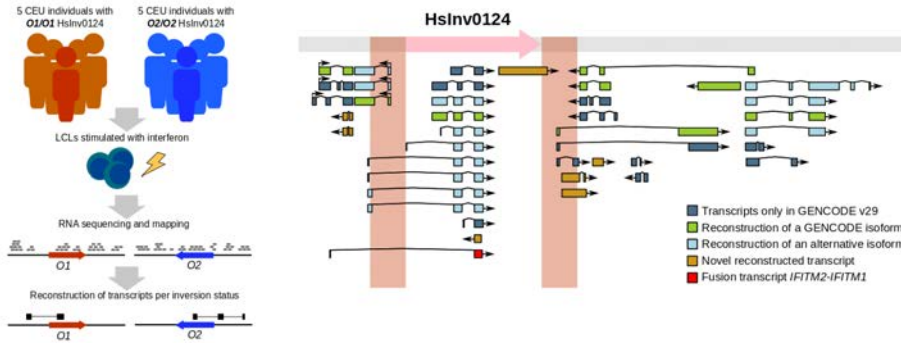


Figure 3.6 – *De novo* transcriptome annotation around inversion **HsInv0124**. (A) Overview of the experimental design to generate high-coverage RNA-seq reads from 5 $O1/O1$ and 5 $O2/O2$ LCL samples stimulated with IFN. (B) Diagram showing the structure of the final set of 40 transcripts present in the *IFITM* locus. Colors depict how each transcript has been annotated (see Methods) and arrows indicate the direction of transcription. A predicted *IFITM2-IFITM1* fusion transcript in GENCODE is also indicated, although it is likely just generated from *IFITM1*.

regulatory domain (CRD) (Delaneau et al. 2019), i.e. a delimited set of chromatin elements with coordinated activity. By analysing the effect of HsInv0124 on histone modifications from 145 non-stimulated LCLs derived from European individuals, we identified that the inversion is the lead variant associated with the global CRD activity (aCRD-QTL) (Figure 3.9). Although some histone modification peaks differ from the general trend, overall chromatin signals are lower in $O2/O2$ individuals, suggesting weaker CRD activity in this orientation. When testing each specific chromatin peak within the CRD, HsInv0124 appeared as chromatin QTL (cQTL) of three chromatin peaks that fall close to its breakpoints (Figure 3.10). To understand the mechanism by which the inversion affects histone modification levels around the breakpoints, we analyzed ChIP-seq profiles of 10 $O1/O1$ and 6 $O2/O2$ experimentally genotyped individuals. The presence of H3K4me1, H3K4me3 and H3K27ac marks at both sides of the inversion was inversely correlated, i.e. with the change of orientation the histone activity of one side increases whereas there is a decrease in the

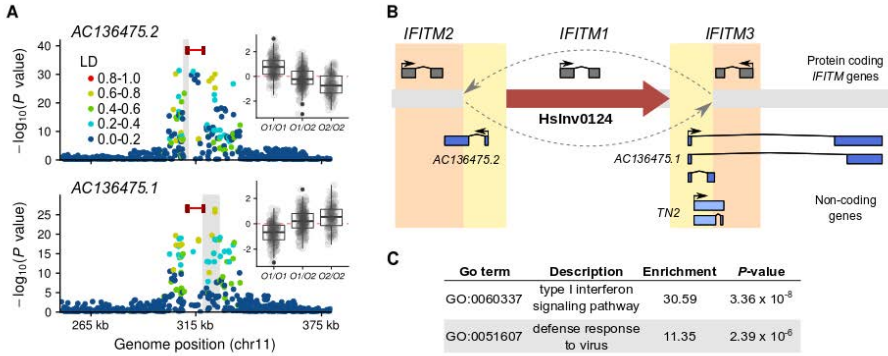


Figure 3.7 – Regulation of gene expression by HsInv0124 in LCLs. (A) Manhattan plots for eQTL associations of non-coding genes *AC136475.1* and *AC136475.2* (represented in gray) with 1000GP variants (dots) and HsInv0124 (red line with terminal rectangles representing the breakpoints) in 454 LCLs from Geuvadis project, together with boxplots of normalized expression and inversion genotype. HsInv0124 is shown as the lead variant. Each neighboring variant is coloured according to LD with the inversion. (B) Organization of the *IFITM* locus, with the structure of the non-coding genes affected by the inversion represented below (blue boxes) and the *IFITM* genes (grey boxes) above. Inverted segmental duplications are depicted in orange, while inversion breakpoints within them are in yellow. (C) Significant GO biological process term enrichment among differentially-expressed genes between 5 *O1/O1* and 5 *O2/O2* LCL samples stimulated with IFN obtained with GOrilla tool (Eden et al. 2009).

other flank (Figure 3.11). This effect could be related to the extension of a strong histone modification peak located within the inversion to the surrounding regions (Figure 3.12).

Interestingly, *IFITM2* and *IFITM3* TSSs are within these peaks and also overlap with H3K4me3 marks, known to tag promoters. In this regard, *O2* orientation is associated to an increase in H3K4me1 levels at *IFITM3*, while this HsInv0124 orientation is linked to and a depletion in the levels of both H3K4me3 and H3K27ac at *IFITM2* region (Figure 3.10). To

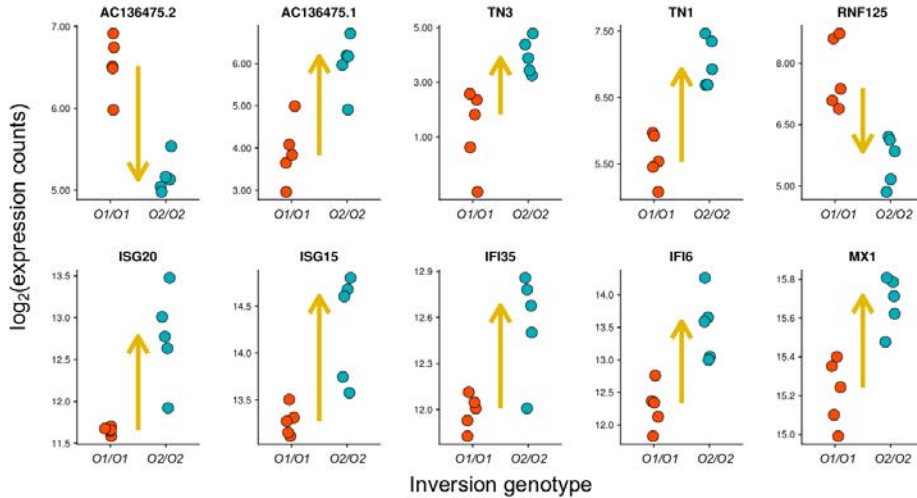


Figure 3.8 – Differentially-expressed genes associated to HsInv0124 in LCLs stimulated with IFN. Expression counts obtained by RNA-seq for genes located close to HsInv0124 (*AC136475.2*, *AC136475.1*, *TN3* and *TN1*) and related to type I interferon pathway (*RNF125*, *ISG20*, *ISG15*, *IFI35*, *IFI6* and *MX1*). Arrows indicate the direction of the expression change between the median values obtained for the homozygotes with each inversion orientation.

characterize the interplay between HsInv0124, epigenetic signals and gene expression, we correlated each of these histone modifications affected by HsInv0124 with all genes at the *IFITM* locus. We identified 12 highly significant associations between chromatin activity and expression ($P < 0.001$) for six genes, including non-coding genes *AC136475.2*, *AC136475.1*, *TN5* and *TN1*. However, the strongest associations with histone peaks were detected for *IFITM2* and *IFITM3* expression levels, showing that HsInv0124 can perturb potential regulatory elements, which in turn results in variable activity of *IFITM* genes. In fact, H3K4me1 changes explained about 17% of the variation in *IFITM3* expression, while those in H3K4me3 and H3K27ac explained $\sim 16\%$ and 32% of *IFITM2* variation, respectively. Since stronger histone modification peaks correlate with higher expression levels, *O2* orientation is therefore linked to higher *IFITM3* and reduced *IFITM2* transcription. Additionally, *IFITM2* expression was strongly as-

Chapter 3. Results

sociated with global CRD activity ($P = 2.55 \times 10^{-8}$), but *IFITM1*, *TN2*, *TN5*, *AC136475.1*, *AC136475.2* and *AC135475.7* were significantly correlated as well ($P < 0.01$), confirming the pervasive effects of HsInv0124 on epigenetic changes and gene expression.

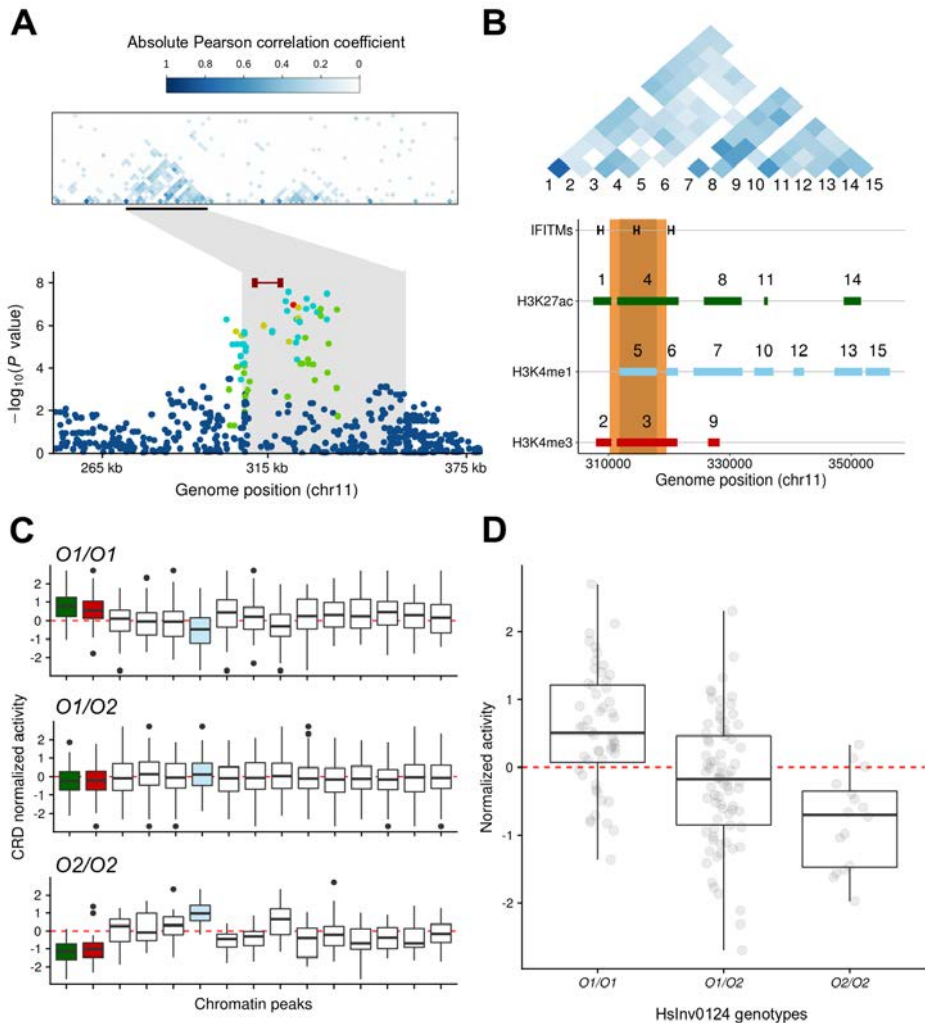


Figure 3.9 – Association of HsInv0124 to a *cis*-regulatory domain (CRD) activity. Caption on the following page.

Figure 3.9 – Association of HsInv0124 to a *cis*-regulatory domain (CRD) activity. (A) HsInv0124 affects a CRD placed at the *IFITM* locus. The upper panel shows the map of the inter-individual correlations between chromatin peaks in a small region of chromosome 11 around the inversion showing the CRD as a blue triangle labeled with a black bar. Below, manhattan plot for CRD-QTL association in LCLs, showing HsInv0124 (red line with terminal rectangles representing the breakpoints), as the top variant. (B) Detail of the previous panel, showing the pairwise correlations between the 15 chromatin peaks that form part of the CRD (represented with the same scale as in A) and the genomic localization of the peaks with respect to the three *IFITM* genes. (C) Distribution of the normalized activity of all chromatin peaks contained within the CRD and stratified by inversion genotype. Coloured boxplots indicate histone peaks where HsInv0124 is the lead variant affecting their activity. (D) Boxplots of normalized global CRD activity by inversion genotype.

3.4.4 HsInv0124 effect on gene expression under infection

As we have seen, different genes related to the interferon pathway change their expression profile under interferon stimulation in association with the inversion orientation. To check the potential effects of HsInv0124 under infection, we retrieved RNA-Seq data collected in non-stimulated and stimulated CD14⁺ monocytes derived from 100 individuals with European origin (Quach et al. 2016). Stimulated monocytes were exposed to bacterial lipopolysaccharide (LPS), Pam3CSK4 (which triggers antibacterial immune responses), R848 compound (responsible for sensing viral nucleic acids), and to human influenza A virus (IAV) responsible of seasonal influenza. We imputed HsInv0124 status in these samples and carried out an eQTL mapping approach as before. Inversion HsInv0124 appear as lead variant of *IFITM3* expression in 4 of the 5 conditions checked, including non-stimulated monocytes and cells exposed to LPS, R848 and IAV (Figure 3.14). In the case of Pam3CSK4, the lead variant was rs7944394, which is in high LD with rs34481144 ($r^2 \approx 0.8$), an SNP already pos-

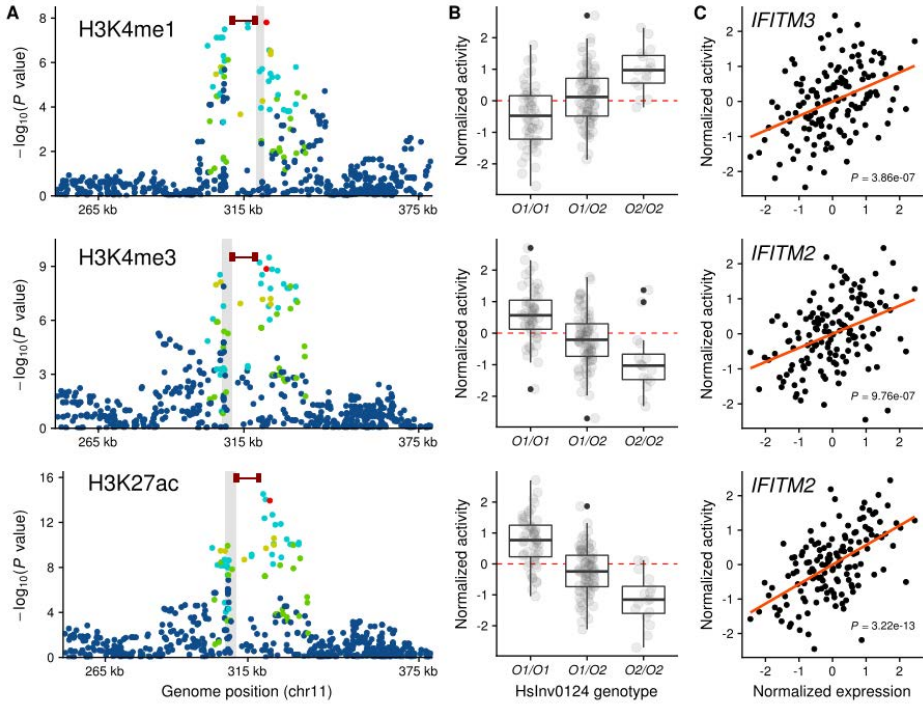


Figure 3.10 – HsInv0124 regulates gene expression in *IFITM* locus through histone modification marks in LCLs. (A) Manhattan plots for hQTL associations of histone modification peaks in LCLs with 1000GP variants (dots) and HsInv0124 (red line with terminal rectangles representing the breakpoints), showing the inversion as lead variant. 1000GP variants are coloured according to LD with the inversion and the peak location is indicated in grey. (B) Boxplots of normalized signal of different histone modifications by inversion genotype. (C) Correlation between the activity of each chromatin peak and gene expression of *IFITM2* and *IFITM3* genes.

tulated to modulate *IFITM3* levels (Allen et al. 2017). Remarkably, inversion HsInv0124 was the lead eQTL of 7 genes under IAV infection, including *IFITM3*, *AC136475.2*, *TN1*, *AC136475.1*, and also new genes not discovered before such as *IFITM5*, *NLRP6* and *AC136475.6* (Figure 3.13). All these genes were up-regulated when exposed to IAV than in non-stimulated conditions. Curiously, in LCLs, *AC136475.1* and *AC136475.2*

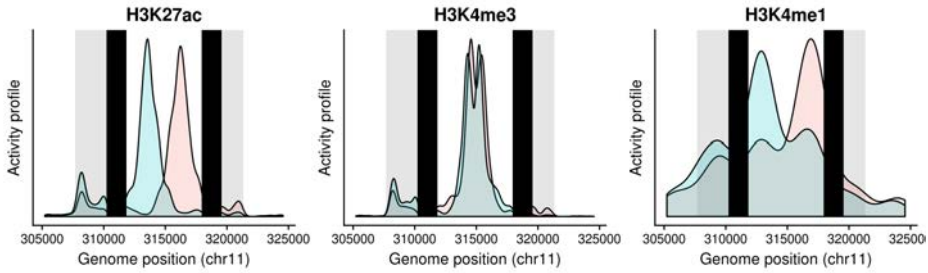


Figure 3.11 – Detail of *Inv0124* affecting histone modification levels. Chromatin activity profiles measured by ChIP-seq reads and mapped in an *O1* genome (blue pattern) and in an *O2* predicted genome (red pattern). Density plots are based on data from 10 *O1/O1* and 6 *O2/O2* experimentally genotyped individuals.

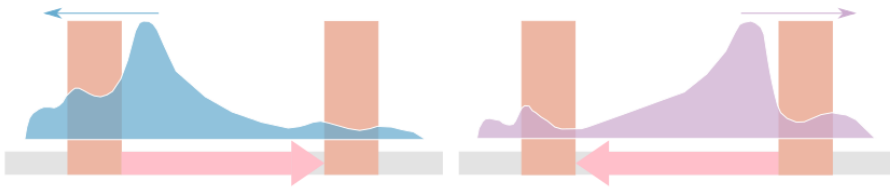


Figure 3.12 – Potential mechanism by which *HsInv0124* is affecting histone modification levels in the surrounding region. Potential mechanism of the effect of the inversion on histone modification peaks. The change of orientation of the inverted region apparently increases the chromatin activity close to one breakpoint due to the proximity of the peak within the inversion, while in the other side the chromatin mark outside the inversion is decreased. The orange rectangles represent the segmental duplications flanked the inversion, shown as a pink arrow.

expression was increased and decreased, respectively, in *O2/O2* individuals, whereas the opposite is happening in $CD14^+$ monocytes following IAV treatment and expression of the two lncRNAs is almost undetectable in non-stimulated monocytes. Interestingly, protein-coding genes *IFITM5*, *NLRP6* and *IFITM3* displayed higher levels of expression in cells with the *O2* allele. *IFITM5* is a member of the *IFITM* family that has not been shown to be interferon inducible. In fact, *IFITM5* has been de-

scribed to play a role in bone mineralization and mutations in this gene has been associated with osteogenesis imperfecta type V (Semler et al. 2012). However, while *IFITM5* remains at low levels in *O1/O1* monocytes treated with IAV comparable similar to those in non-stimulated cells, its expression is highly increased in cells with the *O2* orientation. *NLRP6* encodes an intracellular protein that has been associated with a large number of immunological roles, including antiviral immunity, signaling or host-microbiome interaction regulation (Levy et al. 2017). Additionally, several *IFITM1* and *IFITM2* transcripts appeared to be affected by the inversion. Altogether, these findings indicate a strong cell type and context specificity of gene-expression regulation in this region, and they demonstrate that HsInv0124 has pervasive effects on gene expression, especially under immune response and viral infection conditions.

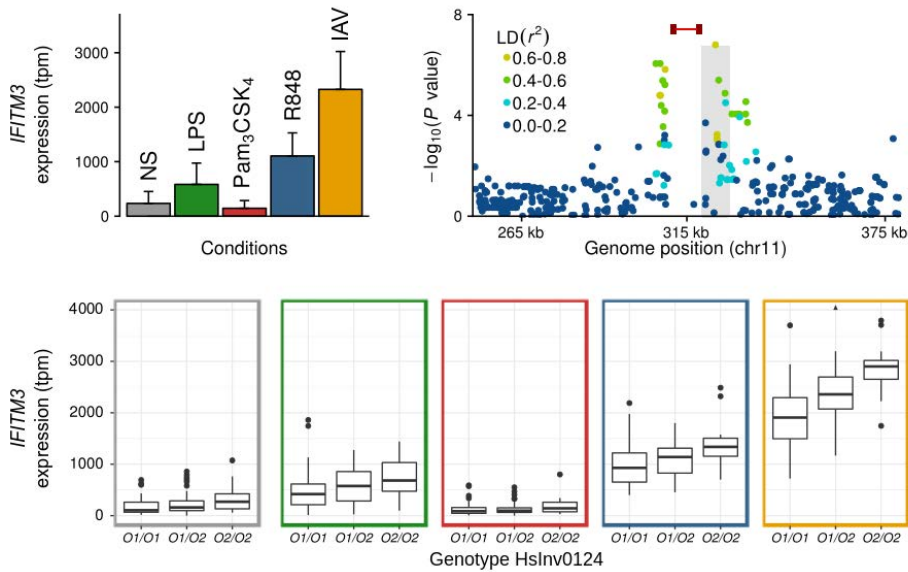


Figure 3.13 – HsInv0124 is associated to *IFITM3* expression in CD14⁺ monocytes. We represent three panels: (i) expression of *IFITM3* measured in transcripts per million (tpm) in non-stimulated monocytes (NS) and under different stimuli (LPS, Pam3CSK4, R848 and IAV) from 100 healthy donors of European origin; (ii) manhattan plot for eQTL association with *IFITM3* in monocytes stimulated with IAV, showing HsInv0124 (red line with terminal rectangles representing the breakpoints) as lead variant; (iii) Boxplots of *IFITM3* expression (tpm) by inversion genotype in non-stimulated and stimulated monocytes with LPS, Pam3CSK4, R848 and IAV.

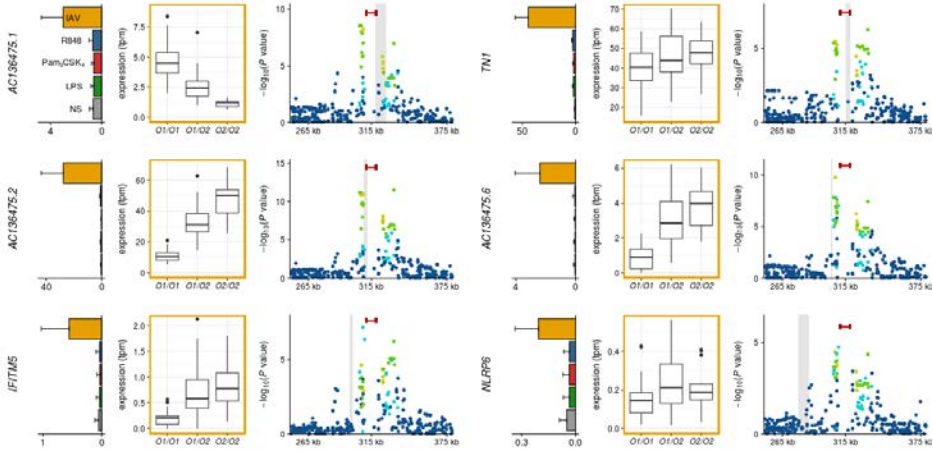


Figure 3.14 – Pervasive effects of HsInv0124 on gene expression in monocytes stimulated with IAV. For each gene, we represent three panels: (i) expression measured in transcripts per million (tpm) in non-stimulated monocytes (NS) and under different stimuli (LPS, Pam3CSK4, R848 and IAV) from 100 healthy donors of European origin; (ii) boxplot of expression levels (tpm) by inversion genotype in stimulated monocytes with IAV; and (iii) Manhattan plot for eQTL associations in stimulated monocytes with IAV, showing HsInv0124 (in red) as lead variant.

3.5 Comprehensive analysis of the influence of human inversions on gene expression, epigenetic changes and phenotypic variation

3.5.1 Inversion set and genotype calling

Inversions are known to be a challenging class of structural variant to detect and genotype accurately. Thanks to different inversion genotyping efforts, in previous studies we analysed the effect of 61 inversions on gene expression mainly in LCLs, but also indirectly in other tissues and conditions through tagging variants (Giner-Delgado et al. 2019; Puig et al. 2019). One of the main problems is that existing inversion predic-

tions tended to include a extremely high rate of false positives (Martínez-Fundichely et al. 2014; Vicente-Salvador et al. 2017;). Therefore, to extend these analysis to the maximum number of possible inversions, we searched for evidences of inverted sequences in publicly available reconstructed sequences from human genome assemblies for already predicted inversions by different methods (Chaisson et al. 2015; Sudmant et al. 2015; Hehir-Kwa et al. 2016; Huddleston et al. 2016). In total, we experimentally validated 50 additional inversions, which were subsequently genotyped in populations from three continents (EUR, AFR and EAS).

From the entire set of 109 autosomal and chr. X validated inversions (excluding two in chr. Y), 63 and 54 have tag variants ($r^2 = 1$), respectively, in genotyped European and African populations included in 1000GP Ph3 (Figure 3.15). Additionally, 45 inversions have variants in perfect LD in both African and European populations and this number increases to 63 if we considered those inversions that are monomorphic in one of these populations. Tag variants make possible to estimate the genotype of these inversions that are untyped in cohorts of interest. For the rest of inversions without a perfect proxy, we evaluated whether they could be accurately imputed using the combined information of neighboring variants. Individuals from 1000GP Ph3 with experimental inversion genotypes were used to generate two reference panels of genomic variation with distinct ancestry: EUR and AFR. Next, we carried out a leave-one out strategy by excluding one genotyped individual at a time and inferring subsequently inversion status from the remainder sample set (see Methods). We found that 13 and 16 (out of 36 and 41) inversions from European and African panels, respectively, can be imputed *in silico* with enough precision (r^2 between imputed and experimental genotypes > 0.8) (Figure 3.15). If we consider the imputation performance globally, only 13 out of 44 inversions can be accurately imputed in both populations, pointing out that the genotypes of approximately two thirds of non-tagged inversions cannot be inferred (Figure 3.15). Moreover, since they are based in whole-genome sequences, these findings give an upper estimate of imputation accuracy, and low-density SNP arrays commonly used in association studies are expected to result in lower imputation performance rates. The maximum LD observed

between inversions and neighbouring variants highly correlated with imputation accuracy ($P = 2.78 \times 10^{-8}$ in EUR; $P = 3.15 \times 10^{-4}$ in AFR) (Figure 3.15). However, there are several other factors that can influence these estimates, including the influence of other variants in moderate to high LD or inversion frequency. There are notable exceptions in which the highest linked variant reports inversion genotypes with more accuracy than imputation, such as in HsInv0072, whereas the opposite happens for HsInv0390 in EUR, which benefits from imputation. Also, HsInv0393 and HsInv0403 have highly linked variants ($r^2 > 0.8$) in EUR and these were used as proxy together with imputable and tagged inversions for downstream analysis.

3.5.2 The impact of human inversions on gene expression and epigenetics

We performed a detailed analysis of inversions acting as eQTLs (INV-eQTLs) by testing associations between inversion genotypes and gene expression in 45 tissues, 11 brain subregions and 34 other body tissues and organs, as well as two cell lines, cultured fibroblasts and EBV-transformed lymphocytes, from the GTEx project (GTEx Consortium 2017). Transcriptome RNA-Seq data was retrieved from the GTEx portal and normalized by population structure and technical confounders (see Methods). We performed a joint eQTL mapping including nearby variants (up to 1 Mb from the TSS) to measure the relative contribution of each polymorphism to gene expression variation. We used a *cis* window of 1 Mb on either side of a gene TSS that included at least one autosomal or chr. X inversion. As shown in Table 3.1, we only report INV-eQTL associations that are lead variants or in high LD ($r^2 > 0.8$) with top markers at 5% FDR.

This analysis confirmed previously-reported effects of HsInv0573 and HsInv0786, which are known to harbour several functional changes that are supposed to be driven by variants within the inversion (de Jong et al. 2012; González et al. 2014, Puig et al. 2019). Interestingly, we also found

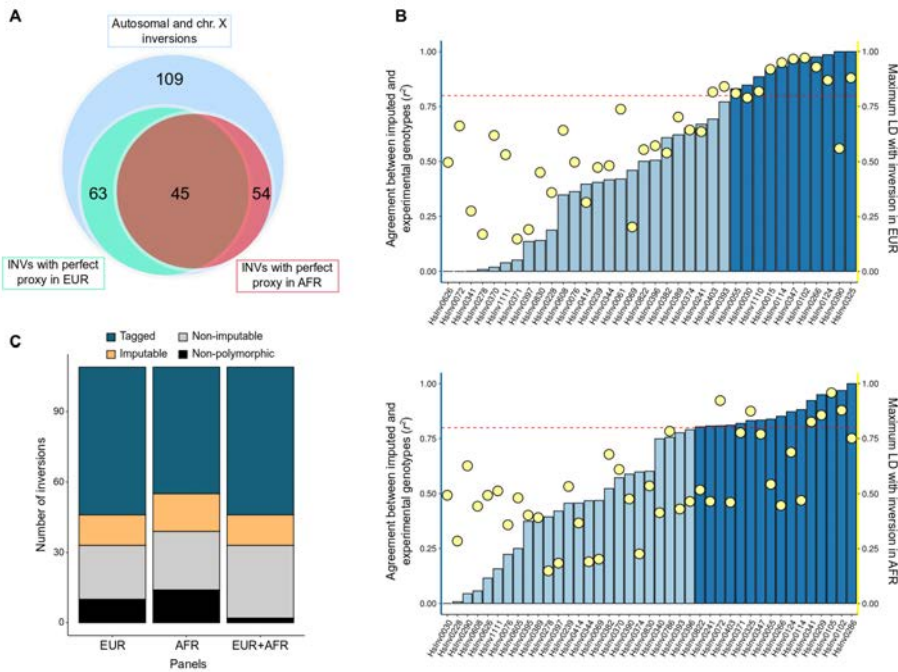


Figure 3.15 – Feasibility of inversion imputation. (A) Total number of polymorphic inversions that are tagged in European (EUR) and African (AFR) populations in the complete set of experimentally-validated autosomal and chr. X inversions. Venn diagram was done with BioVenn (Hulsen et al. 2008). (B) Genotype imputation accuracy of 36 and 41 inversions without perfect tag variants in Europeans and Africans, respectively. Imputation was performed with IMPUTE v2.3.2 based on 1000GP Phase 3 variants for European and African individuals with known inversion genotypes as reference panels (see Methods). Concordance between imputed and experimental genotypes (r^2) were based on masking one of the genotyped individuals and comparing imputed and true genotypes (blue bars). Only inversions that can be inferred with enough precision ($r^2 > 0.8$, red line) are selected for molecular QTL analysis in other datasets. Maximum LD between inversions and neighbouring genomic variants (yellow points) was higher for inversions with better imputation accuracy. (C) Number of inversions classified according to the performance of *in silico* genotype calling. Around 2/3 of non-tagged inversions cannot be reliably imputed.

a novel inversion, HsInv1110, associated with the expression of several pseudogenes and protein coding genes, such as *LIP1*, which has been associated with dyslipidemia (Wen et al. 2003), and *RBM11*, implicated in alternative splicing processing (Pedrotti et al. 2012) (Figure 3.16). Unlike HsInv0573 and HsInv0786, which maintain relatively-divergent haplotypes of highly linked variants (Puig et al. 2019), HsInv1110 does not show high LD with other variants and appears as the clear lead eQTL and likely causal variant of these changes, suggesting that it probably has an impact on chromosomal environment. These three inversions suppose around 80% of the total number of inversion-expression association pairs (53%, 17% and 10% for HsInv0573, HsInv0786 and HsInv1110, respectively). Moreover, we found 46 other inversions also directly associated with the expression of a gene. This represents basically half of the inversions assessed in Europeans and it highlights the substantial potential consequences that inversions can have in the genome. In this sense, there are a number of interesting candidates that may cause important gene-expression changes. For example, the intronic inversion HsInv0174 is associated with the expression of the gene *PTPRF* in testis, which may contribute to insulin resistance (Mander et al. 2005) and cancer (Tian et al. 2018) (Figure 3.16). Other cases in which the inversion is in high LD with the top eQTL include HsInv0191 and *FUT11*, which encodes a fucosyltransferase, or HsInv1075 and *POLA2*, which encodes a DNA polymerase accessory subunit.

Next, we examined inversion effects on histone modification marks, DNA methylation and chromatin accessibility available in a subset of LCL samples. We found epigenetic changes associated with 12 inversions in LCLs (Table 3.1), providing insights into inversions that may influence gene expression through epigenetic modification of regulatory elements. In Giner-Delgado et al. (2019), we already mentioned HsInv0031 as an interesting candidate that is associated with lower levels of *FAM92B* in cerebellum and in almost perfect LD with a suggestive locus associated with Alzheimer's disease (Pérez-Palma et al. 2014). Here, we found that the inverted allele is also associated with a restricted chromatin accessibility region close to HsInv0031 and with higher levels of histone methyla-

tion, but the relationship between these elements and *FAM92B* expression needs to be validated experimentally in cerebellum. Other interesting inversions affecting epigenetic marks and gene expression are HsInv0030, HsInv0124 or HsInv1110, among others.

Table 3.1 – Summary of inversions acting as QTLs of different molecular phenotypes.

Molecular phenotype	Tissue	Reference	Population	Sample size	INV-QTLs or inversions in high LD with QTLs
Gene expression	LCLs (Geuvadis)	Lappalainen et al. 2013	EUR	358	21
Gene expression	45 tissues and 2 cell lines (GTEx)	The GTEx Consortium 2017	EUR	32-421	969
Histone modification	LCLs	Delaneau et al. 2019	EUR	145	30
DNA methylation	LCLs	Moen et al. 2013	EUR+AFR	103	7
Chromatin accessibility	LCLs	Degner et al. 2012	AFR	59	1

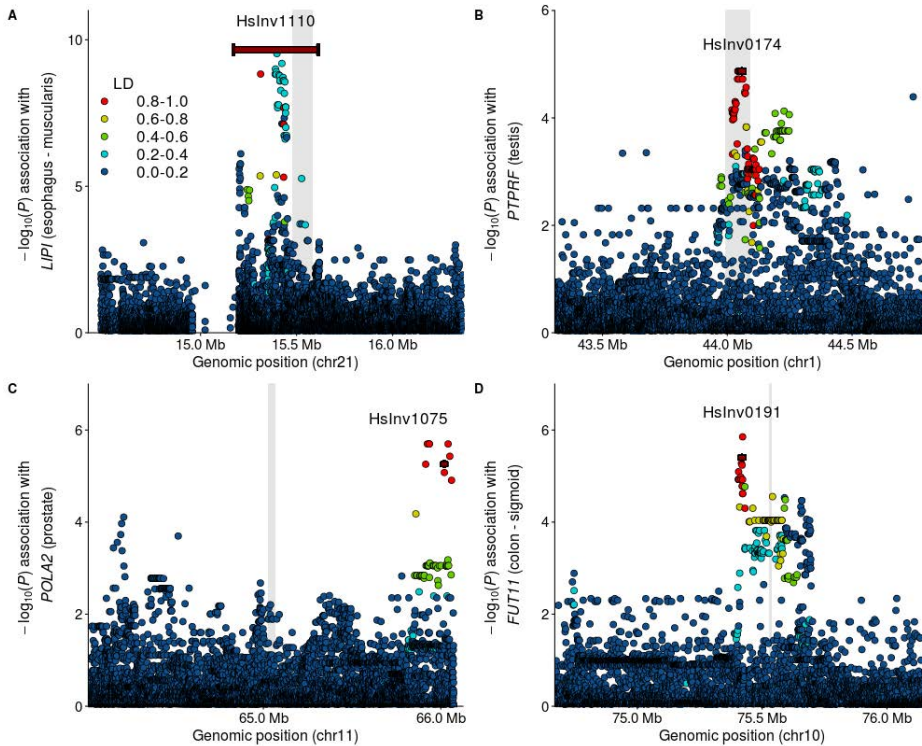


Figure 3.16 – Examples of *cis* INV-eQTLs in different tissues. Manhattan plots for gene expression associations in which an inversion is the sentinel variant (A, B) or is in high LD with the top marker (C, D). Inversions are depicted as dark red rectangles. The grey bar shows the position of the affected gene. Dots indicate 1000GP variants and the color represents the LD with the inversion according to the scale shown.

3.5.3 Inversions and phenotypes

We explored the contribution of polymorphic inversions to phenotypic variation using known GWAS hits collected in the NHGRI GWAS Catalog database (MacArthur et al. 2017). We found a 1.44-fold enrichment ($P = 0.05$) of GWAS signals in the inversion and flanking regions (± 20 kb) (Figure 3.17), which increases to 1.85-fold for inversions >100 kb ($\mathcal{A} = 0.04$). This supports the potential role of inversions on phenotypic traits. Apparently, nine individual inversions were driven these results,

since they showed a higher number of trait and disease-associated signals in their surrounding regions than would be expected by chance (Figure 3.17). For example, GWAS signals detected close to HsInv0124, which is located at the *IFITM* locus, are related to distinct blood cell counts and percentages. Also, there are many hits involving bone mineral density for HsInv0095, mental abilities and psychological conditions (intelligence, attention deficit hyperactivity disorder, educational attainment or adventurousness) for HsInv0174 and IgG glycosylation patterns for HsInv1321.

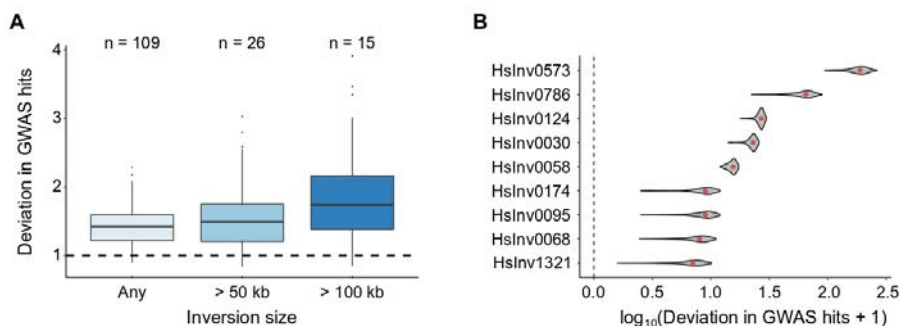


Figure 3.17 – Enrichment of previously reported GWAS signals on inversion regions. Deviation in observed minus expected values of GWAS Catalog trait-associated signals within inversions and flanking regions (± 20 kb), taking into consideration different groups of inversions (A) or individually (B). An stronger enrichment was found for larger inversions (A). Density distributions for nine individual inversions that showed a significant enrichment of GWAS signals (nominal $P < 0.05$, one-tailed permutation test) (B) represent the 95% one-side confidence interval with the median indicated by a red dot.

We next screened for inversions that were in strong LD ($r^2 \geq 0.8$) with significant GWAS variants in the corresponding population in which the association was reported and that were likely to explain prior GWAS results. HsInv0573 and HsInv0786 are associated with many distinct phenotypes, that range from neurodegenerative disorders to immunological conditions, as already demonstrated in earlier works (Puig et al. 2019). In addition, we found several promising candidates with convincing evidence of phenotypic association, and in many cases also of gene expres-

sion effects: HsInv0014 is associated with AKR1C1 blood protein levels and the anthropometric traits height and waist-hip ratio (Figure 3.18); HsInv0030 has been already described to be implicated in chronic pancreatitis susceptibility (Rosendahl et al. 2018); HsInv0395 is linked to a risk allele for educational attainment and is in LD with eQTLs for the gene *MAGEH1*, which has been shown to interact with the p75 neurotrophic receptor (Tcherpakov et al. 2002); HsInv0286 is predicted to affect the expression of the long non-coding RNA *LINC01445* and is in LD with a GWAS hit for smoking initiation; and HsInv0379 is associated with type 2 diabetes in Japanese population and was already described to disrupt the transcription factor *ZNF257* and generate a new fusion transcript (Puig et al. 2015a). Besides, it was already reported that HsInv0191 is in complete LD with GWAS hits related to atrial fibrillation and systolic blood pressure (Christophersen et al. 2017; Kichaev et al. 2019), and we found an association with *FUT11* in colon. Although GWAS and eQTL signals suggest that both traits may be produced by the same underlying causal variant, we identified inversions with strong phenotypic associations but without a clear candidate gene. For example, HsInv0068 is associated with the age at menarche, and the inverted allele of HsInv1075 is linked to a lower performance on intellectual tasks (highest math class taken) and higher insomnia. In total, 14 inversions are associated with GWAS signals, which constitutes a high proportion for this type of variants (13%) taking into account the reduced number of inversions with highly linked SNPs that could be interrogated. In fact, the low LD with SNPs in recurrent inversions excludes around half of our inversion set from this analysis. These findings, combined with the gene expression and epigenetic analysis described above, indicate that a substantial fraction of inversions could play a relevant role to the biology of human traits and disease. Nevertheless, empirical validation is needed to establish causality for trait-associated inversions.

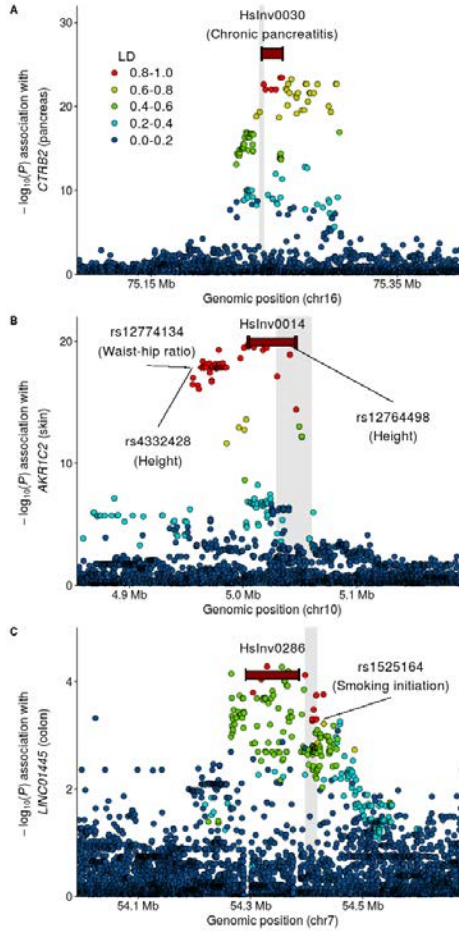


Figure 3.18 – Candidate INV-eQTLs associated to GWAS hits. Manhattan plots for *cis*-eQTL associations, showing inversions as potential lead variants. (A) HsInv0030 creates hybrid transcripts from *CTRB1* and *CTRB2* genes, and has been associated with chronic pancreatitis. (B) HsInv0014 is predicted to be causal eQTL for *AKR1C1* and *AKR1C2* genes and is linked to signals for height and waist-hip ratio. (C) HsInv0286 is associated with the expression of the long non-coding RNA *LINC01445* and is a known risk allele for smoking initiation. Dark-red rectangles represent the predicted causal inversions, and dots other 1000GP variants with the colors indicating its LD (r^2) with the inversion. Reported GWAS risk variants are shown.

3.5.4 Detailed characterization of HsInv0014 and HsInv0030 functional effects

The inversions with the clearest effect on gene expression are those affecting directly gene structure. For instance, HsInv0379 and HsInv1051 have been already reported to disrupt completely a gene and generate new fusion transcripts (Puig et al. 2015; Giner-Delgado et al. 2019). Therefore, we examined in more detail inversions HsInv0014 and HsInv0030, which are associated with both phenotypic traits and strong changes in gene expression levels. As already described (Pang et al. 2014; Giner-Delgado et al. 2019), HsInv0030 exchanges the 5' exon and promoter of chymotrypsinogen precursor genes *CTRB1* and *CTRB2* that partially overlap with the segmental duplications at inversion breakpoints (Figure 3.19). These genes, only expressed in pancreas, have small nucleotide differences at the first and last exons located outside the segmental duplications that allow us to differentiate reference transcripts from the derived *O1* orientation and exchanged transcripts in the ancestral *O2* orientation. In effect, we designated the annotated *CTRB1* and *CTRB2* from human reference genome as *CTRB1-R* and *CTRB2-R*, and the reconstructed forms in the *O2* allele as *CTRB1/2-Q* for the reference version with the promoter and first exon of the paralogous gene. By mapping pancreas RNA-Seq reads from 7 *O1/O1* and *O2/O2* homozygote individuals, we found that, unlike *CTRB2-Q* and *CTRB1-R*, *CTRB1-Q* and *CTRB2-R* show a significant change in expression levels despite sharing the same promoter (Figure 3.19), suggesting the existence of additional regulatory processes.

Similarly, HsInv0014 switches the position of the main isoforms of the aldo-keto reductases *AKR1C1* and *AKR1C2* genes, encoding two main aldo-keto reductases isoforms. We found that in LCLs the *O2* orientation generates a highly-expressed isoform starting outside the inversion with the 3' end corresponding to the *AKR1C2* exons (Figure 3.20). We also investigated the gene-expression profiles in two other different tissues (Figure 3.21). In adipose tissue, both isoforms within the inversions are expressed at high levels with similar profiles, whereas in testis, *AKR1C1*

shows higher expression than *AKR1C2* in the *O1* allele and the other way around in the *O2* orientation. These findings illustrate that HsInv0014 acts through diverse mechanisms on gene regulation, which are tissue specific.

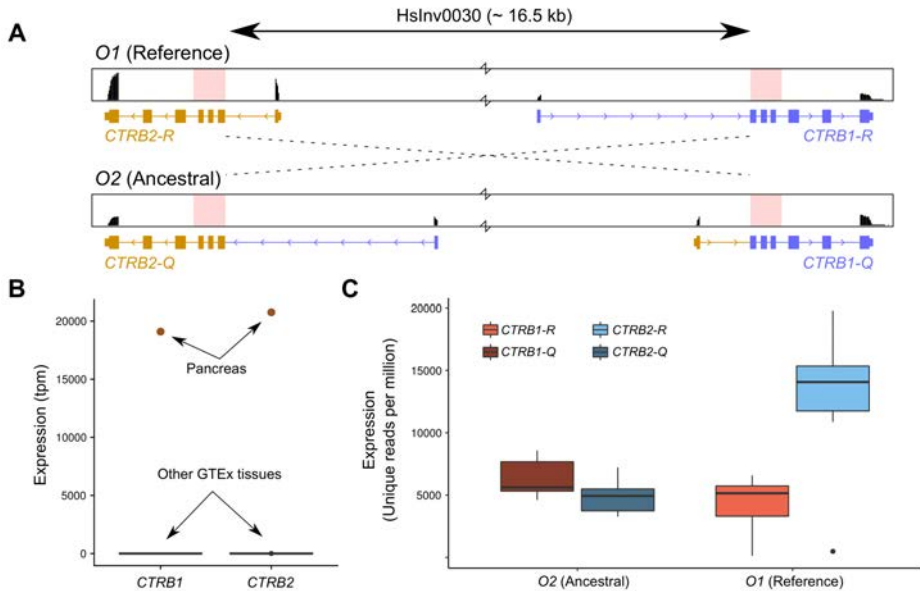


Figure 3.19 – Molecular consequences of inversion HsInv0030.

(A) RNA-Seq profiles from GTEx reads obtained in pancreas for 7 *O1/O1* and 7 *O2/O2* individuals mapped to the first and last exons of *CTRB1* and *CTRB2* genes. Inversion breakpoints are depicted in pink. RNA-seq profiles were visualized on Integrative Genomics Viewer (Robinson et al. 2011). (B) *CTRB1* and *CTRB2* genes are exclusively expressed in pancreas. (C) Expression changes associated to HsInv0030 rearrangement. While *CTRB2-Q* and *CTRB1-R* share the same promoter and have similar levels of expression, *CTRB2-R* shows a strong expression increase compared to *CTRB1-Q*.

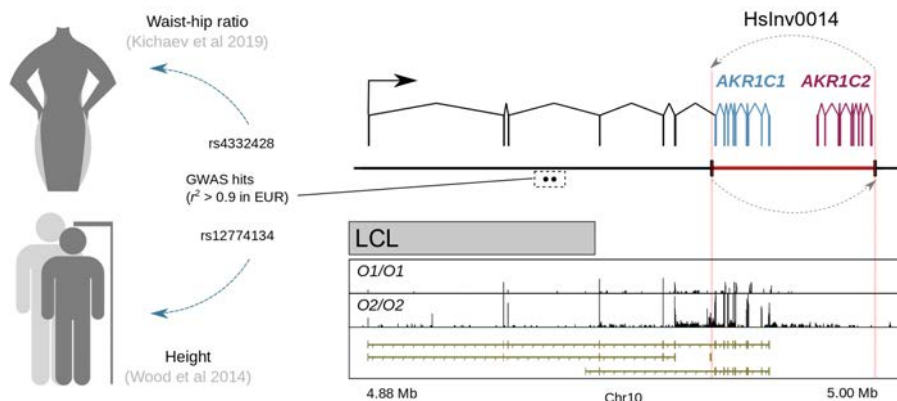


Figure 3.20 – HsInv0014 control of *AKR1C1* and *AKR1C2* expression is dependent on tissue. Diagram of gene *AKR1C1* and *AKR1C2* remodelling by HsInv0014, showing the new extended isoform expressed in *O2* chromosomes and the RNA-Seq profiles from Geuvadis LCL reads. HsInv0014 is in high LD with GWAS variants associated to height and waist-hip ratio that are located within an intron of the new isoform. Inversion breakpoints are shaded in pink.

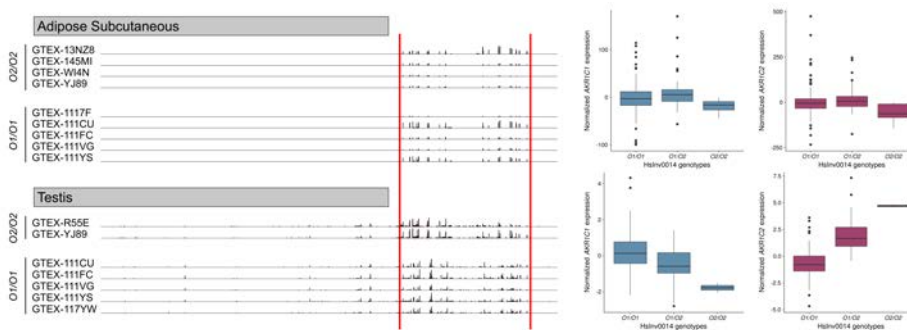


Figure 3.21 – HsInv0014 regulation of *AKR1C1* and *AKR1C2* expression is dependent on tissue. RNA-Seq profiles from adipose and testis tissues in the HsInv0014 region and boxplots of expression levels of *AKR1C1* and *AKR1C2* according to inversion genotype. HsInv0014 appears to have diverse roles on gene regulation depending on the tissue assayed. HsInv0014 breakpoints are indicated in red.

Chapter 4

Discussion

Despite structural variation contributes substantially to genetic diversity, their association to complex traits is just beginning to be understood and is still poorly characterized compared to SNPs. This is even more obvious for inversions, which have often been set aside not only due to their balance nature but also because of the complex repetitive regions in which their breakpoints tend to be located. In this new era of precision medicine, QTL analysis has emerged as a decisive approach in the understanding of how genetic polymorphisms, both single nucleotide and structural variation, influence gene expression and, in turn, human phenotypic traits. Hundreds of studies have associated multiple genetic variants with transcript levels and other molecular phenotypes such as DNA methylation, histone modifications or protein abundance. In fact, QTLs have been useful to interpret GWAS results and have helped to uncover the mechanisms underlying disease. In this regard, when this thesis started, only three inversions -17q21.31, 8p23.1 and 16p11- had been studied in some detail and had been associated to expression changes and traits (Stefansson et al. 2005; Myers et al. 2005; Zabetian et al. 2007; Webb et al. 2008; de Jong et al. 2012; Salm et al. 2012; Okbay et al. 2016; González et al. 2014). There have been other studies designed to investigate structural variation with distinct genotyping efforts, but unfortunately they were restricted to a small number of inversions or limited in the amount of in-

dividuals analysed. The InvFEST Project took on the challenge and tried to address this difficult technical task by developing unique techniques aimed at inversion genotyping. Although there is not yet a single technique able to analyse all the different types of inversions and thousands of individuals, the data generated constitutes an important step in the study of human polymorphic inversions. Here we analyzed the potential consequences of more than 100 inversions that have been accurately genotyped in individuals from diverse human populations with different approaches, which represents the most complete resource of this type of variants so far. Specifically, we have done a detailed analysis of the impact of human inversions on gene expression, epigenetic changes and phenotypic traits. Moreover, we have also characterized particular candidates that should be further investigated.

In the following pages, I will discuss in detail the main topics addressed in this thesis.

4.1 Methodology, data and limitations of this study

In mathematics, a theorem is a not self-evident statement that has been proved. Mathematicians devote themselves to come up with theorems using other established theorems or postulates, which are statements that are assumed to be true without proof. Theorems in mathematics are thereby provable truths, meaning that successfully proven theorems are eternal truths, and will be true forever and ever. This is not the case of natural science and, in particular, life sciences, in which there are many examples of things that were once accepted and got modified afterwards. Scientific theories are established based on the accuracy of the instruments and measurements employed in a specific time period. Therefore, independent evaluation, replication and reproduction of results are basic on scientific research. In contrast to exact science, natural scientists are aware that they can only expect a finite amount of evidence or certainty and they

generate temporary models to approach the reality.

In this sense, we tried to elucidate and understand the functional consequences of many human inversions combining different sources of currently available information, provided mainly by international projects such as GENCODE, Geuvadis, GTEx or multiple GWAS of different traits and diseases. Overall, this study describes the most complete analysis to date of the impact that inversions might have in humans. We can foretell that the publication in the near future of larger and more comprehensive datasets will go in hand with a deeper interrogation of inversions impact. However, before describing the significance of our findings, we will focus on the limitations and pitfalls we encountered during this thesis, and the advantages and disadvantages of the methodology we employed.

4.1.1 Gene expression quantification and QTL mapping

An standard eQTL analysis involves the direct testing of genetic polymorphisms with expression levels measured in tens to hundreds or even thousands of individuals. In this work, we focused on publicly accessible expression datasets that included expression data mainly from LCLs (Lappalainen et al. 2013), but also GTEx tissues (GTEx Consortium 2017) or monocytes exposed to different stimuli (Quach et al. 2016). Despite data from expression arrays exist in LCLs (Stranger et al. 2007a; Stranger et al. 2007b; Stranger et al. 2012) for individuals genotyped by the InvFEST project, all the measurements we used derived from RNA-Seq. Indeed, RNA-Seq has replaced microarrays because the sequencing technology can measure low expressed genes as well as distinct isoforms or specific exons in a pervasive and genome-wide manner. In our case, this has been useful to test inversion effects on alternative transcripts or exons from the same gene. Besides, RNA-Seq reads allowed us to reconstruct fusion transcripts generated by gene disrupting inversions, which would have remained hidden otherwise. One important step in RNA-Seq analysis is the mapping of reads to quantify gene expression. Although there are several RNA-Seq aligners, we preferentially employed STAR (Dobin

et al. 2013), which according to benchmarking studies outperforms other methods in most ways, from mapping accuracy to speed (Engstrom et al. 2013; Baruzo et al. 2017). Moreover, STAR, unlike novel alignment-free tools, generates a BAM file in which aligned sequences are represented and can be visualized with IGV (Robinson et al. 2011), which is useful for describing in detail inversion effects on gene structures or splicing.

As already stated, the majority of our eQTL analyses have been conducted on LCLs provided by the Geuvadis project (Lappalainen et al. 2013). LCLs are human B cells immortalised with the Epstein-Barr virus and are a very useful model system in molecular genetics for a number of reasons, as we will discuss in more detail in the next section. Briefly, these cell lines are commonly used to supply DNA for the HapMap and 1000GP genetic variation studies, and they have also been used to examine genome-wide expression profiles and generate other functional data at the population level. This valuable resource was selected by the InvFEST project for inversion genotyping due to the amount of information accumulated about them and allowed us to assess their influence on gene-expression variation, among other things. On the one hand, experimentally-generated genotyping data can be used for direct inversion eQTL mapping. This is important for recurrent inversions, whose effects cannot be tested otherwise. However, only genes expressed in lymphocytes can be measured. Although many eQTLs are shared across cell types, the expression of a particular gene does not ensure similar regulatory patterns across distinct tissues, thus preventing potential extrapolations (Nica and Dermitzakis 2013). On the other hand, we also analysed other datasets by taking advantage of SNPs in LD with our inversions as a proxy for imputing genotypes. Identifying inversion associations through variants that are inherited together is a powerful and fast approach to scan effects in diverse QTL catalogues very efficiently. However, despite genotyping of the specific tissues or cell lines is not necessary, there are important problems with this approach, in particular for recurrent inversions, since they lack tag SNPs and *in silico* genotyping is not always accurate. Therefore, this has limited the analysis of many of these inversions and their effects remain to be discovered in such contexts. Moreover, many studies tend to omit chromosome X and

such datasets often only recover information from autosomes, in which a significant fraction of our inversions are located.

With regard to the methodology, in Giner-Delgado et al. (2019), we employed distinct commonly-used eQTL mapping methods. Specifically, three strategies for gene-expression analysis from LCLs were compared: the one followed by the GTEx Project (GTEx Consortium 2017), the edgeR-limma workflow (Robinson et al. 2010; Ritchie et al. 2015) and the protocol recommended for the QTLtools software (Delaneau et al. 2017). A common characteristic of eQTL finding methods is that they account for unknown confounding effects by modeling and compiling those factors into a set of variables in decreasing order of estimated impact. For example, PEER (Stegle et al. 2012) and SVA (Leek et al. 2018) programs -employed in the GTEx protocol and edgeR-limma, respectively- implement statistical models to uncover hidden factors that explain expression variability. Similarly, PCA -performed with QTLtools- maximizes variability of the data when projected on each component. Thus, expression values are adjusted by a set of covariates to correct by major effects that can bias our results. Although it is not always clear the number of covariates that should be used, it is not recommended to have neither few nor many, and several tests are often carried out to assess the optimal configuration. After expression correction, linear regressions are applied with FastQTL (Ongen et al. 2016) and QTLtools, whereas edgeR uses negative binomial models for differential expression. In our case, a good degree of consistency was achieved across these pipelines, considering the different expression normalization protocols and analysis methods used. In addition, results were further replicated in other independent datasets such as GTEx, confirming the reliability of our findings and the robustness of the different analysis methods.

In the gene-expression analysis, we predominantly searched for *cis* effects instead of *trans*. In general, eQTL studies focus on proximal regulatory variants, since finding *trans* eQTLs has been less successful so far, with some notable exceptions (Small et al. 2011; Franceschini et al. 2012; Grundberg et al. 2012; Westra et al. 2013; Lee et al. 2014; Fairfax et al.

2014, GTEx Consortium 2017; Bonder et al. 2017; Võsa et al. 2018). The amount of samples needed for interrogating the whole genome for regulatory effects in *trans*, besides the challenge it represents experimentally, is an enormous statistical and computational task that limits this type of analysis. In addition, methods aimed to adjust expression data by technical covariates, such as PEER or PCA, are likely to also capture *trans* eQTL hotspots that are associated with multiple genes regulated by the same transcription factors (Stegle et al. 2012). Therefore, gene expression correction by potential technical confounders can unintentionally remove few potential signals in *trans*, but no correction at all can lead to many false positive signals. Novel strategies designed to discover *trans* eQTLs take advantage of local eQTL mediators of such distal effects (Yang et al. 2017) or the interchromosomal coordination of regulatory elements (De-laneau et al. 2019), and they could prove to be very useful in the future for analysing additional inversion consequences.

In addition to gene expression, the same strategy is increasingly employed to other types of quantitative phenotypes. In this sense, we applied QTL methodology to test genetic effects on DNA methylation (Moen et al. 2013), chromatin state (Degner et al. 2012), histone modification (De-laneau et al. 2019) and protein abundance (Battle et al. 2015). These additional cellular QTLs allowed us to uncover regulatory mechanisms by which variants exerts their effects and achieve a more integrative view of the human genome function.

4.1.2 Tissues assayed and cell-specificity

Studying the regulatory function of the genome by interrogating multiple tissues in large number of individuals is today a reality (Nica et al. 2011; Grundberg et al. 2012; GTEx Consortium 2017). Nonetheless, most of our initial knowledge came from the analysis of whole blood (Bonder et al. 2017; Võsa et al. 2018), specific blood cells such as white cells (Fairfax et al. 2014; Quach et al. 2016), or transformed LCLs (Stranger et al. 2007a; Stranger et al. 2007b; Stranger et al. 2012; Lappalainen et al.

2013). Blood-derived cells have constituted a powerful resource for gene-expression regulation research and continue to be so because they are an easily accessible source to assay transcription or epigenetic changes on a large scale. LCLs are one of the preferred options due to the simple cell maintenance and that they serve as an unlimited supply of human material that can be useful for different types of analysis and replication tests. Estimates indicate that around 50-60% of human genes are expressed in LCLs (Lappalainen et al. 2013; Chiang et al. 2017) and conclusions about gene regulation obtained from these cell-types might be applied to other tissues. In fact, recent data suggest that many eQTLs are shared across tissues and, although there is still a significant proportion of tissue-specific effects, tissue specificity was less prevalent than estimated at first (Nica et al. 2011; Grundberg et al. 2012; Nica and Dermitzakis 2013; GTEx Consortium 2017). Higher rates of tissue-specific eQTLs have been found in those cohorts of individuals with larger samples sizes, but these estimates can be indicating that subtle genetic effects are simply more difficult to replicate in tissues with fewer samples (GTEx Consortium 2017). *trans*-eQTLs are also more restricted to particular cell types or tissues, which could be a consequence of the immense statistical burden to detect these effects. Although interrogating gene expression in the relevant tissue is crucial, shared eQTLs could still be quite useful to decipher the mechanisms behind GWAS results. For instance, eQTLs found in LCLs regulating *ORMDL3* have been useful to interpret significant GWAS association signals in childhood asthma (Libiolle et al. 2007) and *cis*-acting regulatory elements of *PTGER4* in Crohn's disease (Moffatt et al. 2007), two autoimmune inflammatory disorders. LCLs have also helped to determine candidate genes in conditions associated to more distant tissues, such as autism (Nishimura et al. 2007) and bipolar disorder (Iwamoto et al. 2004), which are diseases mostly related with the central nervous system. Therefore, using LCLs for our analysis is a good approach to characterise many eQTL associations that might reflect the basic function of an inversion on gene expression and GWAS outcome.

While biologically-related tissues often tend to share eQTLs, others, like blood or testis, have showed higher rates of tissue-specific eQTLs, and

these findings must be taken into account to estimate the relative contribution of each tissue to a given trait and infer where genetic causality resides (Ongen et al. 2017). We found few inversions with shared effects across different tissues, such as HsInv0041 on *FAM124B* in subcutaneous adipose tissue, visceral adipose tissue and adrenal gland (Giner-Delgado et al. 2019). Of especial interest are inversions 17q21, HsInv0786 and HsInv1110, which seem to accumulate the largest fraction of INV-eQTL associations, and their effects are pervasive across many tissues (see below). Nevertheless, these were the exceptions and most of the inversions in our set were acting as eQTLs in a tissue-specific manner. For example, *CTRB1* and *CTRB2* genes are practically only expressed in pancreas and the effect of HsInv0030 on them must be tested in the corresponding tissue. Moreover, inversion HsInv0124 was lead variant for *IFITM3* expression variation and, depending on the condition to which monocytes were exposed, for other genes such as *IFITM5* and *NLRP6*. This illustrates the necessity of also covering all functional contexts -including cell type, developmental stage or cell-state and environmental conditions-. For instance, detected expression changes on the same transcripts can be the opposite in different cell types, as happens for the non-coding genes *AC136475.1* and *AC136475.2* placed at the segmental duplications surrounding HsInv0124, which behave in opposite manner in LCLs and stimulated monocytes, suggesting tissue-specific regulatory elements at the *IFITM1-3* region. Additionally, inversions can have a broad range of molecular consequences depending on the tissue assayed, as the effect of HsInv0014 on *AKR1C1* and *AKR1C2* expression patterns and isoform usage. This tissue-specific phenomenon extends as well for the rest of intermediate phenotypes, such as histone modifications (Delaneau et al. 2019) or DNA methylation (Gutierrez-Arcelus et al. 2015).

By focusing in some little-studied type of variants, characterizing the regulatory relationships between inversions and molecular processes in diverse cell types will help us fill the substantial gap in our understanding of the mechanisms underlying complex traits, as it is well known that those variants discovered by GWAS tend to be eQTLs (Nicolae et al. 2010). However, we must be cautious about our findings, since there could exist

a regulatory variant that is affecting one gene in one tissue and is linked to another gene in a distinct tissue, giving rise to potential misleading biological interpretations about which gene is responsible of an increment of disease risk (Nica and Dermitzakis 2013). Furthermore, a recent study highlights that the real effect of many variants might remain to be discovered, such as rs7903146, a major GWAS locus for type 2 diabetes, that has now been found to be an eQTL for *TCF7L2* restricted to pancreatic islets (Viñuela et al. 2019). GTEx tissues may not represent a complete catalogue of all tissues in the human body, but it constitutes the most comprehensive set so far. Therefore, we are confident that we have discovered an important fraction of potential inversion effects, excluding those of recurrent inversions.

4.1.3 Sample size, inversion frequency and genotype imputation

In the main paper of the GTEx Project (GTEx Consortium 2017), the authors confirmed that sample size strongly affected eQTL finding. Larger collections of differentially expressed genes were discovered in those tissues with more samples analysed. In fact, the number of significant gene-eQTL pairs kept increasing for bigger sample sizes when the eQTL mapping was repeated using sub-sets of various sizes of the donors of each tissue (GTEx Consortium 2017). This indicates that size is one of the main factors limiting statistical power and may suggest that virtually all genes in the human genome are under genetic control. Remarkably, sample size was a more important contributor to eQTL identification than adding extra tissues due to shared eQTLs (GTEx Consortium 2017; Ongen et al. 2017). An increase of tissues yields few tissue-specific effects, but sample size supposes a bigger boost for eQTL analysis in order to find smaller effect sizes and secondary independent eQTLs. Thus, having a proper number of individuals with genotyped inversions is essential for association studies. Similarly, frequency is another important factor that can limit functional analysis. An inversion with low frequency results in a reduced amount of individuals with different genotypes for statistical

comparison. For example, HsInv1057 inverted allele is present at just 7% frequency in African populations. This means that 0 or 1 homozygotes -actually 0.5- for the inverted orientation would be expected in a cohort of 100 individuals, insufficient to test many molecular changes, whereas ten times more individuals -1,000 samples- would be required for obtaining at least 5 homozygotes.

In this regard, genotype imputation has been successfully used in GWAS and eQTL analysis to obtain large cohorts and enable researchers to detect variants with moderate effects. We used imputation to infer the orientation status of un-genotyped inversions with the purpose of increasing the effective sample size. Imputation is based on the LD that is observed between directly assayed inversions and neighbouring variants, and matching these patterns with a suitable reference population panel (Marchini and Howie 2010). Therefore, it is essential to use a population in which inversions have been experimentally genotyped that is genetically similar to the panel we want to impute, since haplotypes must be comparable. Perfect correlation was found between some inversions and SNP genotypes, and these nearby tag variants can be used directly as proxy for the inversion genotypes, as we did for the identification of GWAS signals in LD with the inversions. In some other cases, no perfect tagging variants are found and probabilistic estimates are given for unknown genotypes. Many imputation programs have been developed, including IMPUTE2 (Howie et al. 2009), Beagle (Browning et al. 2007) or MaCH (Li et al. 2010).

On the other hand, since inversions specifically modify local recombination, nucleotide variation patterns can be used as footprints to uncover inversion status and software packages aimed to impute inversions also exist. Some examples are PFIDO (Salm et al. 2012), inveRsion (Cáceres et al. 2012), invClust (Cáceres and González 2015) and scoreInvHap (Ruiz-Arenas et al. 2019). Most of them classify divergent haplotype groups supported by the suppression of recombination and inversion status is inferred through such haplotype-cluster memberships. However, these last methods have not been benchmarked against other already established software like IMPUTE2 and their real performance has not been clearly

assessed. PFIDO and inveRision were employed to impute HsInv0786 in a previous study (González et al. 2014) with a significant error rate of $\sim 11\%$ according to experimental ddPCR genotyping (Puig et al. 2019), which is also indicative of the presence of a recurrent chromosome with the unexpected orientation. Furthermore, only 20 out of 59 inversions reported in European individuals from the 1000 Genomes Project were suitable for being genotyping with scoreInvHap (Ruiz-Arenas et al. 2019), illustrating the limitations of these approaches. For that, we decided to use IMPUTE2 as one of the most commonly-used and reliable programs for genotype imputation. Another important problem is the high levels of recurrence observed for some inversions. By evaluating the imputation accuracy for several of the NAHR-mediated inversions by masking known genotypes in a leave-one out strategy and comparing the number of correctly and incorrectly inferred genotypes, we found that only $\sim 30\%$ of inversions without tag SNPs are susceptible of being imputed reliably, and the vast majority of them cannot be imputed at all. For most of these inversions, the lack of LD with nearby variants prevents their reliable calling. As a consequence, a notable fraction of potential inversion effects are still missing, and high-throughput experimental techniques in order to genotype those inversions in multiple individuals are needed. Alternative strategies that we used aimed to include the maximum number of inversions in the functional analyses were to focus on specific populations, because many of these inversions have additional tag SNPs that are not present globally or lower levels of recurrence in certain populations, which allow better rates of imputation quality and precision. Increasing the number of genotyped samples in a reference panel can also help to increase imputation accuracy in some cases. Finally, another crucial factor is the source of SNP data, which can include arrays with low to high SNP coverage. As we demonstrated for inversion HsInv0102, only high-density SNP arrays such as OMNI arrays were useful for obtaining reliable HsInv0102 genotypes and could be therefore used in association studies.

Despite these limitations, genotype imputation has been extensively demonstrated to be an efficient strategy in our study of polymorphic inversions by the improvements in INV-eQTL finding in the imputed datasets.

When both experimentally-genotyped and imputed sets in LCLs were available, we verified that this process was correctly implemented through comparison of INV-eQTLs effect sizes, showing highly concordant results and an enrichment of INV-eQTLs located closer to affected genes or transcripts (Giner-Delgado et al. 2019). For example, the analysis of only 59 individuals with expression data genotyped by ddPCR confirmed that only changes with large effect sizes can be detected with this small sample size (Puig et al. 2019). Indeed, a considerable higher number of significant results were found for those inversions that could be imputed in 387 European individuals in the same study. In addition to the gain in statistical power, imputation also provides better localized lead-eQTLs, as illustrated by HsInv0124 and HsInv0030 (Figure 4.1). In these examples, inversion imputation allowed us to uncover the likely causal variant of the expression variation.

4.1.4 Inversions and phenotypic effects

GWAS analysis have identified thousands of trait-associated SNPs genome-wide that have been used as evidence of variant functionality, since they directly link common genetic variation with complex diseases. When gene expression or other molecular phenotypes are measured throughout the whole genome, trans eQTL mapping is similar to performing a GWAS for each gene. However, our analysis of GWAS data suffers from several limitations. First, we did not have any evidence about potential effects of the inversions on phenotypic variation or a particular disease, and this is an exploratory analysis of a wide range of phenotypes. Second, since we do not carry out directly a GWAS -with the exception of the HsInv0102 association analysis with blood cancers-, only inversions with tag SNPs can be tested by crossing these linked variants with GWAS hits. To have higher power to detect potential associations, we focus on the population where the association was reported, avoiding to rely only in global tag SNPs, since more inversions possess population-specific variants in perfect LD. Paradoxically, despite linked variants allow to check inversion potential effects, high LD makes it difficult to identify causal variants. It is reason-

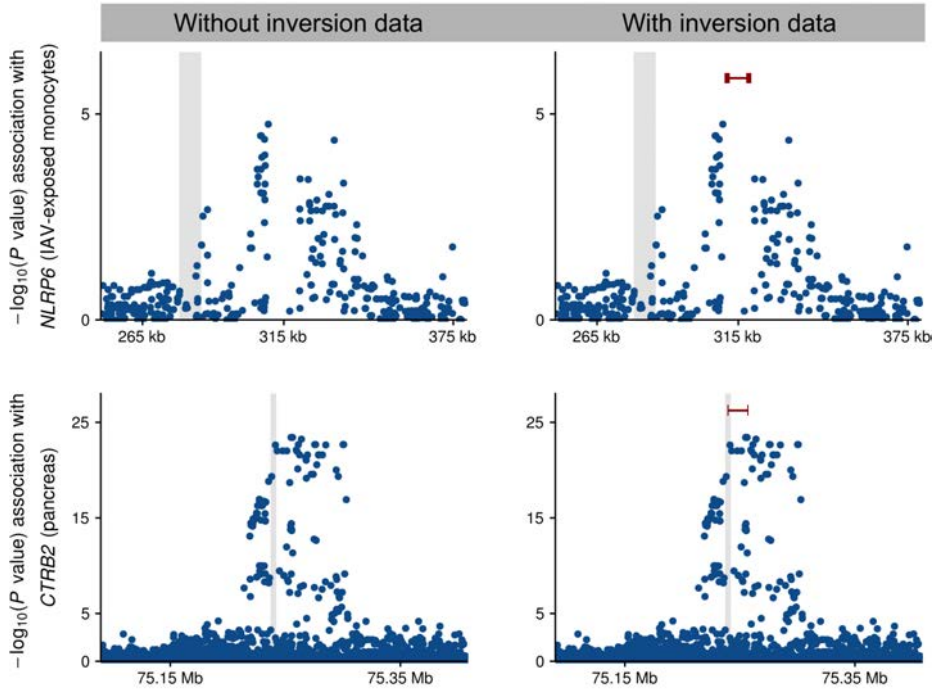


Figure 4.1 – Manhattan plots of *cis* eQTLs of the genes *NLRP6* in IAV-exposed monocytes and *CTRB2* in pancreas with and without inversion data. A map of the association of variants to transcript abundance of these genes shows clearly inversions as lead eQTLs and exemplifies the increase in information when imputed genotypes are included. The gray vertical bar indicates the position of *NLRP6* (chromosome 11) and *CTRB2* (chromosome 16).

able to hypothesize that inversion rearrangements cover a larger fraction of nucleotides and have more chances to affect genes located close to the breakpoints than single SNPs. However, still some additional evidences are necessary to pinpoint the inversion as responsible. Conversely, another limitation of this analysis is that many inversions mediated by NAHR are highly recurrent and those not associated with any SNP cannot be linked to any phenotype. We know that around half of our inversions are originated by NAHR and their contribution to the studied traits is thereby likely missing from GWAS data. Not even all non-recurrent inversions

have tag SNPs and sometimes tagging variants are not covered by typical arrays, which can reduce even more the amount of candidates (Giner-Delgado et al. 2019). Finally, characterizing the molecular functions of the causal variants and how they lead to disease risk is often intricate and require extensive functional analysis that are labour intensive.

An integrative analysis of GWAS and eQTL data is a powerful approach to interpret GWAS results and can compensate in part these limitations, since regulated genes are possible candidates for mediating phenotypic effects. Interesting examples include HsInv0031, where the inverted orientation is linked to lower *FAM92B* expression in cerebellum and is in LD with a variant associated with Alzheimer's disease risk, or HsInv0014, which is associated to height and waist-hip ratio traits, as well as to the expression of *AKR1C1* and *AKR1C2* -implicated in hormone metabolism in testis, LCLs and many other tissues. However, it is undeniable that a complete catalogue of eQTLs across all cell types and contexts has not been achieved yet. Thus, we may merely fail to identify the specific tissue or cell type where genetic causality arises. Examples are HsInv0068 and HsInv1075, which have phenotypic associations -age at menarche and performance on intellectual tasks-, but no clear candidate genes have been found yet. Also, HsInv0191 has been associated with atrial fibrillational and systolic blood pressure, and is linked to *FUT11* expression, but the role of this gene has not been evaluated in these clinical conditions.

We should also take into account that it is possible that many inversions are not associated with a phenotypic trait because such effects have not been uncovered yet. For instance, when Puig and colleagues (2015b) published a study about the inversion HsInv0379, they were unable to detect any association of this inversion with disease markers or other phenotypic traits. Nevertheless, a tag SNP of the inversion, rs148316037, was identified recently as a susceptibility signal for type 2 diabetes in the Japanese population in a meta-analysis of GWAS studies (Suzuki et al. 2019). Although the authors suspected that the effect was driven by a missense variant of the gene *ZNF257* in high LD with the GWAS hit (amino acid change Cys517Arg), the inversion is known to disrupt the gene in these in-

dividuals. According to the linkage with SNPs, HsInv0379 has the largest effect size of all the novel loci discovered in this study, which is consistent with the fact that structural variants have a larger impact (the inversion odds ratio would be 1.26, whereas that of the other signals are between 1.04 and 1.17). Curiously, the inverted orientation, which is found only at low frequency in Asian populations and it was suggested that it is probably deleterious, would act as the protective allele, giving rise to a trade-off scenario. Assuming that long inversions should generate unbalanced gametes in heterozygotes, the beneficial effects on diabetes of this inversion could help overcome slightly the negative impact on fertility to allow its modest rise in frequency.

An important consideration for the association analysis is the selection of a statistically significant cutoff. Given that in this case we are focused exclusively on human inversions, multiple testing burden may be reduced substantially from regular GWAS. In Ruiz-Arenas et al. (2019), the authors performed an association analysis of 15 inversions on breast cancer without taking into account nearby variants. An important question in this case is whether significant signals stem from the inversion as causal variant or are simply due to other variants in LD, giving rise to misleading results. However, if an inversion is known to have clear consequences on gene or protein expression, a targeted association study may be legitimate. We therefore performed an association study on distinct types of blood cancers with HsInv0102, which affects the protein levels of the haematopoietic-specific G-protein RhoH. Even when GWAS include most genomic variants, the suggestive threshold is open to interpretation. We mainly employed the NHGRI Catalog of published GWAS, which stores a curated collection of SNPs highly associated with disease or traits, but there are other databases, such as the GWASdb, which lists less significant genetic variants. Since many true loci involved in disease susceptibility may have moderate P values, we could favor sensitivity at the expense of compromising specificity and expecting some degree of false positive associations. However, we decided to stop using this last database in the next studies after Giner-Delgado et al. (2019) because it is not updated since 2015 and set for conservative criteria that ensure that the observed

associations are reliable.

Another important consideration is that most GWAS analysis do not contain forward conditional regression tests, but a few studies have successfully reported different SNPs located at the same locus and controlling a specific trait (Yang et al. 2012; McLaren et al. 2015; Sun et al. 2018). The routinely use of this approach, in which the association analysis is conditioned on the sentinel variant with the lowest P value to test whether there are any other significant hits, would enable the discovery of additional effects. This opens the door for a more exhaustive analysis of the role of inversions that might be associated to novel independent conditional signals. This is the case of HsInv0201, which is linked to a secondary variant associated to SPINK6 protein levels in human plasma (Sun et al. 2018).

4.2 Significance of findings. The functional impact of human inversions

So far few inversions associated to certain phenotypes in humans have been published. Although the underlying molecular mechanisms remain unclear, these inversions have been involved in expression variation and are supposed to manifest their effects through divergent haplotypes, such as the well-studied 17q21.31 inversion affecting the tau gene *MAPT* and related to neurodegenerative diseases and mental conditions (Stefansson et al. 2005; Myers et al. 2005; Zabetian et al. 2007; Webb et al. 2008; Okbay et al. 2016). Additionally, inversion 8p23 has been related to autoimmune diseases and personality traits (Salm et al. 2012; Okbay et al. 2016) and 16p11 may influence *IL-27* expression and the inverted allele appears to protect against the joint occurrence of asthma and obesity (González et al. 2014). However, a systematic quantification of the functional impact of multiple inversions have been difficult to achieve because technical limitations on genotyping. A joint eQTL analysis by the 1000GP (Sudmant et al. 2015) attempted to uncover the effect of structural variation on LCL expression profiles. This study found 54 structural variants

as lead eQTLs, but none of them were inversions. No inversions were in strong LD with GWAS hits either. Only a 354-bp inversion was in moderate LD ($r^2 = 0.58$) with a lead-SNP for gene *RAP2A* (rs59395497). The most comprehensive study published to date increased considerably whole-genome sequencing depth (from previous low-coverage -median 7.4X- to deep -median 49.9X-), improving SV detection and genotype accuracy, and extended differential expression analysis across 13 tissues (Chiang et al. 2017). In this case, 51 inversions were detected by breakpoint evidence (14 with MAF > 0.05), although just a 198-bp inversion (HsInv0191 in InvFEST) was associated to the expression of two genes (*P4HA1* in skin and *FUT11* in nerve tissue), when using only SVs for eQTL mapping to decrease multitesting burden. According to their genotype data, this inversion was also in high LD ($r^2 = 0.88$) with SNP rs10824026 reported as susceptibility locus for atrial fibrillation. Moreover, the same inversion (predicted as a 193-bp inversion) was reported in another study trying to characterize non-repetitive, non-reference sequence variants in the Icelandic population (Kehr et al. 2017), and was now perfectly correlated with the previous GWAS variant ($r^2 = 0.99$). Remarkably, Chiang et al. (2017) investigated as well rare SVs as the cause of expression changes with large effect sizes, finding a 391-bp intronic inversion increasing *ECHDC1* expression and a 3.6 Mb inversion associated to expression alteration of 3 genes located close to its breakpoints. Nonetheless, in general, the real relevance of inversions on gene-expression heritability has been left out and how much inversions contribute to phenotypic traits in humans has not been addressed successfully.

Given the low success of previous studies at associating inversions with phenotypic variation, we took advantage of the genotyping data generated by the InvFEST project and investigated the potential role of this type of variant on molecular processes and traits at different levels. Although other studies have shown that structural variants have in general higher chances to affect gene expression and be associated with complex traits (Sudmant et al. 2015) and that they show larger effect sizes on gene expression than SNPs (Chiang et al. 2017), it is uncertain whether there is a higher probability that inverted rearrangements affect the measured

molecular phenotypes *a priori*, with the exception of those inversions disrupting gene sequences. However, our findings demonstrate that inversions have an extensive influence on molecular processes compared to other variants and those with interesting functional associations are highlighted for further characterization.

4.2.1 The influence of human inversions on gene expression, epigenetic changes and phenotypic traits

Inhibition of recombination

Although inversions are known to have direct effects at the molecular level, such as altering gene structures (Puig et al 2015a), indirect effects related to the particular role of inversions on recombination are the most intensively studied for these variants. As already mentioned, inhibition of recombination caused by the generation of unbalanced gametes due to single crossovers at the inverted region in heterozygotes could protect allele combinations that are favourable for specific conditions or environments, and inversion alleles could evolve separately and diverge, increasing the number of nucleotide changes that remain linked to the orientation in which they were generated (Kirkpatrick 2010; Puig et al. 2015a). Over time, this leads to the establishment of clearly differentiated haplotypes, with many variants in LD within the inversion that may have functional implications.

A well-known case is the 17q21.31 inversion, which is the one with largest effects discovered to date. We already commented that several human diseases have been associated to the two separated haplotypes maintained by the inversion, H1 and H2, including Alzheimer's disease (Myers et al. 2005), Parkinson's disease (Skipper et al. 2004; Zabetian et al. 2007; Tobin et al. 2008; Setó-Salvia et al. 2011), neuroticism (Okbay et al. 2016), progressive supranuclear palsy and corticobasal degeneration (Baker et al. 1999; Webb et al. 2008), but also increased rates of recombination and higher fertility in females (Stefansson et al. 2005; Fledel-Alon et al.

2011). In this thesis, we performed a phenome-wide analysis by using the GWAS Catalog data, discovering many other associations. For instance, phenotypes reported to be associated with variants that are linked to HsInv0573 include several hematological measurements, ovarian cancer, lung function, type 1 diabetes and bone mineral density, among others. The accumulated nucleotide changes that have appeared as a consequence of the two orientations divergence are supposedly driving such effects by altering gene expression. Consistently, our findings confirmed expression changes already reported by others in blood, cerebellum and cortex (de Jong et al. 2012). Nonetheless, distinguishing the specific causal variant is not possible without empirical evidences due to the high LD block (Figure 4.2).

It is noteworthy that HsInv0573, together with other long inversions like HsInv0786 (~171 kb) and HsInv1110 (~440 kb), cover a major fraction of functional consequences at different levels. These inversions are top eQTLs for many genes and expression changes for HsInv0786 already described in a previous publication were replicated (González et al. 2014). However, unlike HsInv0573, the lower LD patterns of these inversions with neighbouring SNPs indicate that there has been some recurrence that has prevented complete divergence between orientations. Thus, causal variants driving these effects would be partially protected by the inversion only under low recurrence rates. Additionally, we cannot discard that the inversion itself has a direct effect on the phenotype through the reversal of significant portions of the genome that may alter the chromosomal environment or change the position of regulatory sequences. This could be the case of HsInv1110, which often appears as the sentinel variant over the rest of nearby variants. Interestingly, these three inversions tend to affect the expression of several pseudogenes, whose levels are strongly affected in virtually all cell types. Although initially pseudogenes were regarded as non-functional elements of the genome, later studies have revealed that they play important roles in transcriptional and post-transcriptional regulation (Xiao-Jie et al. 2015). Long inversions are expected to have a bigger impact on fertility since there is a higher chance of recombination events causing unbalanced chromosomes and pervasive influence on gene expres-

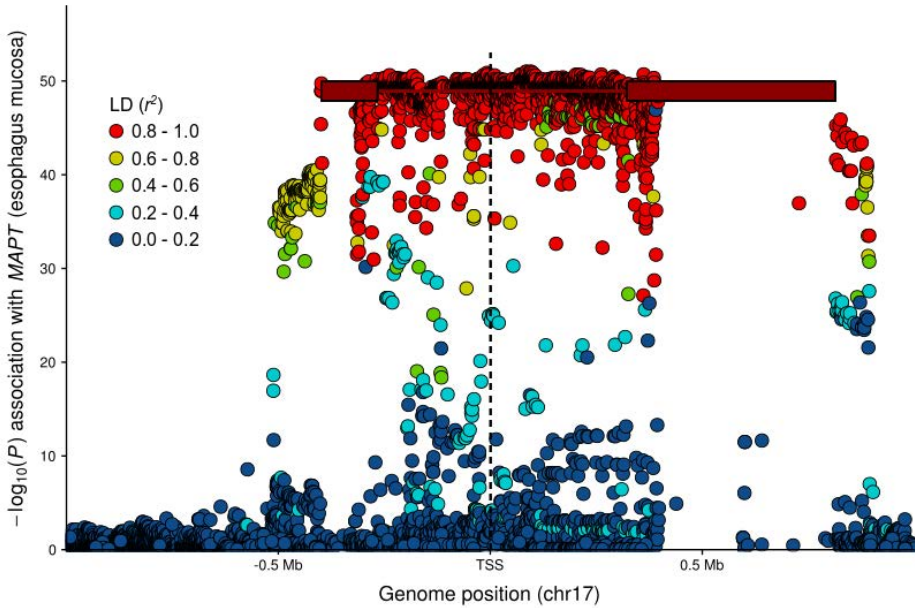


Figure 4.2 – HsInv0573 maintains two separated haplotypes with functional implications. Many variants in high LD located within the inversion define two differentiated haplotypes. Some of these polymorphisms may be associated to gene-expression changes and act as eQTLs of the *MAPT* gene in esophagus mucosa, for example, as shown in the graph. Each dot represents a genomic variant and colors illustrate LD with HsInv0573. Rectangles indicate the segmental duplications at inversion breakpoints.

sion may be indicative of compensatory effects. In this sense, inversions HsInv1051 and HsInv0379, which are also among the longest ones -225 and 415 kb, respectively- are associated with the disruption of a gene and the creation of a novel transcript, reinforcing the idea that potentially beneficial expression changes could help overcome the negative consequences associated to the longer inversions.

Examples of mutational effects caused by inversions

We found several cases of inversions with clear molecular consequences causing direct mutations on gene structure. One of the best examples in this thesis of a gene disruption is caused by inversion HsInv1051, where gene *CCDC144B* spans one of its breakpoints. This gene possesses several potentially coding exons at both sides of one of the segmental duplications implicated in the inversion generation, and the inverted rearrangement moves the promoter and the first two exons more than 200 kb away from the rest (Aguado et al. 2014). *CCDC144B* is part of a gene family with two other members, *CCDC144A* and *CCDC144C*, with ~99% identity and similar exon-intron structure. However, whereas *CCDC144A* encodes a protein of 1,427 amino acids, *CCDC144B* and *CCDC144C* have different frameshift changes that reduce their coding capacity to 725 and 646 amino acids, respectively. There is evidence of *CCDC144A* at protein level in several fetal and adult tissues (Kim et al. 2014), but the possible function of the full-length and truncated proteins is not clear and *CCDC144B* is considered to be a pseudogene. Nevertheless, *CCDC144B* expression is supported by RNA-Seq data and it is expressed at higher levels in cerebellum according to the GTEx project (GTEx Consortium 2017).

In addition, we discovered that HsInv1051 creates a novel fusion transcript derived from the relocated *CCDC144B* exons and promoter (Giner-Delgado et al. 2019). RNA-Seq profiles from LCL reads mapped against a modified version of the human genome with the inversion orientation (created by reversing in silico the HsInv1051 sequence), revealed that the novel fusion transcript is only expressed in heterozygotes and at higher levels in homozygotes for the inverted allele. This example shows that breaking a gene does not always lead to absence of expression. In fact, the novel transcript has much higher expression compared with the typical low levels of *CCDC144B*. The creation of this fusion transcript includes additional 3' sequences from outside the inversion and as a result *CCDC144B* premature stop codon is lost, potentially explaining the increase in the expression of the novel transcript that would not be under the regulation of the

nonsense-mediated decay mechanism anymore. Interestingly, this inversion is only found at low frequency in African populations, which may be related to its effect on this gene. As mentioned above, similarly, HsInv0379 is only present in East Asian populations at very low frequency and is also implicated in gene disruption by breaking the coding gene *ZNF257*, generating in turn a novel fusion transcript in the same manner as HsInv1051 (Puig et al. 2015b). Another experimentally-validated polymorphic inversion with gene-disrupting effects is HsInv0340, which would move apart the first exon and promoter of the long non-coding RNA *LINC00395*, expressed solely in testis (Aguado et al. 2014). In this last case, we could not analyse gene disruption effects since *LINC00395* structure reconstruction from RNA-Seq did not seem to coincide with GENCODE annotations.

Inversions might also exchange gene sequences located within highly similar inverted repeats at breakpoints (Aguado et al. 2014; Puig et al. 2015a; Vicente-Salvador et al. 2017). Considering that genes are likely reconstructed with little or no nucleotide changes, in principle no expression changes or other functional consequences are expected, but that is not always true. A fascinating case is HsInv0030, in which the inverted conformation is the ancestral and most frequent orientation (Pang et al. 2013; Vicente-Salvador et al. 2017; Giner-Delgado et al. 2019). HsInv0030, exchanges the promoter and first exon of the two chymotrypsinogen precursor genes *CTRB1* and *CTRB2* that partially overlap with the 1.5 kb segmental duplications at the inversion breakpoints (Pang et al. 2013). These genes, expressed almost exclusively in pancreas, are differentiated by just a few nucleotide changes in the first and last exon located outside of the repeated elements. Thus, it was believed that this exchange of exons would just reconstruct transcripts with slightly different amino acid combinations at the beginning and the end of the protein sequence, but we found diverse evidence indicating expression changes in these hybrid transcripts.

First, we found that top lead GTEx eQTLs for *CTRB1* and *CTRB2* were the variants in highest LD with the inversion (Giner-Delgado et al. 2019). Due to the presence of the inversion, RNA-Seq reads from GTEx are

mixed in the inverted orientation because all reads for each sample were mapped against the reference genome indistinctly. We hypothesized that the total number of reads coming from both *CTRB1* and *CTRB2* must change in each orientation, because otherwise no differences would have been detected. Second, to determine the exact molecular consequences of the exchange, we imputed HsInv0030 on GTEx samples and mapped reads from pancreas in reference and *in silico* inverted genomes in order to reconstruct *CTRB1-CTRB2* hybrid and reference transcripts. Taking the *CTRB1* and *CTRB2* from the human reference genome (*CTRB1-R* and *CTRB2-R*), the inversion generated a *CTRB2* with the promoter and first exon of *CTRB1* (*CTRB2-Q*) and the same for *CTRB1* with the promoter and the first exon of *CTRB2* (*CTRB1-Q*). We found that in the ancestral inverted orientation both genes are expressed at similar levels, with *CTRB1-Q* having slightly higher expression than *CTRB2-Q*. Theoretically, the promoter change would just alter expression ratios; that is, *CTRB2-R* expression would go up whereas *CTRB1-R* would decrease in the same proportion. However, we discovered that *CTRB2* expression is significantly much higher than *CTRB1* in the reference orientation. Therefore, additional regulatory elements are likely to be involved in *CTRB1* and *CTRB2* regulation, and this control is disrupted by the inversion exchange. Interestingly, the inversion is also associated to changes in an histone modification peak, and this regulatory element may interact with *CTRB1-2* genes and be related to the observed expression differences. Thus, besides the mutational effect, this could be also an illustrative example of a position effect by an inversion.

On the other hand, these genes appear to produce different alternative isoforms, which cannot be discarded to have an effect on phenotype. The proteins encoded by *CTRB1* and *CTRB2* are pancreatic proteolytic enzymes, which cleave aromatic amino acids such as tyrosine, tryptophan or phenylalanine. Nonetheless, there are differences in the primary cleavage specificity and catalytic activity of these two proteins (Szabó and Sahin-Tóth 2012). Thus, expression changes can have consequences in chymotrypsin efficiency. In this regard, we found that the HsInv0030 region is enriched in GWAS hits involving type 1 and 2 diabetes, pancreatic

cancer, insulin secretion, and cholesterol and triglyceride levels. Consistently, the inversion has been discovered as a new risk locus for alcoholic chronic pancreatitis in a recent GWAS analysis (Rosendahl et al. 2018). Another important variant in this region is a low-frequent 584-bp deletion within one of the segmental duplication at the breakpoints that appears in some of the chromosomes with the inverted orientation and removes one of the *CTRB2-Q* exons (Pang et al. 2013; Vicente-Salvador et al. 2017). However, this variant was not detected in the GTEx samples we analysed for hybrid transcript reconstruction.

Other NAHR-mediated inversions possess genes embedded within inverted repeats. HsInv0278, HsInv0344, HsInv0374 and HsInv393 have paralogous genes with identical sequences at the breakpoints, whereas HsInv0209, HsInv0241 and HsInv396 contain genes with small differences. No effects have been detected in any of these cases. One interesting example of a similar case is HsInv0014, which reorganizes sequences of the genes *AKR1C1* and *AKR1C2*, situated inside the inversion. While both genes are lowly expressed in LCLs, the inverted rearrangement creates a highly-expressed isoform that starts outside the inversion and ends in *AKR1C2*. However, we observed that this effect was dependent on the tissue, indicating that the same genetic variant combined with the regulatory landscape of each tissue, can have diverse consequences. In adipose tissue, only *AKR1C1* and *AKR1C2* isoforms within the inversion were strongly expressed, whereas in testis, *AKR1C1* presents higher or lower expression than *AKR1C2* depending on inversion orientation. *AKR1C1* and *AKR1C2* encode aldo-keto reductases, which are enzymes that catalyze the inactivation of male and female sex hormones (Penning et al. 2000). While *AKR1C1* is responsible for the reduction of progesterone to its inactive form (Couture et al. 2003), *AKR1C2* metabolizes dihydrotestosterone (DHT) (Takahashi et al. 2009). Progesterone levels are known to be essential in several processes such as pregnancy, whereas DHT metabolization is key in androgen receptor signaling regulation in the prostate. Moreover, DHT and progesterone play a role in the development of prostate (Bauman et al. 2006; Stanborough et al. 2006) and breast cancer (Ji et al. 2004), respectively. Indeed, the involvement of *AKR1C1* and *AKR1C2*

in progression of diverse types of cancer has been widely documented (Li et al. 2016b; Wengers et al. 2016; Shiiba et al. 2017). Thus, these proteins play a critical role in human traits and disease, and HsInv0014 may modulate some complex phenotypes. In fact, we found that HsInv0014 is in perfect linkage with GWAS variants associated to height and waist-hip ratio in Europeans, which fits well with some of the potential functions of the above hormones.

Another possibility is that inversions affect particular exonic parts. In this sense, HsInv0201, as part of a more complex rearrangement, directly removes one coding exon from the gene *SPINK14*. Specifically, HsInv0201 deletes the third protein-coding exon of one of the two *SPINK14* isoforms. Since another *SPINK14* transcript without the affected exon exists, the inversion might be producing this alternatively-spliced isoform in some individuals. Nevertheless, HsInv0201 disrupts *SPINK14* coding capacity, since a premature stop codon is generated and the protein would lose half of its length. In fact, we found that HsInv0201 *O2* allele is associated with lower levels of *SPINK14* across distinct GTEx tissues, but also HsInv0201 is associated with changes in several nearby genes, being the top eQTL for *SPINK13* and *SCGB3A2* in thyroid and *SPINK6* in Salmonella-infected cells (Nédélec et al. 2016; Alasso et al. 2018). Interestingly, a novel *SPINK6* isoform discovered by the miTranscriptome project (Iyer et al. 2015) would be completely eliminated by HsInv0201, since the inversion deletes its promoter and first exon. In fact, the inverted rearrangement is associated with lower levels of *SPINK6* during immune response, which is consistent with the absence of this putative transcript. Moreover, a secondary pQTL signal linked to the inversion indicates that this change has an impact on *SPINK6* plasma protein measurements (Sun et al. 2018). *SPINK* gene family members contain a Kazal-type serine protease inhibitor domain and are involved in protection against proteolytic degradation of epithelial and mucosal tissues. In particular, *SPINK6* has been shown to play a role in inhibiting the activity of kallikrein-related peptidases in human skin and oral epithelium to maintain homeostasis and normal epithelial barrier functions (Meyer-Hoffert et al. 2010; Plaza et al. 2016). Furthermore, *SPINK6* has been implicated in cancer devel-

opment, acting both as tumor suppressor in hepatocellular carcinoma (Ge et al. 2017) or promoting nasopharyngeal carcinoma metastasis (Zheng et al. 2016). Although we could not link HsInv0201 to any human phenotypic trait, we showed that HsInv0201 presents clear signals of balancing selection in Giner-Delgado et al. (2019). Altogether, these results could suggest that HsInv0201 may play a role in immune response. Another example of a similar effect is HsInv0102, a 2 kb inversion that inverts an alternative 5'-UTR exon of the RHOH gene, which no longer can be part of the final transcript in the inverse orientation (Puig et al. 2015a; Giner-Delgado et al. 2019), as we have studied in more detail in one of the chapters of this thesis and is discussed in an independent section below.

Examples of inversions with position effects

The vast majority of the thousands of genetic loci that contribute to disease susceptibility and other phenotypic traits identified by GWAS are located in non-protein-coding portions of the genome. Interpretation of non-coding variants has been limited by our understanding of the molecular functions of such potential regulatory elements. In our case, we can ask the same question: what might these non-coding inversions be doing? The most part of our inversion set is located in non-coding regions, many of them far from any gene. Nonetheless, we found associations between several of these inversions and molecular phenotypes. Functional SNPs in non-coding regions have been described to disrupt DNA sequence motifs (gain or loss of enhancers, for example), alter miRNA binding or affect splicing in introns (Ward and Kellis 2012; Khurana et al. 2016). In the introduction, we defined the influence of inversions on mRNA levels by mechanisms other than direct mutation of gene sequences as a 'position effect'. Although this term was mainly related to the relocation of genes and regulatory elements to a different context from its original chromosomal environment, here we extend this concept to any inversion located at non-coding regions that has any association with phenotypic variation, independently of its size and the genomic elements altered.

A good example is HsInv0031, where, as already mentioned, the inverted allele is associated with decreased expression of *FAM92B* in cerebellum and is in almost perfect LD in Europeans ($r^2 = 0.98$) with SNP rs2937145 associated with Alzheimer's disease risk (Figure 4.3). FAM92 proteins contain domains that are known to bind and curve the lipid membrane (Ren et al. 2006; Qualmann et al. 2011; Daumke et al. 2014; Suetsugu et al. 2014). Specifically, FAM92B has been seen to facilitate ciliogenesis, likely through membrane remodeling (Li et al. 2016a; Siller et al. 2017), which seems to play an important role in multiciliated cells of the brain ventricles to provide the motive force for cerebrospinal fluid circulation. HsInv0031 is also linked to changes in histone modifications (H3K4me1) and the chromatin accessibility of the surrounding region in LCLs. Despite *FAM92B* is located 50 kb downstream the inversion and the affected histone modification peak is more than 0.5 Mb upstream, they all are located at the boundaries of one TAD according to LCL Hi-C data (Rao et al. 2014) and could be close in the 3 dimensional space. However, to validate our hypothesis, we should confirm that both the histone modification change and the topological domain occur also in cerebellum. Other interesting candidate is inversion HsInv0347. HsInv0347 is located around 92 kb away from *c14orf39* (*SIX6OS1*), the expression of which seems to be regulated by this inversion in skeletal muscle. *SIX6OS1* has been involved in eye development through the regulation of gene expression of the *SIX6* transcription factor (Alfano et al. 2005). Consistently, HsInv0347 area is enriched in GWAS signals associated to glaucoma and optic disc and nerve measurements, but none of these signals were in high LD with the inversion (Giner-Delgado et al. 2019). However, this is not so unexpected given the moderate recurrence levels of the inversion, compared to some of the previous ones. In this regard, it would be interesting to analyse directly gene expression in retina and the possible effect of the inversion. Moreover, *SIX6OS1* encodes a protein product that is a central element of the synaptonemal complex and is essential for fertility (Gómez et al. 2016).

One of the most striking findings in this work is inversion HsInv1110, which seems to regulate a high fraction of genes and, as we have already

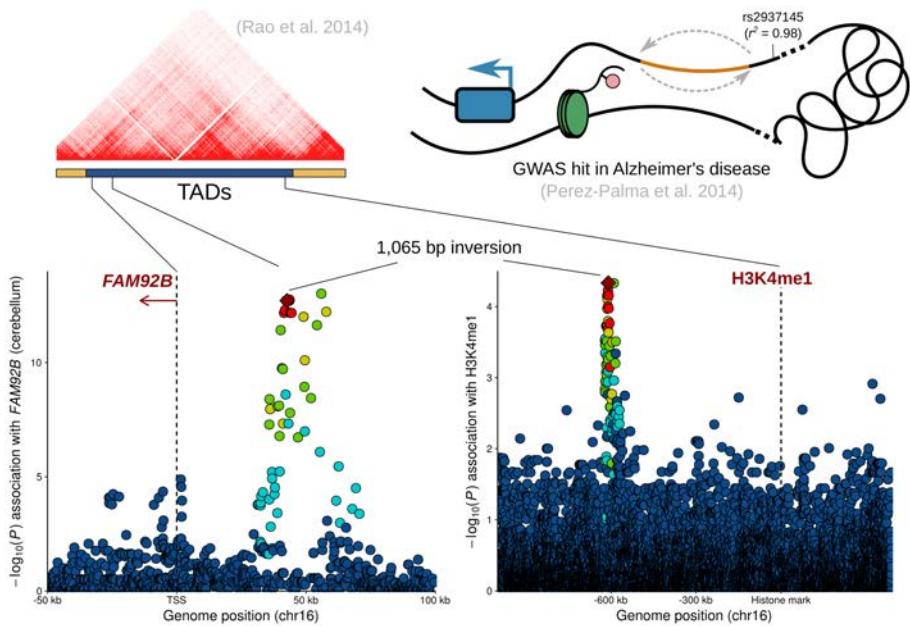


Figure 4.3 – Potential functional impact of HsInv0031. Inversion HsInv0031 is linked to *FAM92B* expression in cerebellum and histone methylation in LCLs. Additionally, HsInv0031 is associated to open chromatin in the surrounding region. All these elements are presumably located close in the three dimensional space according to Hi-C data in LCLs (Rao et al. 2014). HsInv0031 act as a regulatory variant affecting histone marks which in turn may affect *FAM92B* gene expression with potential phenotypic effects (Alzheimer’s disease risk).

commented, its effects do not appear to be driven by different haplotypic configurations. In fact, HsInv1110 is not tagged by nearby variants. With exception of rs71330942 ($r^2 = 0.82$), no more variants are in LD with the inversion ($r^2 < 0.6$) and HsInv1110 may manifest its effects through gene regulatory mechanisms. Besides many pseudogenes and lncRNAs, HsInv1110 is associated with the expression of the *LIP1* and *RBM11* protein coding genes. *LIP1*, which encodes a protein specialized in hydrolyzing phosphatidic acid, has been associated with dyslipidemia, a condition related to an increased risk of heart disease, obesity, and pancreatitis (Wen et al. 2003), whereas *RBM11* is a regulator of alternative splicing (Pe-

drotti et al. 2012). Unfortunately, it is not possible to determine the association of the inversion with phenotypic traits due to its low LD with other SNPs. Indeed, rs71330942 is not included in any commercial SNP array according to the LDLink web portal (Machiela and Chanock 2015). Therefore, this interesting inversion deserves further characterization in future analyses.

In summary, recombination inhibition has been highlighted as the main mechanisms by which inversions can exert functional effects. However, we have demonstrated that there are many cases of reorganization of gene sequences in the human genome due to polymorphic inversions. There are inversions breaking genes, and some generate new transcripts from the promoters and initial exons of these disrupted genes with the incorporation of novel sequences. Inversions can also alter internal gene structures by reversing specific exons, that will be deleted from the final mRNA form, or rearrange genes forming hybrid transcripts. Finally, many inversions located at non-coding portions of the genome may alter regulatory sequences with consequences in gene expression.

4.2.2 HsInv0102 association with Rhoh protein levels and blood cancer susceptibility

As an example of an inversion with a clear molecular effect, we characterized in detail the functional consequences of the relatively small inversion (2 kb) HsInv0102, which removes the alternatively-spliced non-coding exon E8 of the *RHOH* gene from the final mRNA form. In this case, the main question is which could be the consequences of the elimination of this alternative exon. Although E8 is considered a minor exon, we estimated that it is present in ~6% of *RHOH* transcripts in LCLs in absence of HsInv0102 or rs7699141 disrupting alleles. Since gene expression is controlled at each step from DNA to protein, this percentage may not be insignificant. Specifically, we observed that higher levels of E8 inclusion do not seem to affect global *RHOH* transcription, but they correlate with lower levels of the RhoH protein. We also found that E8 together with

constitutive E9 create an upstream ORF of 108 nucleotides, which is the longest and closest to the *RHOH* coding part. The presence of ORFs with these characteristics could affect *RHOH* translation efficiency as it is known to happen for other genes (Chew et al. 2016), although in this case ribosome occupancy correlated well with RNA levels, and HsInv0102 and rs7699141 could act as protein-specific QTLs (Battle et al. 2015). Thus, we hypothesize that E8 inclusion could be related with lower translation rate or higher rates of RhoH degradation. In this sense, further investigation is needed to establish its exact role and understand the possible functional effects of the inversion.

At the evolutionary level, exon E8 is the result of splicing sites provided by Alu and LINE repeat elements, where the C ancestral allele of rs7699141 later derived into the T allele within humans, producing a much more efficient donor splicing site. It is known that selection can operate in alternatively spliced novel exons without strongly harmful consequences (Sorek 2007). In this case, since E8 is only present in a small proportion of mature mRNAs, it may be free to acquire new roles while *RHOH* remains functional. Conversely, if E8 is included at higher rates, it could lead to a defective *RHOH* activity. Therefore, E8 origin could suppose an opportunity to evolve and gain functions, but at the same time it should be maintained under control. Interestingly, the nucleotide variation analysis has detected two quite divergent haplotypes (*Std1* and *Std2*) in a small region around E8 that are relatively old and that are linked to each of the alleles of the SNP involved in the splicing of the alternative exon. This pattern is consistent to the action of balancing selection over the two alleles, and the presence of the HsInv0102 inversion may help to reinforce the effect of one of the alleles by removing completely the exon. For instance, RhoH is associated to cell migration (Tajadura-Ortega et al. 2018), which is an essential process in tissue development or proper immune function, but at the same time it can be responsible for diseases such as cancer or inflammatory disorders. Thus, balancing selection acting on E8 inclusion levels could help to maintain RhoH protein levels at an optimal value. However, it is very difficult to test the existence of these effects.

An extensive body of literature shows the involvement of RhoH in a wide variety of signalling pathways and the presence of potential 5'-UTR regulatory variants affecting the levels of this protein is of great interest. Whereas previous studies failed to include RhoH protein in their analysis (Battle et al. 2015; Folkersen et al. 2017; Suhre et al. 2017, Yao et al. 2018; Sun et al. 2018), we found that rs7699141 and HsInv0102 act as pQTLs of this protein. In fact, the combination of both variants resulted as the most significant signal in the gene region. In particular, we found lower levels of RhoH protein only in individuals with at least one copy of the low expressing E8 haplotype (either *Std2/Std1* or *Std1/Std1*). Thus, the increase in frequency of the inversion could indicate that the presence of the ancestral rs7699141-C allele alone was not enough to reach a threshold of exon exclusion and recover RhoH protein levels. Remarkably, this effect is just seen in the levels of the protein and not of the transcript. In fact, how *RHOH* expression is regulated is not known, as shown in the discrepancies among eQTL studies (Wen et al. 2015; GTEx Consortium 2017; Vösa et al. 2018). We found rs11723134 as lead eQTL for *RHOH* in LCLs, although we could not check its effect at the protein level, since this SNP was not polymorphic in the YRI population. When we looked to the signalling pathways affected, we also confirmed that RhoH has effects in other genes. For example, decreased *CLU* and increased *CD44* gene expression could be the results of compensatory changes in response to the lower levels of RhoH, and further study is needed to investigate other potential changes that might not yet have been found.

As already mentioned, RhoH has been involved in different blood cancers, with different effects. Down-regulation of RhoH is detected in HCL and it is an unfavourable prognostic for AML (Galiegui-Zouitina et al. 2008; Iwasaki et al. 2008), whereas CLL is characterized by high *RHOH* mRNA expression levels (Sanchez-Aguilera et al. 2010). Moreover, RhoH overexpression in hematopoietic progenitor cells resulted in a reduction in cell migration in response to chemokines (Gu et al. 2005), whereas *RHOH* depletion also resulted in a lower cell migration in prostate cancer (Tajadura-Ortega et al. 2018). Although these results seem paradoxical, we should distinguish between the role of RhoH in differentiation or de-

velopment of progenitor cells and its function in the diverse mature blood cells. Additionally, the exposure to cytokines can modify *RHOH* activity and in some tumours *RHOH* overexpression could not be the causal event but a consequence of NF- κ B pathway feedback or act as a compensatory mechanism. For instance, *RHOH* is upregulated during eosinophil differentiation and in patients suffering from hypereosinophilic syndrome, but at the same time its overexpression reduces eosinophil differentiation, acting as a negative regulator of eosinophilopoiesis (Stoeckle et al. 2016). In this regard, we observed that HsInv0102 could be implicated in several blood malignancies, suggesting a possible role during hematopoietic differentiation pathways. In fact, molecular evidences show that RhoH has antitumorigenic activity in hematopoietic progenitor cells (Gu et al. 2006). For instance, RhoH underexpression would reduce apoptosis and promote proliferation and migration, which are aspects that can give rise to cancer. Consistent with RhoH antitumorigenic activity in hematopoietic progenitor cells (Gu et al. 2006), the presence of HsInv0102 correlates with higher levels of RhoH and appears as a protective allele for blood cancer with a moderate odds ratio of 0.75, which is equivalent to that of other discovery loci in complex diseases. Although the largest GWAS carried out for CLL so far did not find any variant linked to HsInv0102 (Law et al. 2017), this could be in part due to the low imputation accuracy of the inversion with common arrays. Indeed, we observed that HsInv0102 cannot be accurately imputed or recovered by some common SNP arrays, highlighting the necessity of using arrays with higher coverage, such as OMNI arrays, or directly genotyping the structural variants for association studies. We expect that the future extension of the analysis to a larger number of cases will confirm our results.

In summary, we hypothesize that HsInv0102 and rs7699141 act as regulatory variants acting on RhoH translation. Although rs7699141 appears as the main variant affecting *RHOH* E8 inclusion levels due to its higher frequency, inversion HsInv0102 seems to play a critical role by enhancing rs7699141 effects to exceed a specific threshold in E8 skipping with clear influences in RhoH levels and possibly cancer susceptibility. Small shifts in the delicate balance of RhoH and other Rho GTPases can modulate

signaling pathways and may likely have implications in leukocyte differentiation and function. Therefore, although more analysis are needed to completely clarify the function of the different *RHOH* isoforms, this is another example of how inversions could have important consequences in the human genome.

4.2.3 HsInv0124 is associated with the expression of immunologically-related genes

Another good candidate inversion with potentially quite important functional and phenotypic consequences is HsInv0124, which overlaps the *IFITM1*, *IFITM2* and *IFITM3* genes. The *IFITM* gene family, with *IFITM3* as the most studied member, is already known to act as an indispensable barrier to viral infection in vivo and in vitro (Everitt et al. 2012; Bailey et al. 2014). Given that HsInv0124 is a moderately recurrent inversion (Giner-Delgado et al. 2019), its effects have been likely missed in GWAS and eQTL studies so far. However, we have demonstrated that inversion genotypes can be accurately imputed with enough SNP coverage, especially in European populations, although the lower precision levels for African populations limit the imputation in cohorts from such ancestry. Thanks to this information, we have found that inversion HsInv0124 is associated with *IFITM2* and *IFITM3* gene expression in LCLs, affecting also the non-coding transcriptional landscape and histone modification patterns at the *IFITM* region. Moreover, in stimulated LCLs and monocytes, we identified additional association of HsInv0124 with other immunologically-related genes and with the interferon signaling and anti-viral response pathways. Therefore, although future experiments should be aimed to discover potential links of this inversion with viral illnesses susceptibility in human cohorts, our findings suggest that HsInv0124 status may have clinical implications in infectious disease outcomes of individuals.

Previous studies have already identified significant associations of other polymorphisms with *IFITM3* gene expression levels and that may af-

fect human influenza disease progression (REFs). Everitt and colleagues (2012) hypothesized that SNP rs12252 might create an alternative splicing site within this gene that would generate a protein with a 21 amino acid deletion, and was associated with more severe flu in patients and higher flu virus infection of cells in vitro (Everitt et al. 2012). Nonetheless, a couple of studies suggested that rs12252 has no real effect on IFITM3 protein length (Randolph et al. 2017; Makvandi-Nejad et al. 2018) and the increase of flu risk associated to the SNP remains controversial (Mills et al. 2014; López-Rodríguez et al. 2016; Randolph et al. 2017; Prabhu et al. 2018; Chen et al. 2018). IAV-exposed monocyte expression data allowed us to confirm that this SNP is the main variant responsible of switching among *IFITM3* isoforms and that it likely affects full-length and truncated protein isoforms (Figure 4.4). In absence of rs12252 alternative allele (G), both spliced *IFITM3.4* and not-spliced *IFITM3.1* isoforms are expressed at similar levels. However, rs12252-G allele produces higher levels of splicing, eliminating completely the *IFITM3.1* transcript in homozygote individuals and expressing just the isoform that encodes for the putative shorter protein form. Another SNP that has been associated with *IFITM3* expression is rs34481144 (Allen et al. 2017). This SNP is in moderate LD with the inversion in CEU individuals ($r^2 \approx 0.62$) and its reference G allele has been described to promote DNA methylation in the *IFITM3* promoter, blocking the binding of a CTCF transcriptional repressor and increasing gene expression. Unfortunately, we could not check if this variant was lead eQTL for any association since it was not included in our monocyte expression dataset. Nonetheless, SNP rs7944394 was the second most significant signal for *IFITM3* expression variation in monocytes exposed to LPS, R848 and IAV after inversion HsInv0124, and this variant can be considered a proxy of rs34481144 due to the high LD between them in Europeans ($r^2 \approx 0.8$). In addition, this SNP was the sentinel variant in cells stimulated with Pam3CSK4 and HsInv0124 was the second most significant variant in this case. Since rs7944394 and HsInv0124 are in moderate LD ($r^2 \approx 0.6$) and there are other less significant variants despite having higher LD with the inversion, this suggests that both rs34481144 and HsInv0124 are strongly affecting *IFITM3* levels, especially in stimulated cells, while rs12252 affects mainly the proportion

of the two alternatively-spliced isoforms.

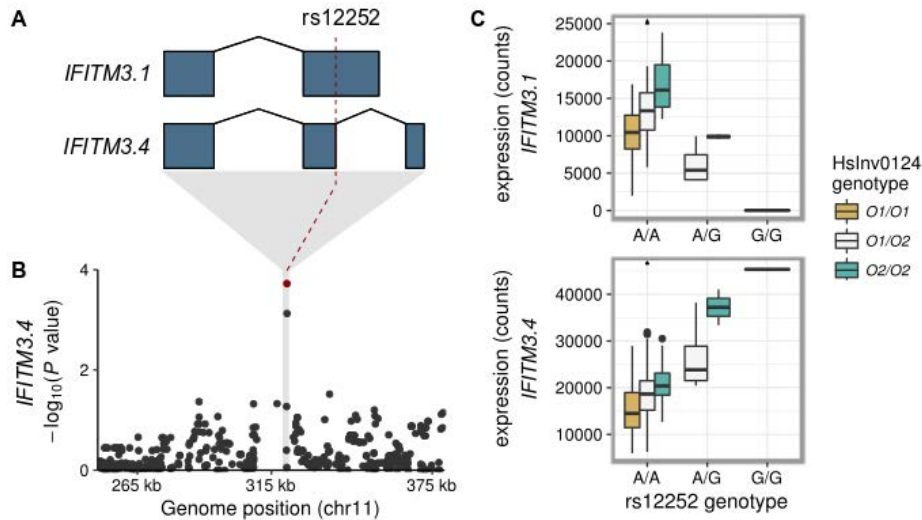


Figure 4.4 – Interplay between SNP rs12252 and HsInv0124 alleles. (A) Structure of *IFITM3.1* (ENST00000399808.5) and *IFITM3.4* (ENST00000526811.4) transcripts. SNP rs12252 is an *IFITM3* splice site polymorphism that creates a new alternative intron and a slightly shorter protein. (B) Manhattan plot for eQTL association to *IFITM3.4* expression in monocytes exposed to IAV. rs12252 position is indicated in red. (C) Boxplots of the expression (estimated counts) of the two *IFITM3* isoforms by rs12252 and HsInv0124 genotype in monocytes stimulated with IAV. Alternative rs12252 allele (G) correlates with higher levels of the *IFITM3.4* spliced isoform, whereas *IFITM3.1* appears to be missing in homozygote individuals. At the same time, the inversion regulates global expression of both isoforms, with higher expression levels associated to the *O2* orientation.

Similarly to the rs34481144-G allele that promotes DNA methylation in the *IFITM3* promoter, thanks to the analysis of the inversion we discovered a novel mechanism that appears to regulate *IFITM* genes through changes in histone modification marks. As we have seen, HsInv0124 is within a CRD of correlated histone modification peaks that seems to be affected mainly by the inversion. Although regulatory elements within this CRD are coordinated, inversion HsInv0124 is associated more intensely

with those marks located at their breakpoints, indicating that not all elements are affected in the same manner within the regulatory domain. We hypothesize that HsInv0124 affects peak activity levels due to the proximity of chromatin marks located within the inverted segment, which would explain that the higher chromatin activity switches from one side to the other with the inversion. Although HsInv0124 is not lead eQTL of *IFITM2* and *IFITM3* in LCLs, the correlation of histone marks with expression levels of these genes suggests that the inversion is affecting gene activity through this regulatory mechanism. Remarkably, *IFITM5* and *NLRP6*, which are regulated by the inversion in IAV-stimulated monocytes, could be also under the influence of these epigenetic regulatory signals. A future more complete characterization of these epigenetic changes across several cells and conditions could be helpful to solve this question.

Other potential mechanisms by which HsInv0124 could be affecting gene expression is through the lncRNAs located within the inverted repeats at the inversion breakpoints (*AC136475.1* and *AC136475.2*). Both non-coding transcripts are differentially expressed between *O1* and *O2* orientations, which is supposedly caused by the exchange of the promoters by the inversion. Since these RNAs overlap with *IFITM2* and *IFITM3*, they could act regulating the expression of the two protein-coding genes. In particular, it has been observed that non-coding genes can act as scaffolds for chromatin modifier enzymes (Davidovich and Cech 2015; Marchese et al. 2017) and the lncRNAs could be responsible of the differential epigenetic marks around the inversion. This agrees well with the fact that the expression of the two overlapping transcripts increases or decreases coordinately with the inversion in LCLs, although the switch of the location of the epigenetic signals could also be the one responsible of the change in the two pairs of genes. However, *AC136475.1* and *AC136475.2* behave in opposite manner in stimulated monocytes and may interfere with *IFITM2* and *IFITM3* transcription. In any case, because HsInv0124 does not disrupt *a priori* any of the genes involved, the observed expression changes constitute a nice example of position effect exerted by an inversion polymorphism in humans.

Despite other studies have focused on SNPs affecting *IFITM3* expression, this is the first study that shows a direct link between a ~ 7 kb inversion at the *IFITM* locus and expression of diverse genes, including also associations for *IFITM5* and *NLRP6*. Interestingly, these three genes are up-regulated in IAV-infected monocytes and individuals with the *O2* orientation, indicating that this orientation may be related with a stronger immune response. We have already commented the important function of IFITM3 in virus infection defense. In addition, the intracellular protein NLRP6, which is a member of the nucleotide-binding oligomerization domain-like receptors, plays a central role in the immune response to diverse microorganisms (Levy et al. 2017). This protein participates in inflammation regulation and host defense, and *Nlrp6*-deficient mice have been observed to be more susceptible to enteric infection (Wang et al. 2015). NLRP6 binds viral RNA and promotes the expression of a large number of genes induced by interferon that are critical for antiviral response (Levy et al. 2017). The fact that HsInv0124 regulates *NLRP6* which, in turn, regulates interferon-stimulated genes is consistent with the differential expression patterns between *O1/O1* and *O2/O2* LCLs stimulated with interferon and the higher expression of genes related to interferon signalling pathway in *O2/O2* individuals. Similarly, *IFITM5* is part of *IFITM* family and, although it was originally identified as a bone formation factor (Moffatt et al. 2008; Hanagata et al. 2011a) expressed only in osteoblasts (Zhang et al. 2012), we found higher expression in monocytes exposed to IAV. Moreover, recent studies have revealed that IFITM5 is involved in immune system activity through the interaction with other proteins to form a complex that induces the expression of several immunologically-related genes (Hanagata et al. 2011b; Tsukamoto et al. 2013). Therefore, the three main genes regulated by the inversion seem to act in the same manner by regulating immunological pathways, which may affect the response to multiple agents and influence the outcome of an infection.

Altogether, the influence of HsInv0124 on the expression levels of *IFITM3* and other genes related to viral defense such as *NLRP6* and *IFITM5* indicates that it may have important phenotypic consequences in the defense

against infections. Interestingly, we have found that the frequency of the inversion shows differences between populations (Giner-Delgado et al. 2019), which could have been shaped by selective processes caused by the multiple viral infections that have afflicted the different continents during human evolution. HsInv0124 *O2* orientation has an intermediate frequency in Europe (43%), whereas it is at much higher frequency in Africa (87%) or East Asia (96%), where it is almost fixed. Because of differences in the allele frequencies among populations, further study is necessary to investigate the interplay among distinct regulatory polymorphism in this region. Indeed, rs12252 minor allele is at much higher frequency in populations from Asian descent, whereas rs34481144 alternative A allele is present more frequently in European populations. Therefore, it is possible that all these elements may have additive effects and give rise to a diverse range of immunological outcomes, explaining some of the apparently contradictory results for the association of these SNPs with flu susceptibility in different cohorts (Everitt et al. 2012; Zhang et al. 2013; Wang et al. 2014; Mills et al. 2014; Zhang et al. 2015; López-Rodríguez et al. 2016; Randolph et al. 2017; Allen et al. 2017; Prabhu et al. 2018; Chen et al. 2018; Makvandi-Nejad et al. 2018). Unfortunately, infection susceptibility is a very difficult trait to study. However, we are planning to genotype experimentally the inversion in influenza-infected patients with different disease severity, as well as imputing the inversion in samples affected by other viral diseases such as AIDS, for which there is an accurate quantification of the HIV viral load (Fellay et al. 2009; McLaren et al. 2015). In addition, we have carried out preliminary experiments of flu virus infection of LCLs of *O1/O1* or *O2/O2* individuals *in vitro*, but as different expression changes have been reported in distinct cell types, it would be interesting to explore other cell lines, tissues and conditions. In fact, monocytes become macrophages, which are related to antiviral activities and could be one of the ideal cell types to test inversion association to viral infection. In summary, HsInv0124 appears as a novel important variant to take into account in infection and immune response, and these and other future studies will contribute to have a better idea of its functional effects and the role of positive selection in its increase in frequency.

4.2.4 Do inversions have more functional effects than expected?

Through the integrative analysis carried out in this thesis, we have found that approximately half of the studied inversions appear to have some type of functional effect. This is quite high when compared to SNPs or indels. It has been estimated that $\sim 0.3\%$ of all tested SNPs and indels in GTEx tissues are lead *cis*-eQTLs (GTEx Consortium 2017; Chiang et al. 2017), demonstrating a disproportionately large role of inversions on gene expression. However, this enrichment of inversions with functional effects might not be so surprising if we take into account the potential negative effects in fertility for long inversions that have been already mentioned. The same negative effects are expected to happen for all inversions with different degrees depending on its genetic size and the frequency to which crossover might happen within the inversion. According to this, it should be possible to predict an inversion size below which the inversions behave as nearly-neutral. This value is difficult to estimate accurately, but depending on different factors such as the local recombination rate and the actual evolutionary cost of the generation of unbalanced gametes could range between 1 and 100 kb (Luca Ferretti, personal communication). Inversions above that size that reach a certain frequency in human populations are expected to have some positive effects that compensate the negative cost in fertility (Giner-Delgado et al. 2019). Therefore, it is not surprising that many of the studied inversions may have functional consequences.

4.3 Future perspectives

Characterizing all the variation in our genome is an essential task to discover potential novel associations and have a complete understanding of genome function. The missing heritability of many human traits may be partially explained by the inclusion of polymorphisms in association studies that have been inaccessible by most existing genotyping methods so

far (Eichler et al. 2010). In this sense, inversions have remained understudied in humans and it is still difficult to estimate the real number of inversions that exist in the human genome and their frequency. Moreover, the almost nonexistent representation of recurrent inversions in GWAS, where most of them cannot be accurately predicted by commonly-used imputation algorithms, means that their effects have been largely missed. Although it is premature to conclude the impact of this type of variants in common diseases, we have shown a few examples of inversions linked to distinct molecular processes and human traits. However, determining the global contribution to complex phenotypes requires specific high-throughput techniques to genotype inversions in large cohorts of thousands of individuals in a relatively easy and affordable way. Therefore, this highlights the necessity to improve current inversion genotyping methods to cover the maximum number of inversions, and especially those recurrent and with low LD to other variants.

GWAS signals are significantly more likely to act as eQTLs (Nicolae et al. 2010). Inevitably, QTL finding is intended to serve as a key tool for translational medicine and is becoming a central component to determine the genetic basis of phenotypic traits and disease susceptibility. In this thesis we identified several inversion candidates associated with gene-expression changes in multiple tissues. Besides, we investigated the molecular mechanisms by which these changes may occur. While additional data from other cell types and conditions is generated, more comprehensive catalogues of QTLs will allow us to get a deeper understanding of the biology underlying disease. Nevertheless, even when the statistical evidence is strong, empirical proof of the predicted effects must be obtained with experimental perturbations in model organisms and relevant cell lines. The CRISPR/Cas9 system has emerged as a powerful toolkit to carry out genome editing and can potentially be used to induce targeted inversions of interest in an efficient manner (Cheong et al. 2018). Furthermore, complementary experimental analysis could also be performed to directly connect inversion alleles to true functional outcomes. Possible examples are quantification of cell death against viral infection, digestive activity of CTRB1 and CTRB2 enzymes, or catalyzation of sex hormones in order

to test HsInv0124, HsInv0030 and HsInv0014 effects, respectively.

Many questions, although important, fall outside of the scope of this thesis, such as the exact role of human inversions on epistasis. Some others have just been started to be elucidated. For instance, a few inversions have been implicated in the formation of chimeric transcripts, but this phenomenon has not been assessed in a systematic manner and the possible function of these new transcripts has not been investigated. Also, genotype imputation has been proven as an efficient approach, but a benchmarking analysis of the advantages and disadvantages derived from using different reference panels, as well as the selection of which inversions have to be experimentally genotyped and which can be reliably imputed from GWAS arrays, is necessary. Overall, this work represents an important contribution to the understanding of these little characterized structural variants in humans, but there is still a lot of work to be done. The future of functional genomics field looks promising and the application of these findings is expected to help in the interpretation of personal genomes, which will be translated in potential improvements of human health.

Chapter 5

Conclusions

The conclusions of this work are the following:

1. By analyzing genotyping data of 109 autosomal and chromosome X human inversions in multiple individuals, we have found that 60% have other variants in perfect LD ($r^2 \approx 1$). Two thirds of inversions not tagged by other variants cannot be accurately predicted with typical imputation algorithms, highlighting the importance of direct genotyping.
2. We have shown that 53 inversions are lead QTLs or in high LD with top QTLs for gene expression and epigenetic changes across different tissues and cell lines, suggesting an enrichment of inversions with functional effects compared to SNPs and indels.
3. Long inversions (>100 kb) are involved in approximately 80% of INV-eQTLs associations (HsInv0573, HsInv0786, HsInv1110, HsInv0379, HsInv1057). This significant number of potential effects of these inversions could help compensate their negative cost in fertility due to production of unbalanced gametes by recombination.

4. We have characterised three inversions (HsInv1057, HsInv0014 and HsInv0030) that disrupt genes and generate fusion or hybrid transcripts, being also associated with gene expression changes.
5. We found an enrichment of GWAS signals in inversion and flanking regions. In addition, 14 inversions are in high LD with some of these trait-associated variants, which constitutes a notable fraction for this type of variation (13%) taking into account the reduced number of inversions with tag SNPs.
6. HsInv0102 inverts an alternative 5'-UTR exon from *RHOH* and is associated with RhoH protein levels, potentiating the effect of a previously existing variant. Potential consequences of this inversion have been likely missed in GWAS due to its low imputation accuracy with common arrays and we estimated that the inverted allele acts as a protective locus on blood cancer susceptibility with a moderate odds ratio of 0.75.
7. HsInv0124 regulates the expression in LCLs of several genes in the *IFITM* region, including *IFITM2* and *IFITM3*, through changes in histone modification patterns. Moreover, in stimulated LCLs and monocytes, the inversion has a more pervasive effect on the expression of these and other genes related with immunity and anti-viral response pathways, like *IFITM5* and *NLRP6*, indicating that it may play an important role in virus infection defense.

Chapter 6

Bibliography

- Aguado, C. et al (2014). Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. *PLoS Genetics*, 10(3):e1004208.
- A.J., B., R., S. and T., K. (2002). Histone methylation: dynamic or static?. *Cell*, 109:801–806.
- Akhtar, A. and Becker, P.B. (2000). Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in *Drosophila*. *Molecular Cell*, 5(2):367–375.
- Alfano, G. et al (2005). Natural antisense transcripts associated with genes involved in eye development. *Human Molecular Genetics*, 14(7):913–923.
- Alkan, C., Coe, B.P. and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–76.
- Allen, E. et al (2017). SNP-mediated disruption of CTCF binding at the IFITM3 promoter is associated with risk of severe influenza in humans. *Nature Medicine*, 23(8):975–983.
- Alves, J.M. et al (2015). Reassessing the evolutionary history of the 17q21 inversion polymorphism. *Genome biology and evolution*, 7(12):3239–48.
- Ameur, A. et al (2017). SweGen: A whole-genome data resource of genetic variability in a cross-section of the Swedish population. *European Journal of Human Genetics*, 25(11):1253–1260.
- Amini-Bavil-Olyaei, S. et al (2013). The antiviral effector IFITM3 disrupts intracellular cholesterol homeostasis to block viral entry. *Cell Host and Microbe*, 13(4):452–464.

Chapter 6. Bibliography

- Anders, S., Reyes, A. and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017.
- Anderson, A.R. et al (2005). The latitudinal cline in the In(3R)Payne inversion polymorphism has shifted in the last 20 years in Australian *Drosophila melanogaster* populations. *Molecular Ecology*, 14(3):851–858.
- Anton, E. et al (2005). Sperm studies in heterozygote inversion carriers: a review. *Cytogenetic and Genome Research*, 111(3-4):297–304.
- Antonacci, F. et al (2009). Characterization of six human disease-associated inversion polymorphisms. *Human Molecular Genetics*, 18(14):2555–66.
- Antonarakis, S.E. et al (1995). Factor VIII gene inversions in severe hemophilia A: results of an international consortium study. *Blood*, 86(6):2206–12.
- Argani, P. et al (2017). RBM10-TFE3 renal cell carcinoma: A potential diagnostic pitfall due to cryptic intrachromosomal Xp11.2 inversion resulting in false-negative TFE3 FISH. *American Journal of Surgical Pathology*, 41(5):655–662.
- Audano, P.A. et al (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, 176(3):663–675.e19.
- Ayala, D. et al (2018). African malaria mosquitoes. 71(3):686–701.
- Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics*, 14(2):113–124.
- Bagnall, R.D., Giannelli, F. and Green, P.M. (2005). Polymorphism and hemophilia A causing inversions in distal Xq28: a complex picture. *Journal of Thrombosis and Haemostasis*, 3(11):2598–9.
- Bagnall, R.D. et al (2002). Recurrent inversion breaking intron 1 of the factor VIII gene is a frequent cause of severe hemophilia A. *Blood*, 99(1):168–174.
- Bailey, C. et al (2014). IFITM-Family Proteins: The Cell's First Line of Antiviral Defense. *Annu Rev Virol*, 1:261–283.
- Bailey, C.C. et al (2012). Ifitm3 Limits the Severity of Acute Influenza in Mice. *PLoS Pathogens*, 8(9):e1002909.
- Bailey, J.A. et al (2002). Recent segmental duplications in the human genome. *Science (New York, N.Y.)*, 297(5583):1003–7.
- Baker, M. et al (1999). Association of an extended haplotype in the tau gene with progressive supranuclear palsy. *Human Molecular Genetics*, 8(4):711–5.
- Bannister, A.J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395.

- Barreiro, L.B. et al (2012). Deciphering the genetic architecture of variation in the immune response to *Mycobacterium tuberculosis* infection. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1204–1209.
- Baruzzo, G. et al (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*, 14(2):135–139.
- Battle, A. et al (2015). Genomic variation. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664–7.
- Battle, A. et al (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, 24(1):14–24.
- Bauman, D.R. et al (2006). Transcript profiling of the androgen signal in normal prostate, benign prostatic hyperplasia, and prostate cancer. *Endocrinology*, 147(12):5806–5816.
- Berg, P.R. et al (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity*, 119(6):418–428.
- Berg, P.R. et al (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports*, 6:23246.
- Besenbacher, S. et al (2015). Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nature Communications*, 6:5969.
- Bibikova, M. et al (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295.
- Boerma, E.G. et al (2009). Translocations involving 8q24 in Burkitt lymphoma and other malignant lymphomas: A historical review of cytogenetics in the light of today's knowledge. *Leukemia*, 23(2):225–234.
- Boettger, L.M. et al (2012). Structural haplotypes and recent evolution of the human 17q21.31 region. *Nature Genetics*, 44(8):881–5.
- Bonder, M. (2017). Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics*, 49(1):131–138.
- Bondeson, M.L. et al (1995). Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Human Molecular Genetics*, 4(4):615–21.
- Bosch, N. et al (2009). Nucleotide, cytogenetic and expression impact of the human chromosome 8p23.1 inversion polymorphism. *PLoS One*, 4(12):e8269.

Chapter 6. Bibliography

- Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33(Suppl):228–37.
- Boutet, E. et al (2016). UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods in Molecular Biology*, 1374:23–54.
- Brass, A.L. et al (2009). The IFITM Proteins Mediate Cellular Resistance to Influenza A H1N1 Virus, West Nile Virus, and Dengue Virus. *Cell*, 139(7):1243–1254.
- Bray, N. et al (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–7.
- Brown, A. et al (2017). Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nature Genetics*, 49(12):1747–1751.
- Browning, S.R. and Browning, B.L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *American Journal of Human Genetics*, 81(5):1084–1097.
- Cáceres, A. and González, J.R. (2015). Following the footprints of polymorphic inversions on SNP data: From detection to association tests. *Nucleic Acids Research*, 43(8):e53.
- Cardoso-Moreira, M. et al (2019). Gene expression across mammalian organ development. *Nature*, 571(7766):505–509.
- Chae, H.D. et al (2010). RhoH regulates subcellular localization of ZAP-70 and Lck in T cell receptor signaling. *PLoS ONE*, 5(11):e13970.
- Chaisson, M. et al (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–11.
- Chaisson, M.J.P. et al (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10(1):1784.
- Chen, T. et al (2018). Association between rs12252 and influenza susceptibility and severity: An updated meta-analysis. *Epidemiology and Infection*, 13:1–9.
- Cheong, T.C., Blasco, R.B. and Chiarle, R. (2018). The CRISPR7Cas9 System as a Tool to Engineer Chromosomal Translocation In Vivo. *Advances in Experimental Medicine and Biology*, 1044:39–48.
- Cherry, L. et al (2004). RhoH is required to maintain the integrin LFA-1 in a nonadhesive state on lymphocytes. *Nature Immunology*, 5(9):961–7.

- Chew, G.L., Pauli, A. and Schier, A.F. (2016). Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nature Communications*, 7(May):11663.
- Chheda, H. et al (2017). Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. *European Journal of Human Genetics*, 25(4):477–484.
- Chiang, C. et al (2017). The impact of structural variation on human gene expression. *Nature Genetics*, 49(5):692–9.
- Choudhury, A. et al (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*, 8:2062.
- Christophersen, I.E. et al (2017). Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nature Genetics*, 49(6):946–952.
- Compton, A.A. et al (2016). Natural mutations in IFITM 3 modulate post-translational regulation and toggle antiviral specificity . *EMBO reports*, 17(11):1657–1671.
- Cooper, G.M. et al (2011). A copy number variation morbidity map of developmental delay. *Nature Genetics*, 43(9):838–46.
- Couture, J. et al (2003). Human 20alpha-hydroxysteroid dehydrogenase: crystallographic and site-directed mutagenesis studies lead to the identification of an alternative binding site for C21-steroids. *Journal of Molecular Biology*, 331(3):593–604.
- Crequer, A. et al (2012). Human RHOH deficiency causes T cell defects and susceptibility to EV-HPV infections Find the latest version : Human RHOH deficiency causes T cell defects and susceptibility to EV-HPV infections. *J Clin Invest*, 122(9):3239–3247.
- Dallery, E. et al (1995). TTF, a gene encoding a novel small G protein, fuses to the lymphoma-associated LAZ3 gene by t(3;4) chromosomal translocation. *Oncogene*, 10(11):2171–8.
- Dana, M. and Stoian, V. (2012). Association of pericentric inversion of chromosome 9 and infertility in romanian population. *Maedica*, 7(1):25–9.
- Daryadel, A. et al (2009). RhoH/TTF Negatively Regulates Leukotriene Production in Neutrophils. *The Journal of Immunology*, 182(10):6527–6532.
- Daumke, O., Roux, A. and Haucke, V. (2014). BAR domain scaffolds in dynamin-mediated membrane fission. *Cell*, 156(5):882–892.
- Davegårdh, C. et al (2018). DNA methylation in the pathogenesis of type 2 diabetes in humans. *Molecular Metabolism*, 14:12–25.

Chapter 6. Bibliography

- Davidovich, C. and Cech, T.R. (2015). The recruitment of chromatin modifiers by long noncoding RNAs: Lessons from PRC2. *RNA*, 21(12):2007–2022.
- Day, F. et al (2017). Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nature Genetics*, 49(6):834–841.
- de Koning, A.P.J. et al (2011). Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genetics*, 7(12):e1002384.
- De Sousa Abreu, R. et al (2009). Global signatures of protein and mRNA expression levels. *Molecular BioSystems*, 5(12):1512–1526.
- Degner, J.F. et al (2012). DNase-I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394.
- Delaneau, O. et al (2017). A complete tool set for molecular QTL discovery and analysis. *Nature Communications*, 8:15452.
- Delaneau, O. et al (2019). Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science*, 364(6439):eaat8266.
- Desai, T.M. et al (2014). IFITM3 Restricts Influenza A Virus Entry by Blocking the Formation of Fusion Pores following Virus-Endosome Hemifusion. *PLoS Pathogens*, 10(4):e1004048.
- Dobin, A. et al (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.
- Dobzhansky, T. and Sturtevant, A.H. (1938). Inversions in the Chromosomes of *Drosophila Pseudoobscura*. *Genetics*, 23(1):28–64.
- Du, P. et al (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11:587.
- Eden, E. et al (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:1–7.
- Edwards, S.L. et al (2013). Beyond GWASs: Illuminating the dark road from association to function. *American Journal of Human Genetics*, 93(5):779–797.
- Elliott, L.T. et al (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*, 562(7726):210–216.
- Engström, P.G. et al (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, 10(12):1185–1191.
- Enroth, S. et al (2014). Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nature Communications*, 5:4684.

- Escaramís, G., Docampo, E. and Rabionet, R. (2015). A decade of structural variants: description, history and methods to detect structural variation. *Briefings in Functional Genomics*.
- Everitt, A.R. et al (2012). IFITM3 restricts the morbidity and mortality associated with influenza. *Nature*, 484(7395):519–23.
- Fairfax, B.P. et al (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, 343(6175):1246949.
- Fang, H. et al (2019). A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nature Genetics*, 51(7):1082–1091.
- Fang, Z. et al (2012). Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics*, 191(3):883–894.
- Fawcett, J.A. and Innan, H. (2013). The role of gene conversion in preserving rearrangement hotspots in the human genome. *Trends in Genetics*, 29(10):561–568.
- Feeley, E.M. et al (2011). IFITM3 inhibits influenza a virus infection by preventing cytosolic entry. *PLoS Pathogens*, 7(10):e1002337.
- Feuk, L. (2010). Inversion variants in the human genome: role in disease and genome architecture. *Genome Medicine*, 2(2):11.
- Feuk, L. et al (2005). Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genetics*, 1(4):e56.
- Finan, C. et al (2017). The druggable genome and support for target identification and validation in drug development. *Science Translational Medicine*, 9(383):eaag1166.
- Flavahan, W. et al (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, 529(7584):110–4.
- Folkersen, L. et al (2017). Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genetics*, 13(4):e1006706.
- Foster, T.L. et al (2016). Resistance of Transmitted Founder HIV-1 to IFITM-Mediated Restriction. *Cell Host and Microbe*, 20(4):429–442.
- Franceschini, N. et al (2012). Discovery and fine mapping of serum protein loci through transethnic meta-analysis. *American Journal of Human Genetics*, 91(4):744–753.
- Francioli, L.C. et al (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46(8):818–825.
- Franke, M. et al (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–269.

Chapter 6. Bibliography

- Fueller, F. and Kubatzky, K.F. (2008). The small GTPase RhoH is an atypical regulator of haematopoietic cells. *Cell Communication and Signaling*, 6:1–13.
- Fuller, Z.L. et al (2016). Genomics of natural populations: How differentially expressed genes shape the evolution of chromosomal inversions in *Drosophila pseudoobscura*. *Genetics*, 204(1):287–301.
- Gaidano, G. et al (2003). Aberrant somatic hypermutation in multiple subtypes of AIDS-associated non-Hodgkin lymphoma. *Blood*, 102(5):1833–1841.
- Galiègue-Zouitina, S. et al (2008). Underexpression of RhoH in hairy cell leukemia. *Cancer Research*, 68(12):4531–4540.
- Gamazon, E.R. et al (2013). Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Molecular Psychiatry*, 18(3):340–346.
- Gao, F. et al (2013). Inversion-mediated gene fusions involving NAB2-STAT6 in an unusual malignant meningioma. *British Journal of Cancer*, 109(4):1051–1055.
- Garge, N. et al (2010). Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells. *Molecular and Cellular Proteomics*, 9(7):1383–1399.
- Gaunt, T.R. et al (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome biology*, 17:61.
- GE, K. et al (2017). Serine protease inhibitor kazal-type 6 inhibits tumorigenesis of human hepatocellular carcinoma cells via its extracellular action. *Oncotarget*, 8(4):5965–5975.
- Gerstein, M. et al (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100.
- Ghavi-Helm, Y. et al (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature Genetics*, 51(8):1272–1282.
- Gibbs, J.R. et al (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in Human Brain. *PLoS Genetics*, 6(5):e1000952.
- Giner-Delgado, C. et al (2019). Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nature Communications*, page In press.
- Gold, L. et al (2010). Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS ONE*, 5(12):e15004.
- Gómez-H, L. et al (2016). C14ORF39/SIX6OS1 is a constituent of the synaptonemal complex and is essential for mouse fertility. *Nature Communications*, 7:13298.

- González, J.R. et al (2007). SNPAssoc: An R package to perform whole genome association studies. *Bioinformatics*, 23(5):644–645.
- Gorello, P. et al (2013). Inv(11)(p15q22)/NUP98-DDX10 fusion and isoforms in a new case of de novo acute myeloid leukemia. *Cancer Genetics*, 206(3):92–96.
- Gouw, S.C. et al (2012). F8 gene mutation type and inhibitor development in patients with severe hemophilia A: Systematic review and meta-analysis. *Blood*, 119(12):2922–2934.
- Gruber, T.A. et al (2012). An Inv(16)(p13.3q24.3)-encoded CBFA2T3-GLIS2 fusion protein defines an aggressive subtype of pediatric acute megakaryoblastic leukemia. *Cancer Cell*, 22(5):683–97.
- Grubert, F. et al (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1051–1065.
- Grundberg, E. et al (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–9.
- GTEX Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (New York, N. Y.)*, 348(6235):648–60.
- GTEX Consortium (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–13.
- Gu, Y. et al (2006). RhoH GTPase recruits and activates Zap70 required for T cell receptor signaling and thymocyte development. *Nature Immunology*, 7(11):1182–90.
- Gu, Y. et al (2005). RhoH, a hematopoietic-specific Rho GTPase, regulates proliferation, survival, migration, and engraftment of hematopoietic progenitor cells. *Blood*, 105(4):1467–1475.
- Gündođdu, M.S. et al (2010). The haematopoietic GTPase RhoH modulates IL3 signalling through regulation of STAT activity and IL3 receptor expression. *Molecular Cancer*, 9:1–13.
- Gurdasani, D. et al (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, 517(7534):327–32.
- Gutierrez-Arcelus, M. et al (2015). Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. *PLoS Genetics*, 11(1):e1004958.
- Hamatani, K. et al (2014). A novel RET rearrangement (ACBD5/RET) by pericentric inversion, inv(10)(p12.1;q11.2), in papillary thyroid cancer from an atomic bomb survivor exposed to high-dose radiation. *Oncol Rep*, 32(5):1809–14.

Chapter 6. Bibliography

- Hanagata, N. and Li, X. (2011). Osteoblast-enriched membrane protein IFITM5 regulates the association of CD9 with an FKBP11-CD81-FPRP complex and stimulates expression of interferon-induced genes. *Biochem Biophys Res Commun*, 409(3):378–84.
- Hanagata, N. et al (2011). Characterization of the osteoblast-specific transmembrane protein IFITM5 and analysis of IFITM5-deficient mice. *J Bone Miner Metab*, 29(3):279–90.
- Hannon, E. et al (2016a). An integrated genetic-epigenetic analysis of schizophrenia: Evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biology*, 17(1):176.
- Hannon, E. et al (2018). Elevated polygenic burden for autism is associated with differential DNA methylation at birth. *Genome Medicine*, 10(1):19.
- Hannon, E. et al (2016b). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nature Neuroscience*, 19(1):48–54.
- Harris, R. et al (2010). Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotechnology*, 28(10):1097–105.
- Harrow, J. et al (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–74.
- Heintzman, N.D. et al (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–8.
- Hellman, A. and Chess, A. (2007). Gene body-specific methylation on the active X chromosome. *Science*, 315(5815):1141–3.
- Hnisz, D. et al (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280):1454–1458.
- Hobart, H.H. et al (2010). Inversion of the Williams syndrome region is a common polymorphism found more frequently in parents of children with Williams syndrome. *American Journal of Medical Genetics*, 154C(2):220–8.
- Hoffmann, A.A. and Rieseberg, L.H. (2008). Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology, Evolution, and Systematics*, 39:21–42.
- Hon, C.C. et al (2017). An atlas of human long non-coding RNAs with accurate 5 ends. *Nature*, 543(7644):199–204.

- Howie, B.N., Donnelly, P. and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529.
- Hsu, L.Y. et al (1987). Chromosomal polymorphisms of 1, 9, 16, and Y in 4 major ethnic groups: a large prenatal study. *American Journal of Medical Genetics*, 26(1):95–101.
- Huang, I.C. et al (2011). Distinct patterns of IFITM-mediated restriction of filoviruses, SARS coronavirus, and influenza A virus. *PLoS Pathogens*, 7(1):e1001258.
- Huddleston, J. et al (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, 27(5):677–85.
- Hulsen, T., de Vlieg, J. and Alkema, W. (2008). BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*, 9(1):488.
- Iafrate, A.J. et al (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9):949–951.
- Iliakis, G. et al (2004). Mechanisms of DNA double strand break repair and chromosome aberration formation. *Cytogenetic and Genome Research*, 104(1-4):14–20.
- International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.
- International Human Genome Sequencing Consortium (2004). International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.
- Iwamoto, K. et al (2004). Expression of HSPF1 and LIM in the lymphoblastoid cells derived from patients with bipolar disorder and schizophrenia. *Journal of Human Genetics*, 49(5):227–231.
- Iwasaki, T. et al (2008). Prognostic implication and biological roles of RhoH in acute myeloid leukaemia. *European Journal of Haematology*, 81(6):454–460.
- Iyer, M. et al (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, 47(3):199–208.
- Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(Suppl):245–54.
- Jansen, I. et al (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer’s disease risk. *Nature Genetics*, 51(3):404–413.

Chapter 6. Bibliography

- Jenuwein, T. and Allis, C. (2001). Translating the histone code. *Science*, 293(5532):1074–80.
- Ji, Q. et al (2004). Selective loss of AKR1C1 and AKR1C2 in breast cancer and their potential effect on progesterone signaling. *Cancer Research*, 64(20):7610–7617.
- Johansson, Å. et al (2013). Identification of genetic variants influencing the human plasma proteome. *Proceedings of the National Academy of Sciences of the United States of America*, 110(12):4673–4678.
- John, S.P. et al (2013). The CD225 Domain of IFITM3 Is Required for both IFITM Protein Association and Inhibition of Influenza A Virus and Dengue Virus Replication. *Journal of Virology*, 87(14):7837–7852.
- Jones, P. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–92.
- Joron, M. et al (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477(7363):203–6.
- Kaneko, H. et al (2014). Inversion of chromosome 7q22 and q36 as a sole abnormality presenting in myelodysplastic syndrome: A case report. *Journal of Medical Case Reports*, 8:268.
- Kasowski, M. et al (2013). Extensive variation in chromatin states across humans. *Science*, 342(6159):750–752.
- Kehr, B. et al (2017). Diversity in non-repetitive human sequences not found in the reference genome. *Nature Genetics*, 49(4):588–593.
- Khera, A.V. et al (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50:1219–1224.
- Khurana, E. et al (2016). Role of non-coding sequence variants in cancer. *Nature Reviews Genetics*, 17(2):93–108.
- Kichaev, G. et al (2019). Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *American Journal of Human Genetics*, 104(1):65–75.
- Kidd, J.M. et al (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64.
- Kilpinen, H. et al (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, 342(6159):744–747.
- Kim, J. et al (2018). KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. *Scientific Reports*, 8(1):5677.

- Kim, M. et al (2014). A draft map of the human proteome. *Nature*, 509(7502):575–81.
- Kim, S. et al (2013). Influence of Genetic Variation on Plasma Protein Levels in Older Adults Using a Multi-Analyte Panel. *PLoS ONE*, 8(7):e70269.
- Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biology*, 8(9):e1000501.
- Kleinjan, D.J. and van Heyningen, V. (1998). Position effect in human genetic disease. *Human Molecular Genetics*, 7(10):1611–8.
- Koolen, D.A. et al (2012). Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nature Genetics*, 44(6):639–41.
- Koolen, D.A. et al (2008). Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. *Journal of Medical Genetics*, 45(11):710–720.
- Koolen, D.A. et al (2006). A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nature Genetics*, 38(9):999–1001.
- Korbel, J.O. et al (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, 318(5849):420–6.
- Kouzarides, T. (2007). Chromatin Modifications and Their Function. *Cell*, 128(4):693–705.
- Krimbas, C.B. and Powell, J.R. (1992). *Drosophila inversion polymorphism*. CRC Press, Boca Raton.
- Kronenberg, Z.N. et al (2018). High-resolution comparative analysis of great ape genomes. *Science*, 360(6393):eaar6343.
- Kunte, K. et al (2014). Doublesex is a mimicry supergene. *Nature*, 507(7491):229–232.
- Lahousse, S. et al (2004). Structural features of hematopoiesis-specific RhoH/ARHH gene: high diversity of 5'-UTR in different hematopoietic lineages suggests a complex post-transcriptional regulation. *Gene*, 343(1):55–68.
- Lakich, D. et al (1993). Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nature Genetics*, 5(3):236–41.
- Lam, E.T. et al (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*, 30(8):771–6.
- Lam, H. et al (2010). Nucleotide-resolution analysis of structural variants using Break-Seq and a breakpoint library. *Nature Biotechnology*, 28(1):47–55.
- Lamichhaney, S. et al (2015). Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nature Genetics*, 48(1):84–88.

Chapter 6. Bibliography

- Lappalainen, T. et al (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.
- Lappalainen, T. et al (2019). Genomic Analysis in the Age of Human Genome Sequencing. *Cell*, 177(1):70–84.
- Law, P.J. et al (2017). Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nature Communications*, 8:14175.
- Lawrence, M., Daujat, S. and Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics*, 32(1):42–56.
- Le Beau, M.M. et al (2010). Association of an inversion of chromosome 16 with abnormal marrow eosinophils in acute myelomonocytic leukemia. A Unique Cytogenetic-Clinicopathological Association. *The New England Journal of Medicine*, 309(11):630–6.
- Lee, M.N. et al (2014). Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. *Science*, 343(6175):1246980.
- Leek, J. et al (2019). sva: Surrogate Variable Analysis.
- Lesurf, R. et al (2016). ORegAnno 3.0: A community-driven resource for curated regulatory annotation. *Nucleic Acids Research*, 44(D1):D126–D132.
- Lev Maor, G., Yearim, A. and Ast, G. (2015). The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, 31(5):274–80.
- Levy, M. et al (2017). NLRP6: A Multifaceted Innate Immune Sensor. *Trends in Immunology*, 38(4):248–260.
- Li, F.Q. et al (2016a). BAR Domain-Containing FAM92 Proteins Interact with Chibby1 To Facilitate Ciliogenesis. *Molecular and Cellular Biology*, 36(21):2668–2680.
- Li, H. et al (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.
- Li, K. et al (2013). IFITM Proteins Restrict Viral Membrane Hemifusion. *PLoS Pathogens*, 9(1):e1003124.
- Li, L. et al (2017a). OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. *Genome Biology*, 18(1):1–19.
- Li, W. et al (2016b). The roles of AKR1C1 and AKR1C2 in ethyl-3,4-dihydroxybenzoate induced esophageal squamous cell carcinoma cell death. *Oncotarget*, 7(16):21542–55.

- Li, X. et al (2002). The Hematopoiesis-Specific GTP-Binding Protein RhoH Is GTPase Deficient and Modulates Activities of Other Rho GTPases by an Inhibitory Function. *Molecular and Cellular Biology*, 22(4):1158–1171.
- Li, X. et al (2017b). The impact of rare variation on gene expression across tissues. *Nature*, 550(7675):239–243.
- Li, Y. et al (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, 34(8):816–34.
- Libbrecht, R. and Kronauer, D.J. (2014). Convergent evolution: The genetics of queen number in ants. *Current Biology*, 24(22):R1083–R1085.
- Libioulle, C. et al (2007). Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genetics*, 3(4):0538–0543.
- Liu, P. et al (2011). Frequency of nonallelic homologous recombination Is correlated with length of homology: Evidence that ectopic synapsis precedes ectopic crossing-over. *The American Journal of Human Genetics*, 89(4):580–8.
- Liu, Y. et al (2015). Quantitative variability of 342 plasma proteins in a human twin population. *Molecular Systems Biology*, 11(2):786.
- López, Rodríguez, M. et al (2016). IFITM3 and severe influenza virus infection. No evidence of genetic association. *Eur J Clin Microbiol Infect Dis*, 35(11):1811–1817.
- Lourdusamy, A. et al (2012). Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Human Molecular Genetics*, 21(16):3719–3726.
- Love, M.I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21.
- Lowry, D.B. and Willis, J.H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology*, 8(9):e1000500.
- Lu, J. et al (2011). The IFITM Proteins Inhibit HIV-1 Infection. *Journal of Virology*, 85(5):2126–2137.
- Lu, X. et al (2008). The effect of H3K79 dimethylation and H4K20 trimethylation on nucleosome and chromatin structure. *Nature Structural and Molecular Biology*, 15(10):1122–1124.
- Lucas-Lledó, J.I. and Cáceres, M. (2013). On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. *PLoS One*, 8(4):e61292.

Chapter 6. Bibliography

- Lundberg, M. et al (2017). Genetic differences between willow warbler migratory phenotypes are few and cluster in large haplotype blocks. *Evolution Letters*, 1(3):155–168.
- Lunnon, K. et al (2014). Methyloomic profiling implicates cortical deregulation of ANK1 in Alzheimer’s disease. *Nature Neuroscience*, 17(9):1164–70.
- Lupiáñez, D.G. et al (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025.
- MacArthur, J. et al (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(Database issue):D896–D901.
- Machiela, M.J. and Chanock, S.J. (2015). LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, 31(21):3555–3557.
- Makvandi-Nejad, S. et al (2018). Lack of Truncated IFITM3 Transcripts in Cells Homozygous for the rs12252-C Variant That is Associated with Severe Influenza Infection. *Journal of Infectious Diseases*, 217(2):257–262.
- Mallick, S. et al (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206.
- Mander, A., Hodgkinson, C.P. and Sale, G.J. (2005). Knock-down of LAR protein tyrosine phosphatase induces insulin resistance. *FEBS Letters*, 579(14):3024–3028.
- Manolio, T.A. (2013). Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics*, 14(8):549–558.
- Manolio, T.A. et al (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53.
- Marchese, F.P., Raimondi, I. and Huarte, M. (2017). The multidimensional mechanisms of long noncoding RNA function. *Genome Biology*, 18(1):206.
- Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511.
- Marioni, J.C. et al (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517.
- Marshall, C.R. et al (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics*, 49(1):27–35.
- Matsumoto, M. et al (2009). Large-scale proteomic analysis of tyrosine-phosphorylation induced by T-cell receptor or B-cell receptor activation reveals new signaling pathways. *Proteomics*, 9(13):3549–3563.

- Maunakea, A.K. et al (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303):253–257.
- McVey, M. and Lee, S.E. (2008). MMEJ repair of double-strand breaks (director’s cut): deleted sequences and alternative endings. *Trends in Genetics*, 24(11):529–38.
- McVicker, G. et al (2013). Identification of genetic variants that affect histone modifications in human cells. *Science*, 342(6159):747–749.
- Mefford, H.C. et al (2009). A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. *Genome Research*, 19(9):1579–1585.
- Melzer, D. et al (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genetics*, 4(5):e1000072.
- Merla, G. et al (2010). Copy number variants at Williams-Beuren syndrome 7q11.23 region. *Human Genetics*, 128(1):3–26.
- Messerschmidt, D.M., Knowles, B.B. and Solter, D. (2014). DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes and Development*, 28(8):812–828.
- Metcalfe, K. et al (2000). Elastin: Mutational spectrum in supravalvular aortic stenosis. *European Journal of Human Genetics*, 8(12):955–963.
- Meyer-Hoffert, U. et al (2010). Isolation of SPINK6 in human skin: Selective inhibitor of kallikrein-related peptidases. *Journal of Biological Chemistry*, 285(42):32174–32181.
- Mills, T.C. et al (2014). IFITM3 and susceptibility to respiratory viral infections in the community. *Journal of Infectious Diseases*, 209(7):1028–1031.
- Ming, R.R. and Moore, P.H. (2007). Genomics of sex chromosomes. *Current opinion in plant biology*, 10(2):123–130.
- Moen, E.L. et al (2013). Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics*, 194(4):987–96.
- Moffatt, M.F. et al (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, 448(7152):470–473.
- Moffatt, P. et al (2008). Bril: A novel bone-specific modulator of mineralization. *Journal of Bone and Mineral Research*, 23(9):1497–1508.
- Mohammadi, P. et al (2017). Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Research*, 27(11):1872–1884.
- Monel, B. et al (2017). Zika virus induces massive cytoplasmic vacuolization and paraptosis-like death in infected cells. *The EMBO Journal*, 36(12):1653–1668.

Chapter 6. Bibliography

- Montesinos-Rongen, M. et al (2004). Primary diffuse large B-cell lymphomas of the central nervous system are targeted by aberrant somatic hypermutation. *Blood*, 103(5):1869–1875.
- Montgomery, S.B. et al (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–7.
- Mozdarani, H., Mohseni Maybodi, A. and Karimi, H. (2007). Impact of pericentric inversion of Chromosome 9 [inv (9) (p11q12)] on infertility. *Indian J Hum Genet*, 13(1):26–29.
- Mudhasani, R. et al (2013). IFITM-2 and IFITM-3 but Not IFITM-1 Restrict Rift Valley Fever Virus. *Journal of Virology*, 87(15):8451–8464.
- Myers, A.J. et al (2005). The H1c haplotype at the MAPT locus is associated with Alzheimer’s disease. *Human Molecular Genetics*, 14(16):2399–404.
- Nagasaki, M. et al (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature Communications*, 6:8018.
- Namjou, B. et al (2014). The Effect of Inversion at 8p23 on BLK Association with Lupus in Caucasian Population. *PLoS One*, 9(12):e115614.
- Natarajan, P. et al (2017). Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*, 135:2091–2101.
- Nédélec, Y. et al (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell*, 167(3):657–669.e21.
- Nica, A.C. and Dermitzakis, E.T. (2013). Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362.
- Nica, A.C. et al (2011). The architecture of gene regulatory variation across multiple human tissues: The muTHER study. *PLoS Genetics*, 7(2):e1002003.
- Nishimura, S. et al (2000). A GATA Box in the GATA-1 Gene Hematopoietic Enhancer Is a Critical Element in the Network of GATA Factors and Sites That Regulate This Gene. *Molecular and Cellular Biology*, 20(2):713–723.
- Nishimura, Y. et al (2007). Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Human Molecular Genetics*, 16(14):1682–1698.
- Nomura, T. et al (2018). Chromosomal inversions as a hidden disease-modifying factor for somatic recombination phenotypes. *JCI insight*, 3(6):e97595.

- Nonaka, T. et al (2019). The analysis of chromosomal abnormalities in patients with recurrent pregnancy loss, focusing on the prognosis of patients with inversion of chromosome (9). *Reproductive Medicine and Biology*, 18(3):296–301.
- Obón-Santacana, M. et al (2018). GCAT—Genomes for life: A prospective cohort study of the genomes of Catalonia. *BMJ Open*, 8(3):e018324.
- Oda, H. et al (2009). RhoH Plays Critical Roles in FcεRI-Dependent Signal Transduction in Mast Cells. *The Journal of Immunology*, 182(2):957–962.
- O’Green, H., Echipare, L. and Farnham, P. (2011). Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods in Molecular Biology*, 791:265–86.
- Okada, Y. et al (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381.
- Okbay, A. et al (2016). Genetic variants associated with subjective well-being, depressive symptoms and neuroticism identified through genome-wide analyses. *Nature Genetics*, 48(6):624–33.
- Oldenburg, J., Pezeshkpoor, B. and Pavlova, A. (2014). Historical review on genetic analysis in hemophilia A. *Semin Thromb Hemost*, 40(8):895–902.
- Ongen, H. et al (2017). Estimating the causal tissues for complex traits and diseases. *Nature Genetics*, 49(12):1676–1683.
- Ongen, H. et al (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485.
- Ono, J.G., Worgall, T.S. and Worgall, S. (2014). 17q21 locus and ORMDL3: An increased risk for childhood asthma. *Pediatric Research*, 75(1-2):165–170.
- Osborne, L.R. et al (2001). A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nature Genetics*, 29(3):321–5.
- Pan, Y.R. et al (2018). STAT3-coordinated migration facilitates the dissemination of diffuse large B-cell lymphomas. *Nature Communications*, 9(1):3696.
- Pang, A.W. et al (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, 11(5):R52.
- Pang, A.W.C. et al (2013). Mechanisms of formation of structural variation in a fully sequenced human genome. *Human Mutation*, 34(2):345–54.
- Partida-Pérez, M. et al (2012). De novo inv(17)(p11.2q21.3) in an intellectually disabled girl: Appraisal of 21 inv(17) constitutional instances. *Journal of Genetics*, 91(2):241–244.

Chapter 6. Bibliography

- Pasqualucci, L. et al (2001). Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature*, 412(6844):341–6.
- Pearse, D.E. et al (2014). Rapid parallel evolution of standing variation in a single, complex, genomic region is associated with life history in steelhead/rainbow trout. *Proceedings of the Royal Society B: Biological Sciences*, 281(1783):20140012.
- Pedrotti, S. et al (2012). The RNA recognition motif protein RBM11 is a novel tissue-specific splicing regulator. *Nucleic Acids Research*, 40(3):1021–1032.
- Penning, T. et al (2000). Human 3alpha-hydroxysteroid dehydrogenase isoforms (AKR1C1-AKR1C4) of the aldo-keto reductase superfamily: functional plasticity and tissue distribution reveals roles in the inactivation and formation of male and female sex hormones. *Biochem J*, 351(Pt 1):67–77.
- Perreira, J. et al (2013). IFITMs restrict the replication of multiple pathogenic viruses. *Journal of Molecular Biology*, 425(24):4937–55.
- Pertea, M. et al (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–5.
- Pickrell, J.K. et al (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–72.
- Pinto, D. et al (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466:368–372.
- Pittman, A.M., Fung, H.C. and de Silva, R. (2006). Untangling the tau gene association with neurodegenerative disorders. *Human Molecular Genetics*, 15(2):R188–95.
- Plaza, K. et al (2016). Gingipains of porphyromonas gingivalis affect the stability and function of serine protease inhibitor of Kazal-type 6 (SPINK6), a tissue inhibitor of human kallikreins. *Journal of Biological Chemistry*, 291(36):18753–18764.
- Prabhu, S. et al (2018). Association between IFITM3 rs12252 polymorphism and influenza susceptibility and severity: A meta-analysis. *Gene*, 674:70–79.
- Puente, X. et al (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*, 526(7574):519–24.
- Puig, M. et al (2015a). Human inversions and their functional consequences. *Briefings in functional genomics*, 14(5):369–79.
- Puig, M. et al (2015b). Functional impact and evolution of a novel human polymorphic inversion that disrupts a gene and creates a fusion transcript. *PLoS Genetics*, 11(10):e1005495.

- Puig, M. et al (2019). Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR (ddPCR). *bioRxiv*.
- Pulit, S.L. et al (2019). Meta-Analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Human Molecular Genetics*, 28(1):166–174.
- Purcell, J. et al (2014). Convergent genetic architecture underlies social organization in ants. *Current Biology*, 24(22):2728–2732.
- Quach, H. et al (2016). Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell*, 167(3):643–656.e17.
- Qualmann, B., Koch, D. and Kessels, M.M. (2011). Let’s go bananas: Revisiting the endocytic BAR code. *EMBO Journal*, 30(17):3501–3515.
- Rada-Iglesias, A. et al (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–83.
- Ramsahoye, B.H. et al (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences*, 97(10):5237–5242.
- Randolph, A.G. et al (2017). Evaluation of IFITM3 rs12252 Association with Severe Pediatric Influenza Infection. *Journal of Infectious Diseases*, 216(1):14–21.
- Ranjbar, S. et al (2015). A Role for IFITM Proteins in Restriction of Mycobacterium tuberculosis Infection. *Cell Rep*, 13(5):874–883.
- Ren, G. et al (2006). The BAR Domain Proteins: Molding Membranes in Fission, Fusion, and Phagy. *Microbiology and Molecular Biology Reviews*, 70(1):37–120.
- Richards, A.L. et al (2012). Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Molecular Psychiatry*, 17(2):193–201.
- Ritchie, M.E. et al (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.
- Robinson, J.T. et al (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–6.
- Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40.
- Rosendahl, J. et al (2018). Genome-wide association study identifies inversion in the CTRB1-CTRB2 locus to modify risk for alcoholic and non-alcoholic chronic pancreatitis. *Gut*, 67(10):1855–1863.

Chapter 6. Bibliography

- Ruiz-Arenas, C. et al (2019). scoreInvHap: Inversion genotyping for genome-wide association studies. *PLoS Genetics*, 15(7):e1008203.
- Saha, A. et al (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Research*, 27(11):1843–1858.
- Salm, M.P.A. et al (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Research*, 22(6):1144–53.
- Sanchez-Aguilera, A. et al (2010). Involvement of RhoH GTPase in the development of B-cell chronic lymphocytic leukemia. *Leukemia*, 24(1):97–104.
- Sato, T., Issa, J.P.J. and Kropf, P. (2017). DNA hypomethylating drugs in cancer therapy. *Cold Spring Harbor Perspectives in Medicine*, 7(5):a026948.
- Savidis, G. et al (2016). The IFITMs Inhibit Zika Virus Replication. *Cell Reports*, 15(11):2323–2330.
- Schoggins, J. et al (2011). A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature*, 472(7344):481–5.
- Schubert, C. (2009). The genomic basis of the Williams-Beuren syndrome. *Cell Mol Life Sci*, 66(7):1178–97.
- Schwanhüusser, B. et al (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.
- Semler, O. et al (2012). A mutation in the 5-UTR of IFITM5 creates an in-frame start codon and causes autosomal-dominant osteogenesis imperfecta type v with hyperplastic callus. *American Journal of Human Genetics*, 91(2):349–357.
- Setó-Salvia, N. et al (2011). Dementia risk in Parkinson disease: disentangling the role of MAPT haplotypes. *Archives of Neurology*, 68(3):359–64.
- Shao, H. et al (2018). npInv: accurate detection and genotyping of inversions using long read sub-alignment. *BMC Bioinformatics*, 19(1):261.
- Sharp, A., Cheng, Z. and Eichler, E. (2006). Structural variation of the human genome. *Annual Review of Genomics and Human Genetics*, 7:407–42.
- Shiiba, M. et al (2017). Mefenamic acid enhances anticancer drug sensitivity via inhibition of aldo-keto reductase 1C enzyme activity. *Oncology Reports*, 37(4):2025–2032.
- Shogren-Knaak, M. et al (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science*, 311(5762):844–847.
- Shukla, S. et al (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371):74–9.

- Siller, S.S. et al (2017). Conditional knockout mice for the distal appendage protein CEP164 reveal its essential roles in airway multiciliated cell differentiation. *PLoS Genetics*, 13(12):e1007128.
- Skipper, L. et al (2004). Linkage disequilibrium and association of MAPT H1 in Parkinson disease. *American Journal of Human Genetics*, 75(4):669–77.
- Small, K. et al (2011). Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature Genetics*, 43(6):561–4.
- Small, K., Iber, J. and Warren, S.T. (1997). Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nature Genetics*, 16(1):96–9.
- Soneson, C., Love, M. and Robinson, M. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res*, 4:1521.
- SoreK, R. (2007). The birth of new exons: Mechanisms and evolutionary consequences. *RNA*, 13(10):1603–1608.
- Sotoudeh, A. et al (2017). Pericentric inversion of chromosome 9 in an infant with ambiguous genitalia. *Acta Medica Iranica*, 55(10):655–657.
- Spitz, F. et al (2005). Inversion-induced disruption of the Hoxd cluster leads to the partition of regulatory landscapes. *Nature Genetics*, 37(8):889–93.
- Stanbrough, M. et al (2006). Increased expression of genes converting adrenal androgens to testosterone in androgen-independent prostate cancer. *Cancer Research*, 66(5):2815–2825.
- Stark, R., Grzelak, M. and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*.
- Startek, M. et al (2015). Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. *Nucleic Acids Research*, 43(4):2188–2198.
- Stegle, O. et al (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–7.
- Steinberg, K.M. et al (2012). Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature Genetics*, 44(8):872–80.
- Stoeckle, C. et al (2016). RhoH is a negative regulator of eosinophilopoiesis. *Cell Death and Differentiation*, 23(12):1961–1972.
- Stranger, B.E. et al (2007a). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–53.

Chapter 6. Bibliography

- Stranger, B.E. et al (2012). Patterns of Cis regulatory variation in diverse human populations. *PLoS Genetics*, 8(4):e1002639.
- Stranger, B.E. et al (2007b). Population genomics of human gene expression. *Nature Genetics*, 39(10):1217–1224.
- Strutevant, A.H. (1913). Linear Arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology*, 14:43–59.
- Sturtevant, A.H. (1917). Genetic factors affecting the strength of linkage in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, 3(9):555–8.
- Sturtevant, A.H. (1921a). A case of rearrangement of genes in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, 7(8):235.
- Sturtevant, A.H. (1921b). Genetic Studies on *Drosophila Simulans*. II. Sex-linked group of genes. *Genetics*, 6(43):43–64.
- Sudmant, P.H. et al (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81.
- Suetsugu, S., Kurisu, S. and Takenawa, T. (2014). Dynamic shaping of cellular membranes by phospholipids and membrane-deforming proteins. *Physiological reviews*, 94(4):1219–1248.
- Suhre, K. et al (2017). Erratum: Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature communications*, 8:14357.
- Sun, B.B. et al (2018). Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79.
- Sun, W. and Hu, Y. (2013). eQTL Mapping Using RNA-seq Data. *Statistics in Biosciences*, 5(1):198–219.
- Suzuki, K. et al (2019). Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nature Genetics*, 51(3):379–386.
- Swamynathan, S.K. and Piatigorsky, J. (2002). Orientation-dependent influence of an intergenic enhancer on the promoter activity of the divergently transcribed mouse *Shsp/αB-crystallin* and *Mkbp/HspB2* genes. *Journal of Biological Chemistry*, 277(51):49700–49706.
- Szabó, A. and Sahin-Tóth, M. (2012). Determinants of chymotrypsin C cleavage specificity in the calcium-binding loop of human cationic trypsinogen. *FEBS Journal*, 279(23):4283–4292.

- Tajadura-Ortega, V. et al (2018). An RNAi screen of Rho signalling networks identifies RhoH as a regulator of Rac1 in prostate cancer cell migration. *BMC Biology*, 16(1):29.
- Takahasi, R. et al (2009). The effect of allelic variation in aldo-keto reductase 1C2 on the in vitro metabolism of dihydrotestosterone. *The Journal of Pharmacology and Experimental Therapeutics*, 329(3):1032–1039.
- Tamehiro, N. et al (2019). Ras homolog gene family H (RhoH) deficiency induces psoriasis-like chronic dermatitis by promoting TH17 cell polarization. *The Journal of Allergy and Clinical Immunology*, 143(5):1878–1891.
- Tan, M.H. et al (2017). Dynamic landscape and regulation of RNA editing in mammals. *Nature*, 550(7675):249–254.
- Tcherpakov, M. et al (2002). The p75 neurotrophin receptor interacts with multiple MAGE proteins. *Journal of Biological Chemistry*, 277(51):49101–49104.
- Teague, B. et al (2010). High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(24):10848–53.
- Team, R.C. (2016). R: A Language and Environment for Statistical Computing.
- The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.
- The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.
- The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.
- The International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8.
- The International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426:789–796.
- The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–320.
- T’Hoen, P.A. et al (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature Biotechnology*, 31:1015–1022.
- Thompson, M.J. and Jiggins, C.D. (2014). Supergenes and their role in evolution. *Heredity*, 113(1):1–8.

Chapter 6. Bibliography

- Tian, X. et al (2018). PTPRF as a novel tumor suppressor through deactivation of ERK1/2 signaling in gastric adenocarcinoma. *OncoTargets and Therapy*, 11:7795–7803.
- Tilgner, H. et al (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*, 111(27):9869–9874.
- Tirado-Magallanes, R. et al (2017). Whole genome DNA methylation: Beyond genes silencing. *Oncotarget*, 8(3):5629–5637.
- Troeger, A. et al (2012). RhoH is critical for cell-microenvironment interactions in chronic lymphocytic leukemia in mice and humans. *Blood*, 119(20):4708–4718.
- Tsukamoto, T. et al (2013). Role of S-Palmitoylation on IFITM5 for the Interaction with FKBP11 in Osteoblast Cells. *PLoS ONE*, 8(9):e75831.
- Tukiainen, T. et al (2017). Landscape of X chromosome inactivation across human tissues. *Nature*, 550(7675):244–248.
- Tuttle, E.M. et al (2016). Divergence and functional degradation of a sex chromosome-like supergene. *Current Biology*, 26(3):344–350.
- Veyrieras, J.B. et al (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics*, 4(10).
- Vicente-Salvador, D. et al (2017). Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. *Human Molecular Genetics*, 26(3):567–81.
- Vogel, M.J. et al (2009). High-resolution mapping of heterochromatin redistribution in a *Drosophila* position-effect variegation model. *Epigenetics & Chromatin*, 2(1):1.
- Võsa, U. et al (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*, page 447367.
- Wagner, J.R. et al (2014). The relationship between DNA methylation, genetic and expression inter-individual variation in untranWagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., & Blanchette, M. (2014). The relationship between DNA methylation, genetic and expression inter. *Genome biology*, 15(2):R37.
- Wahl, S. et al (2017). Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, 541(7635):81–86.
- Walter, K. et al (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90.

- Wang, H. et al (2010). RhoH plays distinct roles in T-cell migrations induced by different doses of SDF1 alpha. *Cell Signal*, 22(7):1022–32.
- Wang, P. et al (2015). Nlrp6 regulates intestinal antiviral innate immunity. *Science*, 350(6262):826–830.
- Wang, Y. et al (2017). Recurrent Fusion Genes in Leukemia: An Attractive Target for Diagnosis and Treatment. *Current Genomics*, 18(5):378–384.
- Ward, L. and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology*, 30(11):1095–106.
- Warren, C.J. et al (2014). The antiviral restriction factors IFITM1, 2 and 3 do not inhibit infection of human papillomavirus, cytomegalovirus and adenovirus. *PLoS ONE*, 9(5):e96579.
- Waszak, S.M. et al (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell*, 162(5):1039–1050.
- Webb, A. et al (2008). Role of the tau gene region chromosome inversion in progressive supranuclear palsy, corticobasal degeneration, and related disorders. *2008*, 65(11):1473–8.
- Weber, M. et al (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37(8):853–62.
- Weckselblatt, B. and Rudd, M.K. (2015). Human Structural Variation: Mechanisms of Chromosome Rearrangements. *Trends in Genetics*, 31(10):587–599.
- Wellenreuther, M. and Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, 33(6):427–40.
- Wen, X., Luca, F. and Pique-Regi, R. (2015). Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLoS Genetics*, 11(4):1–29.
- Wen, X.Y. et al (2003). Identification of a novel lipase gene mutated in *lpd* mice with hypertriglyceridemia and associated with dyslipidemia in humans. *Human Molecular Genetics*, 12(10):1131–1143.
- Wenners, A. et al (2016). Stromal markers AKR1C1 and AKR1C2 are prognostic factors in primary human breast cancer. *International Journal of Clinical Oncology*, 21(3):548–56.
- Westra, H.J. et al (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243.

Chapter 6. Bibliography

- Wilkins, C. et al (2013). IFITM1 is a tight junction protein that inhibits hepatitis C virus entry. *Hepatology*, 57(2):461–469.
- Williams, D.E.J. et al (2014). IFITM3 polymorphism rs12252-C restricts influenza A viruses. *PLoS ONE*, 9(10):e110096.
- Wilson, C., Bellen, H. and Gehring, W. (1990). Position effects on eukaryotic gene expression. *Annual Review of Cell Biology*, 6:679–714.
- Wong, L.P. et al (2013). Deep whole-genome sequencing of 100 southeast Asian Malays. *American Journal of Human Genetics*, 92(1):52–66.
- Wu, L. et al (2013). Variation and genetic control of protein abundance in humans. *Nature*, 499(7456):79–82.
- Xiao-Jie, L. et al (2015). Pseudogene in cancer: Real functions and promising signature. *Journal of Medical Genetics*, 52(1):17–24.
- Yang, F. et al (2017). Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Research*, 27(11):1859–1871.
- Yang, X. et al (2014). Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell*, 26(4):577–590.
- Yao, C. et al (2018). Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications*, 9(1):3268.
- Ye, C.J. et al (2018). Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of ERAP2 transcripts under balancing selection. *Genome research*, 28(12):1812–1825.
- Yi-Xiang, J. et al (2019). Up-Regulated AKR1C2 is correlated with favorable prognosis in thyroid carcinoma. *J Cancer*, 10(15):3543–3552.
- Yu, M. et al (2012). Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nature Protocols*, 7(12):2159–70.
- Zabetian, C.P. et al (2007). Association analysis of MAPT H1 haplotype and subhaplotypes in Parkinson’s disease. *Annals of Neurology*, 62(2):137–44.
- Zhang, B., Zhang, Y. and Shacter, E. (2004). Rac1 Inhibits Apoptosis in Human Lymphoma Cells by Stimulating Bad Phosphorylation on Ser-75. *Molecular and Cellular Biology*, 24(14):6205–6214.
- Zhang, F. et al (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics*, 41(7):849–53.

- Zhang, Y. et al (2015). Interferon-induced transmembrane protein-3 rs12252-C is associated with rapid progression of acute HIV-1 infection in Chinese MSM cohort. *AIDS*, 29(8):889–894.
- Zhang, Y.H. et al (2013). Interferon-induced transmembrane protein-3 genetic variant rs12252-C is associated with severe influenza in Chinese individuals. *Nature Communications*, 4:2–7.
- Zhang, Z. et al (2012). Evolutionary Dynamics of the Interferon-Induced Transmembrane Gene Family in Vertebrates. *PLoS ONE*, 7(11):e49265.
- Zheng, L.S. et al (2017). SPINK6 promotes metastasis of nasopharyngeal carcinoma via binding and activation of epithelial growth factor receptor. *Cancer Research*, 77(2):579–589.
- Zhernakova, D.V. et al (2017). Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics*, 49(1):139–145.
- Zhong, H. et al (2010). Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genetics*, 6(5):32.
- Zhou, V., Goren, A. and Bernstein, B. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, 12(1):7–18.
- Zody, M.C. et al (2008). Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nature Genetics*, 40(9):1076–83.
- Zollino, M. et al (2012). Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. *Nature Genetics*, 44(6):636–8.

