Universitat Autònoma de Barcelona

# Large-scale evolutionary analysis of polymorphic inversions in the human genome

**Carla Giner Delgado**

Doctoral thesis

# Large-scale evolutionary analysis of polymorphic inversions in the human genome

*Author*
Carla Giner Delgado

*Director*
Mario Cáceres Aguilar

Universitat Autònoma de Barcelona

Departament de Genètica i de Microbiologia
Facultat de Biociències
Universitat Autònoma de Barcelona
2017

# Large-scale evolutionary analysis of polymorphic inversions in the human genome

Memòria presentada per Carla Giner Delgado
per a optar al grau de Doctora en Genètica
per la Universitat Autònoma de Barcelona

*Autora* *Director*

Carla Giner Delgado Mario Cáceres Aguilar

Bellaterra, 28 de Setembre de 2017

**Abstract**

Chromosomal inversions are structural variants that invert a fragment of the genome without usually modifying its content, and their subtle but powerful effects in natural populations have fascinated evolutionary biologists for a long time. Discovered a century ago in fruit flies, their association with different evolutionary processes, such as local adaptation and speciation, was soon evident in several species. However, in the current era of genomics and big data, inversions frequently escape the grasp of current technologies and remain largely overlooked in humans. During the last few years, the InvFEST Project has aimed to address the missing knowledge about human inversions by validating and genotyping a large fraction of predicted polymorphisms. In particular, it has generated one of the most useful data sets on human inversions, consisting of 45 common inversions (with sizes from 83 bp to 415 kbp) genotyped at high-quality in 550 individuals of seven populations of diverse ancestry. This thesis takes advantage of the available population-scale information, combined with whole-genome sequences available from the 1000 Genomes Project, to carry out the first detailed analysis of the evolutionary properties of human polymorphic inversions. The methods used combine theoretical models, simulations and empirical comparisons with other mutation types. Besides the complete characterization of the data set, the results confirm fundamental differences between inversions created by different mechanisms. The frequency distribution of the 21 inversions originated by non-homologous mechanisms (NH) is similar to that expected for neutral variants when controlling for detection biases, which indicates that they are not subjected to strong negative selection. Recombination is completely inhibited across the whole inversion length, with no clear genetic exchange found, and possibly over a few kbp beyond the breakpoints. As a result, NH inversions strongly affect local genome variation levels, as predicted by computer simulations, with older inversions increasing total nucleotide diversity, while younger ones at very high frequency could have the opposite effect. In contrast, most inversions created by non-allelic homologous recombination (NAHR) (19/24) have appeared independently in different haplotypes in the sample. These high recurrence levels are reflected in several measures: they are enriched in intermediate frequencies, share multiple nucleotide polymorphisms between orientations, and have little linkage disequilibrium with neighbouring variants, which limits their detection by tag SNP strategies. Finally, in order to find inversions that are functional candidates, different signatures of selection on inversions were explored based on their frequencies, population differentiation and sequence variation patterns. Ten candidates were revealed, with three of them found to be >1.5 million years old and maintained at intermediate frequencies, possibly by

balancing selection. One of these was also found in archaic hominins. Other candidates seem to have reached high frequencies in a short period of time in some populations, consistent with positive selection. Notably, over half of the candidates are located within gene regions, which suggests that they may have functional effects. Thus, this work offers an overview of inversion dynamics and their role as genomic modifiers, opening interesting avenues of investigation.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Human genome hidden complexity

More than 15 years have passed since the sequencing of the (first) human genome (Lander et al. 2001). However, that reference sequence represents only a sample out of all diversity within the human genome. Individuals differ in many genome positions. Some of the differences produce normal healthy phenotypic diversity, while other are responsible of increased health risks or genetic diseases. During this time, some types of variation have been remarkably well described and analysed. Others are far more difficult to detect and their contribution to phenotypic variation remain largely unexplored. Among them, inversions are probably the most elusive.

### 1.1.1 Types of genomic variation

New mutations can modify single positions in the sequence or change large regions in one event. There are different classifications of new mutations, that usually reflect the outcome of the change and its magnitude, although sometimes also the underlying mutational mechanisms or the techniques required to detect them. In general terms, there can be changes that modify the amount of sequence (additions or deletions) and changes in the location or the content, but keeping the same amount of sequence (Figure 1.1).

The simplest change is a single nucleotide variant (SNV), generally called single nucleotide polymorphism (SNP), which results in a new base pair (Figure 1.1 A). The insertion or deletion of few base pairs is usually referred with the shortened word indel (Figure 1.1 A), given that the direction of the change is frequently unknown when first detected. Special repetitive regions


**Figure 1.1: Types of genomic variation** Overview of some basic types of simple genomic variation (A) and structural variation (B) classified according to the outcome.

of 2 to 5-base-pair motifs with recurrent indels are called microsatellites (and similar but longer motifs are called minisatellites) or short tandem repeats (STRs).

Mutation events that involve many nucleotides are usually known as structural variants (SV) (Figure 1.1 B). When the total amount of sequence is not altered, they are called balanced events. This is the case of inversions, where the orientation of a sequence of DNA is turned 180 degrees with respect to the flanking regions but remains in the same position, and translocations, where a sequence is moved from one position in the genome to another. Unbalanced structural variants involve the addition or deletion of sequence and can be referred as copy number variants (CNV). This includes more specific events, such as duplications –insertions that are a copy of another region– or insertions from specific mechanisms like the mobilization of a transposable element.

Especially for the insertion of deletion of sequence, the threshold between indel (*few* base pairs) and CNV (*many* base pairs) can be arbitrary and the concepts overlap. Sometimes it just reflects practical reasons, such as the sensitivity ranges of the techniques used. For instance, indels had been initially defined as $< 10$ kbp (Mills et al. 2006) but later generally lowered to <50-100 bps, the size detectable with sequencing reads from next-generation platforms (Carvalho and Lupski 2016). On the other hand, CNV definition has expanded to include smaller variations excluded from the new indel concept. Inversions and translocations are always considered structural variants,

and the minimum size is limited by our ability to recognize it in the sequence (an inversion can only be ambiguous at a very small scale like, for instance in the sequences `5'-C|GTAAT|C-3'` and `5'-C|ATTAC|C-3'` where there could be either three SNPs –`G>A`, `T>A` and `T>C`– or a single 5-bp inversion).

According to recent variation surveys, a typical human genome has between 4.1 and 5.0 million positions that differ from the HG19 version of the reference genome (The 1000 Genomes Project Consortium 2015). 96 to 99% of this variation is shared among many individuals (at frequencies $> 0.5\%$), and only a small fraction (less than 0.4% of the positions) is unique to one individual. However, if we add up all known variants from all the analysed samples, $\sim 75\%$ of the known variation is at low frequencies ($< 0.5\%$).

Our current knowledge about the different variant types is uneven. SNPs are by far the most studied variant, followed by indels. With the development of the cost-effective SNP arrays capable of genotyping many known SNPs in large sample sizes, SNPs were the preferred genetic marker. A particularly relevant work was the International HapMap Project that run for several years and finished in 2010 (The International HapMap Consortium 2005; The International HapMap 3 Consortium 2010). In the final phase, around 1.6 million common SNPs were genotyped in 1184 individuals from 11 diverse-origin populations. Among many contributions, SNP array-based projects allowed big improvements in our understanding of the human genome and genetic relations between populations. And are still a fundamental tool for applications where large sample sizes are required, importantly in genome-wide association studies.

Later, high-throughput sequencing technologies (HTS or NGS for next generation sequencing) became widely available, with good power to detect both SNPs and indels. The 1000 Genomes Project launched in 2008, as a natural progression of the HapMap Project, with the aim of provide a complete catalogue of human genome sequence variation through low-coverage sequencing with newly developed techniques (The 1000 Genomes Project Consortium 2010). In initial stages it was mostly focused on SNPs and indels, but in the final phase (phase 3, main release) it also covered more complex types of genome variation in a total of 2,504 individuals from 26 populations, including HapMap individuals (The 1000 Genomes Project Consortium 2015; Sudmant et al. 2015). This and other sequencing projects focused on single populations (Wong et al. 2013; The Genome of the Netherlands Consortium 2014; The UK10K Consortium 2015) or on diversity panels (Gurdasani et al. 2015; Mallick et al. 2016) are improving our picture of human genome variation.

Nevertheless, detecting and genotyping structural variants is challenging with HST (Huddleston and Eichler 2016). Despite our still limited power to detect them (that we will discuss in the next section), their large potential

effects as modifiers of organism functions is clear. These more recent studies report that structural variants are estimated to account for less than 0.1% of the variants (The 1000 Genomes Project Consortium 2015; Chiang et al. 2016). However, since each of them spans a larger region, they affect more base pairs of the genome that SNPs and indels together. Also, the average functional impact of structural variants is expected to exceed that of shorter variants (Sudmant et al. 2015; Hehir-Kwa et al. 2016). And indeed, they are repeatedly found to be enriched in functional associations (Sudmant et al. 2015; Chiang et al. 2016). Specifically for inversions, some have been associated to changes in expression of nearby genes (Jong et al. 2012; Salm et al. 2012; González et al. 2014; Puig et al. 2015a). Therefore, structural variants are getting more attention and slowly catching up.

### 1.1.2 Detection of structural variants

Structural variants and aneuploidies in humans were known for a big part of the last century, but during many decades they were thought to be rare and mostly related to disease. Most of the the knowledge came from microscopically visible variants (of at least several Mbp) through cytogenetic studies, usually investigating the origin of diseases and syndromes (Escaramís, Docampo, and Rabionet 2015) (Table 1.1). It was not until 2004, with the development of techniques such as BAC and oligonucleotide array comparative genomic hybridization (aCGH), that an unexpected amount of structural variation was found in healthy individuals (Sebat et al. 2004; Iafrate et al. 2004) (Table 1.1). However, the nature of the strategies applied based on the intensity of hybridization restricted their application to unbalanced variants.

After that, more powerful technologies followed. Paired-end mapping (PEM) was soon explored as an alternative to survey all types of structural variation (Tuzun et al. 2005) (Table 1.1). Briefly, in PEM the genome of a target sample is randomly fragmented and sequences of a set size are chosen. Then, the extremes of the selected fragments are sequenced and mapped to a reference sequence. Unexpected distances or orientations between paired reads reveal the presence of structural differences between the reference and the target genomes (Figure 1.2). PEM is powerful to detect structural variation, as long as repetitive sequences at variant breakpoints are not longer than the fragments (last example in Figure 1.2). Initial PEM applications used Sanger method to sequence the extremes of fragments cloned in fosmid vectors (Tuzun et al. 2005; Kidd et al. 2008). Later, HTS imposed as the preferred low-cost technique, despite normally using shorter fragment sizes. Also, the *de novo* assembly of human genomes offered another opportunity to detect polymorphic structural variation (Levy et al. 2007; Cao et al. 2015), as it had already been done for the fixed structural differences between the chimpanzee and human genomes (Feuk et al. 2005). A main limitation of

assembly comparison is the completeness of the genome sequences, that often present gaps in locations with structural variation, that are difficult to resolve (Table 1.1).

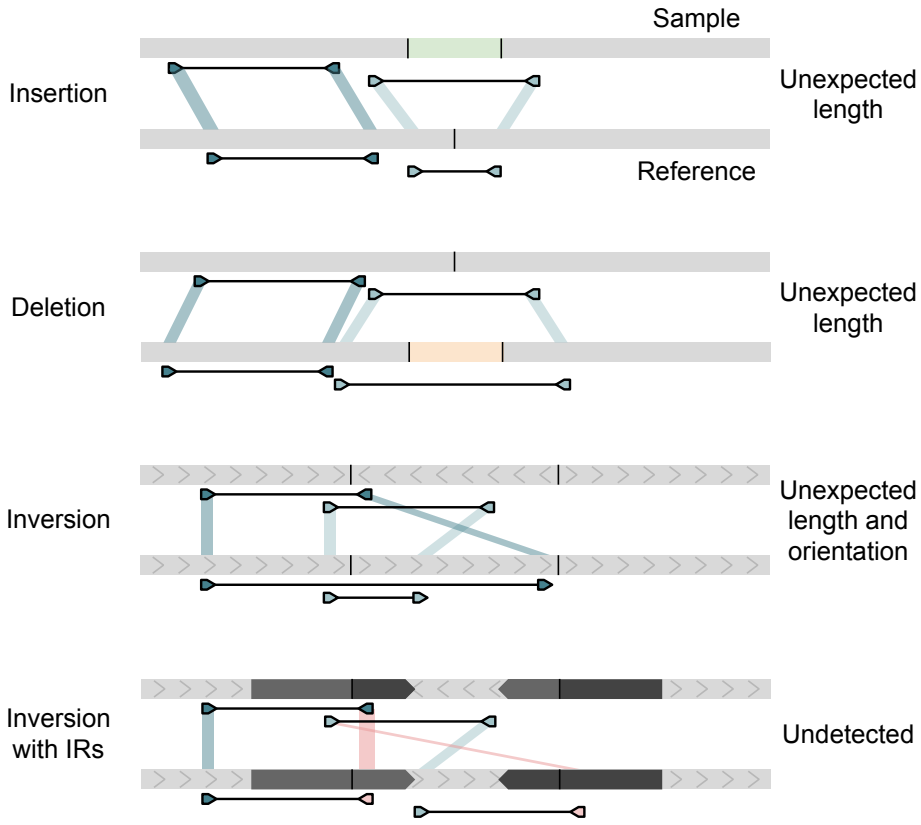


**Figure 1.2: Paired-end mapping signatures**. Insertions and deletions can be detected by unexpected distance between fragment ends. Inversions may be detected by the unexpected mapping orientation (mapping to the alternative strand) and distance. The fourth example represents a complex inversion with inverted repeats (IRs) at the breakpoints, where paralogous mapping leaves the inversion undetected.

With the development and wide availability of HTS, many strategies have been developed to detect large genomic variants from short reads. Common signatures used are: amount of DNA from a specific sequence measured as read depth, discontinuous sequence highlighted by reads with split mapping and, as already mentioned, inconsistent distance or orientation of paired reads. Although the amount of DNA is not sensitive to balanced rearrangements, the other two signatures can potentially be used to detect inversions with relatively simple breakpoints (Table 1.1). In addition, beyond detecting the presence of a variant, HTS are well suited to simultaneously genotype them in a large sample panel to obtain population frequency and haplotype estimates. The main limitation of HTS is that reads are usually short and they rely on mapping on a reference sequence. Therefore, it depends on both the completeness of the reference and the absence of repetitive sequence to

be able to map the reads unambiguously. To overcome the dependence on the reference, some methods also use a *de novo* local assembly of the target genome reads, although the repetitive sequence is still problematic in short reads. Some HTS-based projects, such as the 1000 Genomes Project (Sudmant et al. 2015) or the Genome of the Netherlands (Hehir-Kwa et al. 2016) have successfully detected and genotyped several types of structural variants using a combination of these approaches. However, inversions are systematically the type of variation with poorer performance. The validation rates and sensitivity estimates are always the lowest of all types (Sudmant et al. 2015; Hehir-Kwa et al. 2016). For instance, the overall sensitivity for inversions in 1000GP phase 3 is 32% (versus 65-88% of CNV) and false discovery rate between 9 and 17% (versus 1-4% of CNV) (Sudmant et al. 2015).

Previous strategies with higher power for inversions, such as fosmid-based PEM or assembly comparison, have the disadvantage of being more costly. As a consequence, they usually require alternative targeted methods to genotype inversions in larger samples (Aguado et al. 2014; Puig et al. 2015a; Vicente-Salvador et al. 2016). Commonly used strategies are regular PCR with allele-specific amplification through the breakpoints, and modified protocols to avoid amplifying through long inverted repeats at the breakpoints, like the inverse PCR (iPCR) (Aguado et al. 2014) (Table 1.1). Once some individuals have been genotyped for an inversion as well as for other nearby SNPs, correlation between SNP and inversion genotypes can be assessed to determine if some nearby variant can be used as proxy for the inversion genotype (called tag SNP). Inversions modify the local recombination patterns (discussed later in section 1.2.2), so the idea of nearby variation as footprint of an inversion has also been explored to genotype known inversions as well as to detect new ones (Table 1.1). Some examples are the PFIDO algorithm (Salm et al. 2012), the inveRsion package (Cáceres et al. 2012) or invClust (Cáceres and González 2015), that use linkage disequilibrium (LD) patterns and haplotype clustering.

Additionally, in the last few years different teams have explored several new approaches to detect complex structural rearrangements, also applicable to balanced variants (Table 1.1). Optical mapping is a technique based on genome-wide nicking with restriction enzymes coupled to fluorescent labelling of the nicks, so that restriction patterns can be read with optical microscopy (Teague et al. 2010). Variants are detected as pattern differences between samples, so the resolution and sensitivity depends on the restriction fragment size. Long-read sequencing is another popular technology that takes advantage of similar signatures than those used by HTS methods, with increased power to sequence through longer repetitive regions. Projects such a the sequencing of the CHM1 and CHM13 haploid genomes (Chaisson et al. 2015; Huddleston et al. 2017) use long-read technology as main strategy. Other recent HTS projects use long-read methods mainly to validate

predictions from short reads (Sudmant et al. 2015), given the higher cost of long-read technologies. As a cheaper alternative, linked reads approaches aim to gain long-read-like span using highly optimized HTS platforms by labelling short reads coming from a same long fragment (Eslami Rasekh et al. 2017). Strand-seq was recently developed (Falconer et al. 2012) and latter applied to detect inversions (Sanders et al. 2016). The main idea behind the approach is the sequencing one specific strand of each chromosome in a cell. Despite requiring some extra steps to prepare the samples and several dividing cells per individual to complete an entire genome, it is a very promising single-cell application for detecting inversions. Finally, the method to analyse DNA three-dimensional architecture Hi-C, which quantifies interactions between distant genomic regions, has been also applied to detect known and novel rearrangements in cancer cells (Harewood et al. 2017).

**Table 1.1: Methods to detect structural variants**. Overview of some of the available methods to detect structural variants, with emphasis in the limits and their application to the detection of inversions.

| Method | Detected inversions | Cost | Mode* | Example of application |
|---|---|---|---|---|
| *Microscopic* | | | | |
| Trad. cytogenetics | Inv > 3 Mbp | $$$ | ◯ | Carr (1962) |
| FISH | Inv > 1 Mbp | $$$ | ↓ | Feuk et al. (2005) |
| *Pioneers submicroscopic* | | | | |
| aCGH | - | $ | - | Iafrate et al. (2004) |
| Sanger paired-end | IR < fragment | $$$ | ◯ | Tuzun et al. (2005) |
| Assembly comparison | Assembly quality | $$$$ | ◯ | Levy et al. (2007) |
| *High-throughput sequencing* | | | | |
| Read depth | - | $ | - | Sudmant et al. (2015) |
| Split reads | IR << read | $ | ◯ | Sudmant et al. (2015) |
| Paired-end/mate-pair | IR < fragment | $ | ◯ | Sudmant et al. (2015) |
| *Targeted* | | | | |
| PCR | IR < 1 kbp | $$ | ↓ | Vicente-Salvador et al. (2016) |
| iPCR | IR < 25 kbp | $$ | ↓ | Aguado et al. (2014) |
| Tag SNPs | Presence of tag SNPs | $ | ↓ | Alves et al. (2015) |
| Linkage diseq. | Diverged haplotypes | $ | ↓/◯ | Cáceres and González (2015) |
| *Alternative* | | | | |
| Nanochannel mapping | Inv > restrict. frag. | $$$ | ◯ | Teague et al. (2010) |
| Long-read | IR < read | $$$ | ◯ | Chaisson et al. (2015) |
| Strand-seq | Inv > 1 kbp | $$$ | ◯ | Sanders et al. (2016) |
| Linked reads | IR < 100 kbp | $$ | ◯ | Eslami Rasekh et al. (2017) |
| Hi-C | ND | $$ | ◯ | Harewood et al. (2017) |

* ◯ = genome-wide technique; ↓ = targeted technique (to detect pre-ascertained inversions); - = does not detect inversions. ND: not determined.

Because of the complexity of structural variation and its detection, specialized databases have been created to collect and analyse the increasing number of variants described in the literature. The Database of Genomic Variants (`http://dgv.tcag.ca`) (MacDonald et al. 2014) is a curated reference resource for structural variation that started with the seminal works of 2004 (Iafrate et al. 2004; Sebat et al. 2004) and currently hosts more

than 500,000 structural variants of different sizes and frequencies. Most of the entries are CNVs and inversions represent less than 1% of the variants (3164). Since inversion prediction methods have high false discovery rates, InvFEST database (`http://invfestdb.uab.cat`) (Martínez-Fundichely et al. 2014) was created more recently to exclusively deal with these elusive variants, trying to identify the different inversions and refine as precisely as possible their breakpoints. InvFEST database, through its merging engine and reliability scoring system, aims to offers the most accurate overview of human polymorphic inversions at the moment. It currently contains 1092 candidate inversions, 85 of which have been validated and 51 are predictions or reference genome errors.

### 1.1.3 Mutational mechanisms

Structural variation is a complex category that includes a wide range of events with different underlying molecular mechanisms of generation. Current mutational models are based on the sequence signatures at the breakpoints together with evidences from experimental studies in model organisms, such as yeast, and human cells under stress (Gu, Zhang, and Lupski 2008; Conrad et al. 2010; Pang et al. 2013; Abyzov et al. 2015; Carvalho and Lupski 2016). There are at least three general processes that can lead to the formation of a structural variant: DNA recombination (through non-allelic homologous recombination (NAHR)), repair (such as non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ)), and replication (as in fork stalling and template switching (FoSTeS) or microhomology-mediated break-induced replication (MMBIR)). Also, the mobilization of transposable elements creates itself new insertions (mobile element insertions, MEI) that can be used as substrate for other mechanisms that require homology or microhomology. Each mechanism is characterized by different sequence signatures in and around the breakpoints. The main mechanisms of each type are summarized below.

**NAHR (recombination-based).** In NAHR, recombination happens between two paralogous copies of the same sequence. Depending on the location and relative orientation of the copies, the resulting structural variant could be a deletion (direct copies in the same chromosome), duplication and deletion (direct copies in homologous chromosomes), inversion (copies in the same chromosome but in inverted orientation) or translocation (copies in non-homologous chromosomes). Segmental duplications (also called low-copy repeats or LCR) are typically the substrate for NAHR, although other types of repeats can be also involved (Escaramís, Docampo, and Rabionet 2015; Carvalho and Lupski 2016). SV mediated by NAHR have been shown to appear recurrently

in the population, including inversions (Flores et al. 2007; Aguado et al. 2014).

**NHEJ/MMEJ (repair-based).** NHEJ is the most common method to repair double-strand breaks in mammals, together with homologous recombination, and does not require sequence homology (Escaramís, Docampo, and Rabionet 2015). MMEJ is a more error-prone alternative that requires microhomology at broken ends. It mostly happens when NHEJ machinery is unavailable and is thought to be an important source of genomic instability (McVey and Lee 2008). The creation of structural variants by end-joining mechanisms generally results in clean (blunt) breakpoints, or with short stretches of microhomology.

**FoSTeS/MMBIR (replication-based).** If a replication fork gets stalled or broken, it can invade a nearby fork with or without microhomology and re-initiate DNA synthesis (Weckselblatt and Rudd 2015). This process can lead to to complex rearrangements, including different types of SV together, ranging from few kilobases to several megabases (Escaramís, Docampo, and Rabionet 2015), and is known as FoSTeS/MMBIR.

Several studies have attempted to measure the relative importance of the different mechanisms in normal genomic variation, as well as in pathogenic rearrangements. Most studies of CNVs found that the majority of non-recurrent variants have blunt ends or microhomology (Weckselblatt and Rudd 2015; Pang et al. 2013), suggesting that non-homologous mechanisms are more prevalent. In contrast, inversions appear to have a higher proportion derived from NAHR, in around 50% of the cases (Pang et al. 2013). In any case, the relative proportions observed are very affected by the power of the methods used to detect the different types of variants. The most used read-based methods have important limitations to access repetitive regions and could partly exaggerate the importance of non-homologous mechanisms (Lucas Lledó and Cáceres 2013).

In a sense, the abundance of structural variation in the human genome should not be surprising. Many of the proposed mechanisms of formation of structural variation involve homology or microhomology. And ours is a specially repetitive genome, with a 50% of its sequence composed by repetitive sequence (**Lander2001s** ).

## 1.2 Inversions: a special mutation type

Among the different structural variation types, inversions are probably the least well understood and studied. Ironically, inversions have been known

for a longer time. They were first described by Alfred Sturtevant at the beginning of the twentieth century while studying genetic linkage in *Drosophila* (Sturtevant 1917; Sturtevant 1921), before any other variant type. At that time they were easier to detect than other variants, dues to the giant polytene chromosomes of the salivary glands in insects, that allow a direct observation of the karyotypes with optical microscopy. Since then, they have attracted the attention of evolutionary biologists because of their unique properties as genetic markers, as well as their apparent key role in many evolutionary processes, such as adaptation, evolution of sex chromosomes or speciation (Hoffmann and Rieseberg 2008; Kirkpatrick 2010).

Our knowledge about inversions comes in great part from the exhaustive studies in *Drosophila* continued by Dobzhansky and colleagues and followed by many others (Dobzhansky 1970), that identified thousands of inversions both within and between species (Krimbas and Powell 1992) With the improvement of cytogenetic techniques, inversions were studied in other species, including humans (Carr 1962). Like structural variants in general, most of the initially known inversions in humans were either associated to reproductive problems (Gardner, Sutherland, and Shaffer 2011) or discovered studying some disease locus (Small, Iber, and Warren 1997). Later on, more submicroscopic inversions have been detected with the high-throughput methods described earlier. However, only a handful of inversions have been studied at a population level, the most well-studied being the 4.5-Mbp inversion in 8p23.1 (HsInv0501 in InvFEST) and the 835-kbp inversion in 17q21.31 (HsInv0573).

Therefore, without doubt, chromosomal inversions are important actors in the evolution of species and genomes throughout taxa. What does make them so special? A key characteristic seems to be that they limit genetic sharing between sequences in the ancestral and the inverted orientation, through the inhibition of recombination in heterozygotes (Kirkpatrick 2010).

### 1.2.1 Inhibition of recombination

In humans and other diploid organisms, homologous chromosomes pair and recombine during meiosis I. Meiotic recombination starts with a programmed double-strand break that can be repaired as a crossover or a non-crossover product. Non-crossovers result in an unidirectional copying of a small region from one chromosome to the other (known as gene conversion) and are estimated to outnumber the crossover products. Crossover products imply an exchange of large chromosomal regions between homologues and are required for correct homologue orientation and accurate segregation (see Baudat, Imai, and Massy (2013) for a recent review of meiotic recombination in mammals).

Inversions challenge normal pairing of homologue chromosomes (synapsis) from the loss of linear homology in the inverted region. Indeed, from sequence and cytogenetic analyses, inversions are known to inhibit recombination in heterozygotes (Hoffmann and Rieseberg 2008; Kirkpatrick 2010). However, there are two possible mechanisms that could lead to suppression of recombination, each of them with different consequences and reproductive costs (Figure 1.3). Disentangle the effects of inversions on recombination is a key aspect to understand their evolutionary role.

#### 1.2.1.1 Possible mechanisms of recombination inhibition

In the simplest scenario, the local lack of homology in heterozygous chromosomes may just prevent the homologous synapsis in the inverted region (Figure 1.3 A). A physical impediment of recombination excludes crossovers and non-crossovers. Under this model, there is no reproductive cost for the heterozygote carrier. Both paracentric inversions (those with both breakpoints in the same chromosome arm) and pericentric (that include the centromere) could in theory physically inhibit recombination.

In an alternative scenario, inversions can be long enough to create a loop that allows homologues to locally pair along the inverted region (Figure 1.3 B and C). If a single crossover event happens between homologues within the inversion, the recombinant chromosomes will be unbalanced. Balanced chromosomes can only result from an even number of crossovers between the same pair of chromatids within the inverted region. The probability of multiple crossovers leading to balanced chromosomes is nevertheless decreased by the fact that two sister chromatids are available for each homolgue (Navarro and Ruiz 1997). In paracentric inversions (that affect only one chromosomal arm), an odd number of crossovers results in one chromosome without centromere (acentric) and another with two (dicentric) (Figure 1.3 B). Conversely, recombinant chromosomes in a pericentric inversion have one centromere each, but with deletions and duplications of non-inverted arm fractions (Figure 1.3 C). Unbalanced chromosomes generally cannot give rise to viable offspring. Therefore, the reproductive cost will depend on how far the unbalanced chromosome progress through gametogenesis and embryonic development.

Male individuals from most *Drosophila* species are an exception, since they do not recombine in meiosis and therefore can not produce unbalanced gametes (Hoffmann and Rieseberg 2008; Krimbas and Powell 1992). In female flies, where only one of the four daughter cells in each meiosis will become the mature egg, the slower migration of the acentric and dicentric chromosomes ensures that they are relegated to the polar bodies (Krimbas and Powell 1992; Hoffmann and Rieseberg 2008). This system conveniently avoids the

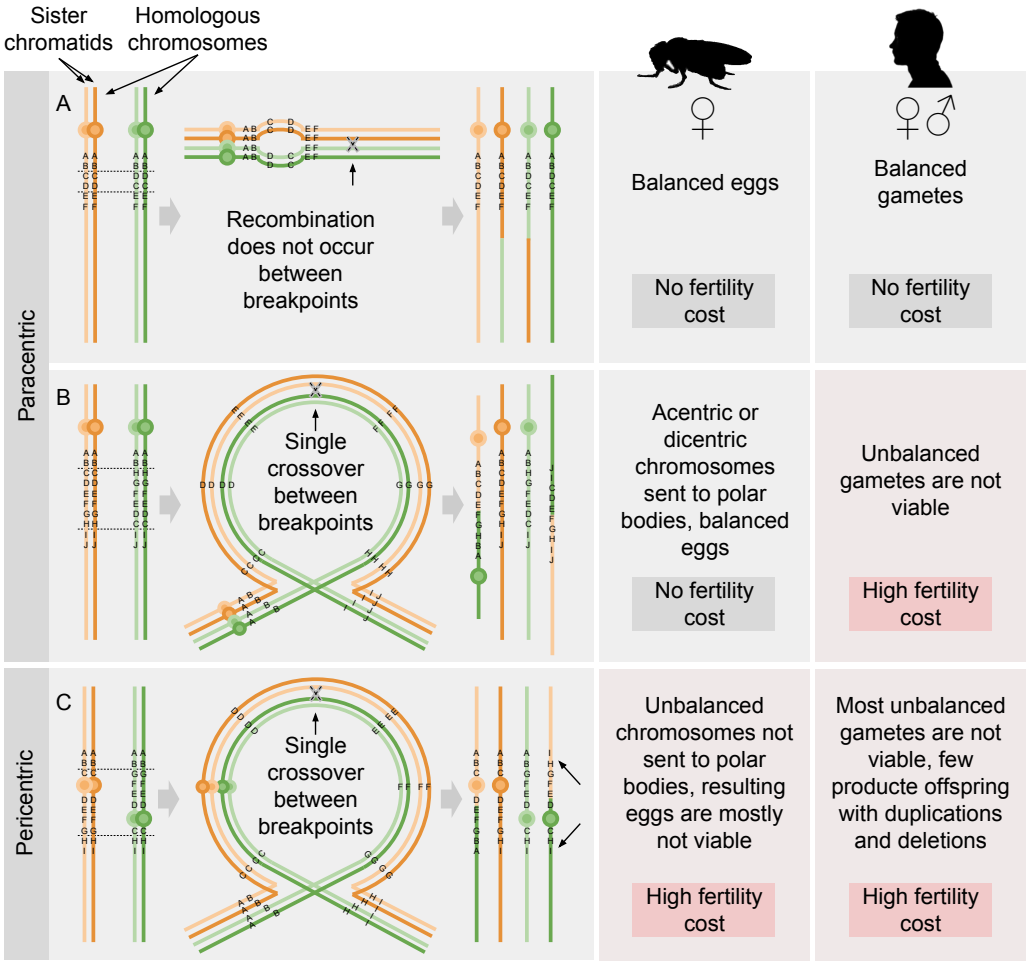


**Figure 1.3: Models of inhibition of recombination.** Simplified models for inversion consequences on recombination. (A) Small paracentric inversion with physical inhibition of synapsis, and long paracentric (B) and pericentric (C) inversions with a single crossover that results in unbalanced recombinant chromosomes. Dotted lines indicate the position of the inversion and circles represent centromeres. Arrows in the last example highlight duplicated regions. Right columns indicate the reproductive consequences in *Drosophila* females (males usually do not recombine) and in humans.

potential reproductive cost of crossover within paracentric inversions. Recombinant chromosomes from pericentric inversions have each a single centromere and all migrate at the same rate, reducing fertility later on. This is consistent with the observation that in *Drosophila* big citologically-visible paracentric inversions are very abundant both as polymorphisms and fixed differences, whereas pericentric are more limited (Krimbas and Powell 1992).

In humans, no such system exists, so sperm cells and oocites can carry unbalanced chromosomes resulting from recombination within inversions. Acentric and dicentric chromosomes are likely to create problems early, during gametogenesis, limiting the fertility impact of recombinant products of paracentric inversions (Gardner, Sutherland, and Shaffer 2011). In contrast, the recombinant chromosomes from a pericentric inversion may only cause problems later in development, increasing the fertility cost. In extreme cases of pericentric inversions including a big fraction of the chromosome, it has been reported the birth of children with recombinant unbalanced chromosomes (Gardner, Sutherland, and Shaffer 2011). This means that some recombinant products are even viable throughout pregnancy, although they lead to children with birth defects. The resulting duplication and deletion in those cases were small and only one recombinant type is ever viable (generally that with the smaller deletion) (Gardner, Sutherland, and Shaffer 2011). Therefore, in less extreme pericentric inversions, recombinant products are expected to have an intermediate viability.

#### 1.2.1.2 Experimental evidences for each model

Classic long polymorphic inversions in *Drosophila* (typically inverting around a third of the chromosome arm) are known to create a loop structure during chromosome pairing, and accordingly, show noticeable levels of recombination between orientations (Andolfatto, Depaulis, and Navarro 2001). Theoretical models of genetic flux caused by double crossover and gene conversion between orientations predict a non-uniform recombination rate in heterozygotes, with increased levels in the middle of the inverted region and decreased recombination near the breakpoints (Navarro et al. 1997; Navarro and Ruiz 1997). In addition, the suppression of recombination is expected to be stronger in smaller inversions. These patterns fit well with empirical observations (Andolfatto, Depaulis, and Navarro 2001) and confirm that classical *Drosophila* inversions indirectly inhibit recombination by the generation of unbalanced chromosomes.

The two scenarios of recombination inhibition in inversions have very different consequences in humans. The first can be mostly neutral while the second can strongly reduce fertility in heterozygotes (Figure 1.3). What are the factors that lead to one or the other situation, and how common are they

in human inversions?

While most inversions in human populations are thought to be smaller than 1 Mbp (Puig et al. 2015b), our direct knowledge about inversion effect on human meiosis comes mainly from cytogenetics, so from a few Mbp-long inversions. For several decades studies concerned about inversion effect on fertility and miscarriage risk have studied the conformation of the affected chromosomes in meiosis, as well as recombinant meiosis products (gametes) (Morin et al. 2017). We know that in humans many cytogenetically-visible inversions do create loops and also recombine (Cheng et al. 1999; Morel et al. 2007), even some of the pericentric inversions involving heterochromatin that nevertheless are regarded as innocuous (Ferfouri et al. 2009). Although it has been suggested that only inversions longer than 100 Mbp and encompassing 50% of the chromosome show significant levels of unbalanced gametes (Anton et al. 2005), recombination has been detected in smaller inversions (e.g. in the same study, the smallest one showing recombinant gametes covered 20% of the chromosome and was 49-Mbp long). Therefore, inversions smaller than 100 Mbp may not have an impact on fertility to be relevant at individual-level, but they can have consequences at population level.

Studies in mice have offered some clues to understand the pairing and recombination process in large inversions (Torgasheva and Borodin 2010; Torgasheva, Rubtsov, and Borodin 2013). In order to create a loop, at least one point of synaptic initiation has to be set in the inverted region. For that, inversion position and size (relative to the chromosome) play an important role (Torgasheva and Borodin 2010). When synapsis takes place within the inverted region, recombination can happen and lead to a crossover. It has also been described the process where created loops progressively unwind until homologues are co-lineal, called synapsis adjustment (Moses et al. 1982). Only in cells where an internal crossover takes place between homologues, loops are stopped from completely untangling in later prophase stages and are thus visible by microscopy (Torgasheva, Rubtsov, and Borodin 2013). These observations could explain in part why only a small fraction of inversion-carrier human cells seem to form a visible loop.

Meiosis direct assays are more difficult to apply to small submicroscopic inversions, since they are not visible to cytogenetic technologies. Instead, if they are frequent in general population, we can in theory detect the footprints left by past double crossovers and gene conversion events in the sequence variation patterns. Unfortunately, none of the best-studied polymorphic inversions in our species offer a clear picture.

Inversion 8p23.1, spanning 4.5 Mbp, has been shown to lack any variant in complete linkage disequilibrium (perfectly correlated genotypes) (Antonacci et al. 2009; Salm et al. 2012). This observation implies the presence of some genetic flux (exchange of genetic information) between orientations.

However, the pattern could also be explained by recurrence instead of by recombination, given that NAHR is the most likely mechanism and it is associated to recurrence (Aguado et al. 2014). Even a low recurrence rate (Salm et al. 2012) would result in some degree of genetic flux. A similar pattern is found in four extra inversions, all longer than 1 Mbp, described in Antonacci et al. (2009), and in 300-kbp inversion 16p11.2 (HsInv0786) (González et al. 2014). Since no variant in complete disequilibrium is found and sometimes several independent haplotypes are present in inverted chromosomes, recurrence cannot be ruled out and therefore no clear patterns of recombination can be obtained.

A slightly clearer picture emerges with the 835-kbp inversion 17q21.31. In that case, there are variants in complete linkage disequilibrium with the inversion spanning all the region (Alves et al. 2015). This observation excludes recurrence and makes unlikely high genetic flux between orientations. However, few shared variants were identified between orientations (Zody et al. 2008) and a 30-kbp region with low divergence between orientations was attributed to a double-crossover event (Steinberg et al. 2012). Both observations suggest that some pairing and recombination do take place within inversions smaller than 1 Mbp.

### 1.2.2 Effect on nucleotide diversity patterns

Recombination is one of the factors strongly influencing nucleotide variation patterns in genomes (Duret and Arndt 2008). It has been positively correlated with polymorphism within genomes and also between species (Casillas and Barbadilla 2017). Inversions, as modulators of recombination, are expected to alter neutral variation patterns, and we have already seen some examples in human inversions above. The reduced recombination will create a local population stratification that will condition the frequencies of variants within.

Several models have been developed to understand inversion effect on nucleotide variation. Most of them are based on *Drosophila* observations, with the corresponding population sizes, mutation and recombination rates. Many of the proposed scenarios assume that inversions are maintained at an equilibrium frequency (presumably by strong natural selection), following what seems to happen in some *Drosophila* inversions (Navarro et al. 1997; Navarro, Barbadilla, and Ruiz 2000; Guerrero, Rousset, and Kirkpatrick 2012). Both analytical expressions and coalescent simulations have been used to describe expectations under different parameters. While this balanced-polymorphism model is useful to understand some old *Drosophila* inversions, it is unrealistic for more recent ones, as it is probably the case for putatively neutral humans inversions. Guerrero, Rousset, and Kirkpatrick (2012) relaxed the equilib-

rium assumption and also explored the patterns obtained by an inversion evolving neutrally by random drift, modelling independent loci in the two orientations as two populations that are subject to migration, representing recombination between orientations.

More recently, Peischl et al. (2013) developed an efficient algorithm to compute ancestral recombination graphs with inversions. In this case, inversions can be simulated to follow different evolutionary trajectories and therefore could be suitable for studies of human inversions. With a similar purpose but implemented in a forward-in-time manner, the software InvertFREGENE allows to simulate inversions in SNP data to replicate linked-variation patterns (O'Reilly, Coin, and Hoggart 2010). There are several difference between the two strategies. While the former is focused on inversions allowing a small recombination rate in heterozygotes, the latter restricts recombination to homozygotes. Additionally, InvertFREGENE models recombination as a hierarchical process to realistically simulate human recombination hotspots and allows for simple demography changes.

Overall, expected patterns depend strongly on the underlying selective model, inversion's age and genetic flux between orientations (Andolfatto, Depaulis, and Navarro 2001; Peischl et al. 2013). Models of balanced polymorphisms predict decreased nucleotide variation at the breakpoints for young inversions ($10^5 - 10^6$ generations with *Drosophila* parameters) and increased nucleotide variation for older inversions at equilibrium (Navarro, Barbadilla, and Ruiz 2000). In all cases, when inversions rise fast in frequency, derived chromosomes have very little variation, consistent with a sweep intensified by the limited recombination (Navarro, Barbadilla, and Ruiz 2000; Guerrero, Rousset, and Kirkpatrick 2012; Peischl et al. 2013).

The most recent approximations offer the tools to explore specific situations. However, all discussed models assume a unique inversion origin, whereas at least in humans recurrence seems to be common (Cáceres et al. 2007; Flores et al. 2007; Aguado et al. 2014; Antonacci et al. 2009). The variation patterns in more complex situations, as in the presence of recurrence, are still difficult to predict.

### 1.2.3 Evolutionary importance

As we have seen, inversions can be deleterious from the generation of unbalanced chromosomes. Alternatively, if inversions do not reduce fertility in heterozygotes (likely for small ones), they could be expected to evolve neutrally in most cases, given that the genetic content stays the same. However, they may still have a wide range positive and negative effects. And indeed, many inversions seem to be unambiguously evolving in a non-neutral way

(as reviewed in Hoffmann and Rieseberg (2008) and Kirkpatrick (2010)), and there are different examples of inversions with adaptive effects.

What are the ways inversions can make a difference? One option is a direct effect on the breakpoints: a coding sequence or functional element could be disrupted. Another subtler effect could be related to the spatial position of elements within the inversions (traditionally called position effects): for example, a gene and its promoter could be split. In addition to the position effects, inversion special recombination characteristics give them an extra potential advantage. Reducing recombination can be useful in some settings, as when two variants in the same haplotype work well together (Dobzhansky 1970). A complete reduction of recombination in a long genomic region is usually undesirable, given that it reduces effective population size and selection efficiency (Brandvain and Wright 2016). Inversions, suppressing recombination only in certain individuals, scape the drawbacks of a indiscriminate reduction of recombination (Otto and Lenormand 2002; Kirkpatrick 2010). It is mostly their role as recombination modifiers that has linked them to different evolutionary processes like speciation, the evolution of sex chromosomes and local adaptation (Hoffmann and Rieseberg 2008; Kirkpatrick 2010).

#### 1.2.3.1 Speciation

Although initially inversion-related speciation models were proposed on the basis of the reduction of fertility in heterozygotes (underdominance) resulting in reproductive isolation (White 1978), alternative models where inversions act as introgression barriers are thought to be more common in animals (Coyne and Orr 2004). The main problem of the underdominant model is that, in order to explain inversion fixation in a population, it is necessary to assume strong structure or drift to counterbalance its deleterious effect (Hoffmann and Rieseberg 2008; Kirkpatrick 2010). The model may be relevant in plants, that show reduced fertility in hybrids and can meet the conditions of extreme drift and inbreeding (Hoffmann and Rieseberg 2008). The alternative models rely on the sequence divergence between orientations, that help speciation in different ways (Hoffmann and Rieseberg 2008). Genes that cause incompatibility may be captured by an inversion and be linked to a long region protected from introgression, or they can be accumulated after the inversion. Also, inversions may link variants under divergent adaptation. These second class of models are supported by evidences in multiple species. For instance, in the case of yellow monkeyflower, two different forms with a fixed inversion are adapted to different climates and flower at different time. A reduced survival of hybrids due to climate is added to the pre-mating isolation from a different flowering time (Lowry and Willis 2010). In addition, chromosomal rearrangements have been suggested to protect from introgres-

sion between modern humans and Denisova or Neandertal archaic humans (Rogers 2015). However, these effects may not be universal, since inversions have also been proposed to facilitate the transfer of long well-adapted haplotypes instead (Kirkpatrick and Barrett 2015).

#### 1.2.3.2 Evolution of sex chromosomes

Inversions are also key elements in the evolution of sex chromosomes (Hoffmann and Rieseberg 2008; Kirkpatrick 2010). In the establishment of a system of sex chromosomes, it is necessary to suppress recombination to maintain a sex-determining locus and sex-beneficial alleles together. Inversions are an effective mechanism to stop recombination between evolving sex chromosomes, and models show that it should be favoured. The evolution of different sex chromosome systems involves inversions (Bachtrog 2013). For example, during the evolution of mammal chromosomes X and Y, a series of overlapping inversions have extended the non-recombining region between them (Lahn and Page 1999).

#### 1.2.3.3 Local adaptation

Finally, some polymorphic inversions in different species are thought to evolve under strong selection (Krimbas and Powell 1992; Hoffmann and Rieseberg 2008). One classical evidence is that they display striking frequency patterns that are difficult to explain under neutrality. Sometimes, inversions follow reproducible geographical frequency gradients (clines), as seen in several *Drosophila* species and also in *Anopheles* mosquitoes (Kapun et al. 2016; Ayala et al. 2017). In other occasions, they change rapidly in frequency in cage and natural populations of *Drosophila*, and after perturbations, frequencies sometimes quickly return to their original values (Krimbas and Powell 1992). These patterns have been interpreted as local adaptation to spatially-varying selection (Kapun et al. 2016) and as overdominance maintaining a balanced polymorphism. The mystery of inversions is that, although many traits have been related to selected inversions, the molecular target of selection is generally unknown (Hoffmann and Rieseberg 2008; Kirkpatrick and Kern 2012). Thus, it is not clear if selection acts on a favourable combination of independent alleles trapped within the inversion, some epistasis between them or perhaps just a direct effect at the breakpoints. In humans, no common inversion polymorphism has a clear selective pattern comparable to those found in other species, but in some inversions selection acting on one orientation has been suggested (Puig et al. 2015b). The most well-known example is the inversion 17q21.31, associated to an increased recombination and fertility in Europeans (Stefansson et al. 2005).

## 1.3 Inference of evolutionary history

Genomic sequences can tell us about the past of the species as a whole and specific genomic regions. Within the last 100,000 years, humans were subject to a wide array of selective pressures: they expanded from Africa to new continents with new environments, replaced other archaic human groups and transitioned from hunter-gatherers to agriculturist and pastoralist. Therefore, there is a great interest in detecting genomic regions that have been important for recent evolution, which could have a potential impact on human health and disease risk (Fan et al. 2016). Many methods have been developed to infer past and current demography, as well as to identify regions evolving under natural selection (Vitti, Grossman, and Sabeti 2013). However, in most cases the approaches have been designed to work with the most widely available genetic variation: SNPs. In this section, we briefly highlight important events in human evolution, methods available to detect different types of natural selection in genomic data, and possibilities and limitation for their application to inversions.

### 1.3.1 Human evolutionary history

The human closest extant relatives are chimpanzees and bonobos (*Pan* genus), followed by gorillas (*Gorilla* genus). Estimates of *Pan-Homo* and *Gorilla-Homo* split times have varied considerably between studies, partly because of the close relatedness that leads to incomplete linage sorting (different regions of the genome infer different species tree topologies) (Rogers and Gibbs 2014). Generally accepted estimates are around 5-9 (Rogers and Gibbs 2014) and 6-10 million years ago (Scally et al. 2012), respectively, although in divergence analyses it is important to be aware of the large uncertainty around the estimates.

Today we have a general good idea of the genetic relationships of present-day human populations, and the details about the demographic history increase with the continued genetic studies of new populations and archaic humans (Figure 1.4) (see Nielsen et al. (2017) for a recent review). Recent estimates show that all present populations share a common demographic history before ∼150,000-200,000 years ago (The 1000 Genomes Project Consortium 2015). Current models also estimate that all non-African populations originated from a common out-of-Africa wave, which occurred around 50,000-100,000 years ago and replaced previous modern human groups (from possible previous waves) and other hominin lineages (Gravel et al. 2011; Gazave et al. 2014). Later, the populations split into an European branch (together with western Asian) and an Asian one. Oceania was soon peopled from the Asian branch, while the Americas were only colonized more recently (al-

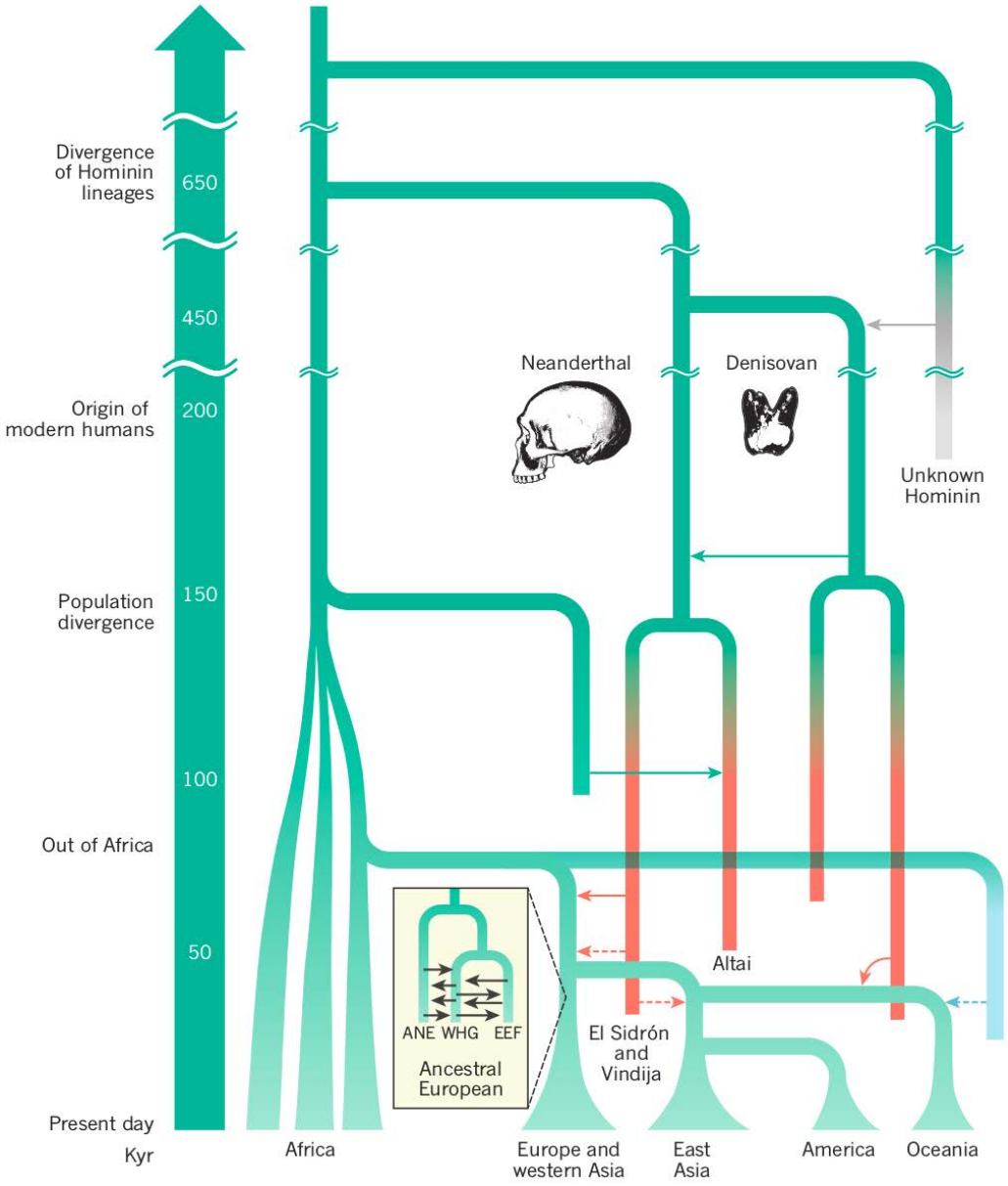


**Figure 1.4: Human evolutionary history**. Simplified model of recent human evolution. Horizontal solid lines indicate well established admixtures and dashed lines possible admixtures still under debate. Times (in thousands of years, kyr) are approximate. Reprinted by permission from Macmillan Publishers Ltd: Nature Nielsen et al. 2017, copyright 2017.

though exact times are still unclear). In sub-Saharan Africa is where we find the highest diversity and deepest population subdivisions, although with extensive admixture too (Nielsen et al. 2017). Populations that left Africa underwent at least one strong bottleneck. Although less severe, genomes of African populations have also the footprints of past bottlenecks (The 1000 Genomes Project Consortium 2015). Consistent with the advent of agriculture ∼10,000 years ago, populations underwent important expansions, although that was also linked to poorer health (worse nutrition and more pathogens).

Thanks to the genome sequencing of members of Neanderthal and Denisovan extinct human groups (Meyer et al. 2012; Prüfer et al. 2014), it has been possible to identify evidences of past admixture among *Homo* species. The divergence time between modern humans on the one hand, and Neanderthals and Denisovans (that were related hominin groups) on the other, has been estimated to be around 550,000 and 750,000 years ago, depending on the method used (Prüfer et al. 2014) (Figure 1.4). However, it has been measured that all present non-African populations have around 2% of DNA of Neanderthal origin, from at least one early interbreeding soon after the out-of-Africa (Wall and Yoshihara Caldeira Brandt 2016; Nielsen et al. 2017). Additionally, Melanesians in Oceania carry about 3-6% of Denisovan-like genome, whereas the south-east Asians carry Denisovan DNA in lower amounts (Wall and Yoshihara Caldeira Brandt 2016; Nielsen et al. 2017). Although it is likely that modern humans admixed with additional extinct groups, the absence of genetic sequence from these groups makes it difficult to determine.

### 1.3.2 Detecting selection in humans

New mutations can be deleterious, neutral or beneficial for an individual, which in the population evolve under negative (or purifying) selection, neutrally or under positive selection. Positive and negative selection are sometimes called directional selection, contrasting with other more complex situations found in diploid organisms, such as balancing selection. Most of the polymorphisms found in a population are expected to be neutral variants experiencing random frequency trajectories, governed by the intrinsically arbitrary nature of sampling gametes in a finite population (Kimura 1983), as well as by indirect effects of selection acting on neighbouring linked mutations (Maynar Smith and Haigh 1974). Strongly deleterious mutations do not contribute much to the polymorphism observed in a population, since they are quickly removed from the gene pool. Background selection is the term used to denote recurrent removal of deleterious mutations in functionally important genomic sites with little tolerance for changes. On the other hand, beneficial mutations are the substrate of adaptation and are expected

to be functional, since selection acts on a phenotypic level. Not surprisingly, much effort has focused on developing strategies to quantify and identify them in the genome (Nielsen 2005; Vitti, Grossman, and Sabeti 2013).

#### 1.3.2.1 Models of adaptive selection

The most widely-used model of positive selection is known as selective sweep or hard sweep (Maynar Smith and Haigh 1974). According to this model, a strongly beneficial mutation is expected to become fixed in few generations, increasing the frequency of neighbouring variants sitting nearby in the same chromosome and sweeping the population diversity of the genomic region (which is also known as genetic hitch-hiking) (Maynar Smith and Haigh 1974). Yet, it has been estimated that the classic sweep model has not been frequent in adaptations shaping the current human variation (Hernandez et al. 2011; Schrider and Kern 2017). Instead, soft sweeps are thought to be the norm in the recent human history. A soft sweep has a less obvious signature in the local genetic diversity and can originate from different situations. For example, a change in the environment can favour an allele already present in the population (that is, selection on standing variation). In this situation, the original haplotype has had time to recombine and several haplotypes can carry the mutation, so when the selected allele increase in frequency, some of the surrounding variation is retained. An alternative is that different mutations with similar effect are favoured and increase in frequency at the same time. In this case, when all individuals of the population carry one or the other mutation, selection would stop, leaving a compound signature.

Balancing selection maintains favourable genetic variability in population and is another important player in human evolution and adaptation (Key et al. 2014). There are different underlying mechanisms that can cause it, including overdominance (or heterozygote advantage), frequency-dependent selection and fluctuating environments. In humans, it was thought to be restricted to few well-known loci, such as the $\beta$-globin allele protecting from malaria and causing sickle cell anemia (Pasvol, Weatherall, and Wilson 1978) or variants in the major histocompatibility complex (Hughes and Nei 1988). In the last few years new targets have been identified (Andrés et al. 2009; DeGiorgio, Lohmueller, and Nielsen 2014; De Filippo et al. 2016), bringing renewed interest to the topic. However, despite its importance, it is thought to be less prevalent than other types of selection (Key et al. 2014). It has been suggested that balancing selection can be a source for soft sweeps, where alleles maintained by long-term balancing selection become favoured and fixed by directional selection (De Filippo et al. 2016).

Finally, adaptive introgression has been proposed as an important alternative model in humans (Racimo et al. 2015). In this case, admixture with

archaic humans acts as a resource of adaptive mutations. By the time modern humans expanded to other non-African regions, archaic humans like Neandertals and Denisovans had already had time to adapt to the new environments, and the adaptations could have been transferred in posterior introgressions. For example, data suggests that one of the haplotypes that helped Tibetan populations to adapt to high altitude hypoxia had Denisovan origin (or sister archaic group) (Huerta-Sánchez et al. 2014).

#### 1.3.2.2 Selection footprints in genomic variation

Past and ongoing selective events can leave strong footprints on genomic variation. Thus, a wide range of popular methods are designed to detect or quantify selection from genomic diversity. Selective sweeps can cause a strong local reduction in variation, whereas balancing selection produces an increase. A variety of distinctive patterns resulting from selection are the basis of the strategies reviewed below (see Vitti, Grossman, and Sabeti (2013) for an extensive review).

While there are methods to detect general rates of adaptation in the genome, specific genomic elements or regions, others are better suited for pinpointing individual recently-selected variants. Methods in the first group are able to detect older and recurrent events and typically make use of comparative data from different species (Nielsen 2005). Examples of these methods are the classic substitution rate comparison between putatively neutral sites (e.g. synonymous) and functional ones (e.g. non-synonymous), or McDonald-Kreitman and Hudson-Kreitman-Aguadé tests, that detect regions were levels of polymorphism and divergence differ from the expected correlation under neutrality. The second group is in theory capable of detecting individual mutations or regions under selection, although they can be also used to infer overall selective patterns in categories of elements. This different methods are limited to the detection of more recent or ongoing selection. They focus on altered frequency spectrum, linkage disequilibrium (LD) patterns, population differentiation or a combination of signals, which are explained in more detail below.

A shift in the allele frequency spectrum of the nucleotide variation of the surrounding genomic region is one of the characteristic patterns of a selective sweep. During the sweep, linked derived alleles get hitch-hiked and fixed or taken to high frequencies, while all the rest of variation is removed. Variation slowly recovers with the arrival of new mutations, that are initially at low frequency. Classic statistics such as Tajima's D (Tajima 1989) and similar tests exploit some of these signatures. In short, they compare two diversity estimators with different weights for each component of the frequency spectrum, and the difference between them measures the frequency distribution

shift. Site frequency spectrum-based statistics tend to have the strongest power to detect recently fixed mutations (Vitti, Grossman, and Sabeti 2013). This type of methods can also be used to detect old ongoing balancing selection, which shows a different pattern than directional selection. A region under balancing selection will display an excess of polymorphisms close to the frequency of the selected position (Key et al. 2014).

Linkage disequilibrium (LD) patterns are also altered with a rapid increase in frequency of a selected variant. The haplotype of the selected variant is expected to be longer that normal, given that recombination does not have time to break it down. Statistics such as the extended haplotype homozygosity (EHH) (Sabeti et al. 2002) are based on this second pattern. The most popular approach that uses the EHH idea is the integrated haplotype score (iHS) (Voight et al. 2006). Since recombination rate is variable across the genome, iHS uses the haplotype diversity levels in the alternative allele to correct for local recombination levels. As a result, iHS is well-powered to detect ongoing incomplete sweeps, but power drops quickly after fixation. A variant of the same concept is the cross-population extended haplotype homozygosity statistic (XP-EHH) (Sabeti et al. 2007), that uses haplotype lengths in different populations as a correction for local recombination levels. This alternative strategy allows XP-EHH to detect near-complete sweeps in one population. However, LD-based methods can incorrectly predict positive selection in regions with introgressed haplotypes, given that they result in similar extended haplotype signatures (Racimo et al. 2015). Therefore, care must be taken in contrasting alternative scenarios.

Another approach is based on unusual population differentiation in a genomic region, that can also indicate the action of selection. Adaptation is likely to act on specific environments, thus strong differences between populations suggest the presence of a locally beneficial variant. In contrast, low population differentiation could indicate balancing selection acting on multiple populations in the same way. Population differentiation methods are usually more robust to the presence of introgressed haplotypes, and can help discerning between neutral and selected introgressions (Racimo et al. 2015). The most widely-used statistic to measure population differentiation is Wright's fixation index $F_{ST}$, that compares the variance of allele frequencies within and between populations (Holsinger and Weir 2009). The first application of the statistic to detect positive selection was in the Lewontin-Krakauer test (LKT) (Lewontin and Krakauer 1973). Variations that include information about the demography and distances between populations have been developed, such as the bayesian approximations BayesFst (Beaumont and Balding 2004), or hapFLK (Fariello et al. 2013), that additionally focuses on differences between haplotype frequencies instead of that of alleles. There are also other simple statistics used for genomic scans with similar power to classic $F_{ST}$, such as the population branch statistic (PBS) (Yi et al. 2010)

and the difference in derived allele frequency between pairs of populations ΔDAF (The 1000 Genomes Project Consortium 2012).

Finally, there are composite methods, that use a combination of signals to improve resolution and specificity. Some integrate few types of patterns over multiple positions, while others integrate patterns from the three types in a single region. The composite likelihood ratio test (CLR) (Kim and Stephan 2002) and later modifications are examples of the first class, and the more recent composite of multiple signals (CMS) (Grossman et al. 2010) of the second. In addition, machine-learning strategies have been implemented in methods such as the hierarchical boosting strategy in Pybus et al. (2015) or S/HIC (Schrider and Kern 2016), offering extra information about the age or the strength of a selective sweep event.

#### 1.3.2.3 Discerning selection from neutral processes

In order to distinguish the patterns of a selected position from a neutral one, it is necessary to have an expectation. Classic tests are based on the rejection of neutrality expectations for simplified population models (i.e. a panmictic, constant-sized population). This is a convenient setting, since allow to derive analytic of the statistic expectations. However, natural populations are rarely in equilibrium and in certain situations neutral processes can mimic the signatures described above, such as changes in population size or structured populations (Simonsen, Churchill, and Aquadro 1995).

A simple approach to control for demographic effects relies on the comparison of the region of interest with genome-wide patterns (Haasl and Payseur 2016). The strategy assumes that all the genome is affected by past demography in the same way and that selection must be responsible for the regions with outlier values. Nevertheless, some demographic changes, such as a subdivided population, can increase the variance of the statistics, inflating the distribution tails with false positives and hiding real signals. It also assumes that most of the genome evolves neutrally. Humans have a small effective population size and the vast majority of the genome may be evolving by drift (Ohta 1972). However, we can not assume neutrality is the norm in organisms like *Drosophila*, with effective population sizes two orders of magnitude larger than ours (Vitti, Grossman, and Sabeti 2013; Haasl and Payseur 2016).

Alternatively, explicit models of demographic changes and linked selection can improve power, although they rely on uncertain demographic parameters (usually estimated from putatively neutral sites in the same sequences) (Bank et al. 2014). There models can be incorporated to scans or used to test alternative hypothesis in candidate regions. When models get too

complex, deriving analytical formulas to predict expectations can become intractable. Fortunately, technological advances have allowed the application of computationally-demanding simulation-based approaches. Popular applications are either based on forward-in-time simulations or coalescence simulations. While the former are more flexible and allow a wide range of scenarios, the efficient coalescent approach (only models the genealogy of the present-generation chromosomes) allows for computationally-intensive applications, such as approximate bayesian computation (ABC), that facilitates the estimation of underlying demographic parameters and model assessment (Sunnåker et al. 2013).

Another source of potential systematic biases that could be confused with selection is the source of the variants used in the tests (Nielsen 2004). This is an important concern in SNP array data, where SNPs are discovered in a panel of individuals and later genotyped in a larger sample in order to analyse frequencies and haplotypes (The International HapMap 3 Consortium 2010). In sequencing data, variation detection and genotyping are a single step, so the ascertainment biases are generally regarded as negligible, although they do have other biases (Crawford and Lazzaro 2012). The complex process in SNP arrays, and in any other two-step detection-genotyping scheme (including some used in inversions and other SVs), affects from the LD patterns to the frequency spectrum. Nevertheless, this kind of data can still be used for population genetic inferences, as long as the ascertainment process is modelled, including sizes and ethnicities of individuals used for the detection step, as well as any other filtering criteria (Nielsen 2004).

#### 1.3.2.4 Alternative strategies to detect selection

The above strategies assume a strong effect on fitness from a single or few positions. However, with the newest results from genome-wide association studies (GWAS) it is becoming more and more clear that many complex trait are polygenic (Boyle, Li, and Pritchard 2017). It has been suggested that selection may act also in a distributed way, leaving correlated footprints in regions of the same polygenic network. Turchin et al. (2012) found evidences of widespread weak selection on height, a highly polygenic trait, and Berg and Coop (2014) suggested a general framework to test selection from GWAS outputs. Other studies are following and testing for polygenic selection will probably become an important strategy in the future (Vitti, Grossman, and Sabeti 2013; Fan et al. 2016; Boyle, Li, and Pritchard 2017).

Overall, footprints of selection on genomic variation have been effective to detect selected variants in the human genome (Fan et al. 2016). Nevertheless, sequence signatures can be ambiguous and complementary analyses can help to convincingly determine the presence or absence of selection (Key et

al. 2014). A direct measuring of the fitness effect of a mutation is costly, but it has been employed successfully in some cases (Pasvol, Weatherall, and Wilson 1978). Correlation of genetic variants with potential selective pressures (e.g. malaria protective alleles and disease prevalence) can also improve selection estimates (Haasl and Payseur 2016), although they can be confounded by migrations and dispersal and need to be properly modelled (Frichot et al. 2013; Günther and Coop 2013).

Finally, another direct way to infer selection is by measuring allelic population frequency at different time points, since frequency trajectories can be very informative of the underlying selection coefficient. In natural populations multi-time point data can be difficult to obtain, especially in species with long-generation times, such as humans. Fortunately, the possibilities to do these studies are increasing with the development of technologies to analyse ancient genomes. Some authors have already explored the possibilities of the current limited data (Mathieson et al. 2015; Key et al. 2016). For now, frequency estimates are limited to either one (Key et al. 2016) or few individuals of similar geographical and temporal origin (Mathieson et al. 2015), assumed to represent ancestral frequencies. The lack of population samples of multiple individuals does not allow for the identification of individual selected sites yet, but it is already powerful enough to detect general trends.

### 1.3.3 Neutrality tests applied to human inversions

Most of the methods described above were designed with SNPs in mind, given that they are currently the best studied type of variation. Yet, structural variants could have greater potential to drive adaptation, since they are expected to have larger functional effects on cell and organism biology than smaller variants (Iskow, Gokcumen, and Lee 2012; Radke and Lee 2015).

As discussed earlier in Section 1.2, inversion special recombination characteristics create a strong population structure locally in the inverted region. As a result, nucleotide variation null models designed for SNPs are unlikely to represent the neutral scenario for inversions. Subdivided populations have longer times to the most recent common ancestor between groups (or in inversions, between orientations). Similar to introgressions (Racimo et al. 2015), neutrality tests based on genome-wide empirical distribution of local variation and linkage disequilibrium are likely to be misled by the presence of an inversion and the most robust methods are going to be those based on population differentiation. Distinguishing selection from neutrality in inversion nucleotide variation patterns requires more specific models. In that sense, some authors have suggested that models of linked variation (in absence of recombination) could be applied to inversions (Ferretti et al. 2017).

Their approach assumes absence of recombination also within chromosomes in the same orientation, that probably represents some of the patterns seen in inversions. However, accurate representation of inversion neutral variation patterns is likely to require explicit recombination modelling. Unfortunately, simulation options developed until now are not flexible enough to also incorporate complex demography (O'Reilly, Coin, and Hoggart 2010), or are not implemented as software packages (Peischl et al. 2013). For recurrent inversions, the available neutral null models are even more limited. Besides the special recombination patterns, high mutation rates violate the common unique-origin assumption. In particular, these high mutation rates may make them theoretically more similar to microsatellite models of evolution, although with only two alleles.

The best-known examples of selected inversions, together with similarly difficult structural variants, have been detected with strategies other than the analysis of nucleotide variation patterns. The 7q21.31 inversion was claimed to be under positive selection from its association to increased recombination rates and fertility in female carriers (Stefansson et al. 2005). Besides the direct measure of the fitness effects, the selective hypothesis was also supported by a low nucleotide diversity within the inverted chromosomes (Stefansson et al. 2005) (although higher diversity has been reported in later studies (Alves et al. 2015)). In other cases the correlation with environmental variables is a decisive evidence, such as in the classic *Drosophila* clines (Kapun et al. 2016) or the CNV example of the amylase gene duplication correlated with diets rich in starch (Perry et al. 2007).

Despite the increased difficulty posed by their effect on recombination, inversions offer some advantages for population genetics inference. The net divergence between orientations can be used as molecular clock to estimate the age of the inversion event, as explained in Hasson and Eanes (1996). However, a more recent study obtained very different age estimates when revisiting the same inversions, and even a negative estimate in others (Corbett-Detig and Hartl 2012), highlighting the high variance of the estimate and its dependence on the sampled individuals. Suggested alternatives include a minimum age estimate that assumes inversion is maintained at current frequency by balancing selection (Andolfatto, Wall, and Kreitman 1999) and an ABC estimate of age from intra-allelic variation assuming inversion exponential growth (Corbett-Detig and Hartl 2012). Similarly, reduced recombination rates can also be helpful to reconstruct local genealogy from long, informative haplotypes (Steinberg et al. 2012; Alves et al. 2015).

## 1.4 The InvFEST Project

The InvFEST Project started in 2010 with the aim of improving our understanding of the functional and evolutionary role of human polymorphic inversions, and ultimately fill the knowledge gap about this type of variation in the human genome (Cáceres 2010). The four specific objectives of the project are the following:

- Catalogue the precise location of all common polymorphic inversions in the human genome
- Determine the population distribution and the evolutionary history of these inversions
- Investigate inversion functional consequences and their effects on gene expression of human inversions
- Assess the effect of inversions on nucleotide variation patterns and the role of natural selection in their maintenance

This project has already contributed to much of the knowledge available today about human inversions with the creation of a unified non-redundant inversion database (Martínez-Fundichely et al. 2014), development of methods to analyse and interpret inversion predictions (Lucas Lledó and Cáceres 2013; Lucas-Lledó et al. 2014), optimization of targeted genotyping and validation techniques (Aguado et al. 2014), functional and evolutionary analyses of inversions of interest (Puig et al. 2015a) and systematic inversion validation studies (Vicente-Salvador et al. 2016).

Probably the most ambitious study within the InvFEST Project is focused on the exhaustive characterization of 45 common inversions by experimentally genotyping them in a large, diverse human sample (Figure 1.5). The study has been highly collaborative and had different stages. In the first one, a set of validated and already well annotated inversions was selected and the experimental methods to efficiently genotype them in multiple individuals were developed. The inversions included were originally predicted in early studies using strategies of paired-end mapping and genome assembly comparison, and all inversions had been validated and annotated within the InvFEST Project, with many of them already described in different publications (Korbel et al. 2007; Pang et al. 2013; Aguado et al. 2014; Lucas-Lledó et al. 2014; Puig et al. 2015a; Vicente-Salvador et al. 2016; Martínez-Fundichely et al. in prep.). In order to make such large-scale genotyping possible, a new experimental technique for high-throughput genotyping of inversions based prove hybridization was developed and optimized (Cáceres, Villatoro, and Aguado

2015). Then, in the second stage, the developed assays were used to genotype a large sample of 550 HapMap individuals (The International HapMap 3 Consortium 2010). Alternative low-throughput methods were also used in order check a large fraction of the genotypes and ensure the high quality of the data. Finally, in the third stage the newly generated data was combined with SNP and sequence data available for the samples (The International HapMap 3 Consortium 2010; The 1000 Genomes Project Consortium 2012; The 1000 Genomes Project Consortium 2015), to characterize the populations genetics patterns and functional effects of the inversion data set in a level of detail previously unreachable.

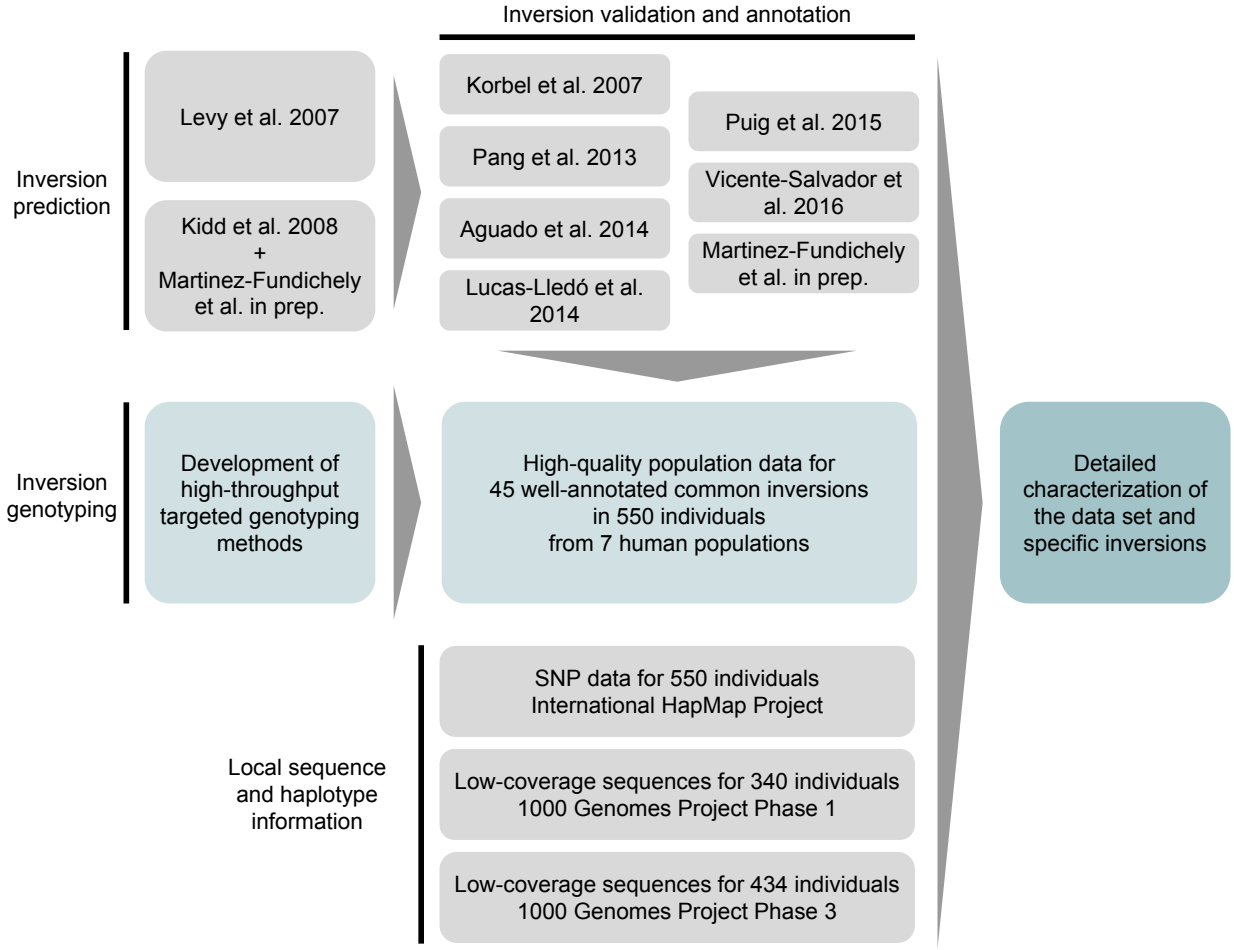


**Figure 1.5: Overview of InvFEST study to genotype and characterize common inversion polymorphisms in humans.**

## 1.5 Objectives

The aim of this thesis is to contribute to the complete characterization of the 45-inversion population data set generated within the InvFEST Project to determine the evolutionary importance of inversions and their effect on genome nucleotide variation and recombination. The specific objectives are the following:

1. **Complete the basic annotation of the 45-inversion data set and identify any systematic biases**
   The InvFEST data set offers a unique opportunity to understand different aspects of human inversions. However, the origin of the data is diverse and the level of information uneven. Thus, first it is necessary to complete existing annotations and to identify biases in the study design that could affect posterior analyses. The improvement of the basic information would also allow the application of the data set for other biologically relevant questions beyond this work.

2. **Determine if the frequency distribution of the studied inversions fits the expected neutral patterns**
   Inversions can create of unbalanced gametes if there is recombination within the inverted region in heterozygotes. Inversions in the data set are relatively small and common, so their impact on fitness of heterozygotes is hypothesised to be small but possible. Given the important evolutionary and medical implications of a reduction in fertility, an objective of the thesis is to infer their potential deleterious effect from the frequency distribution. Also, higher frequencies could be associated to the action of positive or balancing selection.

3. **Measure the inhibition of recombination between orientations and its impact on genomic variation**
   The inhibition of recombination between orientations could be complete or some genetic flux could take place between orientations as a result of gene conversion or double crossovers. Additionally, the consequences of the special recombination patterns on local genomic variation in humans are unclear. Therefore, the study aims to describe the recombination details and its impact on genomic variation for the diverse range of inversion frequencies and sizes in the data set, using both simulations and real sequence data with inversion genotypes.

4. **Estimate age and evolutionary history of individual inversions**
   An important fraction of inversions in the data set are expected to have a unique origin, meaning that all inverted chromosomes derive from a single event, while others may be recurrent and have appeared multiple times in the population. The aim is to use the sequence data to infer

their age and identify inversion recurrence using the large number of individuals from diverse populations in the data set.

5. **Identify inversions with patterns suggestive of natural selection**
   The functional impact of the different inversions in the data set is expected to be heterogeneous. The last objective is to highlight those inversions that could have been favoured by natural selection, as candidates to have an important functional effect in the human genome.

# Chapter 2

# Materials and methods

## 2.1 Data set description

This thesis is based on the most complete population-level inversion data set available. The data set includes 45 common inversions experimentally genotyped by targeted methods in 550 individuals from seven worldwide populations. What follows describes the origin and available annotations of the inversions, the composition of the genotyping panel, and an overview of the experimental methods employed.

### 2.1.1 Origin of the studied inversions

All inversions included in the final data set were originally detected in one or both of the following studies:

- The comparison of the HuRef genome and reference genome HG18, reported in Levy et al. (2007). The authors described all classes of differences between the two independently assembled genomes, including 90 inversions.
- A paired-end mapping (PEM) survey of nine individual fosmid libraries (one described in Tuzun et al. (2005), and the other eight added in Kidd et al. (2008)) designed to detect different types of structural variants in individuals from different populations. The data was then processed within the InvFEST Project using the specially developed inversion-detecting software GRIAL (Martínez-Fundichely et al. in prep.), obtaining 636 predictions, as well as reliability scores for each of them.

The human HG18 reference genome, used in the two studies, was built from sequences of different donors, although more than two thirds of the reference genome derive from the BAC library RPCI-11 (Lander et al. 2001). The anonymous male donor of RPCI-11 has been latter reported to be likely of admixed West African - European ancestry (Green et al. 2010). The HuRef genome is of European origin and the individual haplotypes were resolved in more than half of the assembly. This makes it partially diploid (Levy et al. 2007) and implies that variation in the two copies of each chromosome can potentially be accessed. For the second study, the nine analysed individuals (eight female and one male) have diverse ancestry: four African (YRI population), two East Asian (CHB and JPT populations) and three European (two CEU and individual NA15510, presumably European (Korbel et al. 2007)). In this case, the libraries were built from diploid cells, and therefore the variation found comes from 18 sets of autosomes, 17 chromosomes X and one chromosome Y. The characteristics of the original fosmid libraries are shown in Figure 2.1.

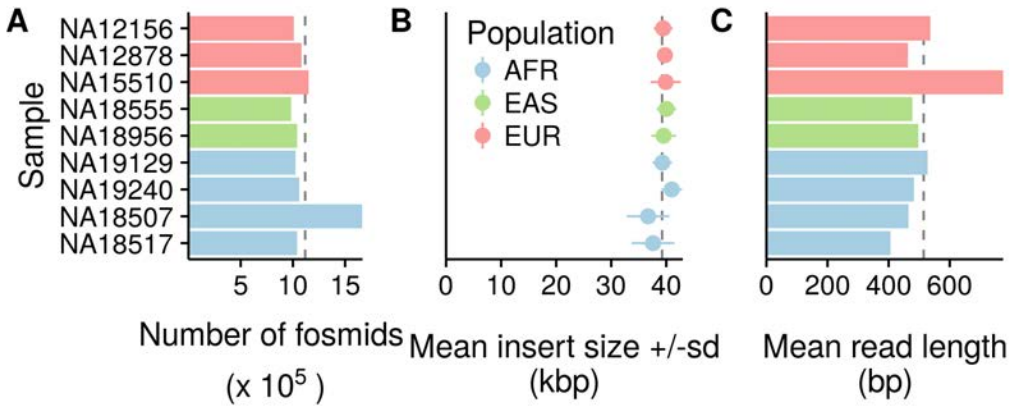


**Figure 2.1: Characteristics of the nine fosmid libraries used in the PEM analysis.** A. Number of fosmids per library. B. Fosmid insert sizes. C. Read length. NA18507 is the only male in the panel and the authors prepared two libraries to ensure a minimum coverage in chromosome Y (Kidd et al. 2008). NA15510 was the original library in Tuzun et al. (2005) and it has longer reads on average. Dotted lines indicate mean values for all libraries. Values obtained from Kidd et al. (2008) table 1 and Tuzun et al. (2005).

From all the putative inversions, there was a posterior experimental validation (or invalidation) step performed as part of the InvFEST project (Martínez-Fundichely et al. 2014). In order to validate an inversion prediction, a complementary experimental assay is necessary to confirm that the original samples where it has been detected do indeed carry the inversion. If the original samples are not available, validation consists in the confirmation that some fraction of the general population have both the reference and the alternative orientation. Each region needs a tailored PCR-based method, which can be more or less complicated depending on the complexity of the

breakpoints. Therefore, only a fraction of the best-scoring inversions can be validated (e.g. with more fosmid support or less mapping in repetitive sequence, that can generate false positives).

For the first study, 62 out of the 90 HuRef-HG18 inversion predictions were classified as errors in one of the assemblies or in the comparison (Vicente-Salvador et al. 2016). Of the rest, 22 were validated and only 6 remained as possible real inversions. From the 22 validated, 16 ended up in the population genotyping study. For the second study, the 636 predictions were ranked according to length, complexity and prediction reliability. The top-scoring inversions undergone a manual inspection step to identify prediction errors and select those predictions with characteristics consistent with real inversions and all of those for which a PCR assay could be set up were validated. Finally, at the end of this process, a set of 39 validated inversions were selected. In total, 45 inversions were included in the large-scale genotyping project, given that 10 of them were predicted by both studies.

### 2.1.2 Available inversion annotations

The 45 studied inversions are paracentric and located in different regions of the genome (Figure 2.2). One inversion is located in chromosome Y, seven in chromosome X and the remaining are autosomal. Detailed breakpoint annotations in reference genome HG18 were already available for all of them. Sequence annotation for 16 inversions from the assembly-comparison study were refined in Vicente-Salvador et al. (2016). The remaining 29 inversion annotations were described in the literature (Pang et al. 2013; Lucas-Lledó et al. 2014; Aguado et al. 2014; Puig et al. 2015a; Martínez-Fundichely et al. in prep.) and refined by other members of the group for the present study. The corresponding annotations and sources can be found in the InvFEST database (Martínez-Fundichely et al. 2014) and a summary is available in Tables B.1 and B.2 in Appendix B.

Inversion sizes range from 83 bp to 415 kbp, with a median of 4.1 kbp (Figure 2.3). Only three inversions have clean breakpoints (HsInv0092, HsInv0102, HsInv0379), whereas all the rest have other indels or repeats associated to the inversion breakpoints. More than half of the inversions (24) are flanked by inverted repeats (from 654 bp to 24.2 kbp, with a median of 5.9 kbp) present in the two orientations and with identities higher than 90%. The remaining 18 inversions are accompanied by small duplications, deletions or insertions in the derived allele, likely created in the same mutational event as the inversion itself (Figure 2.3) Three of them (HsInv0031, HsInv0045 and HsInv0098) have also inverted repeats in the ancestral orientation, although they have lower identity (from 83.2% to 86.2%) and are shorter ($< 300$ bp) than in the previous group. In addition, in all three cases one of the repeats

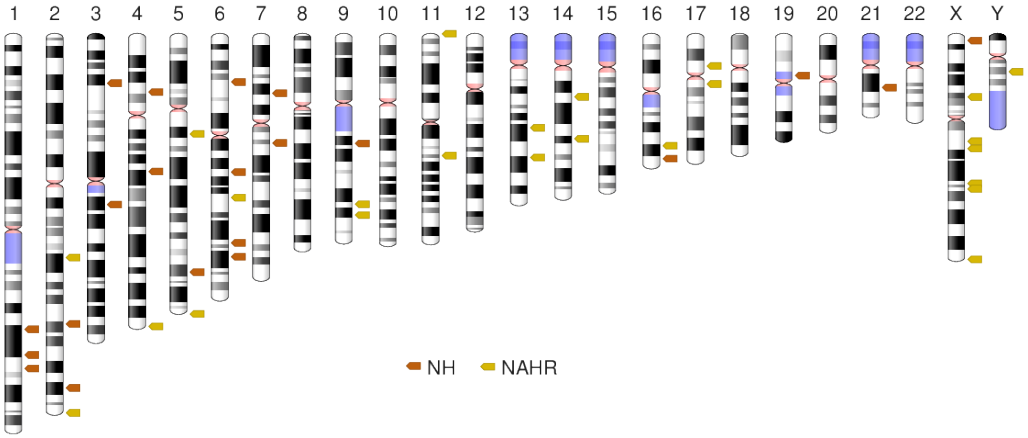


**Figure 2.2: Location of the 45 genotyped inversions in the human genome**. The ideogram was created with inversion positions (Tables B.1 and B.2) and NCBI Genome Decoration Page service (`https://www.ncbi.nlm.nih.gov/genome/tools/gdp`). Orange arrows indicate the position of inversions created by non-homologous mechanisms (NH) and yellow arrows indicate the position of inversions created by non-allelic homologous recombination (NAHR).

is partially deleted in the derived orientation.

The 24 inversions with highly-identical repeats are likely created by non-allelic homologous recombination while the remaining by other non-homologous mechanisms that can leave similar sequence signatures and are associated with other rearrangements. Since it has been shown that inversions mediated by NAHR are prone to invert recurrently (Cáceres et al. 2007; Aguado et al. 2014), it is convenient to treat them as a separate class. Thus, inversions in the project are classified in two groups: NAHR inversions, likely created by NAHR; and NH inversions, created by other non-homologous mechanisms, and likely to be single-event mutations. Additionally, for some inversions other information such as ancestral orientation and direct effect on genes is described in previous publications (Aguado et al. 2014; Puig et al. 2015a; Vicente-Salvador et al. 2016). In the frame of the genotyping study, other members of the group experimentally genotyped most inversions in a panel of 23 chimpanzees and 7 gorillas (Giner-Delgado et al. in prep.). Contributions of this thesis to the inversion annotation of the data set are described in results section 3.1.1 and methods section 2.2.1

### 2.1.3 Genotyping panel

The genotyping panel consists of 550 individuals from seven populations from four ancestry groups (here called super-populations, following the nomenclature used in 1000GP (The 1000 Genomes Project Consortium 2015)) Each population has between 45 and 100 individuals, with African and European

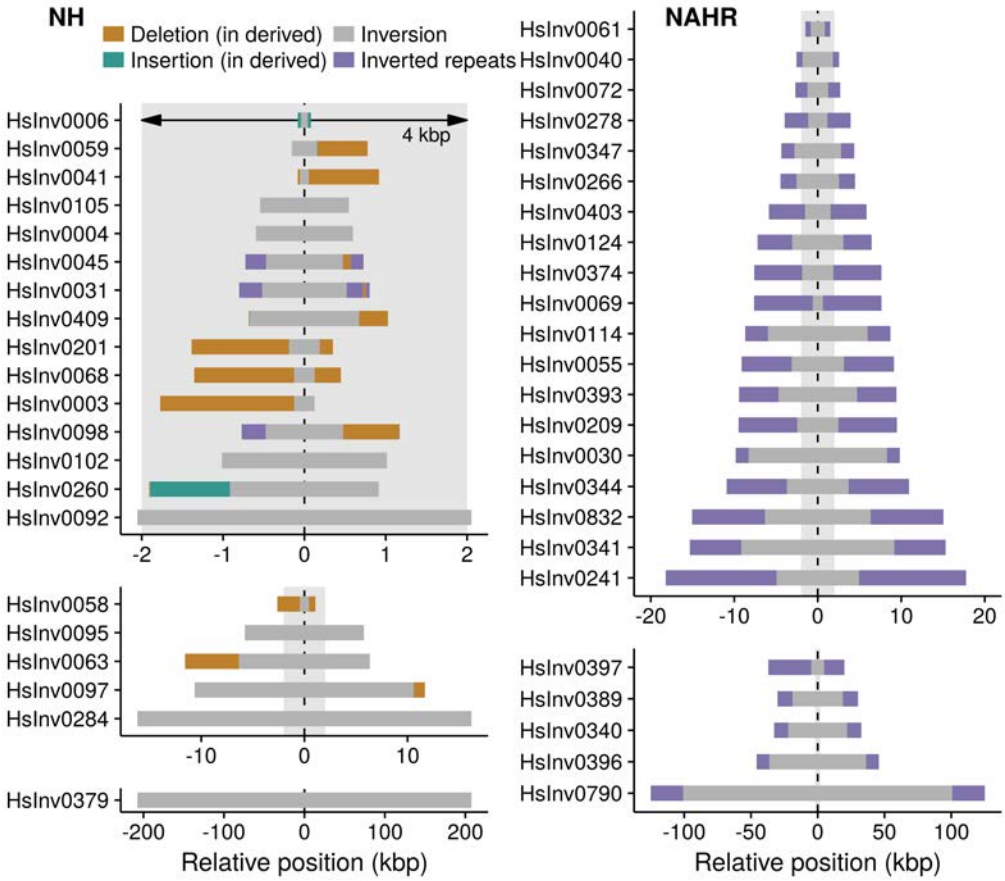


**Figure 2.3: Size and breakpoint complexity of the 45 genotyped inversions.** NH: inversions created by non-homologous mechanisms. NAHR: inversions created by non-allelic homologous recombination. In the representation of NH inversions, deletions are sequences present in the original sequence that are deleted in the derived orientation, and insertions are sequences gained.The HG18 reference genome has the derived orientation for some inversions and the ancestral for others (discussed in section 3.1.1.1). Grey shaded area in all panels corresponds to 4 kbp.

super-populations being the best represented (Table 2.1). Most individuals (480) do not share recent ancestors among them and therefore can be used to estimate population frequencies. The 70 remaining individuals are either children of mother-father-child trios (30 in YRI and 30 in CEU) or individuals with first and second degree of relationship unknown at the sample collection time and estimated based on sequence data (nine from LWK and one from GIH populations) (The 1000 Genomes Project Consortium 2012; The 1000 Genomes Project Consortium 2015). All tested DNA isolates come from lymphoblastoid cell lines, commercialized by Coriell repository and were provided directly by Coriell or extracted at the laboratory (Aguado et al. 2014).

**Table 2.1: Populations analysed**.

| Pop. code | Description | Super-pop. code* | Unrelated female/male | Offspring and related | Total Indiv |
|---|---|---|---|---|---|
| LWK | Luhya in Webuye, Kenya | AFR | 41/40 | 9 | 90 |
| YRI | Yoruba in Ibadan, Nigeria | AFR | 33/37 | 30 | 100 |
| CHB | Han Chinese in Beijing, China | EAS | 23/22 | 0 | 45 |
| JPT | Japanese in Tokyo, Japan | EAS | 22/23 | 0 | 45 |
| CEU | Utah residents with Northern and Western European ancestry from the CEPH collection | EUR | 30/30 | 30 | 90 |
| TSI | Toscani in Italia | EUR | 45/45 | 0 | 90 |
| GIH | Gujarati Indians in Houston, TX, United States | SAS | 45/44 | 1 | 90 |

* AFR: African; EAS: East Asian; EUR: European; SAS: South Asian.

The populations were chosen because of the numerous resources and information available regarding the samples, which have been used in different population genetics and functional studies. All individuals were included in the last phase of the International HapMap Project (The International HapMap 3 Consortium 2010). In addition, 82% of the individuals of the panel are part of the 1000 Genomes Project (1000GP) phase 1 (340) or 3 (434) (The 1000 Genomes Project Consortium 2012; The 1000 Genomes Project Consortium 2015). Therefore, the genome-wide variants of those individuals have already been characterized and are available. Figure 2.4 shows the number of individuals from each population that are included in each 1000GP phase. Some individuals in the 1000GP phase 1 were dropped from phase 3. Notably, 15 individuals from LWK population, some of them because of the cryptic relationships found later (The 1000 Genomes Project Consortium 2015). Overall, phase 3 includes 94 more individuals from our data set than phase 1, mainly due to the addition of the GIH population. For that reason, when phase 3 was officially released in 2015, we repeated most of the analyses with the new data.

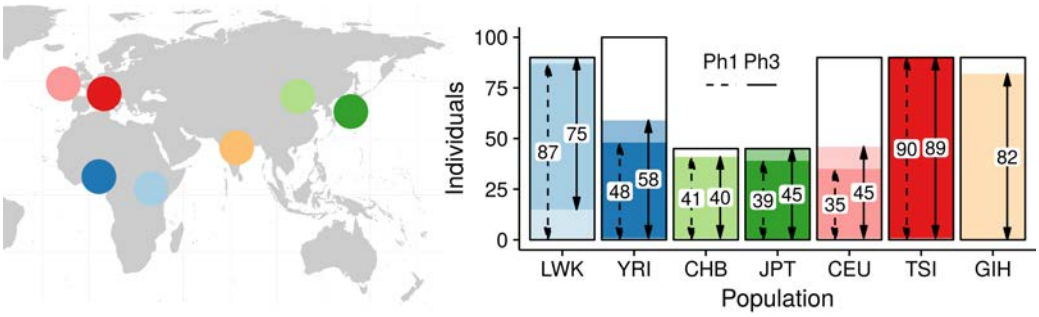


**Figure 2.4: Genotyped individuals also in 1000GP.** Overview of the individuals from the genotyping panel present in 1000GP. 340 individuals were included in phase 1 (62%) and 434 in phase 3 (79%), represented by dashed and solid lines and population colours.

### 2.1.4 Experimental methods overview

Most of the inversions were genotyped using multiplex ligation-dependent probe amplification (MLPA) (Schouten 2003) or inverse MLPA (iMLPA) (Cáceres, Villatoro, and Aguado 2015), a high-throughput technique that allows the interrogation of multiple inversions simultaneously in a sample. The method is based on the amplification of targeted probes with fluorescent labelling (Figure 2.5). For each region of interest two adjacent probes are designed that will ligate and amplify only if both are present together in the tested genome. The probe pair has three components: 1) a sequence that will hybridize with the target region, 2) a primer sequence that does not hybridize with the genome and allows the amplification of multiple probe-pairs with common primers, and 3) a stuffer sequence that modifies the final length of the amplified fragment, allowing for size identification by capillary electrophoresis.

While each orientation of inversions with non-repetitive breakpoints can be directly genotyped with regular MLPA probes designed at the breakpoints, more complex inversions with inverted repeats need some extra processing in order to obtain a orientation-specific unique target sequence. The strategy is similar to that of inverse PCR (Aguado et al. 2014), and it is based on the restriction of the sequence at both sides of the inverted repeats, followed by a self-circularization and ligation of resticted ends (Figure 2.5). Then specific probe pairs complementary to the ligation site can be used to detect the circular molecules from the two orientations.

Of the 45 inversions, 17 were genotyped in a direct MLPA experiment, 24 in an iMLPA experiment, and four additional inversions were added later and genotyped independently by multiplex regular PCR or inverse PCR (Giner-Delgado et al. in prep.). All inversion genotypes have passed quality control measures such as correct trio transmission and expected Hardy-Weinberg

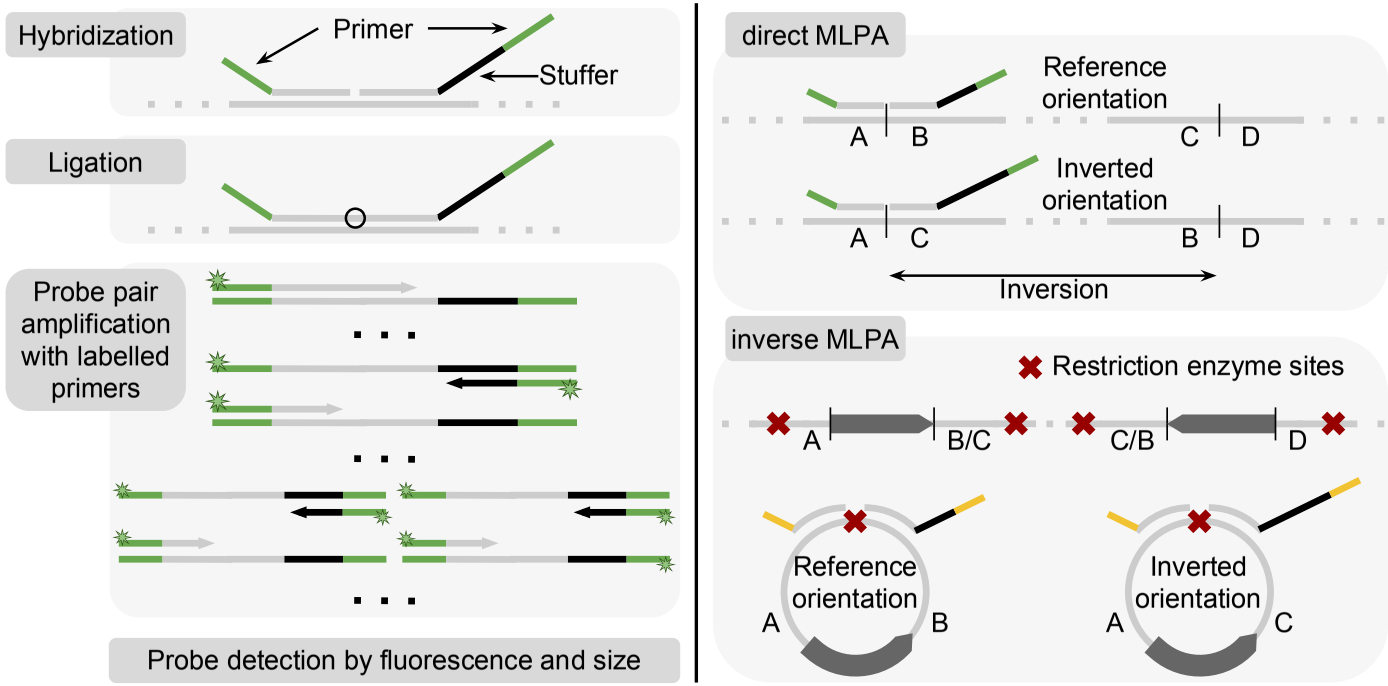


**Figure 2.5: Experimental genotyping methods** Overview of the strategy used for high-througput genotyping of inversions in the data set. Left panel shows main steps of the MLPA technique and right panel the application to inversions with simple breakpoints (top) and breakpoints with inverted repeats (bottom).

proportions. In addition, many were confirmed by other PCR techniques. The work in this thesis uses the version 4.7 of the genotype file, that includes several genotype additions and corrections.

## 2.2 Characterization of the inversion data set

### 2.2.1 Improvement of inversion annotation

#### 2.2.1.1 Inversion position and orientation in other genome assemblies

UCSC liftOver tool (Kent et al. 2003) was used to convert inversion coordinates, indels and inverted repeats from HG18 into HG19, the reference genome used in 1000GP. In order to estimate inversion orientation in other primate assemblies, as well as in newer human assemblies, we used an automated strategy based on `blat` tool (Kent 2002). Primate assemblies used are the following: chimpanzee panTro4 and panTro5 (Mikkelsen et al. 2005), bonobo panPan1 (Prüfer et al. 2012), gorilla gorGor4 and gorGor5 (Scally et al. 2012; Gordon et al. 2016), orangutan ponAbe2 (Locke et al. 2011), and rhesus macaque (Gibbs et al. 2007). All assemblies were downloaded from UCSC Genome Browser website in 2bit format. For each inversion, three separate sequences were extracted from genome HG18 in fasta format using `twoBitToFa` UCSC utility: the 10-kbp flanking region preceding the first breakpoint, the sequence between breakpoints and the 10-kbp flanking region after the second breakpoint. We excluded breakpoint regions and their associated inverted repeats and indels to avoid ambiguous mappings. For inversions were the region between breakpoints is longer than 20 kbp, two separate 10-kbp sequences internally adjacent to each breakpoint were extracted instead. Then, each sequence was aligned to the genome of interest using the command-line `blat` (v35x1) (Kent 2002). The longest hit was kept as the likely homologous region in the target assembly. Orientation was defined as reference if all best hits mapped in the same strand and as alternative if internal best hit(s) mapped in opposite strand than the external. To accept an orientation as valid, all best hits were required to be in the same scaffold or chromosome and the overall region span in the target assembly had to be between half and two times the HG18 sequence span. The process was automatized in a bash script. For result exploration and validation, sequences spanning the entire region were retrieved from each assembly and aligned with `Gepard` dotplot application (Krumsiek, Arnold, and Rattei 2007) using default parameters.

#### 2.2.1.2 Inversion position effect on genes

To determine the direct effect on genes, inversion positions in assembly HG19 were coded as 0-based bed files and overlapping RefSeq gene annotations (O'Leary et al. 2016) were retrieved from refGene table in UCSC Table Browser service (Karolchik et al. 2004), that had been last updated on 06-07-2017. Annotations include protein coding genes, non-protein coding genes and pseudogenes. Effect was classified according to the relative position of the genes with each breakpoint interval and checked manually using the Integrative Genomic Viewer (IGV) (Thorvaldsdóttir, Robinson, and Mesirov 2013). To estimate the distance to the closest gene, we downloaded the overlapping RegSeq genes with extended regions of 100 kbp at each side of the inversion. For each gene, distance was estimated as the smallest difference between the start or the end of the gene annotation and the outer positions of the breakpoints. Relative orientation of the inversion with respect to the nearby genes was annotated as upstream or downstream of de gene (i.e. the inversion is located in the 5' or 3' region flanking the gene).

### 2.2.2 Simulation of inversion frequency ascertainment bias

In order to estimate and reproduce the frequency biases introduced by the study design, we simulated the detection process in biallelic SNP from the 1000GP phase 3. We divided the process in three steps: small detection panel, limitations of the detection method, and inversion validation and inclusion criteria.

#### 2.2.2.1 Step 1: Use of a small detection panel

Inversions included in the study were originally detected in a reduced number of individuals and later on genotyped in a larger panel in this project. The detection of variants in a fraction of the population always introduces a certain frequency bias, given that mutations at high frequencies will be more likely to be detected. The ancestry of the panel also affects the variation that is going to be accessible –e.g. if a panel was mostly European, we would expect an over-representation of European variants.

For our model, we considered a random sample of 1000GP phase 3 biallelic SNPs as a proxy of the frequency distribution present in real populations (it is estimated that the 1000GP phase 3 includes 95% of variants over 0.5% in frequency and 99% of variants over 1% (The 1000 Genomes Project Consortium 2015)). Additionally, we considered only SNPs located in areas defined

as accessible for the 1000GP sequencing technologies, as indicated by the 1000GP strict accessibility mask.

To simulate the fraction of variation missed in the small detection panel, we filtered out SNPs where the alternative allele was absent in a same-sized detection panel. In principle, the number of chromosomes involved in the detection of inversions in the HuRef-HG18 comparison study were three for HuRef haplotype-resolved regions (one diploid and one haploid genome), although not for the entire genome (Levy et al. 2007). Thus, we opted by a conservative bias and kept those SNPs that were heterozygous in a randomly chosen CEU male sample (NA12872). For the paired-end mapping study, we used the original individuals of the fosmid libraries when possible. In two cases the original individual is not part of the 1000GP phase 3 study, and we replaced them with another individual of the same gender and population (NA18502 for NA19240 and NA12717 for NA15510).

#### 2.2.2.2 Step 2: Detection method limitations

The comparison of independent assembled genomes is in theory one of the least biased methods to detect structural variation, because it should not be directly affected by read mapping errors or library insert size limits. Comparing two complete assemblies should in theory allow to detect all inversion polymorphisms present, independently of their characteristics. Thus, we did not add any extra filter. However, in practice very complex repetitive regions that are too difficult to resolve are not included in the assembly, so inversions located in those areas will not be detected. Looking at predicted inversion sizes in Figure 2.6, we can observe that inversions from the assembly comparison study tend to be smaller than paired-end mapping predictions, indicating that the inversions detectable by assembly comparison are probably restricted to inversions shorter than 100 kbp. Nevertheless, the relationship between frequency and size is unknown and we cannot rule out independence.

For the fosmid paired-end mapping strategy some limitations had been described in the past (Lucas Lledó and Cáceres 2013). Paired-end mapping consists in sequencing the extremes of a fragment of known size and map them to a reference genome. Sequence read pairs that map in unexpected ways (also regarded as discordant) are indicative of either the presence of structural variation or errors in the process, for example errors in the reference genome or in the mapping step. To detect an inversion, one end has to map inside and the other outside of a chromosomal region with an orientation different from the reference.

We modelled the probability of detecting an inversion that is present in

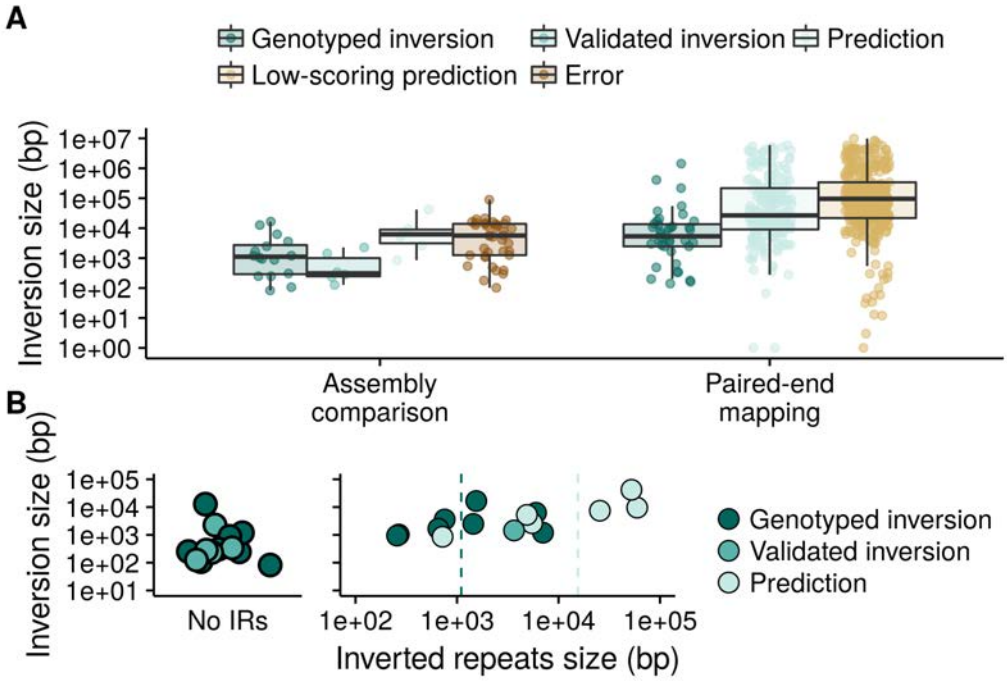


**Figure 2.6: Size and breakpoint characteristics of predicted inversions.** A. Sizes of all predicted inversions by each method. B. Detail of the size of the inversions and inverted repeats at the breakpoints in the 28 non-error predictions from assembly comparison. Dashed lines indicate median inverted-repeat sizes for inversions included in the population genotyping project and for those putative inversions not validated. Most of the inversions that have not been validated have long inverted repeats, which are challenging for the experimental assays.

the detection panel as a Poisson distribution with a $\lambda$ parameter equal to the expected number of discordant read pairs. A Poisson distribution in this genomic context is a good approximation to the binomial, given that the regions of interest are small compared to the whole genome. Although mapping biases are likely to exist (Lucas Lledó and Cáceres 2013), here we assumed that concordant and discordant fosmids are able to map with the same probability. GRIAL algorithm predicts an inversion if there are at least a minimum number of discordant pairs in the panel supporting it. For the generation of the predictions used in this study it was set to two. Therefore, we modelled the probability of detecting an inversion with two or more discordant pairs ($X$) as:

$$P(X > 1) = 1 - P(0) - P(1) = 1 - \frac{\lambda^0 e^{-\lambda}}{0!} - \frac{\lambda^1 e^{-\lambda}}{1!} = 1 - (1 + \lambda)e^{-\lambda} \quad (2.1)$$

The expected number of discordant read pairs, $E(d) = \lambda$, can be estimated as the number of reads that will map if uniformly distributed times the fraction of discordant chromosomes in the panel:

$$E(d) = \frac{min(inv - read, insert - 2read - ir)}{g} \times n \times f \quad (2.2)$$

Where $g$ is the sequenced haploid genome size, approximated to 3 Gbp for humans, $n$ the total number of fosmids sequenced in the nine original fosmid libraries, and $f$ is the fraction of chromosomes carrying the mutation in the nine diploid individuals of the study. The area where a reads can map depends on the library and inversion physical characteristics: fosmid insert size ($insert$, here 39.4 kbp, the mean size in the study), read size ($read$, 524 bp, again the mean), inversion size without repeats ($inv$) and size of the inverted repeats at the breakpoints ($ir$).

Inversions of any sizes are in principle detectable by paired-end mapping. However, it is not possible to detect inversions with inverted repeats at the breakpoints longer than the fosmid insert size. This is because the paired reads would not be able to span the entire repetitive element, making the inverted orientation indistinguishable. Figure 2.7 illustrates mappable area and probability of detection for the mean library characteristics and different inversion sizes. Small mappable areas, either because of a small inversion or because of long inverted repeats, reduce the probability of detection (Figure 2.7 A). On the other hand, if more individuals carry the inversion, then it is more likely that there are at least two discordant fosmids overall (Figure 2.7 B). Nevertheless, the probability of detection of inversions longer than ∼10 kbp is not very affected by their frequency in the panel, given that the

chances of having two or more discordant read pairs with just one carrier chromosome are high with the fosmid coverage used in the study.

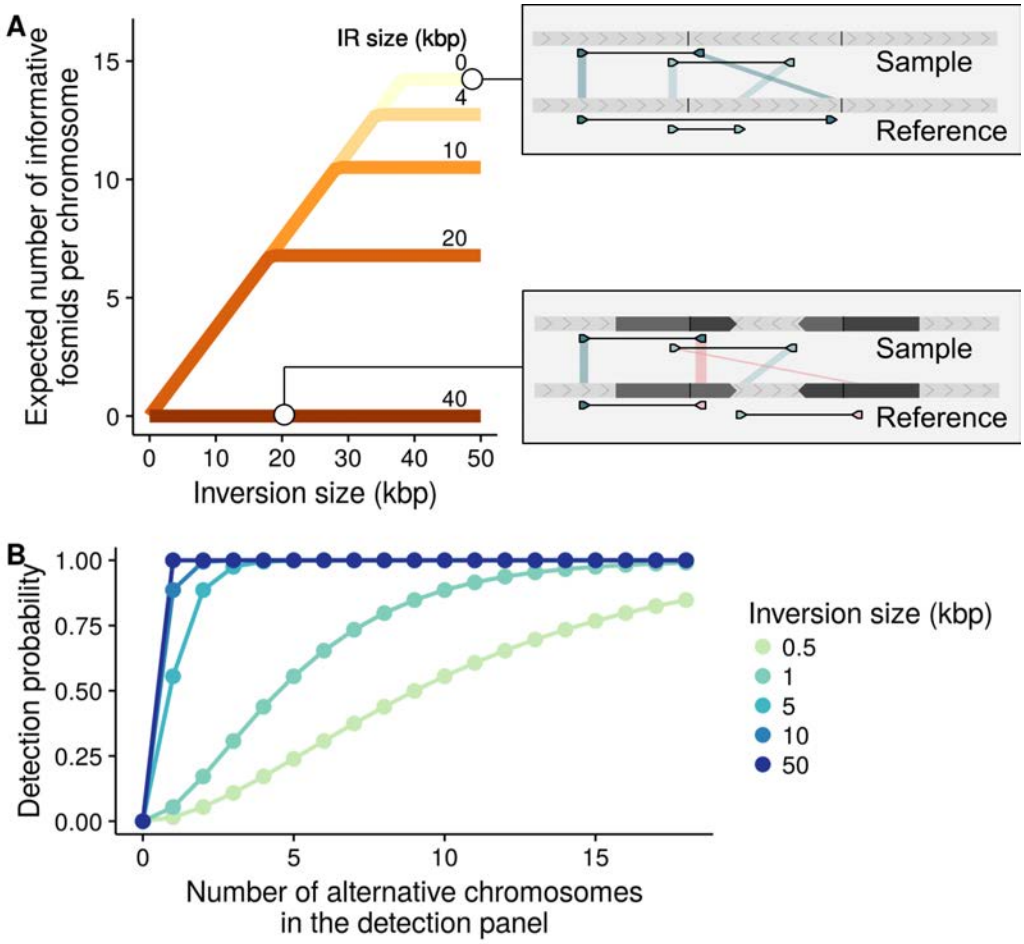


**Figure 2.7: Probability of inversion detection by paired-end mapping.** A. Effect of the inversion and inverted repeats size on the number of informative fosmids coming from one chromosome in the study. B. Effect of the number of inverted chromosomes on the overall detection probability in the study of inversions of different size. Parameters used here represent the fosmid library characteristics of analysed individuals.

In order to estimate the fraction of variation missed with the method, each random SNP was kept with the probability obtained from their panel frequencies and the physical characteristics of one of the inversions in the study. This process was repeated until a given number of SNPs was accepted for each simulated inversion characteristics, so that the final SNP set is representative of characteristics of the genotyped inversions. It is worth noting that this assumes that the inversion and inverted repeat lengths of the genotyped inversions are representative of the lengths of the entire set of human inversions.

For inversions in chromosome X, only SNPs from the same chromosome

were used to account for different sample size and effective population size. Pseudoautosomal regions were excluded from the analysis, since alleles in chromosome X and Y are jointly called in males and all our chromosome X inversions are located outside these regions. Autosomal inversions were matched with autosomal SNPs, independently of the specific chromosome. Inversion in chromosome Y was excluded from the analysis.

#### 2.2.2.3 Step 3: Inversion validation and inclusion criteria

Although an effort was made to include as many inversions as possible, given the cost of the validation and posterior genotyping, the selection of inversions to validate probably introduced additional biases to the fraction of inversions included in the large-scale genotyping project. This involuntary tendencies could favour inversions affecting genes, inversions easier to analyse, or inversions in well-resolved genomic regions.

By looking at basic characteristics such as inversion and inverted repeats size (Figure 2.6), we observed that selected inversions from the paired-end mapping analysis are significantly smaller than the high-scoring predictions (5.3 kbp against 26.6 kbp, P=$1.8 \times 10^{-5}$, Kolmogorov-Smirnov test). That suggests an additional bias towards small sizes in the selection step. In the assembly comparison data there are not big differences in inversion size. However, some of the inversions with longer inverted repeats have been excluded from the large-scale genotyping project, presumably because they are difficult to genotype. Again, since the relationship between size and frequency is unknown, we did not model any further bias.

#### 2.2.2.4 Simulation of the frequency bias in SNPs

For the analysis of the detection and selection process, we matched 1,000 SNPs per inversion and recorded the characteristics of both the rejected and the accepted (i.e. detectable) variants. In total 702,577 random SNPs were used. In all cases we also discarded SNPs with a GERP score higher than two (Davydov et al. 2010), in order to avoid positions under strong selection. The information was obtained from the functional annotation files available for 1000GP main release at `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/supporting/functional_annotation/unfiltered/`.

For later frequency-related analyses that use bias-matched SNPs as empirical expectation, a total of 10,000 were matched for each inversion. In this case, we recorded only the information about detectable variants and discarded that about undetected because of storage constrains. Also, two additional filters were applied for convenience: SNPs had to have the ancestral allele

defined in 1000GP files to allow derived allele frequency comparisons and they needed an assigned SNPdb ID for filtering purposes.

To avoid undesired biases from the additional filters, we assessed their effect on the smaller data set. From the 702,577 SNPs used, 7.6% did not pass the additional filters. However, 97.9% of those were not present in the detection panel, so they would have been missed in any case. The fraction that was in the panel and could have been detected corresponds to 1.5% of the variants in the panel. They are mostly discarded because of the absence of defined ancestral allele and have the same frequency distribution than SNPs that passed the filters. Therefore, we considered the effect of the additional filters negligible.

### 2.2.3 Frequency estimates and population distribution

The 480 unrelated individuals in the genotyping panel were used for population inversion frequency estimates, both globally and in each population and superpopulation (Table 2.1) The children from the 30 YRI and 30 CEU trios were excluded. Additionally, NA19313, NA19382, NA19470, NA19469, NA19352, NA19373, NA19396, NA19444, NA19311 and NA20871 were excluded from the frequency estimates, given that they were identified in the 1000GP phase 1 and phase 3 as being first and second degree relatives of other samples also present in our genotyping panel (The 1000 Genomes Project Consortium 2012; The 1000 Genomes Project Consortium 2015). The same unrelated individuals were used for the estimate of the median numbers of inversions and total inverted sequence per sample. The length of each inversion was measured as the span between breakpoint ranges, without including them.

### 2.2.4 Comparison to 1000GP inversion data

We compared the inversions in our study with those predicted and genotyped in the full structural variants release of 1000GP phase 3 (Sudmant et al. 2015). To identify those predictions that were detecting an inversion also present in our data set, we took the coordinates from 1000GP structural variant vcf file available in their ftp site and selected the inversions with overlapping breakpoints $\pm 5$ kb.

The inversion genotype error was estimated as the number of samples with discordant genotypes divided by the 434 total samples shared between our experimental genotypes and 1000GP predictions. We represented the genotypes according to the number of alternative alleles. In order to record the

direction of the error, we count an error as type *Alternative in 1000GP* when the experimental diploid genotype is 0 or 1 and the 1000GP genotype is 1 and 2, respectively. And *Reference in 1000GP* when the experimental diploid genotype is 1 or 2 and 1000GP genotype is 0 or 1, respectively. Frequencies in InvFEST were estimated using the 480 unrelated individuals in the study, and frequencies in 1000GP use all individuals in the main release of each population. Accuracy in the breakpoint position was assessed by comparing the POS and END attributes of the 1000GP vcf file with our breakpoint annotations.

### 2.2.5 Tag variant analysis

We measured the correlation between inversion genotypes and the genotypes of all nearby biallelic SNPs and indels of 1000GP phase 3 up to a 2 Mbp apart with `plink v1.90` (Purcell et al. 2007). The analysis was conducted for each population independently, as well as by superpopulation and globally, with the 434 samples together. BCF files from the 1000GP data portal (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`) were filtered using `bcftools` v1.2 (`hstlib` 1.2.1) (Li 2011) and recoded to plink format using `vcftools --plink` option (v0.1.15) (Danecek et al. 2011). Inversion genotypes were added with a custom bash script as mutations at the inversion breakpoints. Variants located within the breakpoint interval, in associated deletions, or in inverted repeats were excluded to avoid possible SNP or indel genotyping errors.

In order to assess the tag SNP coverage of 76 commonly-used commercial SNP arrays, we interrogated our SNPs of interest through the SNPChip web service from LDLink portal (Machiela and Chanock 2015). The implementation of LDlink references dbSNP build 142 and only accepts input for biallelic variants. Information provided for each SNP was then crossed with $r^2$ values to obtain inversion coverage per array. We also downloaded the specifications of the UK Biobank SNP array from Affymetrix website (`http://www.affymetrix.com/analysis/downloads/na34/genotyping/Axiom_UKB_WCSG.na34.annot.csv.zip`). Positions of both global and European-specific tag SNPs were crossed with the array marker list to select those present.

## 2.3 Inversion frequency patterns

In analyses comparing inversions with SNPs, we used the 434 individuals included in both the 1000GP phase 3 and the inversion genotyping panel. This way, sample size and population composition stays the same in both inversion and SNP measures. Inversion in chromosome Y was excluded

from the analyses. Measures from super-populations use all individuals of the super-population together, irrespective of the proportions of the composing populations. Empirical distributions of mean frequencies and mean $F_{ST}$ were estimated by sampling 10,000 sets of SNPs without replacement, one matched-SNP per inversion analysed at time, to preserve any underlying frequency structure. Section 2.2.2 details steps and filters used in the SNP matching process. Inversion classes were compared separately to their matched SNPs. Empirical P-values were estimated as twice the fraction of samples with values more extreme or equal than the observed.

$F_{ST}$ values were estimated with `vcftools --weir-fst-pop` function (v0.1.15) (Danecek et al. 2011) for pairs of populations within the same super-population and for each pair of super-populations. A global $F_{ST}$ was estimated using the four super-populations. Weir and Cockerham's $F_{ST}$ estimator (Weir and Cockerham 1984) can give negative values close to zero for little differentiated variants. Since conceptually $F_{ST}$ can only take values between zero and one, we substituted the few negative values for zero. To determine unusual $F_{ST}$ values of specific inversions, the value for each inversion and population combination was compared to the $F_{ST}$ distribution of SNPs in the same chromosome type (autosome or chromosome X), given that $F_{ST}$ in chromosome X tend to be higher.

## 2.4 Effect on local nucleotide diversity

### 2.4.1 Inversion simulation

We used `InvertFREGENE` software (O'Reilly, Coin, and Hoggart 2010) to simulate the effect of inversions on nucleotide diversity. The version used was modified after publication by one of the authors, Clive Hoggart, to label inverted chromosomes and allow to stop the simulations at a specified time instead of at a specified frequency. The algorithm simulates forward-in-time the evolution of a neutral inversion in a population of a specified size. Inversion-specific simulation works as follows. At each generation, parent chromosomes are sampled and recombination events are proposed according to the recombination map. If the event is inside the inverted region and the two sampled chromosomes are in opposite orientations, the recombination is rejected and proposed somewhere else.

The scaling factor used in the simulations was ten. That means that only with only 10,000 generations and 1,000 individuals, we are simulating the process of 100,000 generations in 10,000 individuals. To keep the recombination rate and the mutation rate constant in the population, the values are multiplied by the same factor (in this case we used the scaled parame-

ters suggested in the manual). A total of 10,000 independent simulations of 500-kbp were run for 10,000 generations without inversion, to reach variation levels expected in equilibrium, each of them using a different random seed. Then, for each simulation at equilibrium, a 300-kbp inversion or a 1-bp mutation were introduced in the center of the region and simulated for 10,000 generations more. Recombination seeds were kept the same between the equilibrium simulation and the inversion simulation in order to ensure the same recombination map, but simulation seeds were changed.

The modification of the algorithm, that allows stopping the simulation at a specified number of generations, results in an unbiased representation neutral end frequencies and ages. However, the process is less efficient than the default alternative of simulating an inversion until it reaches certain frequency (and therefore is enriched in young inversions). In the time-limited version used here, `InvertFREGENE` simulates during 10,000 the frequency changes of an inversion. If the inversion is lost from the population, another inversion is introduced. If the frequency of the previous simulation exceeds 0.1, it starts with a new random seed. During the course of the 10,000 generations many inversions are lost and only the one segregating at the last generation is analysed. The age and frequency of the last inversion were recorded for later analyses. The haplotype files of the last generation were analysed with a custom R script to estimate nucleotide diversity measures (Nei and Li 1979) and Tajima's D statistic (Tajima 1989). Only the 300 kbp in the middle of each simulation were used, representing the inversion region (or the inversion-free control).

### 2.4.2 Inversion nucleotide diversity

We downloaded the vcf file slices of the inversion regions plus 2 Mbp to each side from the breakpoints using bcftools v1.2 (hstlib 1.2.1) (Li 2011), from the 1000GP data portal (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`). Statistics were calculated individually for 1-kbp windows and for the inside regions using the R package `PopGenome` (Pfeifer et al. 2014). Then, we discarded windows with less than 80% of accessible positions, according to 1000GP pilot accessibility mask. In this analysis, the pilot accessibility mask was chosen, because only 11 inversions have at least 80% of the internal region accessible according to the strict accessibility mask. Accessibility of the inverted region and the flanking 2 Mbp can be found in Figure A.3. Regions comprising the breakpoint interval (including small insertions and deletions and inverted repeats) were not considered.

Average pairwise nucleotide differences within and between chromosome types were estimated with the nucleotide diversity method of the `PopGenome` package (Pfeifer et al. 2014). Standard chromosomes (carrying the ancestral

or major allele) and inverted chromosomes (carrying the derived or the minor allele) were defined as separate populations. To read each chromosome as a separate *individual* (so that it could be assigned to different *populations* of reference and alternative orientations), we generated pseudo-diploid vcfs, as suggested by the authors in case of haploid sequences. Only homozygotes or hemizygotes were included. Four inversions do not have homozygotes or hemizygotes for the inverted orientation and were excluded in the inversion diversity measures (Hsnv0061, HsInv0097, HsInv0379, HsInv0790). Five inversions do not contain polymorphisms to estimate diversity values inside the inverted region or have less than 80% accessible (HsInv0003, HsInv0006, HsInv0068, HsInv0278, HsInv0340 and HsInv0790).

In order to obtain relative measures, each statistic was divided by the same statistic estimated in the flanking regions with the same sequences, which also corrects for possible sample-size biases. This approach assumes that the flanking regions are not affected by the presence of the inversion. Although for the regions closer to the inversion breakpoints that may not be true, it is likely that the great majority of the 2 Mbp at each side will not be affected and the effect will be diluted. In that case, a possible effect on the neighbouring area would lead to conservative sub-estimate of inversion influence on nucleotide diversity.

Tajima's D (Tajima 1989) was estimated using the inversion regions with a custom R script. Only inversions with more than 80% of sequence accessible were considered. In order to separate the effect of the inversion from the effect of population structure, it was estimated independently for each population. Estimates were required to be based on at least five polymorphisms in the population.

## 2.5 Linked variation and recombination

### 2.5.1 Generation of same-frequency SNP data set

In order to have an expectation of normal linkage disequilibrium levels and types of linked variation, we selected 1,000 random genome-wide SNPs for each inversion and population with exactly the same allele frequency as the inversion in the population, using `bcftools` v1.2 (`hstlib` 1.2.1) (Li 2011) and custom bash script. For inversions with ancestral orientation known we matched the derived allele frequency, while for the remaining, the minor allele was used instead. Only target SNPs accessible according to the 1000GP strict accessibility mask were considered. Each inversion was compared to SNPs in the same chromosome type, autosomes or chromosome X, excluding pseudoautosomal regions. Inversion in chromosome Y was excluded from the

analysis, since it does not have recombination.

Statistical significance of differences between SNP-based values and inversion values was assessed by a strategy of sampling without replacement, as applied earlier for SNPs matched by ascertainment bias. Each summary statistic estimated for an inversion was compared to distribution of the statistic in the 1,000 SNPs at the same frequency. When the statistic is estimated for several inversions together, then 1,000 samples of one matched SNP per inversion are used to estimate the joint distribution, keeping any possible underlying structure. Empirical P-values were estimated as twice the fraction of samples with values more extreme or equal than those observed.

### 2.5.2 Linkage disequilibrium patterns

Linkage disequilibrium between inversions (or frequency-matched SNPs) and nearby SNPs was estimated with $r^2$ for all individuals together and for each population. As earlier, we used `plink` v1.90 (Purcell et al. 2007) to calculate $r^2$ for all biallelic SNPs and indels, but this time using only 500 kbp (instead of 2 Mbo) and keeping all variants with $r^2$ over 0.1 with the inversion.

### 2.5.3 Linked variant classification

In the first classification, biallelic SNP and small indels in the inverted region and the flanking 100 kbp for each inversion were classified using unphased genotypes from 1000GP. Only variants accessible according to the 1000GP strict accessibility mask were considered. Variants were classified as private to one or the other orientation, fixed between orientations (in complete linkage disequilibrium) or shared, according to its polymorphic state within each orientation. To consider a variant polymorphic in an orientation the two alleles need to be unambiguously present in the orientation after taking into account the genotypes of both homozygotes for the two orientations and heterozygotes. The classification was done as follows: if a variant is polymorphic in the two orientations, it is classified as shared; if it is only clearly polymorphic in one, it is classified as private to that orientation; if it is monomorphic in both orientations and the orientations have different alleles, it is classified as fixed. When only present in heterozygosis in individuals heterozygous for the inversion, it remains unclassified.

For the second classification we took into account the direction of the mutations to have higher power to detect allele combinations that require recombination to happen. We followed the classification of linked variants used in Ferretti et al. (2017), adapted to inversions. The definitions applied to

inversion polymorphisms used here are as follows, with mutation meaning the derived allele of a variant.

- STRICTLY NESTED: mutation carried by a subset of the chromosomes with the derived orientation and absent in the ancestral orientation.
- CO-OCCURRING: mutation carried by all the chromosomes with the derived orientation and absent in the ancestral orientation.
- ENCLOSING: mutation carried by all the chromosomes with the derived orientation and a subset of those with the ancestral-orientation.
- COMPLEMENTARY: mutation absent in the derived orientation and carried by all the chromosomes with the ancestral orientation.
- STRICTLY DISJOINT: mutation absent in the derived orientation and carried by a subset of chromosomes with the ancestral orientation.

Additionally, in the presence of recombination (or genotyping errors), we define two more types:

- SWITCHED: mutation carried by all the chromosomes with the ancestral orientation and also by a subset of the chromosomes with the derived orientation.
- SHARED: mutation carried by subsets of chromosomes with the derived and with the ancestral orientation (same as in the previous classification).

All the biallelic SNPs within the inversion region and in the 20 kbp after the breakpoints were classified. Since the classification requires phased haplotypes, we incorporated the inversion alleles in the haplotype data, when possible. For inversions with perfect tag SNP defined from genotypes ($r^2 == 1$), we were able to assign the orientation of each haplotype of heterozygous individuals using the tag SNPs. For inversions with more than one tag SNP, all inversion-associated alleles where located in the same chromosome, giving some confidence about the phase quality. For inversions without tag SNPs but with known ancestral orientation, we limited the analysis to the chromosomes from homozygous individuals. Inversions without ancestral orientation known, likely to be recurrent, were excluded from the analyses.

The same classification was also applied to the 20 kbp flanking regions of the same-frequency SNPs. In this case, the phase provided by the 1000GP was used for the target SNP and its linked SNPs.

NCBI 1000 Genomes Project Browser v3.6 was used in order to analyse the short read alignments of selected individuals for shared or fixed variants of interest (`https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/phase3/`).

### 2.5.4 Definition of non-recombining regions

We tested three different criteria to define the extended regions without evidences of recombination between orientations (visually compared in Figure A.2). All of them were implemented in R (R Core Team 2017). The inverted region was always considered as non-recombining and all exchanged variants were assumed to be genotyping or phasing errors.

In an initial simple criterion, the first switched or shared variant (globally called here it exchanged variant) is considered an evidence of recombination. We saw that in many cases the first shared variant was isolated from others and in some cases there were still fixed variants extending past that point. Since a single event of recombination by crossing-over should turn all fixed variants into shared, they are likely to be a consequence of genotyping or phasing errors. In order to overcome this potential errors, we initially thought of taking into consideration the frequency of the exchanged variants. However, since some inversions are in low frequency, setting a frequency threshold for reliable exchanged variants looked too inflexible. Instead, in the second criterion we set minimum fraction of fixed divided by informative variants (exchanged plus fixed) in 5 kbp non-overlapping windows. Despite solving the initial problem in some cases and covering most of the regions with fixed variants, it lacked of precision. Finally, we opted for a more flexible sliding-window approach, more tolerant to the presence of exchanged variants. Specifically, if the same fraction as before decreases from one window to the next one, it indicates a possible recombination events. The possible event is only considered an evidence of recombination if the fraction does not fully recover in further windows. We selected a window size of ten polymorphisms with step size of one. This third criterion was finally applied using the individuals of each population independently and also globally with all the 434 individuals in the 1000GP phase 3. The same criterion was applied to the same-frequency 1,000 SNPs per inversion and population.

## 2.6 Evolutionary history inference

### 2.6.1 Age estimate from divergence

Inversion age was estimated with a divergence-based approach similar to that Hasson and Eanes (1996). The differentiation accumulated since the creation of the inversion is estimated from the mean pairwise nucleotide differences between sequences in opposite orientations after subtracting the average differences expected in the original population (with free recombination). For that, the intra-allelic variation within the ancestral orientation is usually employed, results in the formula

$$T_{inv} = \frac{d_{inv} - d_{intra-allelic}}{2k}$$

where $d_{inv}$ and $d_{intra-allelic}$ are the pairwise differences between and within orientations and $k$ is the local substitution rate. However, ancestral orientation variability can be strongly reduced in inversions where the derived allele has reached high frequency (as seen in InvertFREGENE simulations, Figure 3.14).

In order to assess the precision of the estimate, we first applied the formula to the previous simulated inversions. As approximation to the expected variation between chromosomes in the original population (the subtracting factor in the formula) we considered three alternatives: a) the expected differences at equilibrium $4N_e\mu$, b) the average pairwise differences within ancestral chromosomes, and c) the largest average pairwise differences within chromosomes with the ancestral or the derived orientations (Figure 2.8). As general patterns, ages tend to be overestimated, although age estimates for recent inversions can be negative with the three alternatives. Simulations confirmed that the second alternative exacerbates the age overestimation in high-frequency inversions and that it is preferable to use the largest of the observed intra-allelic differences, so we applied that third option to real data. Since correspondence between simulated age and real age is unclear, because of non-simulated factors (e.g. complex demography) and uncertain parameters (e.g. generation time or substitution rate), we decided not to attempt any additional correction but instead consider the likely age overestimation when interpreting the results.

In the real data, we estimated pairwise differences between phased haplotypes from 1000GP phase 3 for the 434 individuals with inversion genotypes, taking all populations together and only using SNP variants. Inversion phased was incorporated as earlier using tag variants. For inversions without tag SNPs, only homozygous individuals were used. We used the inverted sequence and, in inversions with fixed SNPs, we were able to include extra flanking region until the first evidence of recombination (as de-

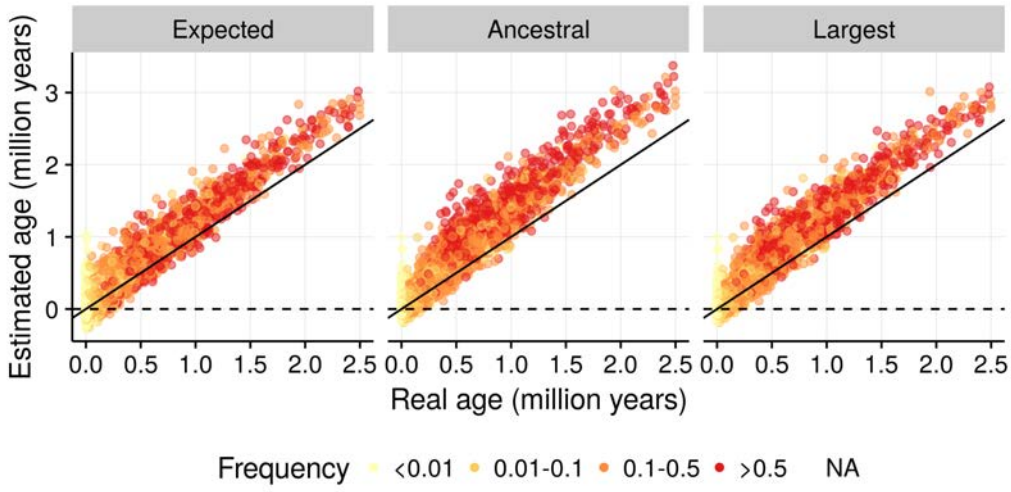


**Figure 2.8: Performance of age estimator in simulations.** Simulated age against that predicted subtracting the expected number of differences in equilibrium (*Expected* panel), the observed average pairwise differences between chromosomes with the ancestral orientation (*Ancestral*) or the largest average pairwise differences of the observed within ancestral or derived chromosomes (*Largest*). Ages shown assume generation time of 25 years.

fined by the third criteria in 2.5.4), with a maximum distance of 20 kbp. That allowed us to have enough information in some short inversions. In all cases, breakpoint intervals including inverted repeats, microhomology or inversion-associated deletions were excluded to avoid errors created by incorrect mapping of short reads. In two low-frequency inversions, HsInv0061 and HsInv0790, divergence could not be estimated because all inversion carriers are heterozygous and we could not phase the chromosomes with confidence. Inversion HsInv0832 in chromosome Y was also excluded from this analysis. In order to control for the noise in the mean pairwise difference values, we repeated the estimates by sampling the same number of total individuals but with replacement (i.e. bootstrap). A first age estimate was obtained by subtracting the largest intra-allelic average pairwise differences $d_{intra-allelic}$ to $d_{inv}$ and using a constant substitution rate of $1 \times 10^{-9}$ changes per base-pair per year.

Then, in order to account for local differences in substitution rate, we estimated local substitution rates using the divergence with chimpanzee and gorilla genomes, according to the following formula

$$k = \frac{d_{outgroup} - d_{intra-allelic}}{2T_{split}}$$

where $d_{outgroup}$ is the divergence with the outgroup and $T_{split}$ is the split time between humans and the outgroup. For that, we retrieved the pairwise LASTZ alignments (Harris 2007) of human with chimpanzee (assem-

bly CSAC 2.1.4/panTro4) and with gorilla (assembly gorGor3.1) from ENSEMBL GRCh37 portal (Yates et al. 2016), using the compara Perl API (Herrero et al. 2016). We recorded the substitutions between outgroup assemblies and HG19/GRCh37 in the same region used above, after removing alignment gaps and non-syntenic alignment blocks. Kimura's 2-parameter substitution model (Kimura 1980) was used to convert observed differences into genetic divergence per bp. We set a minimum alignment length of 1 kbp to estimate the substitution rate. Three inversions did not pass the criteria for neither outgroup alignment (HsInv0041, HsInv0055 and HsInv0069). Only one outgroup alignment was left for five inversions (HsInv0072 only chimpanzee, and HsInv0045, HsInv0241, HsInv0278 and HsInv0344 only gorilla). Additionally, the alignment with chimpanzee for inversion HsInv0374 was also discarded because it harboured an unusually high substitution rate (around 5 times the average value) and it was shorter than the alignment with gorilla.

Estimated substitution rates from chimpanzee and gorilla genomes showed an overall Pearson's correlation coefficient of $r = 0.6$, showing that there is high heterogeneity in local divergence times (time to the most recent common ancestor for the specific genomic region) and possibly noise in the estimates based in short sequences (Figure 2.9). When using the split time ranges 5-7 and 8-10 million years for chimpanzee-human and gorilla-human, substitution rates are within the range $0.3 \times 10^9 - 7 \times 10^9$ substitutions per bp per year. We observed that in our regions the speciation times of 6 and 8 million years age fit better the generally accepted divergence-based substitution rates of around $1 \times 10^{-9}$ (Figure 2.9), and used them to convert genetic divergence into time.

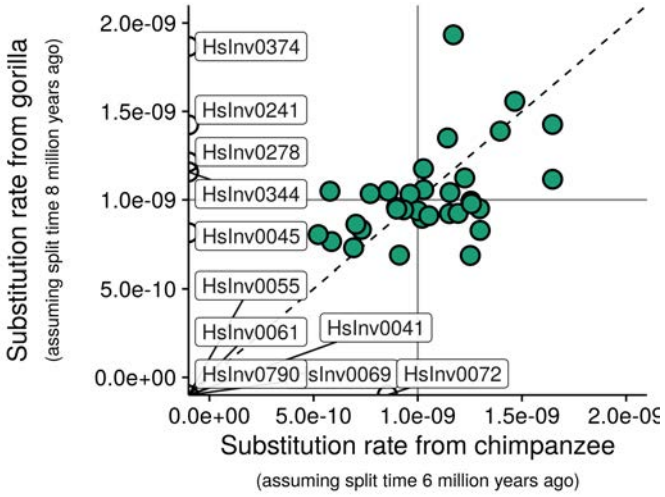


**Figure 2.9: Local substitution rate estimates**. Inversion regions without alignments or with only one are labelled and shown in white. Lines at $1 \times 10^{-9}$ indicate generally used substitution rate. Dashed line indicates perfect correspondence between two estimates.

### 2.6.2 Haplotype analysis

We used 1000GP phase 3 phased haplotypes of the 434 individuals shared between projects. In order to reduce haplotype complexity, we only considered SNP variants and that are present in at least two chromosomes. To avoid misleading noise from possible genotype errors, we also discarded SNPs within inverted repeats or insertions and deletions associated to the breakpoints, and those SNPs not accessible to 1000GP sequencing technologies according to the pilot criteria. We analysed the variants within the inverted region for all inversions. For inversions where we had previously identified the non-recombining region between orientations, that was also included. Inversion in chromosome Y does not have any SNP within the inverted region reported by 1000GP phase 3, since it is located outside the regions defined as callable (The 1000 Genomes Project Consortium 2015). However, since it is located outside the pseudoautosomal regions of chromosome Y and there is no possible recombination between haplotypes, we used variants in the flanking area (within 200 kbp from the breakpoints) to have enough information. All other inversions have at least nine variants, except inversion HsInv0041 that only has one SNP and was excluded from the analysis.

Distances between simplified haplotypes were computed as the number of pairwise differences. Simplified haplotypes were then clustered with UPGMA method implemented in R base function `hclust` (method average) using computed distances. The hierarchical clustering was represented in a dendrogram using `ggdendro` R package (de Vries and Ripley 2016) coupled to information about the population and the orientation of the chromosomes carrying each haplotype, the haplotype alignment and the distance matrix. Information about each haplotype is tracked in a three-table system: (i) sample information with the population, inversion genotype and haplotypes; (ii) haplotype information with the count of chromosomes from each orientation and population carrying the haplotype; and (iii) position information with details about SNPdb ID, ancestral allele, variant type and alternative allele count. The system facilitates manual identification of individual inversion events or potential genotype errors.

# Chapter 3

# Results

The analysis of the 45-inversion data set has been the principal project in my thesis. The article of the study is currently in preparation (Giner-Delgado et al. in prep.) and describes both the generation of the genotype data and the different types of analyses performed. In this chapter I present my contribution to the characterization of the data set and its evolutionary analysis, which forms part of the article.

**Carla Giner-Delgado***, Sergi Villatoro*, Jon Lerga-Jaso*, Magdalena Gayà-Vidal, Meritxell Oliva, David Castellano, David Izquierdo, Isaac Noguera, Bárbara Bitarello, Iñigo Olalde, Alejandra Delprat, Antoine Blancher, Carles Lalueza, Tõnu Esko, Paul O'Reilly, Aida Andrés, Luca Ferretti, Lorena Pantano, Marta Puig, Mario Cáceres (in prep.). "Functional and evolutionary impact of polymorphic inversions in the human genome".
* Equal contribution

Additionally, during de development of the thesis, we have published three related studies with a smaller scope. In them, some of the inversions in the large study were described in detail in specific populations and methods were set up and tested in a small scale, as a natural first step to the work presented here. The three published articles are:

(i) The validation and genotyping in a European population of 17 inversions with inverted repeats predicted with GRIAL algortithm (Martínez-Fundichely et al. in prep.) using paired-end mapping data from Kidd et al. (2008). I contributed to the analysis of the SNPs and indels from the 28 unrelated individuals in the 1000GP phase 1.

C. Aguado, M. Gayà-Vidal, S. Villatoro, M. Oliva, D. Izquierdo, **C. Giner-Delgado**, V. Montalvo, J. García-González, A. Martínez-Fundichely, L. Capilla, A. Ruiz-Herrera, X. Estivill, M. Puig and M. Cáceres (2014). "Validation and genotyping of multiple human poly-

morphic inversions mediated by inverted repeats reveals a high degree of recurrence". *PLoS Genetics* 10.3, e1004208.

(ii) A single-inversion study analysing the evolutionary properties and functional impact of inversion HsInv0379. This inversion is also part of the 45-inversion data set used here. This inversion is characterized by being the longest inversion in the data set and one breaking a gene, and it is only found in East Asian populations at an average frequency of 4.7%. I contributed to the evolutionary characterization of the inversion and the *in silico* genotyping of individuals in the 1000GP phase 3 (The 1000 Genomes Project Consortium 2015). In particular, I estimated the nucleotide diversity and age of the inversion with a high-coverage sequencing resource (Wong et al. 2013), and simulated different selective scenarios to determine inversion's likely selection coefficient.

M. Puig, D. Castellano, L. Pantano, **C. Giner-Delgado**, D. Izquierdo, M. Gayà-Vidal, J. Lucas-Lledó, T. Esko, C. Terao, F. Matsuda, and M. Cáceres (2015). "Functional impact and evolution of a novel human polymorphic inversion that disrupts a gene and creates a fusion transcript". *PLoS Genetics* 11.10, e1005495.

(iii) The validation and characterization of the 90 inversions predicted in Levy et al. (2007) from the comparison of HuRef and HG18 assemblies. I contributed to the characterization of 17 inversions that had been experimentally genotyped in European samples. Specifically, I updated the tag SNP analysis and compared inversion genotypes and frequencies with those of 1000GP phase 3 (Sudmant et al. 2015). All inversions in that analysis, except one (HsInv0052), are also included in the extended work presented here with 29 additional inversions.

D. Vicente-Salvador, M. Puig, M. Gayà-Vidal, S. Pacheco, **C. Giner-Delgado**, I. Noguera, D. Izquierdo, A. Martínez-Fundichely, A. Ruiz-Herrera, X. Estivill, C. Aguado, J. I. Lucas-Lledó, and M. Cáceres (2016). "Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution". *Human Molecular Genetics* 26.3, pp. 567–581.

## 3.1 Inversion data set characterization

Inversions in the human genome are still largely understudied because of the technological challenges in their detection. The high-quality population data for 45 inversions created within InvFEST Project represents a unique resource to understand different aspects of inversions, from their impact on health and disease and evolutionary role, to the improvement of detection algorithms and analysis methods. The results in this section have a double purpose: 1) to complete required information for the evolutionary analyses in later sections, and 2) to improve the usability of the data set for other applications. In the following pages, basic annotation is completed in a unified and systematic way, the biases of the study design are identified, genotype information is compared to predictions in external projects, and tools are provided to facilitate the study of the data set in more populations.

### 3.1.1 Inversion annotation

#### 3.1.1.1 Ancestral orientation

Information about the directionality of the mutation event is key to understand inversion evolution, as well as to understand the generation mechanisms. The ancestral orientation had been defined for a fraction of the inversions in the data set in Aguado et al. (2014), Puig et al. (2015a) and Vicente-Salvador et al. (2016). Several methods were used in these studies: sequence signatures indicating mutation direction (interrupted genes or transposable elements in the breakpoints), experimental genotyping in chimpanzee and gorilla, and alignments to chimpanzee assembly panTro4, gorilla assemblies gorGor3 and gorGor4, and rhesus macaque assembly rheMac8. In order to complement and expand the published information, other members of the group experimentally genotyped most inversions of the study in a larger panel of 23 chimpanzees and 7 gorillas (Giner-Delgado et al. in prep.) (represented together with published information as *Experimental* and *Assembly* comparison methods in Figure 3.2).

To complete the assembly-based information as well as to survey the newest reference assemblies, we aligned HG18 inversion region against different chimpanzee, bonobo, gorilla, orangutan and rhesus macaque assemblies (panTro4, panTro5, panPan1, gorGor4, gorGor5, ponAbe2 and rheMac8). A blat-based automated pipeline was designed to infer the most likely orientation in each assembly for which enough information was available (see Materials and Methods).

The 21 inversions likely created by non-homologous mechanisms (NH) had

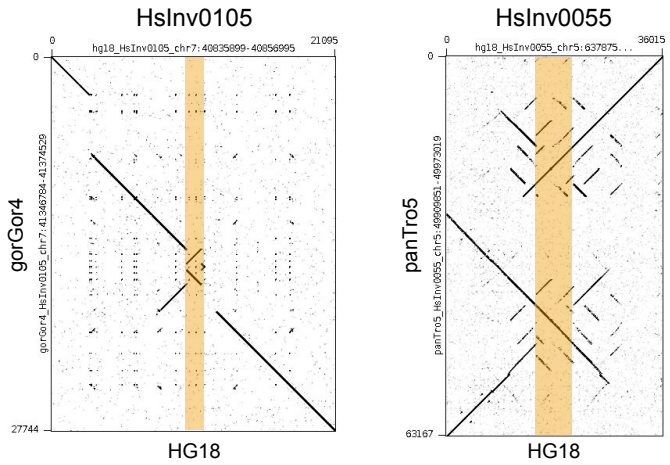


**Figure 3.1: Unclear dotplot alignments.** Alignments of inversions HsInv0105 and HsInv0055 between human reference genome HG18 (x-axis) gorilla assembly gorGor4 and chimpanzee assembly panTro5 (y-axis). Inversion region is highlighted in orange.

consistent orientations in all assemblies assessed, and were also consistent with published data. The only exception was inversion HsInv0105, where gorGor4 assembly seemed to have the same orientation as HG18, while all other assemblies had the alternative orientation (including gorGor5, from another sample of the same species). A dot plot alignment of the gorGor4 and HG18 assemblies revealed an inverted tandem duplication of the inversion region (Figure 3.1). Since the duplication is absent in gorGor5 assembly, it is likely an error of gorGor4 assembly. Figure A.1 shows the dot plot alignments of HG18 with the best-hit regions in the last assembly of each species.

For the 24 NAHR inversions, orientation in other assemblies proved more difficult to determine. On average, we obtained an orientation prediction for less than four assemblies out of the seven surveyed and most inversions (14) there is at least one assembly with a discordant orientation. In inversions HsInv0030 and HsInv0040 gorGor4 is the only outgroup assembly with the same orientation as HG18. Also, assembly gorGor4 failed at a higher rate than other assemblies in the rest of inversions. Together, it suggests that the quality of gorGor4 (Sanger and short-read-based assembly) is not high enough for the analysis and is biased towards human reference allele. Therefore, we excluded gorGor4 predictions from the ancestral orientation analysis. For two inversions (HsInv0396 and HsInv0790) orientation could not be predicted in any of the assemblies.

A consensus ancestral orientation was defined merging the information from all sources (ancestral row in Figure 3.2). Experimental genotypes were prioritised when assembly comparisons were unclear. Over half of NAHR inversions (14 out of 24) are polymorphic in chimpanzee or gorilla, have different orientations in different species or both (Figure 3.2). Such pattern indicates that they have probably inverted multiple times in primates. For instance, inversion HsInv0389, here polymorphic in chimpanzee, had been previously reported to be recurrent in mammals (Cáceres et al. 2007). Thus, those in-

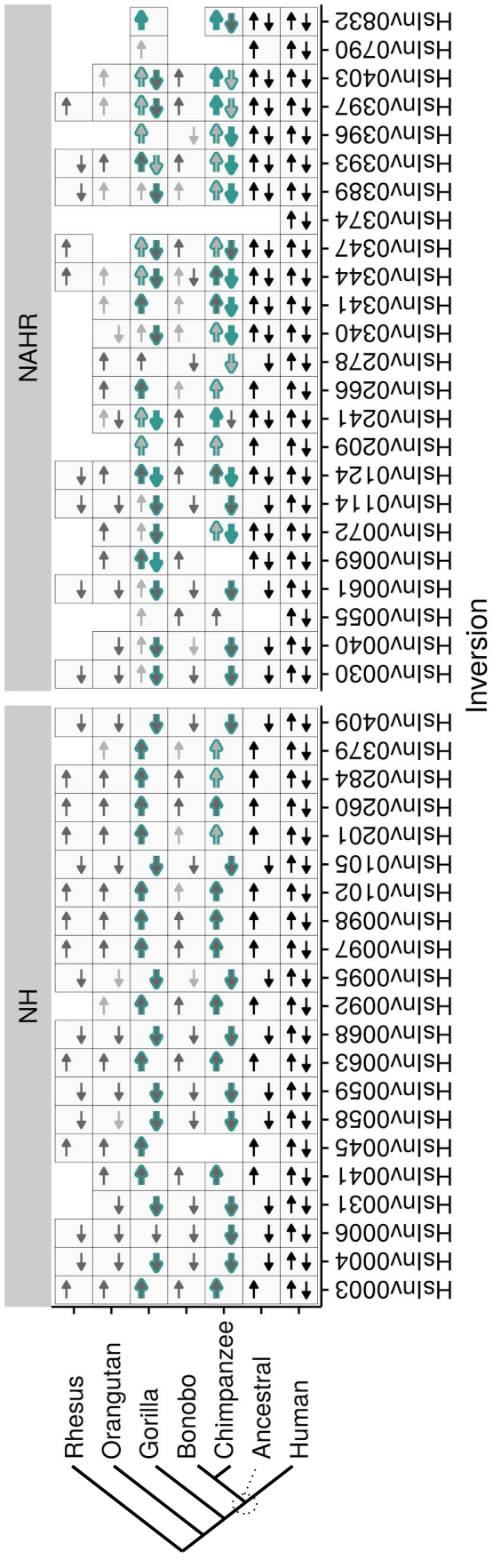


**Figure 3.2: Inversion orientation in different primates.** Right arrows represent orientation in HG18 and left arrows the alternative orientation. Green and light grey are the orientations defined in the literature and in the research group using experimental genotyping and assembly comparison, respectively. Dark grey are the orientations defined by the blat-based automated pipeline. When an orientation has been defined both experimentally and by assembly comparison, it is represented with a green-grey combined arrow. Black arrows represent final ancestral allele definition and the observed polymorphism in humans. HG18 orientation of inversion HsInv0241 was initially considered the ancestral and most analyses presented here assume so but it was later updated to ancestrally recurrent.

versions have been considered ancestrally recurrent and no ancestral status can be reliably established. No final ancestral orientation could be defined either for inversion HsInv0055 and HsInv0374. In both cases experimental genotypes were unavailable. For inversion HsInv0055, panTro5 and panPan1 were automatically predicted to have the HG18 orientation. However, the dotplot alignment showed an inverted duplication in panTro5, creating a complex structure (Figure 3.1). In total, we obtained a clear ancestral allele for 29 inversions, including all NH inversions. The reference genome HG18 carries the ancestral allele in 15 of them, and the derived allele in the remaining 14.

Additionally, HG18 sequences were aligned to newer HG19 and HG38 human reference assemblies to record possible changes in the orientation represented in human reference genomes. Most inversions keep the same allele represented in all thee human genome assemblies, with only two having changed. Inversion HsInv0403 is reversed in both HG19 and HG38 assemblies and inversion HsInv0072 in HG38 assembly. This is caused by the use of sequences of different clones in the newer assemblies.

Classically the ancestral or first known orientation of an inversion is called *standard* and the derived or newly discovered *inverted* (Krimbas and Powell 1992). For recurrent inversions, the allele nomenclature is sometimes assigned arbitrarily (Cáceres et al. 2007) or referring to the orientation in the reference genome and the alternative allele (regardless (Aguado et al. 2014). However, this usage of the word *inverted* can be misleading, leading to the incorrect assumption that the *inverted* orientation is the derived. Here, to avoid confusion, we use the words *reference* and *alternative* orientations, always using orientation in HG18 as reference. When the direction of the mutation is important, the words *ancestral* and *derived* orientations are used if know, or *major* and *minor* orientations otherwise (defined equally for all populations, considering the 480 unrelated samples).

#### 3.1.1.2 Functional effect of inversions

Knowing the potential functional effect of inversions can help interpreting their overall estimated effect on fitness and any possible signatures of selection. Inversions do not modify the total content of DNA but they can have a direct effect of functional elements at the breakpoints or within the inverted region. The effect of some of the inversions in the data set over genes had been already annotated (Aguado et al. 2014; Puig et al. 2015a; Vicente-Salvador et al. 2016), but missing for others. To complement and update the information found in the literature, we analysed the direct effect of each inversion in nearby protein coding genes, non-protein coding genes and pseudogenes as annotated in RefSeq.

We found that 18 out of 45 inversions affect directly or indirectly one or more genes (Table 3.1). Since most of the inversions in the data set are small compared to human genes, only six contain complete genes or pseudogenes that are inverted, and four of them have some additional effect from the described below. Seven inversions are completely included within an intron. In addition, within the inverted repeats of four NAHR inversions there are pairs of paralogous genes in which some of their sequence gets exchanged in the inverted allele. Finally, five inversions disrupt in some way a gene or pseudo-gene: two inversions remove an exon (by inverting it or from a deletion associated to the breakpoints) and three inversions separate some exons from the rest of a gene or pseudogene.

**Table 3.1: Functional effect of inversions**.

| Inversion | Gene effect |
|---|---|
| Inverts a gene | |
| HsInv0124 | Inverts protein coding gene *IFITM1* |
| HsInv0389 | Inverts protein coding genes *FLNA* and *EMD* |
| Intronic | |
| HsInv0006 | Within protein coding gene *DSTYK* intron |
| HsInv0059 | Within protein coding gene *GABRR1* intron |
| HsInv0061 | Within long intergenic non-protein coding RNA *LINC02532* intron |
| HsInv0098 | Within protein coding gene *ULK4* intron |
| HsInv0105 | Within protein coding gene *SUGCT* intron |
| HsInv0374 | Within pseudogene *LOC107133515* intron<br>Inverts pseudogene *SH3GL1P2* |
| HsInv0409 | Within protein coding gene *NLGN4X* intron |
| Exchanges genic sequence | |
| HsInv0030 | Exchanges sequences of protein coding genes *CTRB2* and *CTRB1* |
| HsInv0069 | Exchanges sequences of protein coding genes *FAM225B* and *FAM225A* |
| HsInv0241 | Exchanges sequences of protein coding genes *AQP12B* and *AQP12A* |
| HsInv0396 | Exchanges sequences of protein coding genes *PABPC1L2B* and *PABPC1L2A* and antisense RNA *PABPC1L2B-AS1* |
| Deletes or inverts an exon | |
| HsInv0102 | Inverts non-coding exon of protein coding gene *RHOH* |
| HsInv0201 | Associated deletion of coding exon of protein coding gene *SPINK14* |
| Breaks a gene | |
| HsInv0340 | Breaks long intergenic non-protein coding RNA *LINC00395*<br>Inverts pseudogene *OR7E156P* |
| HsInv0379 | Breaks protein coding gene *ZNF257*<br>Inverts protein coding genes *ZNF100*, *ZNF43*, *ZNF208*, and pseudogenes *LOC400682* and *LOC641367* |
| HsInv0790 | Breaks pseudogene *CCDC144B*<br>Inverts protein coding genes *TBC1D28*, *ZNF286B*, *TRIM16L*, *FBXW10* and *TVP23B*, and pseudogene *FOXO3B* |

The remaining 24 inversions do not affect directly any gene. In order to estimate their potential to affect regulatory regions, we measured their distance to the closest gene. Inversion HsInv0031 and HsInv0058 have genes within 10 kbp, in both cases the inversions are located downstream relative to the transcription direction. Within 20 kbp, there are genes near six

more inversions: HsInv0003, HsInv0004, HsInv0097, HsInv0114, HsInv0341 and HsInv0403. The other inversions have some gene within 100 kbp, except HsInv0040, HsInv0063, HsInv0260, HsInv0284 and HsInv0832, where the closest genes are more distant.

### 3.1.2 Inversion frequencies

#### 3.1.2.1 Frequency ascertainment bias

Inversions in the human genome are still poorly understood, so any insight obtained from this study may be used as a reference for future projects involving inversion polymorphisms. In order to be able to generalize any of the findings found here, it is important to know what fraction of the entire human inversion landscape we are having access to. Every study design has their own limitations. Therefore, the best way to extrapolate the observations is to clearly identify and, when possible, control the biases introduced.

To measure and control the effect of the study design in the observed frequency and frequency-related statistics, we simulated the detection and genotyping process in biallelic SNPs of the 1000GP phase 3 release (see Materials and Methods section 2.2.2 for model details) and measured the proportions of variants missed at different frequencies. We classified the potential frequency biases in three steps: small detection panel, limitations of the detection method, and inversion validation and inclusion criteria (as illustrated in Figure 3.3).

Briefly, in the first step, we simulated the detection of variants present in the small panel of two or nine individuals, depending on which study the inversion was detected, and measure the proportion of low-frequency variants in the population that are missed. In the second, we tried to control the additional biases introduced by the detection method. For inversions detected in the genome assembly comparison, limitations are assumed to be mostly linked to size and complexity of the genome, and therefore do not introduce any biases to the frequency distribution. Paired-end mapping detection is limited by the frequency in small inversions and by the size of inverted repeats at the breakpoints. This, we modelled that effect by using the inversion sizes observed in the selected inversion data set and the frequencies of SNPs present in the detection panel. Finally, biases introduced in third of inversion selection for experimental validation and genotyping were considered independent of inversion frequencies and not simulated.

The model allowed us to estimate the fraction of variants lost at each step with each detection method (Figure 3.4). We found that the detected in-

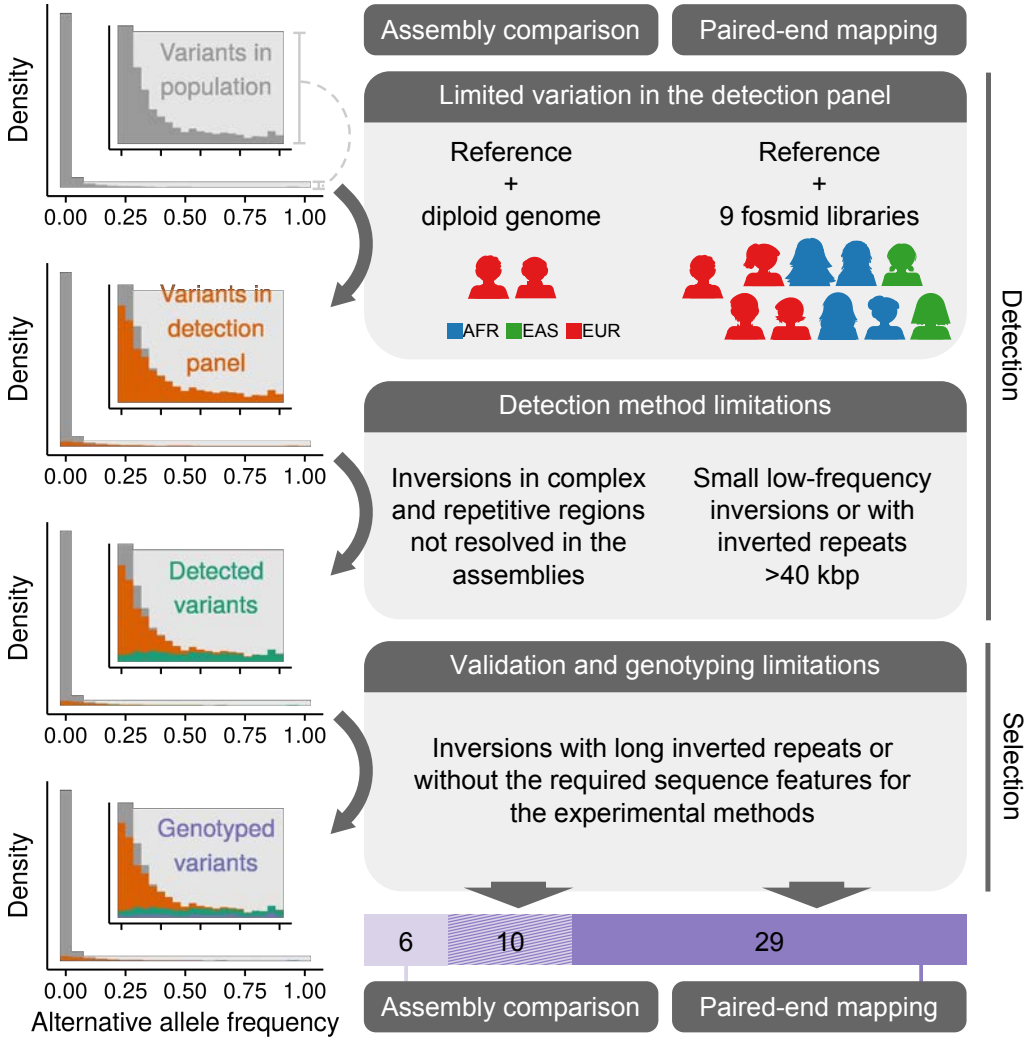


**Figure 3.3: Overview of the detection and selection of analysed inversions.** All steps limit the fraction of variants we can access. Left panels show the frequency distribution of alternative allele frequencies in African populations using random biallelic SNPs from 1000GP phase 3. The detection frequency bias shown is for an inversion of ∼1,000 bp with ∼250-pb inverted repeats at the breakpoints detected by paired-end mapping. Frequency panels from top to bottom: (1) variants present in the population; (2) fraction of variants that are carried by a sample of individuals used for the variant detection; (3) fraction of variants detected by the methods employed (fosmid paired-end mapping and genome assembly comparison); and (4) fraction of variants included in the genotyping assay. Shaded area in the main plot is enlarged in an inset plot to appreciate the proportions in less abundant frequency intervals.

versions are expected to have very high frequencies. For instance, according to the simulated SNPs, the average frequency of detected variants in Africa is expected to be 46% using the assembly comparison method, and 34% using the paired-end mapping method, while the average frequency of variants without any filtering in 1000GP African populations is only 3.6%.

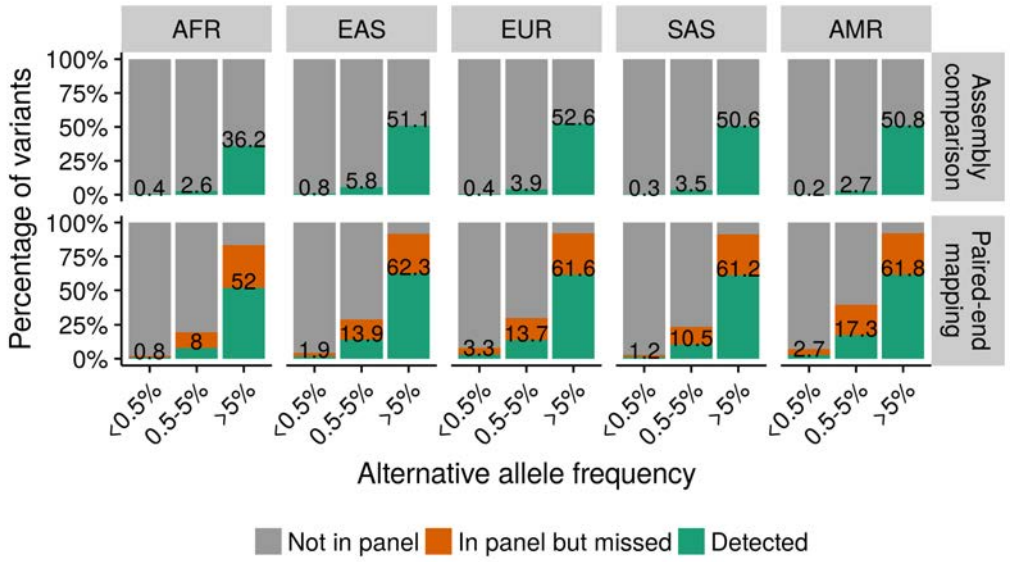


**Figure 3.4: Estimated proportion of variants missed in each step.** Proportions have been estimated through the simulation of the model described in the text using 1000GP random genome-wide SNPs. For the inversions detected by paired-end mapping, the detected SNPs match the probability of detection of each of the inversions included in the genotyping assay (1000 SNPs per inversion). For inversions from the assembly comparison, we have not modelled the fraction missed. However, the fraction missed could be less dependent on the variant frequency. The percentage of detected variants in each category is annotated on the green bars. Note that variants are not equally distributed in the three categories, more than half of 1000GP variants are in the lowest frequency bin (<0.5%). Super-population codes: populations of African (AFR), East Asian (EAS), European (EUR), South Asian (SAS) and American (AMR) ancestry.

The reason of these expected high frequencies is because most low-frequency variants in the population are missed in the detection process. The estimated fraction of variants missed at different frequency ranges in each of the five super-populations of 1000GP is shown in Figure 3.4. Specifically, at frequencies below 5%, we expect that at least 99% and 90% of the variants are missed in the assembly comparison and paired-end mapping studies, respectively, just because they are not present in the samples of the detection panel (first step). In contrast, nearly half of the variation at higher frequencies is represented in the assembly comparison panel and up to 90% in the paired-end mapping panel. Overall, since low-frequency variants outnumber high-frequency ones, the fraction of super-population variation missed is

87% and 70% for each method. The loss seems to be more intense in super-populations not represented in the detection panel. In the second step, we expect to detect 40% of the variants present in the paired-end mapping panel. And the loss is again more extreme in low-frequency variants: we lose more than half of variants below 5%, while just a third of the more frequent variants.

#### 3.1.2.2 Inversion frequency and geographical distribution

As expected from the described detection biases, inversions in the data set tend to be quite frequent. Half of the inversions are at more than 35% of frequency in some population and 40 inversions (out of 45) are present at least in one population with a frequency higher than 10% (Figure 3.5). Most inversions in the data set are spread worldwide and are polymorphic in all super-populations (Figure 3.5). Therefore, the data set is an important resource for human populations in all continents analysed. Only four inversions are polymorphic in a single super population (HsInv0097, HsInv0284 and HsInv0790 in Africa; HsInv0379 in East Asia), which coincides with those with lower frequencies. Eight more are absent or fixed in one super population: HsInv0003, HsInv0030, HsInv0068, HsInv0072 and HsInv0340 in East Asia; HsInv0061 in Africa; HsInv0209 in South Asia and HsInv0832 in Europe.

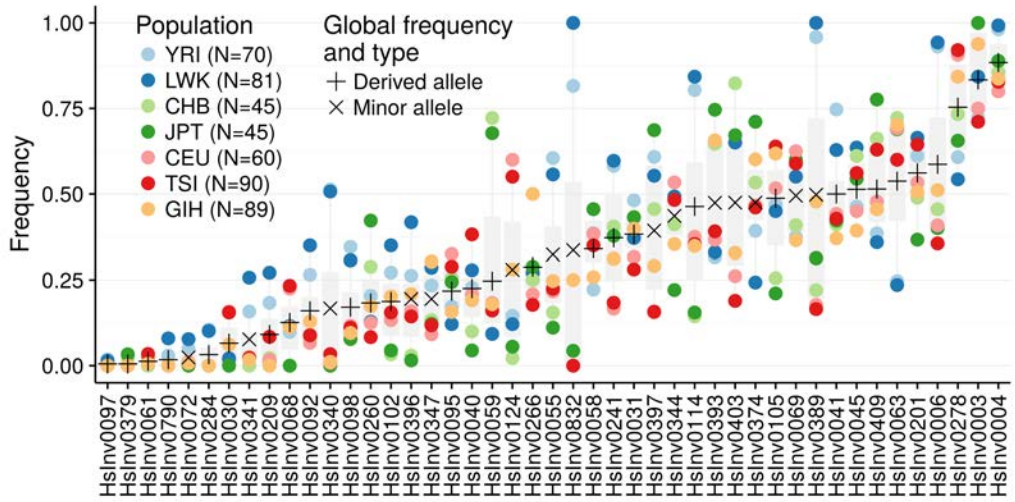


**Figure 3.5: Inversion frequency overview.** Inversion frequencies in the seven populations studied. Global frequency in the 480 unrelated individuals is indicated with a cross. Vertical crosses (+) indicate that the ancestral orientation is known and the frequencies represented are from the derived allele. Diagonal crosses (×) indicate that the ancestral orientation is unknown and the frequencies represented are from the global minor allele.

If we focus at individual level, we can see that all the genotyped samples

carry at least three inversions in heterozygosis and up to 20, with an average of 12. That translates into 3 kbp to 539 kbp of the genome having opposite orientations in pairs of homologous chromosomes (Figure 3.6, detailed in Figure 3.7). In addition, if we compare each genome to the reference HG18, there are on average 13.5 extra inversions with the alternative orientation in homozygosis or hemizygosis (Figure 3.6). The typical African individual has 96 kbp in heterozygosis and 117 kbp of alternative orientation in homozygosis or hemizygosis, coming from 13 and 15 inversions in the data set (median values). In a non-African individual, the typical number is slightly lower, 51 kbp in heterozygosis and 66 kbp with the alternative orientation in homozygosis or hemizygosis, spread in 11 and 13 inversions, reflecting the decreased genome diversity outside of Africa.

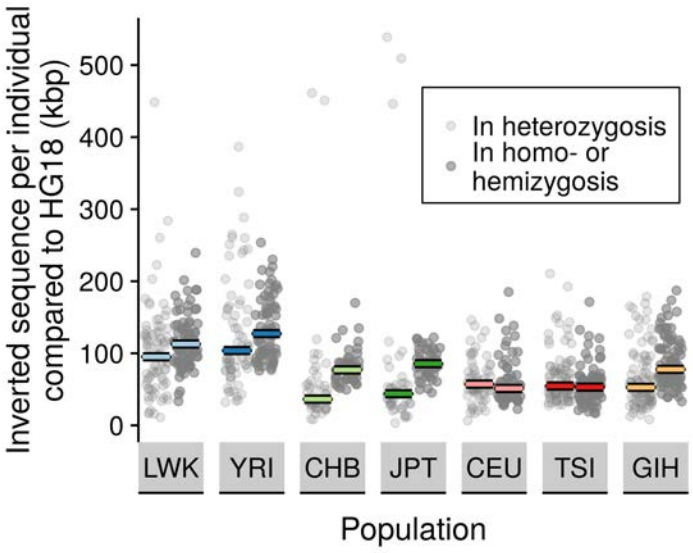


**Figure 3.6: Inverted sequence per individual compared to HG18**. Colour lines represent median values per population.

### 3.1.3 Comparison with 1000GP inversion data set

To assess the relative performance of the experimental method employed in the analysis against other strategies, we compared our 45 inversions with the 786 predicted and genotyped in the structural variant release of the 1000GP phase 3 (Sudmant et al. 2015). We found that only 14 (31%) of the inversions in our study were also detected by their methods and included in the final structural variation release, even though all the 45 inversions are polymorphic in the 1000GP data set (see Methods for comparison details). Of these, only two are NAHR inversions. This illustrates the limitations of the available inversion data, especially for those mediated by inverted repeats.

Since 434 individuals in the inversion-genotyping panel were also included

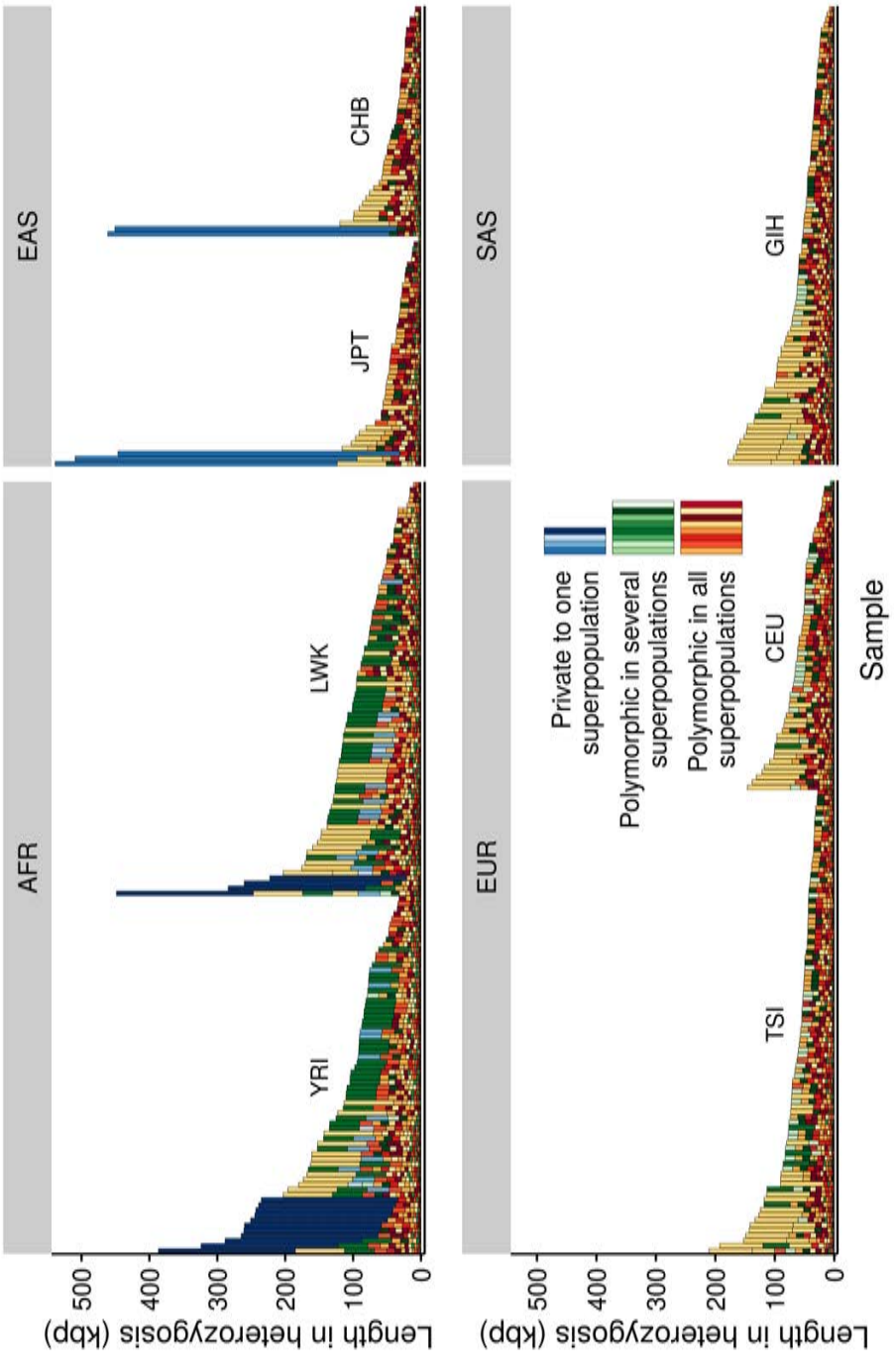


**Figure 3.7: Cumulative length of inversions in heterozygosis in the genotyped samples.** Each bar represents one individual and length of inversions in heterozygosis is represented as coloured rectangles. Main color indicates the distribution range of each inversion, and different shades are only to differentiate individual inversions.

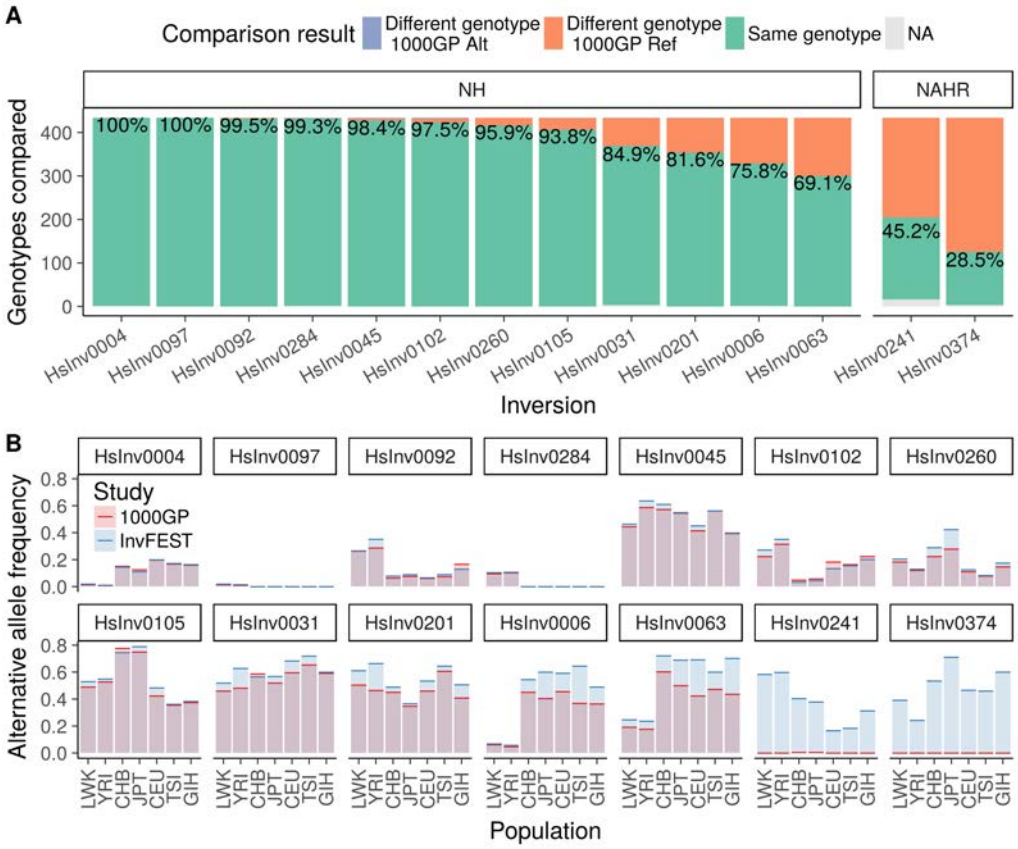


**Figure 3.8: Genotype and frequency comparison with 1000GP inversions.** A. Genotype disagreement between the 14 inversions in common with 1000GP structural variant release. Genotypes that differ between studies is further classified according to the allele that was assigned in 1000GP, in most of the cases the reference allele (in orange). B. Frequency estimates for the seven populations using all the unrelated individuals available in each study (480 in our study, InvFEST, and the full release in 1000GP). Some frequencies could vary slightly because of differences in the sample composition. Two inversion polymorphisms (HsInv0241 and HsInv0374) have very low frequency estimates in 1000GP predictions, while show high frequencies in our genotyping study.

in the 1000GP phase 3, we were able to further compare the genotypes in detail. For inversions created by non-homologous mechanisms, genotypes are the same in 91.31% of the comparisons (Figure 3.8 A). Only two inversions, HsInv0004 and HsInv0097, have a 100% genotype agreement for all individuals. The remaining inversions have lower values, ranging from 99.5% in inversion HsInv0092 to 69.1% in inversion HsInv0063. The observed errors are nearly always in favour of the orientation in the reference genome (Figure 3.8 A), and thus lead to underestimates of alternative allele frequency (Figure 3.8 B). The two inversions with inverted repeats at the breakpoints (HsInv0241 and HsInv0374) show much higher error rates and their frequencies have been largely underestimated, with all chromosomes in the 434 compared individuals with the alternative allele except one having been assigned the reference orientation. The overall low genotype agreement in these two inversions (45.2% and 28.5%) is therefore mostly due to the homozygous individuals with the reference orientation, which seems correctly genotyped just by chance.

We then looked into the accuracy of 1000GP breakpoint predictions. Interestingly, the four inversions with better genotype agreement have breakpoint predictions within or at less than 10 bp from our annotated breakpoint ranges. In contrast, the predicted size of the two NAHR inversions is noticeably smaller with 1000GP predicted breakpoints than those determined by us. Deletions and insertions at the breakpoints may also play an important role in their accuracy, because the alternative orientation of the three NH inversions with lower agreement (HsInv0063, HsInv0006 and HsInv0201) have 5216-bp, 80-bp and 1200-bp deleted or duplicated, respectively. However, inversion HsInv0097, with 100% of agreement has also a 1102-bp deletion in the alternative orientation, so other factors are probably involved. These results highlight the difficulty not only to detect complex inversions, but also to genotype them accurately from low-coverage short-read data. It also reveals a bias towards reference allele in the 1000GP data.

### 3.1.4 Inversion tag variants

Inversions can be very difficult to genotype even when their breakpoints have been fully characterized. For that reason, indirect genotyping methods that interrogate an associated variant (tag variant) are very convenient, especially when large sample sizes are required. Here, we took advantage of the high-quality inversion genotyping data together with 1000GP sequences for 434 individuals to identify variants capable of tagging inversion genotype.

We selected biallelic SNPs and indels within the inversion region or in the flanking region, located up to 2 Mpb from the breakpoints. Then, we estimated genotype correlation or linkage disequilibrium with the inversion

using the $r^2$ measure and kept those over a 0.8 threshold, considering it the lower limit for a variant to be used as inversion tag. All individuals in common with the 1000GP were included in order to estimate global level tag variants. Additionally, correlations were also performed using individuals of each super-population and population separately, to find variants that can be used to genotype the inversion in a specific region. All tag variants are listed in Table B.4 in Appendix B with their global $r^2$ value as well as for each population and super-population.

We found that 21 inversions have at least one perfect tag variant ($r^2 = 1$), with four inversions having only one, and up to 43 variants in inversion HsInv0040. Six more inversions are tagged by some variant at $r^2 > 0.8$ (Figure 3.9). The most obvious pattern is that NAHR inversions tend to have less tag variants, probably because they could be recurrent and multiple inversion events would carry different alleles from the nearby variants. This contrasts with NH inversions, in which all but two have perfect tag variants ($r^2 = 1$).

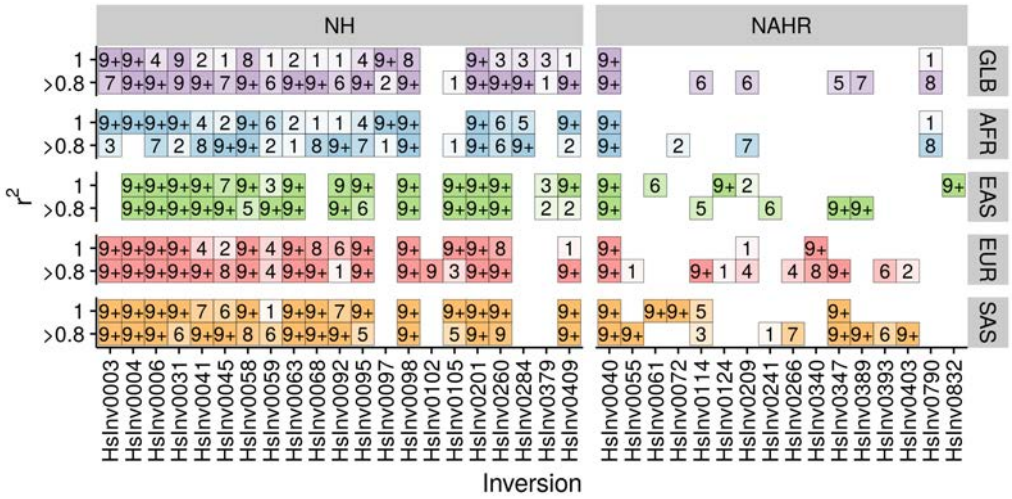


**Figure 3.9: Inversions with tag variants**. The number of tag variants with $r^2 = 1$ or between 0.8 and 1 ($r^2 > 0.8$) is indicated for each inversion, when considering the 434 individuals (GLB) and considering individuals in each super-populations. 9+ indicates that more than 9 tag variants. The nine inversions not present in the figure do not have any tag variant at any of the levels shown. Tag variants are only estimated in super-populations where the inversion is polymorphic.

Overall, the amount of inversions with some tag variant increases when focusing in individual populations or super-populations (Figure 3.9), possibly reflecting population-specific haplotypes associated with the inversion. From the two NH inversions without global perfect tag variants ($r^2 = 1$), inversion HsInv0105 has tag variants both at population and super-population level in all non-African groups, whereas inversion HsInv0102 has only population-

specific tag variants in East Asia and Europe. In the NAHR group, 11 out of the 22 inversions without perfect global tag variants have some at a population level (Figure 3.10 A) and five of them also at a super-population level. However, inversion genotypes could be inferred with more or less confidence in samples from populations with some specific tag variant. For example, inversions at low frequency in a population have a high number of SNPs in complete LD (e.g. HsInv0209 in CEU, HsInv0341 in JPT or HsInv0832 in LWK). In those cases, the detected associated variants are more likely to be spurious signals, and probably would not remain as tag variants in a larger sample.

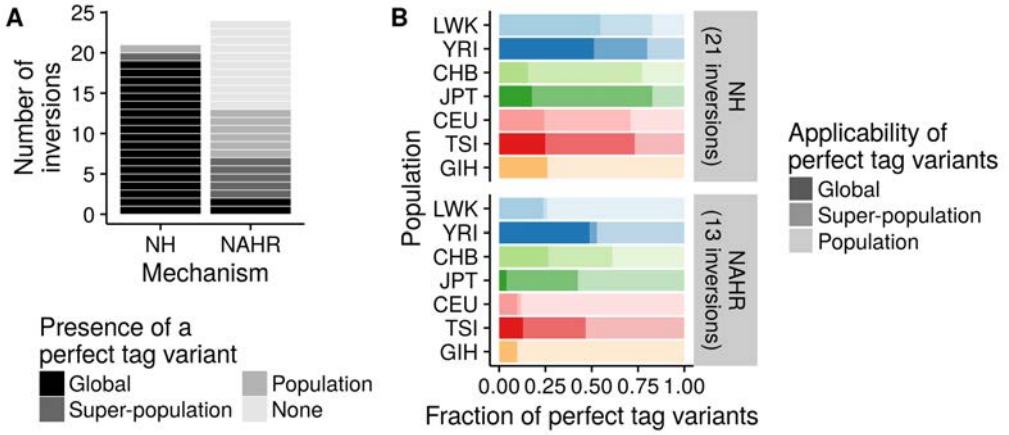


**Figure 3.10: Tag variants at population and super-population level**. A. Number of inversions with some perfect tag variant at a global, super-population and population scale or without perfect tag variants at any level. B. Scope of perfect tag variants associated with the corresponding tagged inversion in each population. Only inversions with some perfect tag variant are considered. Colour intensity represents scope or applicability of the variants (global, super-population or population specific).

#### 3.1.4.1 Tag variants in other populations

In some inversion studies, tag variants are detected in one experimentally-genotyped population and then used to infer the genotype in individuals from other populations. Thus, we next measured how often a perfect tag variant ($r^2$) found in one population could be extrapolated to another. Every perfect tag variant for each inversion found in a population was classified into one of the three categories: a tag variant the inversion only in that population (limited applicability), a tag variant also in all the individuals of the super-population or global tag variants, always tagging the inversion, irrespective of the population (wider applicability). Then, we obtained the fraction of each type of tag variant per population that represents how often tag variants in that population are going to be reliable in other populations (Figure 3.10

B).

As suspected, population tag variants of NH inversions are slightly more robust than those of NAHR inversions. On average, 78% of tag variants of NH inversions are applicable in populations within the same super-population, while only the 40% of tag variants of NAHR inversions applicable out of the same population. However, in most of the cases for all inversions, a tag variant found using one population is not going to be applicable to other continents.

There are also different patterns depending on the super-population the tag variant is initially found in. In NH inversions, tag variants associated in African populations have more than 50% chance of accurately tagging their inversion also in other super-populations, while in non-African populations the probability is always lower (between 16 and 26%). The explanation for that could be higher diversity and recombination levels found in Africa. Since there is a wider diversity, only variants found in all haplotypes within African populations will appear as perfect tag variants. And then it is likely that less diverse populations have maintained also those variants associated to the inverted orientation.

#### 3.1.4.2 Power to genotype inversions in common SNP arrays

Genome-wide association studies (GWAS), that detect the association of genetic variants with different phenotypes, need very large sample sizes to detect the subtle effects of individual variants. SNP arrays are still preferred for genotyping large cohorts, given that the cost is much lower and genotype qualities are higher than the obtained from low-coverage sequencing. In order to detect the effect of variants not included in the array, there is usually an imputation step, that uses a fully sequenced reference panel such as 1000GP to infer the genotypes of low frequency and not represented variants. However, genotypes from SNPs in the array are always more reliable than those imputed, so we were interested in knowing how many of the inversions in the data set could be indirectly genotyped with SNPs present in commonly used arrays.

To measure the power of commonly-used genotyping platforms to capture possible inversion effects, we systematically checked the presence of inversion global tag SNPs ($r^2 > 0.8$) in 76 commercial SNP arrays available in the LDLink web portal (Machiela and Chanock 2015). Inversions are on average covered in half of the SNP arrays analysed, and there is not any array that captures all the 26 inversions that have some tag SNP (Figure 3.11). The performance depends greatly on the array and two of them could not genotype any inversion. The best performing arrays assessed, HumanOmni5-

4v1 and HumanOmni5Exome-4v1, could detect up to 23 inversions, 85% of the inversions with some global tag SNP and 51% of all inversions in the data set. Of those, only seven inversions would be represented by perfect global tag SNPs ($r^2 = 1$), being the remaining 16 tagged by variants with lower $r^2$ values. Therefore, it is very likely that the effects of most inversions have been (and will continue to be) missed in studies using SNP arrays.

As an example, one of the most complete cohorts with genetic and phenotypic data is the UK Biobank (Sudlow et al. 2015), that recently completed the genotyping of more than 500,000 individuals with a SNP array of over 800,000 markers. In order to assess the inversion tag SNP coverage of the array, we downloaded the array variant list and compared it with our tag variant list. We found that 15 inversions could be detected with a global tag SNP in the array, although only four cases the tag SNP has a $r^2 = 1$. Therefore, we are missing 30 inversions: 19 simply do not have global tag SNPs and 11 more have global tag SNPs, but are not present in the array. If we assume that most of the cohort is of European ancestry, we could use European-specific tag SNPs. In that case, we would be able to genotype four more inversions. Thus, the genotype of most of inversions in the data set can not be inferred in the UK Biobank genetic data.


**Figure 3.11: Commercial array coverage of inversion tag SNPs.** The $r^2$ of the best global tag SNP in each array for each inversion is indicated by different colours. White means that the inversion does not have any tag SNP or that it is not present in the array.

## 3.2 Analysis of inversion frequencies

Population frequencies can be very informative about the role of natural selection on the evolution of mutations. A deleterious mutation will be usually kept at low frequencies and removed from the population. If a mutation is beneficial, frequency will increase in a short time, and the more advantageous, the fastest. Also, a mutation that is favourable at intermediate frequencies in the population, can be maintained by balancing selection. However, other factors may affect the frequency patterns too, such as demographic events, recurrent mutations and the study design. In this section, we take into account the different factors to understand inversion evolution and determine if natural selection is likely to be affecting globally inversion frequencies. Inversions with special frequency patterns are also highlighted as candidates to be under positive or balancing selection.

### 3.2.1 Inversion allele frequency spectrum

One first question was to determine if inversions in the data set are enriched in lower or higher frequencies than expected under neutrality. In order to control for the frequency bias introduced in the study design and the effect of past demographic events, we simulated the detection process (see Materials and Methods section 2.2.2 for details) in a large number of 1000GP phase 3 SNPs, and then estimated the frequency in the 434 unrelated individuals from 1000GP present in our study. We obtained 440,000 *detectable* SNPs in total, 10,000 SNPs matching each inversion-specific bias (excluding the inversion in chromosome Y). Highly conserved SNPs were discarded in the process, and therefore we assume most of the SNPs are neutral.

We considered that inversions generated by NAHR may be recurrent in humans, so the frequency distribution could be less comparable to that of neutral SNPs. For that reason, we separated the inversions by mechanism of formation. Also, since most NAHR inversions are ancestrally recurrent and the ancestral orientation cannot be determined, we compared the minor allele frequency, defined with all the 434 individuals. For NH inversions we were able to use derived allele frequencies.

First, we compared global frequency distributions of inversions and selected SNPs (Figure 3.12 A). The small number of inversions in each category (21 for NH and 24 for NAHR) makes the distributions noisy, but it is clear that inversions are not less frequent than the SNPs under the simulated detection model. Therefore, although power is suspected to be low for a common inversion sample such as this one, we do not find evidences of a distribution enriched in low frequencies that could derive from pervasive

negative selection in inversions.

Indeed, when compared with the SNP values, NH inversion distribution is found to be as expected according to the detection process (two-tailed Kolmogorov-Smirnov test, $D = 0.163$, $P = 0.64$), while NAHR inversion distribution may be different (two-tailed Kolmogorov-Smirnov test, $D = 0.28$, $P = 0.059$). We obtained similar results by comparing the observed mean frequencies of the inversion with the means of 10,000 samples of one matched SNP per inversion (see methods for an explanation about the sampling strategy). The mean derived allele frequency in NH inversions (0.36) is expected from the SNP distribution (empirical $P = 0.57$), but the mean minor allele frequency in NAHR inversions (0.27) is higher than expected (empirical $P = 0.006$), showing that the detection bias alone is unlikely to explain the observed distribution. However, what attracted our attention is that in both inversion classes there seems to be a clear excess of inversions with frequencies close to 0.5: 33.3% of NH inversions are in the frequency range 0.4-0.6, while expected proportion is 17.4%; and in the same folded frequency ranges there are 30.4% of NAHR inversion, while the expected proportion is 12.6%. An increased proportion of inversions at intermediate frequency could be expected if some inversions in the data set were under balancing selection, and an overall increase in frequencies if some inversions are under positive selection. For NAHR inversions, an alternative explanation would be a high bidirectional recurrence rate that brings some inversions at frequencies near 0.5.

In order to understand the origin of the excess of intermediate frequencies, we separated the four super-population frequency distributions (Figure 3.12 B). For NH inversions, the four distributions are mostly as predicted by the model, and the excess of intermediate frequencies is diluted and only visible in East Asian (EAS) super-population. And even in that case both the mean frequency and the distribution is well within expected values (two-tailed Kolmogorov-Smirnov test, $D = 0.140$, $P = 0.72$, and empirical mean difference, $P = 0.58$). Therefore, we cannot rule out that the excess of frequencies around 0.5 in the global distribution is just due to the small number of inversions and chance. In addition, we can conclude that the model largely captures the frequencies observed.

For NAHR inversions, the separated distributions make it clear that the model explains poorly their frequencies. For instance, inversion frequency distribution in Africa is largely different than that of the SNPs (two-tailed Kolmogorov-Smirnov test, $D = 0.378$, $P = 0.003$) and the mean frequency is significantly higher (empirical $P = 0.0008$). South Asia and European inversion frequency distributions show similar results, with less clear differences with respect to SNPs and but still clear increase in average frequency. The origin of these differences does not seem to come from an enriched 0.4-0.5 frequency range, but instead from a depleted 0 frequency category. That

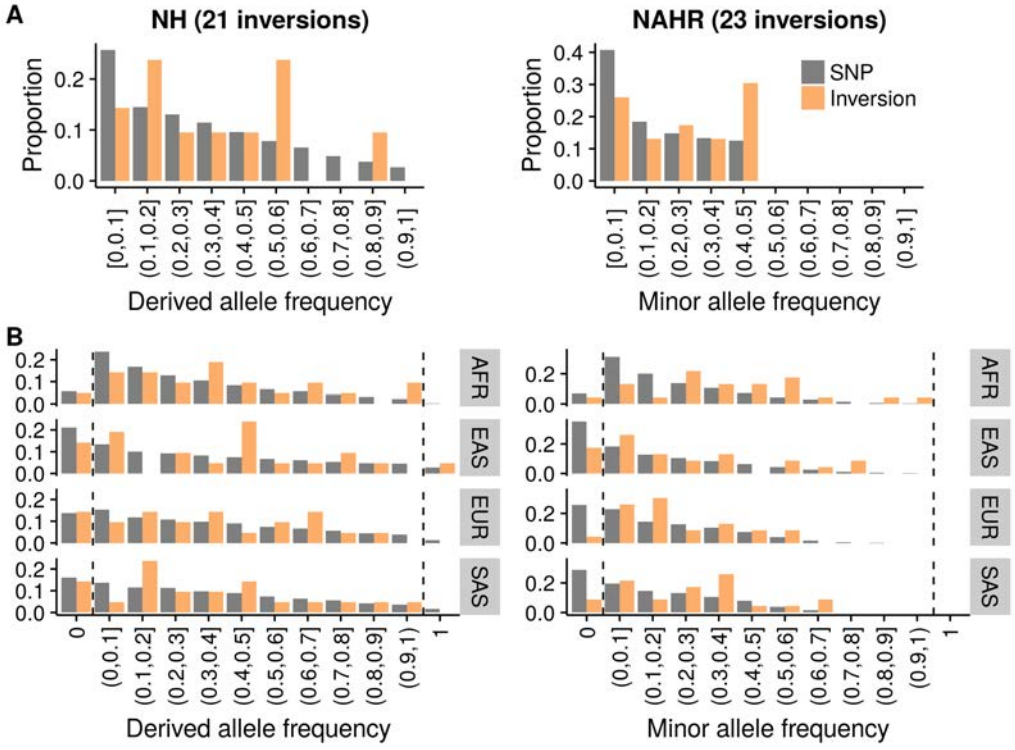


**Figure 3.12: Observed and expected frequency distribution of inversions.** A. Distribution of global allele frequencies in the 434 unrelated individuals in 1000GP phase 3. Frequencies of the derived allele are shown for NH inversions (all with known ancestral orientation), and of the minor allele for NAHR inversions (most are ancestrally recurrent). Expected distributions obtained from the simulation of the detection process in SNPs. B. The same frequency distributions separated by super-population. Dashed lines separate polymorphic frequency categories from the absent or fixed (frequency 0 or 1) categories for each super-population, that represent variants polymorphic in other super-populations. Note that minor alleles are defined globally and therefore they can be the most frequent allele in some super-population.

means that we would expect more NAHR inversions to be absent or fixed in some super-populations, so NAHR inversions are more cosmopolitan than predicted by the model. Although this pattern could fit well with the action of balancing selection in old polymorphisms, it is also consistent with some amount of recurrence.

### 3.2.2 Population differentiation

Populations with shared history are expected to have similar allele frequencies, and the differences between them increase with time and some demographic events (e.g. bottlenecks or migrations). Alleles under local selection can experience unusual frequency differences and therefore population differentiation is commonly used as a measure to detect selected variants (Vitti, Grossman, and Sabeti 2013). When a mutation is beneficial in one environment and it increases in frequency locally, the difference between populations may become unusually high. Conversely, if balancing selection acts on a mutation in several populations, it can show unusually similar frequencies across populations. So we next analysed frequency differences between populations to assess whether inversions show the expected differentiation patterns according to demography and to detect potential outliers.

Using the SNPs from the detection process simulation, we estimated the expected joint allele frequency spectra of pairs of populations and super-populations, represented as two-dimension distributions in Figure 3.13 A and B. Then, we compared frequency combinations observed in the inversion data set. Inversion frequency combinations fall within the ranges observed in the sampled SNPs and seem to follow similar correlation trends: frequencies within the same super-population (Figure 3.13 A) are more correlated than frequencies between super-populations (Figure 3.13 B). And related super-populations, such as European and South Asian, show stronger correlations than more distant ones, such as African and non-African super-populations (Figure 3.13 B).

In order to quantify the frequency differences between populations and super-populations, we calculated average population differentiation levels estimated by the $F_{ST}$ statistic in pairs of populations and super-populations, as well as globally between super-populations (table 3.2). When compared to the values in the SNP model, NH inversions show the expected populations differentiation levels (all comparisons empirical $P > 0.25$). Although NAHR inversions seem to have increased population differentiation levels in eight out of ten comparisons, none of the differences are significant and can be explained by the small sample size.

Even though average population differentiation levels are as expected, some

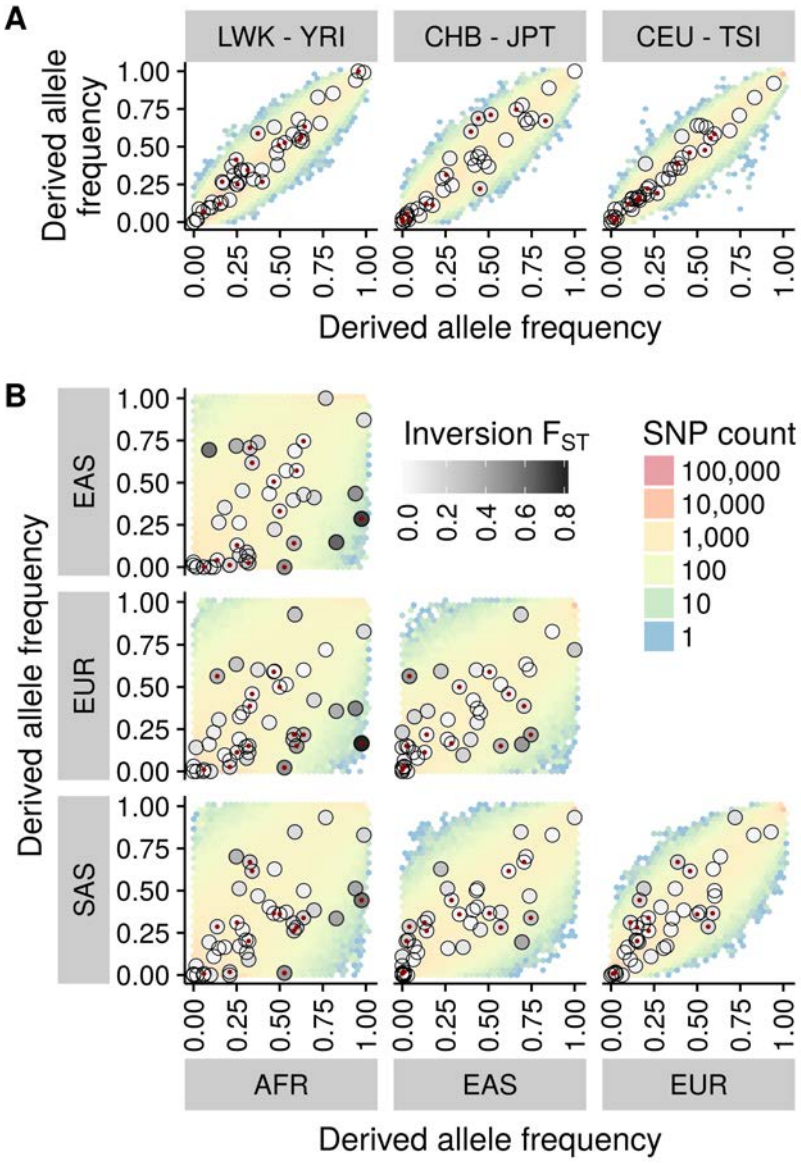


**Figure 3.13: Inversion population differentiation levels.** A. Inversion derived or minor allele frequency combinations for pairs of populations within the same super-population. First population in the x-axis and second in the y-axis. B. Inversion derived or minor allele frequency combinations for every pair of super-populations. Global minor allele frequencies are used for inversions with unknown ancestral orientation and are indicated by a red dot. Pairwise $F_{ST}$ values indicated in the circle shading. Joint allele frequency spectra of SNPs under the ascertainment model are shown as a two-dimension distribution in the background, with the colour representing the number of SNPs with each value.

**Table 3.2: Mean $F_{ST}$ values within and between super-populations**. Weir and Cockerham estimates are used (Weir and Cockerham 1984). None of the inversion $F_{ST}$ values is significantly different than that of SNPs under the same ascertainment bias.

| | | inversions (empirical $P$) | |
|---|---|---|---|
| Comparison/$F_{ST}$ | SNPs | NH | NAHR |
| *Within super-populations* | | | |
| LWK-YRI (AFR) | 0.010 | 0.008 (0.651) | 0.012 (0.665) |
| CHB-JPT (EAS) | 0.012 | 0.005 (0.267) | 0.022 (0.189) |
| CEU-TSI (EUR) | 0.007 | 0.006 (0.760) | 0.007 (0.935) |
| *Between super-populations* | | | |
| AFR-EAS | 0.132 | 0.151 (0.596) | 0.151 (0.510) |
| AFR-EUR | 0.116 | 0.092 (0.430) | 0.181 (0.068) |
| AFR-SAS | 0.108 | 0.086 (0.443) | 0.143 (0.245) |
| EAS-EUR | 0.088 | 0.082 (0.894) | 0.102 (0.586) |
| EAS-SAS | 0.064 | 0.069 (0.789) | 0.061 (0.988) |
| EUR-SAS | 0.032 | 0.021 (0.411) | 0.056 (0.111) |
| Global | 0.107 | 0.092 (0.551) | 0.150 (0.115) |

specific inversions may have unusual frequency patterns. And indeed, some inversions seem to have frequency combinations rare according to the SNP joint allele frequency spectra, as shown in inversion circles near the edges of the distribution in Figure 3.13. So we further explored that possibility. Individual $F_{ST}$ values were compared against the SNP distribution of the same population or super-population comparison and from SNPs in the same chromosome type (autosome or chromosome X).

A total of 15 inversions were found to have at least one $F_{ST}$ value within the highest 5% of the corresponding distribution, and some of them have also values within the highest 1% (Table 3.3). In four of them African populations have the most differentiated frequencies, with high $F_{ST}$ values in all or nearly all combinations involving African super-population. In particular, the derived orientation of inversions HsInv0006 and HsInv0114 shows higher frequency in Africa than in non-African populations. For inversions HsInv0340 and HsInv0389 African populations have higher frequency of the minor allele, but considering that the ancestral orientation is unknown, the increase in frequency could have been in the non-African populations instead. Inversion HsInv0059 has a similar pattern but in East Asia. The derived orientation is in much higher frequency in CHB and JPT populations than the rest, with the difference between South and East Asia being the most unusual (in the top 0.7% of the distribution). Four inversions more have high $F_{ST}$ between two non-African super-populations. In inversion HsInv0124 the clear outliers with higher frequencies are European populations, and in inver-

sion HsInv0266, South Asian populations (although the ancestral orientation is unknown in both). Finally, six more inversions have high differentiation between populations in the same super-population. The most extreme is inversion HsInv0003 that has a frequency of 0.2 in CEU and nearly double in TSI (0.39).

**Table 3.3: Inversions with high $F_{ST}$ values.**

| Inversion | Class | Top 1% $F_{ST}$ | Top 5% $F_{ST}$ |
|---|---|---|---|
| *African* | | | |
| HsInv0006 | NH | - | Global, AFR-EAS, AFR-EUR, AFR-SAS |
| HsInv0114 | NAHR | - | Global, AFR-EAS, AFR-SAS |
| HsInv0340 | NAHR | Global | AFR-EAS, AFR-EUR, AFR-SAS |
| HsInv0389 | NAHR | AFR-EUR | Global, AFR-EAS, AFR-SAS |
| *East Asian* | | | |
| HsInv0059 | NH | EAS-SAS | Global, AFR-EAS, EAS-EUR |
| *Between non-African super-populations* | | | |
| HsInv0003 | NH | - | EUR-SAS |
| HsInv0105 | NH | - | EAS-SAS |
| HsInv0124 | NAHR | - | EAS-EUR, EUR-SAS |
| HsInv0266 | NAHR | EUR-SAS | - |
| *Within super-populations* | | | |
| HsInv0040 | NAHR | CEU-TSI (EUR) | - |
| HsInv0045 | NH | - | CEU-TSI (EUR) |
| HsInv0069 | NAHR | - | LWK-YRI (AFR), CHB-JPT (EAS) |
| HsInv0344 | NAHR | - | CHB-JPT (EAS) |
| HsInv0374 | NAHR | - | CHB-JPT (EAS) |
| HsInv0397 | NAHR | - | CHB-JPT (EAS) |

While unusually high differentiation is relatively easy to detect, the abundance of low $F_{ST}$ values makes more difficult to detect possible inversions maintained at similar frequencies by balancing selection. Around 10% of the SNPs show no or very little differentiation between African and East Asian populations, and around 60% between CEU and TSI. Therefore, there are no unusually low differentiation values. However, we were interested in highlighting those inversions that are little differentiated among all the super-populations and at intermediate frequencies, as candidates to be under long-term balancing selection (if any). Using an arbitrary threshold of within the lower 20% of the global differentiation distribution and a global frequency higher than 0.3, we selected three inversions: HsInv0031, HsInv0045 and HsInv0058 (table 3.4). The three candidates have a global $F_{ST}$ around 0.02 and population frequencies ranging from 0.22 (HsInv0058 in LWK) to 0.63 (HsInv0045 in YRI). Relaxing the criteria to those inversions within the 25% of the lower tail of the distribution, we would also include inversions HsInv0069, HsInv0201 and HsInv0344, with global $F_{ST}$ values up to 0.03. The rest of inversions have $F_{ST}$ values over 0.06, with higher variation in frequency among populations.

**Table 3.4: Inversions with low global $F_{ST}$ and intermediate frequencies**.

| | | | Derived/*minor* allele frequency | | | |
|---|---|---|---|---|---|---|
| Inversion | Class | $F_{ST}$ | AFR | EAS | EUR | SAS |
| HsInv0031 | NH | 0.0198 | 0.439 | 0.435 | 0.289 | 0.402 |
| HsInv0045 | NH | 0.0199 | 0.538 | 0.571 | 0.515 | 0.372 |
| HsInv0058 | NH | 0.0221 | 0.284 | 0.453 | 0.347 | 0.268 |
| HsInv0069 | NAHR | 0.0290 | *0.456* | *0.506* | *0.603* | *0.365* |
| HsInv0201 | NH | 0.0303 | 0.636 | 0.428 | 0.600 | 0.506 |
| HsInv0344 | NAHR | 0.0264 | *0.490* | *0.318* | *0.503* | *0.354* |

## 3.3 Effect on local nucleotide diversity

The inhibition of recombination between orientations can have different effects on the nucleotide diversity within the inverted region. The effect is likely to depend on several factors, including inversion frequency, age and population demographic history. We first used simple simulations to obtain an expectation of the effect of the special recombination patterns, and then compared them with the real values for the two classes of inversions. Finally, we explored the effect of altered nucleotide diversity levels on statistics used for neutrality tests, such as Tajima's D.

### 3.3.1 Inversion simulation

In order to investigate the relationship between the different factors mentioned above and the expected variation, we simulated neutral segregating inversions with human levels of genomic variation and recombination in a constant sized population. We used `InvertFREGENE`, a forward-in-time simulation software that is adapted to recreate the particular recombination patterns of inversion polymorphisms (O'Reilly, Coin, and Hoggart 2010). In `InvertFREGENE` model, recombination can only take place through crossing overs and double recombination events are not allowed either.

Ten thousand simulations of a 300-kbp inversion were enough to obtain a wide range of neutral trajectories covering all frequencies and with ages up to 100,000 generations (2.5 million years old, assuming 25-year generation time). We then measured the nucleotide diversity levels present in each simulation, considering the global diversity and the private diversity of each chromosome orientation, as well as the sequence divergence between orientations. The same number of simulations were run simulating a 1-bp inversion centred in the middle of the 300-kbp region, representing a reference scenario of free recombination between orientations. Simulation parameters were chosen to resemble human mutation rates and recombination properties. Since inversions are simulated under neutrality, in most of the cases they are young and at low frequency, but we obtained a representation of different combinations of frequency and age. Figure 3.14 A shows nucleotide diversity measures between sequence types ($\pi$), that represents average pairwise differences, in the inversion region for different final frequencies and age ranges.

Observations are the following. Total nucleotide diversity increases with the frequency of the inversion, with a peak around frequency 0.5 (Total diversity panel, Figure 3.14 A). The magnitude of the increase is related to the age of the inversion, with older inversions having larger effect. Nearly-fixed

inversions cause a reduction of the total nucleotide diversity. The nucleotide diversity within sequences with ancestral orientation is similar to the free recombination case, negatively correlated with frequency, although inversion presence slightly reduces diversity at all frequencies with a stronger effect at high frequencies (Anc diversity panel, Figure 3.14 A). Nucleotide diversity in inverted sequences is generally lower than it would be with free recombination, although in both cases diversity is positively correlated with frequency (Der diversity panel, Figure 3.14 A). Also, diversity is more severely reduced for young inversions. Finally, divergence between orientation increases with age and it is mostly independent of inversion frequency (Divergence panel, Figure 3.14 A). In summary, inversions of all frequencies and ages affect some component of the local nucleotide diversity, with the strongest effect seen in old inversions.

### 3.3.2 Observed values

With a clear model of the effect of recombination inhibition in inversion regions, we next wanted to measure the effect of real inversions and its consistency with the simulation-based model. Nucleotide diversity levels were estimated both inside for those inversions with enough variants in the internal region and also in the flanking regions up to 2 Mbp from the breakpoints. Three inversions with high global frequencies did not have enough variants and were excluded (HsInv0003, HsInv0006 and HsInv0278), making the high-frequency range poorly represented. Ancestral/major and derived/minor nucleotide diversity, and divergence between them were estimated only from homozygote or hemizygote individuals to avoid the potential phase errors. Here we assume the major is the ancestral for inversions with unknown ancestral orientation. Four inversions do not have homozygotes or hemizygotes for the inverted orientation and no diversity measures could be calculated (Hsnv0061, HsInv0097, HsInv0379, HsInv0790). To account for local differences in basal diversity levels between regions, inside values were divided by the mean observed value in the flanking regions using the same individuals.

#### 3.3.2.1 NH inversions

The 17 NH inversions analysed match the general trends observed in simulated data. Although the diversity units are not directly comparable, the magnitude of the differences are higher in real inversions than in the simulations, perhaps because of an increase in the variance due to the real complex demography and the smaller inversion size. Average relative change in diversity are summarised in table 3.5, divided in inversions with global low frequency ($< 0.25$) and intermediate frequency (0.25 to 0.75). Relative vari-

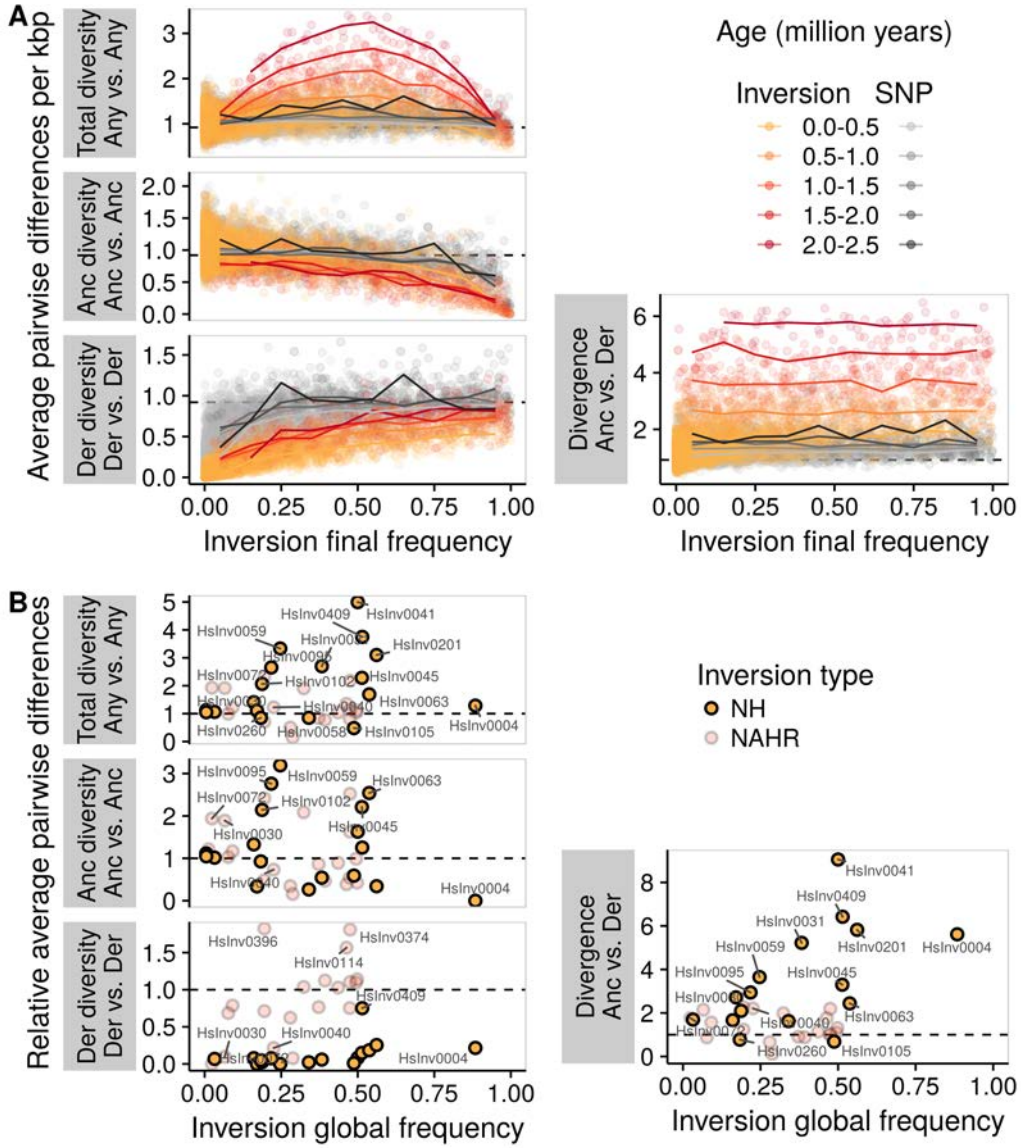


**Figure 3.14: Effect of inversions on nucleotide diversity levels**. A. `InvertFREGENE` simulations. Each point represents the average pairwise differences between different types of chromosomes (Anc: ancestral orientation, Der: derived orientation) of a 300-kbp simulation. Solid lines show mean values in bins of 0.1 frequency ranges, grouped by age category. Dashed lines indicate the approximate equilibrium diversity levels in the simulations. Orange values represent simulations with inhibition of recombination between orientations (inversions) and grey values equivalent simulations of 1-bp mutation in the center that does not restrict recombination (SNPs). B. Real inversions. Average pairwise differences between types of chromosomes relative to the mean values observed in the flanking regions (2 Mbp). Recurrent NAHR inversions are expected to follow different patterns from the ones simulated above and are shown as background lighter points.

ation values for each inversion are shown in Figure 3.14 B and absolute values in table 3.6. Results for each measure are explained in detail below.

**Total nucleotide diversity.** On average, total diversity is higher inside the inversion than in the flanking areas, with a 63% and 148% increase relative to the flanking regions for low and intermediate frequency inversions, respectively. Thus, the inversion effect is stronger at intermediate frequencies, as predicted. There are three inversions with a lower diversity inside than in the flanking regions: HsInv0058, an inversion located in a highly variable region, and HsInv0105 and HsInv0260, small inversions with limited information inside (they span 1092 and 1831 bp, respectively). Inversion with the highest increase is HsInv0041, with a global frequency of 0.5 and population frequencies ranging between 0.37 and 0.74.

**Anc nucleotide diversity.** Ancestral chromosomes show an average diversity increase of 54% and 17% for low and intermediate global frequencies, although individual values are very variable. Ancestral chromosomes of the most frequent inversion, HsInv0004, show no variation, consistent with the strong reduction of diversity observed in high-frequency inversions in the simulations.

**Der nucleotide diversity.** All inversions have a reduction of diversity levels within the inverted chromosomes. The reduction is stronger for low frequency inversions, which on average only have a 4% of the variation observed in the flanking regions. Inversion HsInv0409, with a global frequency of 0.52, shows the mildest reduction, having 75% of the variation outside the inversion. It is the only NH inversion in chromosome X and perhaps it is related to the lowest basal nucleotide diversity found in that chromosome.

**Divergence between orientations.** Most NH inversions show some net divergence between orientations with respect to the flanking regions. The only exceptions being the small inversions HsInv0105 and HsInv0260, that also have reduced total diversity. Frequency and divergence are positively correlated (Pearson's $r = 0.55$, $P = 0.027$), which is expected from the correlation between frequency and age in neutral variants.

#### 3.3.2.2 NAHR inversions

Overall, nucleotide diversity measures for NAHR inversions are more similar to the flanking regions, especially in intermediate frequency inversions (table 3.5). Nevertheless, the effect of NAHR inversions is more heterogeneous

**Table 3.5: Average change in nucleotide diversity in inversions**. Average difference in variation levels within the inversion region relative to the 2 Mbp flanking regions, using the same individuals. Inversions are divided by class and global frequency range. Only one inversion is in the frequency range 0.75 - 1 and it is not included in the summary table (see table 3.6 for full values).

| Class | Frequency | Total | Anc/Maj | Der/Min | Divergence |
|---|---|---|---|---|---|
| NH | 0.00 - 0.25 | 1.63 | 1.54 | 0.04 | 2.22 |
| | 0.25 - 0.75 | 2.48 | 1.17 | 0.19 | 4.32 |
| NAHR | 0.00 - 0.25 | 1.44 | 1.36 | 0.62 | 1.72 |
| | 0.25 - 0.75 | 1.07 | 0.94 | 1.01 | 1.18 |

(Figure 3.14 B). Patterns of some inversions, such as HsInv0030, HsInv0040 and HsInv0072, are more similar to those of NH inversions: strong reduction of diversity in Inv chromosomes and moderate increase in divergence between orientations. Other inversions show high diversity levels within derived/minor chromosomes (HsInv0114, HsInv0374, HsInv0396). HsInv0374 and HsInv0396 have increased diversity levels in all measures, that could indicate higher genotyping error rates or just an increased mutation rate. However, HsInv0114 has reduced diversity within chromosomes considered ancestral, what may indicate that the other orientation is actually the ancestral in the human lineage. Although NAHR inversions should also inhibit recombination in heterozygotes, their recurrence nature may soften the strong population structure observed in NH inversions. As a result, no clear patterns are seen (Figure 3.14 B).

### 3.3.3 Neutrality tests

An important class of neutrality tests are based on the altered frequency distribution of the nucleotide diversity in a region of interest. So next we explored the impact of the inversion on the statistics in order to determine if they can be informative for detecting inversions under selection. The most well known statistic is Tajima's D (Tajima 1989), that compares different estimates of the population-scaled mutation rate $\theta = 4N_e\mu$. Under neutrality and ideal conditions, total nucleotide diversity estimated as average pairwise differences between sequences, $\pi$, should be an unbiased estimator of $\theta$. And the same would hold for the Watterson estimator $\hat{\theta}_w$, that is based only on the number of segregating sites. Tajima's D measures the difference between the two estimators. An excess of variants at low frequencies makes $\hat{\theta}_w$ larger than $\pi$ and can indicate positive selection (selective sweep signature). An excess of variants at intermediate frequencies makes $\hat{\theta}_w$ less than $\pi$ and can indicate balancing selection. The first case results in a negative value of Tajima's D and the second in a positive one.

**Table 3.6: Nucleotide diversity measures for inversions**. Nucleotide diversity is represented as average pairwise differences per kbp in the inverted region. In parenthesis, average values from the 2 Mbp flanking regions using the same individuals. Global frequency of the derived allele is shown when ancestral orientation is known, and minor allele otherwise.

| Inversion | Class | Freq. | Total | Anc/Maj | Der/Min | Divergence |
|---|---|---|---|---|---|---|
| HsInv0004 | NH | 0.88 | 0.99 (0.77) | 0 (0.56) | 0.17 (0.78) | 4.27 (0.76) |
| HsInv0031 | NH | 0.38 | 4.43 (1.64) | 0.87 (1.61) | 0.1 (1.69) | 8.66 (1.66) |
| HsInv0041 | NH | 0.50 | 4.61 (0.92) | 1.32 (0.81) | 0.08 (0.98) | 8.6 (0.95) |
| HsInv0045 | NH | 0.51 | 2.41 (1.05) | 2.31 (1.05) | 0.16 (1.05) | 3.52 (1.07) |
| HsInv0058 | NH | 0.34 | 2.67 (3.14) | 0.83 (3.11) | 0.07 (3.03) | 5.18 (3.19) |
| HsInv0059 | NH | 0.25 | 2.77 (0.83) | 2.75 (0.86) | 0 (0.63) | 2.97 (0.81) |
| HsInv0063 | NH | 0.54 | 1.46 (0.87) | 2.42 (0.95) | 0.14 (0.76) | 2.23 (0.91) |
| HsInv0092 | NH | 0.16 | 1.32 (0.94) | 1.19 (0.9) | 0.08 (1.03) | 1.76 (1.05) |
| HsInv0095 | NH | 0.22 | 2.31 (0.87) | 2.45 (0.89) | 0.06 (0.79) | 2.55 (0.87) |
| HsInv0097 | NH | 0.01 | 0.98 (0.88) | 0.98 (0.87) | - | - |
| HsInv0098 | NH | 0.17 | 1.01 (0.91) | 0.29 (0.86) | 0 (1.01) | 2.84 (1.05) |
| HsInv0102 | NH | 0.19 | 2.08 (1.01) | 2.09 (0.97) | 0.03 (1.07) | 2.3 (1.1) |
| HsInv0105 | NH | 0.49 | 0.39 (0.81) | 0.47 (0.79) | 0.01 (0.78) | 0.55 (0.8) |
| HsInv0201 | NH | 0.56 | 2.82 (0.91) | 0.3 (0.87) | 0.23 (0.91) | 5.48 (0.94) |
| HsInv0260 | NH | 0.18 | 0.7 (0.82) | 0.76 (0.82) | 0.03 (0.67) | 0.62 (0.8) |
| HsInv0284 | NH | 0.03 | 0.79 (0.74) | 0.74 (0.73) | 0.05 (0.75) | 1.47 (0.87) |
| HsInv0379 | NH | 0.01 | 1.34 (1.28) | 1.34 (1.28) | - | - |
| HsInv0409 | NH | 0.51 | 2.82 (0.75) | 1.01 (0.8) | 0.53 (0.7) | 4.94 (0.77) |
| HsInv0030 | NAHR | 0.07 | 2.06 (1.07) | 2.04 (1.08) | 0.09 (0.93) | 2.25 (1.05) |
| HsInv0040 | NAHR | 0.23 | 1.16 (0.94) | 0.68 (0.93) | 0.2 (0.93) | 2.14 (0.97) |
| HsInv0055 | NAHR | 0.32 | 1.54 (0.81) | 1.53 (0.73) | 0.91 (0.88) | 1.74 (0.87) |
| HsInv0061 | NAHR | 0.01 | 0.97 (0.81) | 0.99 (0.82) | - | - |
| HsInv0069 | NAHR | 0.49 | 0.97 (0.94) | 0.93 (0.94) | 1.02 (0.93) | 1.03 (0.94) |
| HsInv0072 | NAHR | 0.02 | 1 (0.52) | 1 (0.51) | 0 (0.56) | 1.09 (0.62) |
| HsInv0114 | NAHR | 0.46 | 0.94 (0.69) | 0.23 (0.58) | 1.24 (0.79) | 1.33 (0.75) |
| HsInv0124 | NAHR | 0.28 | 0.63 (1.24) | 0.43 (1.24) | 0.73 (1.17) | 0.84 (1.26) |
| HsInv0209 | NAHR | 0.09 | 1.11 (0.91) | 1.02 (0.87) | 0.87 (1.11) | 1.64 (1.05) |
| HsInv0241 | NAHR | 0.37 | 1.04 (1.23) | 1.02 (1.19) | 0.98 (1.28) | 1.15 (1.26) |
| HsInv0266 | NAHR | 0.29 | 0.21 (1.42) | 0.22 (1.42) | 0.1 (1.34) | 0.17 (1.39) |
| HsInv0341 | NAHR | 0.08 | 0.89 (0.89) | 0.9 (0.87) | 0.64 (0.93) | 0.85 (0.96) |
| HsInv0344 | NAHR | 0.44 | 1.01 (0.97) | 0.84 (0.95) | 1.02 (0.99) | 1.12 (0.98) |
| HsInv0347 | NAHR | 0.19 | 0.6 (0.85) | 0.39 (0.81) | 0.6 (0.84) | 1.12 (0.91) |
| HsInv0374 | NAHR | 0.47 | 1.41 (0.65) | 1.72 (0.68) | 1.06 (0.58) | 1.44 (0.66) |
| HsInv0389 | NAHR | 0.50 | 0.53 (0.48) | 0.15 (0.34) | 0.63 (0.55) | 0.69 (0.53) |
| HsInv0393 | NAHR | 0.47 | 0.41 (0.51) | 0.3 (0.54) | 0.32 (0.43) | 0.52 (0.53) |
| HsInv0396 | NAHR | 0.19 | 0.93 (0.4) | 0.93 (0.38) | 0.77 (0.42) | 0.96 (0.42) |
| HsInv0397 | NAHR | 0.39 | 0.28 (0.36) | 0.14 (0.3) | 0.48 (0.42) | 0.34 (0.39) |
| HsInv0403 | NAHR | 0.47 | 0.7 (0.61) | 0.63 (0.39) | 0.67 (0.61) | 0.75 (0.71) |

We estimated Tajima's D statistic in the previous neutral inversion simulations (Figure 3.15 A). Simulations show that old inversions increase Tajima's D statistic, giving a signal expected under balancing selection or population structure, especially for intermediate frequencies. And inversions at high frequencies can also produce negative Tajima's D values, that could be interpreted as positive selection. Therefore, inversions can create false positives in neutrality tests based on frequency distribution of variants. Nevertheless, their effect could be potentially corrected by controlling inversion frequency and age.

We then estimated Tajima's D for real inversions from the total nucleotide diversity and segregating sites in order to see if observations follow the simulations and if there are some unexpected high or low values. Since population structure can have a strong effect on Tajima's D, we estimated it independently for each population. As before, NH inversions recapitulate general trends observed in the simulations (Figure 3.15 B). Inversions at intermediate frequencies have the highest values of the statistic. This is the case of three inversions: HsInv0031, with values higher than two in LWK, East Asian populations and GIH; HsInv0102 in all populations; and HsInv0409 in Africa, Europe and South Asia. Inversion HsInv0004, the NH inversion with higher frequencies in the analysis and fixed in East Asia, has low Tajima's D everywhere, consistent with the simulations. However, inversion HsInv0063 has less than 75% of frequency in all populations, but it has in CHB population the lowest Tajima's value in the data set, -2.43. That suggests that there could be some other factor affecting HsInv0063 low values.

NAHR inversions tend to have negative values (Figure 3.15 B). Some inversions have values above two in specific populations but lower elsewhere: HsInv0374 in JPT, or HsInv0209 and HsInv0341 in CEU. Among those with low Tajima's D values, inversions HsInv0266 and HsInv0397 have consistently low values in Europe and South Asia. Inversion HsInv0114 in CHB has the lowest Tajima's D of NAHR inversions, -2.35. And although Tajima's D in JPT is also low, it has more moderate values in other populations.

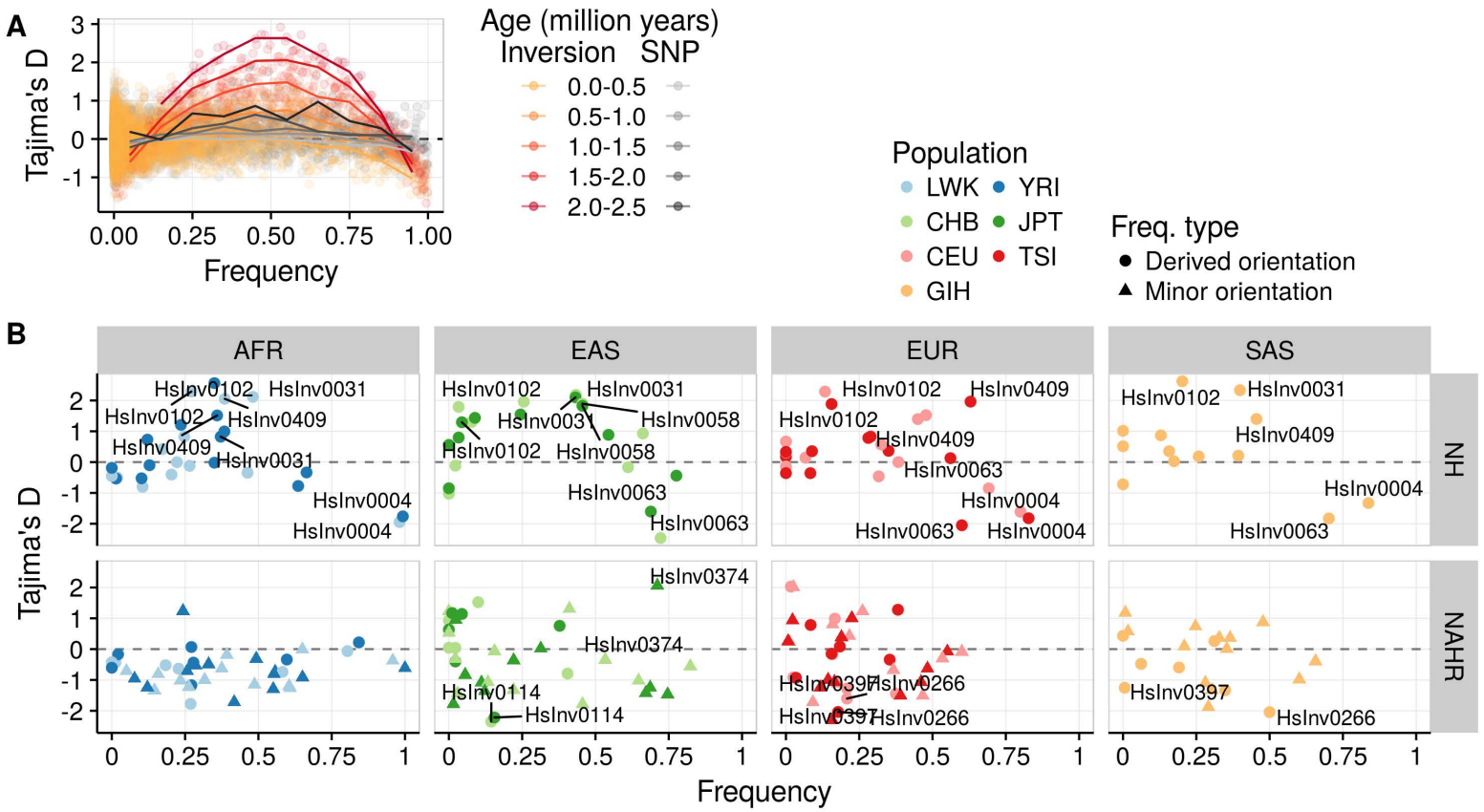


**Figure 3.15: Inversion effect on Tajima's D**. A. `InvertFREGENE` simulations. Each point represents Tajima's D value in a 300-kbp simulation. Solid lines show mean values in bins of 0.1 frequency ranges, grouped by age category. Orange values represent simulations with inhibition of recombination between orientations (inversions) and grey values equivalent simulations of 1-bp mutation in the centre that does not restrict recombination (SNPs). B. Real inversions. Tajima's D was estimated from the total variation in the inverted region for NH and NAHR inversions separately.

## 3.4 Linked variation and recombination

The effects of the recombination inhibition between orientations are clearly visible on the total nucleotide diversity levels of NH inversions. However, some levels of recombination (as double crossovers or gene conversion) could still be taking place within the inverted region. Also, recombination may happen normally after the inversion breakpoints or instead the inhibition could extend over a longer sequence because of pairing impediments between chromosomes with the two orientations during meiosis. In order to clarify the recombination patterns, we analysed its footprints in the linked variation. We also wanted to get some insight about the interplay of recombination inhibition and recurrence taking place in NAHR inversions.

We were concerned about the possible confounding effect of errors in the sequencing data, so we first analysed the data in a conservative way without trusting the phases and focusing on reliable variants according the the strict accessibility criteria in the 1000GP phase 3. Later we used the phased information in order to detect possible recombination events.

### 3.4.1 Linkage disequilibrium patterns

We had seen in the tag variant analysis that NH inversions generally have SNPs and indels in complete linkage disequilibrium (perfect tag variants), while NAHR inversions mostly lack of them. We further explored those patterns and estimated the maximum $r^2$ values between each inversions and nearby variants using all populations together (Figure 3.17 A). As already mentioned, NH inversions have almost always other variants in perfect linkage disequilibrium (LD), $r^2 = 1$. The only exceptions are HsInv0105, with a maximum value of 0.98, and HsInv0102 with a lower maximum value, 0.60. NAHR inversions are markedly different, and low $r^2$ values seem to be the norm, with the most extreme inversion being HsInv0061 with a maximum $r^2$ of just 0.14.

In order to determine if the observed $r^2$ values are unusual in the genome, we selected a sample of 1,000 genome-wide SNPs for each inversion with the same population frequency as the inversion. SNPs also were sampled from the same type of chromosome as the inversion (autosomes or chromosome X), to control for the effect of the different effective population size and recombination differences. Then, we estimated the maximum $r^2$ value between the SNPs and other variants located within 500 kbp. Therefore, we had up to seven sets of 1,000 SNPs for each inversion, representing the usual LD values in each population.

We compared the proportions of inversions and SNPs tagged by other vari-

ants at different levels of $r^2$ and found different patterns for the two types of inversion (Figure 3.16). As expected from the tag SNP analyses, NAHR inversions have a strong depletion of associated variants, presumably as a consequence of recurrence (Figure 3.16). For instance, the proportion of NAHR inversions perfectly tagged by another neighbouring variant ($r^2 = 1$) is 12% in African populations and 24% in non-African populations, whereas the proportion for SNPs is 66% and 84%, respectively.

On the other hand, NH inversions have more often other variants in complete LD than equivalent SNPs (96% of inversions against 78% of SNPs, Fisher exact test using mean values across populations, $P = 0.096$, testing populations per separate, LWK $P = 0.018$ and TSI $P = 0.020$). This result could be expected from the inhibition of recombination within the inverted region between orientations. However, a similar pattern holds when considering only variants outside the inverted region (89% of inversions with variants at $r^2 = 1$ against 78% of SNPs), suggesting that inhibition of recombination could affect LD levels past the breakpoints. Another alternative explanation could be a difference in recombination background levels (e.g. from inversions being removed from high-recombination regions).

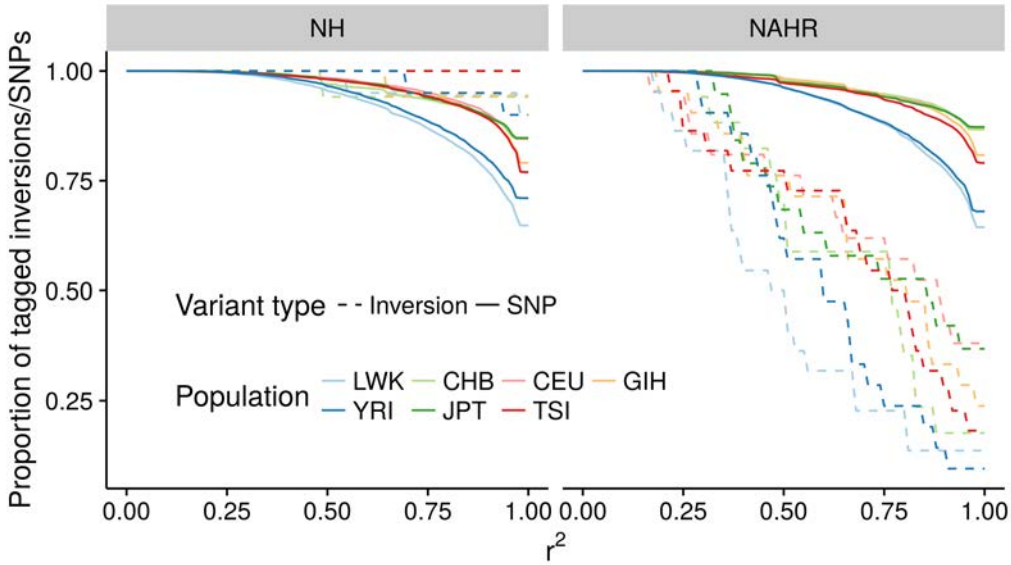


**Figure 3.16: Maximum $r^2$ with nearby variants**. Proportion of inversions or SNPs with linked nearby biallelic variants (within 500 kbp) at decreasing pairwise linkage disequilibrium values ($r^2$). The patterns observed for the two types of inversions are compared to random genomic SNPs with frequencies matched to each inversion/population combination.

## 3.4.2 Types of linked variants

### 3.4.2.1 Classification from genotypes

We next wanted to understand the low linkage values in NAHR inversions and also check if NH inversions showed any evidence of recombination inside the inverted regions. For that, we looked at the distribution of different types of polymorphisms, including those shared between orientations that would only be expected if recombination between chromosomes takes place. In order to rule out phasing errors that could create false shared polymorphisms in heterozygous individuals, we used only the unphased genotypes of 1000GP phase 3. Also, to minimize genotyping errors, that could also create false shared polymorphisms, we only considered variants accessible according to 1000GP strict accessibility mask (accessibility overview in Figure A.3). We followed a conservative criteria where we only regard a variant as polymorphic in one orientation if it is unambiguously polymorphic (e.g. an individual Std/Std for the inversion carries the two alleles of the variant, see Methods). We classified the nearby polymorphisms as private of one orientation or the other, fixed or shared between orientations.

Most NH inversions have no shared variant within the inverted region (Figure 3.17 B), consistent with a complete inhibition of recombination across the inverted region. However, we could only test 13 NH inversions, because the other eight inversions do not have accessible variants inside. The only exception is inversion HsInv0063, that has the SNP rs74405082 with the alternative allele in three individuals, each of them with a different genotype for the inversion. We checked the reads of each of the samples from the 1000GP read alignments, and all of them had support for the two alleles. Thus, genetic flux between orientations could be in theory responsible for the shared variant, although a single shared variant seems unlikely to come from a gene conversion or double recombination event. It is worth noting that inversions HsInv0102 and HsInv0105, without perfect tag variants, have enough sequence and do not have any shared variant. When looking at the variant proportion in the flanking regions, shared variants slowly increase with distance from the inversion until a constant proportion of around 25% (Figure 3.17 B).

On the other hand, NAHR inversions have on average a 19% of variants shared between orientations, which are distributed across all the inverted region. Importantly, that seems to be approximately the same proportion of shared variants found in flanking regions (Figure 3.17 B). These patterns suggest that the shared variants within NAHR inversions are created by recurrence rather than double crossovers or gene conversion. However, NAHR inversions show high heterogeneity in the levels of linkage disequilibrium with nearby variants and the proportion of shared polymorphisms between

orientations. In one extreme we find inversions with patterns similar to NH inversions. Two inversions, HsInv0040 and HsInv0790, have global fixed variants and do not have any shared polymorphism (Figure 3.17), suggesting that all inverted chromosomes derive from the same inversion event. Inversion HsInv0061, although lacks fixed variants, does not have any shared polymorphism. In the other extreme, inversion HsInv0069 shows very low linkage disequilibrium levels with nearby variants and has the highest proportion of shared polymorphisms (50%). The remaining inversions show intermediate characteristics, and the two measures, LD and proportion of shared variants, have a clear negative correlation in NAHR inversions (Pearson's $r = -0.53$, $P = 0.009$), probably related with recurrence rate.

#### 3.4.2.2 Classification from phased haplotypes

In order to obtain as much detail as possible into the recombination patterns associated to inversions, we used phased data as well as information about the ancestral and derived alleles. We also relaxed the accessibility criteria, and used all the polymorphisms accessible according to the pilot mask, less restrictive than the strict mask used previously (overview in Figure A.3 in Appendix A). However, we restricted the analysis to SNPs and, since the ancestral state is required, only those with ancestral information in 1000GP data were used. Here inversion phase needs to be incorporated to the phased haplotypes, so we analysed homozygotes and heterozygotes only when perfect tag SNPs are available..

Assuming phases as correct, we refined the classification of variants within and surrounding inversions using the classification of linked variants described in Ferretti et al. (2017). We classified mutations into the five categories without recombination (table 3.7), with two additional categories (switched and shared) that are only explicable with recombination. Note that switched variants also indicate recombination and in the previous classification we were considering them as private and not taking them into account.

With the new classification we found 21 shared and 3 switched variants within the inversion region of NH inversions that, if phase and genotype are correct, would require recombination between haplotypes in opposite orientations. However, 75% of them (16 and 2) are located in inaccessible areas according to the strict mask, but accessible with the pilot mask. Given that the total proportion of variants accessible to the pilot but not to the strict mask is much lower (regardless of the variant classification), a 44%, it is likely that most exchanged SNPs are indeed genotyping or phasing errors. The six SNPs located inside the strict accessibility areas, that would indicate recombination, are within two inversions: two shared and one switched

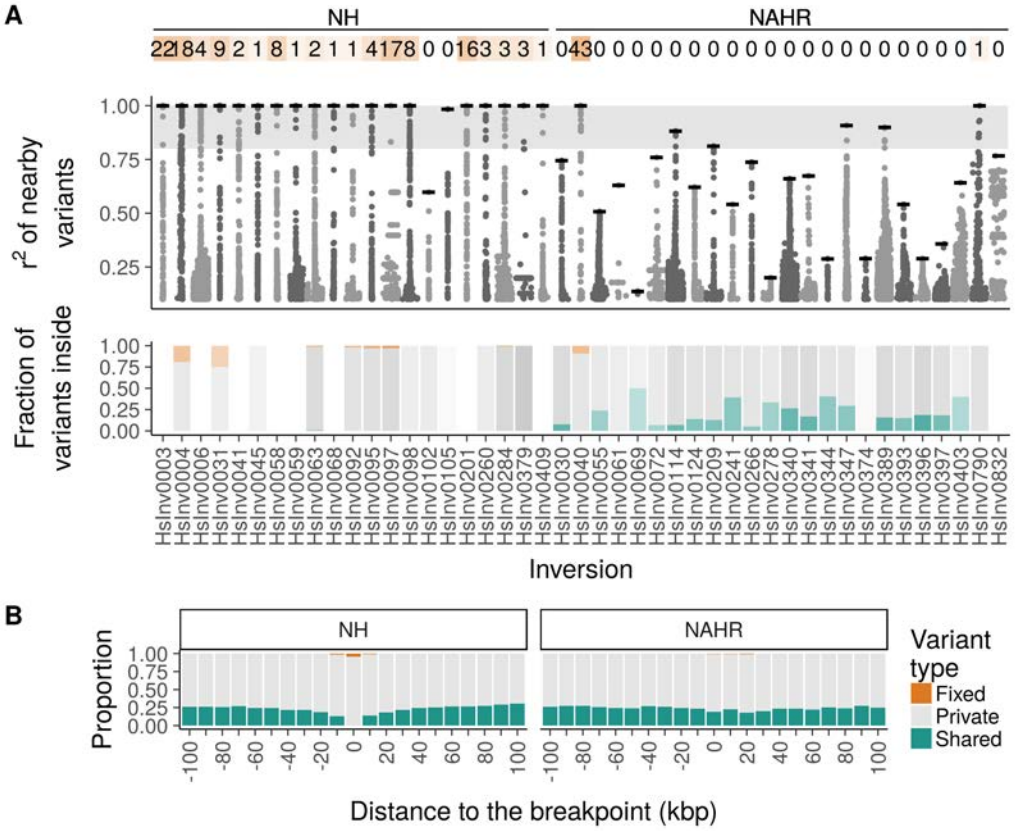


**Figure 3.17: Linkage disequilibrium patterns.** A. Distribution of nearby SNPs and small indels and their linkage disequilibrium ($r^2$) with each inversion. Relative position to the breakpoints of each inversion is represented in the x axis. Maximum $r^2$ level is indicated with a horizontal line. Variants in the shaded area are classified here as inversion tagging variants. The total number of variants in perfect linkage disequilibrium is indicated above. Below, proportion of different types of variants inside the inverted sequence. Overall colour intensity is proportional to the number of variants in the region. Empty bars are inversions without accessible variants. B. Average proportion of different types of variants inside the inverted sequences (distance 0) and in flanking regions grouped in bins of 10 kbp. Both flanking regions should be equivalent and sign just indicates position with respect to the breakpoint. Here they are represented separately for clarity.

**Table 3.7: Correspondence between classifications of nearby variants**. The genotype-based classification uses the number of alleles present in each orientation while the phase-based also incorporates information about ancestral and derived alleles and can detect private derived variants that are only possible in the presence of recombination (or sequence errors).

| Genotype classification | Haplotype classification |
|---|---|
| Fixed | Co-occurring<br>Complementary |
| Private ancestral | Enclosing<br>Strictly disjoint |
| Private derived | Strictly nested<br>Switched (requires recombination) |
| Shared | Shared (requires recombination) |

in HsInv0063 (one of them the same shared SNP found before), and three shared in HsInv0379. We further inspected the putatively exchanged variants. The new shared SNP in inversion HsInv0063, rs6960637, is likely to be a phasing error, given that no homozygote for the inversion carries the derived allele T, and only one heterozygote carries it in heterozygosis in the same phase as the inversion tag SNPs. The switched SNP, rs546825992, is a singleton only present in a heterozytgote, that has probably been assigned to the wrong haplotype because of the lack of information from other chromosomes. In inversion HsIn0379, the longest in the data set and only present in five individuals in heterozygosis, the three shared SNPs are polymorphic in the ancestral orientation and are in heterozygosis in inversion carries. The derived alleles of SNPs rs12977333 and rs568739874 are assigned to the inverted chromosome in individual NA18956. A phase error could explain the pattern, although the two SNPs are separated by only 562 bp and an event of gene conversion could be possible. The third SNP, rs76319388, can be also explained by a phase error in another individual, NA18632.

We found 22 shared or switched SNPs in the seven NAHR inversions, 15 of them in the accessible according to the strict mask, and that had been already found in the first classification. Inversion HsInv0040, the only NAHR inversion in this analysis with tag SNPs and without shared variants in the genotype classification, still does not have any shared variant when relaxing the accessibility criteria. Two inversions have only one shared SNP: HsInv0030 in an inaccessible region according to the strict mask and HsInv0266 in the accessible region. Inversion HsInv0209 has six shared SNPs, all within accessible areas. Finally, inversion HsInv0114 has ten shared and two switched, half of them accessible and half inaccessible to the strict mask.

Considering also the other types of SNPs, individual inversions show different

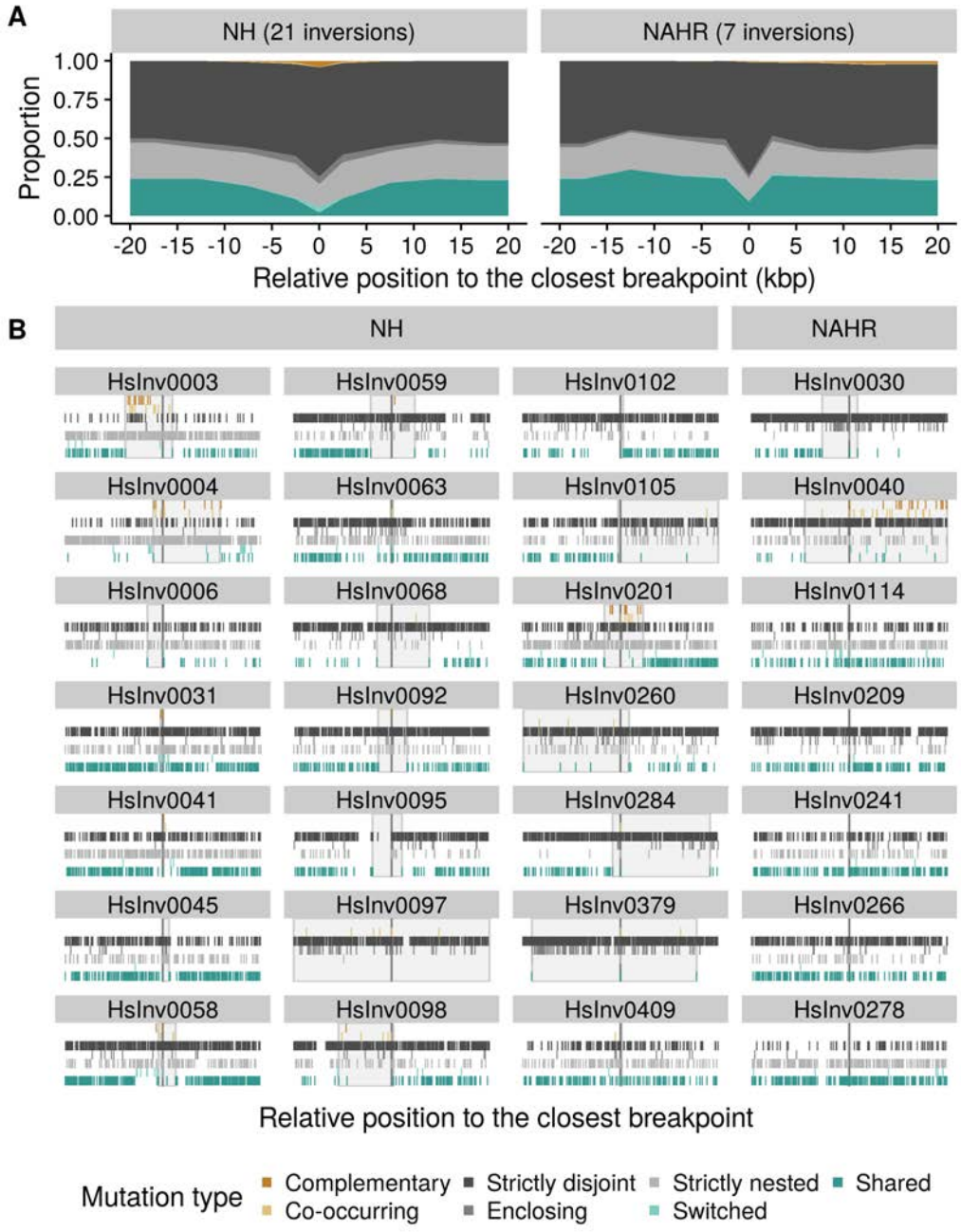


**Figure 3.18: Physical position of different classes of variants linked to inversions.** Only inversion with known ancestral orientation have been analysed. A. Summary of average variant type proportion inside and in 5-kbp windows outside the inversions. B. Detailed positions of the different variants per inversion in the flanking regions. Shared area indicates extension of the non-recombining region as defined in section 3.4.3.2.

proportions and distribution of each type of variant, but some general trends are observed (Figure 3.18). Fixed mutations between orientations are mostly located either inside the inversions or in the flanking regions close to the breakpoints (central positions in Figure 3.18 B). In some cases, a region lacking shared and switched variants overlaps with the region with fixed variants (e.g. HsInv0003, HsInv0004, HsInv0201), while in others, there is a lack of shared and switched variants without the presence of fixed ones (e.g. HsInv0068, HsInv0284). Most inversions have more variation private to the ancestral allele (mostly strictly disjoint), although inversions with high derived allele frequency can outnumber the derived-private mutations strictly nested (like HsInv0003 and HsInv0004). NAHR inversions in the analysis are just those with known ancestral orientation, that are probably less recurrent, and show similar patterns to those seen with NH inversions (Figure 3.18 A). Notably, inversion HsInv0040 has the highest number of fixed SNPs extending though a large region.

### 3.4.3 Recombination outside the breakpoints

Inversions are expected to affect the variation within the inverted area, but it is less clear if they have an effect on the surrounding areas. If they do affect linked variation in the flanking regions, it could be interpreted as an extended inhibition of recombination past the breakpoints. Alternatively, the physical impediments for pairing within the inverted regions could increase the recombination events in the flanking regions. In order to test these alternative scenarios, we compared the values of different measures with those observed in SNPs at the same frequencies across the genome. All comparisons are made at a population level, in order to match meaningful population frequencies.

#### 3.4.3.1 Variant types outside the inversion region

We took advantage of the increased resolution gained with the new linked variant classification to further dissect the possible effect on the flanking areas. We used again 1,000 random same-frequency SNPs for each inversion and population and classified the variants in the flanking 20 kbp of each of them. Then, we estimated the proportions of each type of variant at four different non-overlapping distance ranges of 5 kbp from the target SNP. The average proportion observed in each distance window for each type of inversion was compared against the 1,000 average proportions obtained from the same-frequency SNPs. Figure 3.19 shows the spacial change in the proportion of the variants in complete linkage disequilibrium (fixed variants, both co-occurring and complementary), as well as variants evidencing recombina-

tion (switched and shared, globally referred as exchanged).

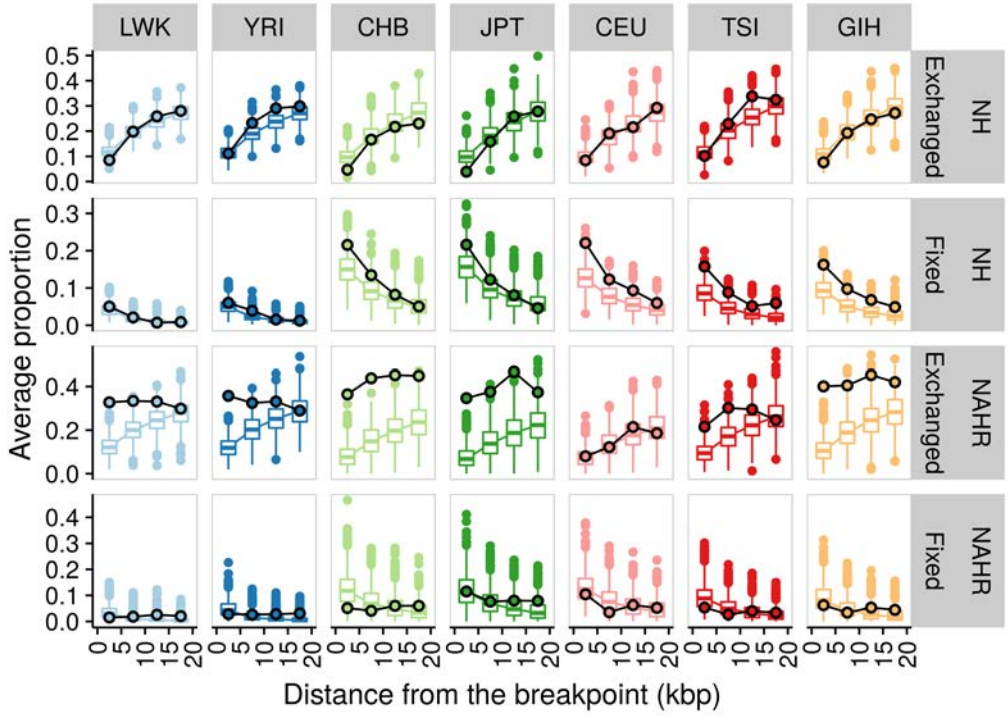


**Figure 3.19: Changes in variant types close inversion polymorphisms**. Points linked by a black line represent the average proportions of exchanged and fixed variants in bins of 5 kbp from the inversion breakpoints at both sides. Inversions are grouped by mechanism of formation, meaning that the averages are made from up to 21 NH and 7 NAHR inversions. Boxplots show the distribution of average proportions estimated from 1,000 samples of 21 and 7 same-frequency SNPs.

NH inversions have very similar proportion of exchanged variants in distance ranges between 5 kbp and 20 kbp. If anything, it could be argued that the proportion of exchanged variants in the first 5 kbp is systematically in the lower part of the distribution for all populations. The proportion of fixed variants in the first 5 kbp also seems to be higher for inversions than for SNPs with the difference decreasing with the distance. However, all inversion proportions fall within the SNP-based distributions, so we do not have any strong evidence of NH inversions having altered linked variant patterns outside the breakpoints.

The low linkage disequilibrium for NAHR inversions seen earlier is visible again in the proportions of variant types. The main difference seems to be in the exchanged variants, that is clearly higher than expected close to the breakpoint and remains nearly constant across all distance ranges. In contrast, the proportion of fixed variants is very similar to the one obtained from SNPs at the same frequency.

#### 3.4.3.2 Non-recombining region

We used the distribution of the mutation types to define the region outside the inversion breakpoints that does not show evidences of recombination between chromosomes in opposite orientations. The criteria used tolerates some exchanged variants as possible genotype or phase errors, as long as there is an important proportion of fixed variants after them (regions highlighted in Figure 3.18 B). The extension of the region without recombination after the breakpoints (and combining the two flanking regions) has a median size of 15.0 kbp (mean of 16.5 kbp), additional to the inverted region. However, individual inversions have very different lengths of extra non-recombining regions, probably reflecting their age and frequency, which have affected the number of recombination events that have occurred between chromosomes in opposite orientations in the flanking regions. Inversions such as HsInv0031 or HsInv0063 have evidences of recombination right after the breakpoints. In contrast, low-frequency inversions HsInv0097 and HsInv0379 do not show recombination in most of the region studied.

In order to know what is the usual length of non-recombining regions linked to other variants, we applied the criteria to the frequency-matched SNPs. Again, the comparison was made at population level, in order to have meaningful frequencies matched and account for differences in recombination rate between populations. We found that the observed median lengths of 15.0 kbp are quite common: 90% of the 1,000 same-frequency SNP samplings have median non-recombining distances between 11.1 kbp and 17.5 kbp. Therefore, it does not seem to be any systematic extension of the recombination inhibition after the breakpoints or increase of recombination near the breakpoints.

Nevertheless, if inhibition of recombination was taking place past inversion breakpoints, it would be difficult to detect it in recent inversions, where few recombination events would have had time to occur anyway. Thus, we took a closer look at high-frequency inversions, where the power should be higher. Inversions HsInv0004 and HsInv0006 seem to have excessively long non-recombining regions for their high frequency in TSI and African populations, respectively (in the top 5% of the empirical distributions). However, other several high-frequency inversions, such as HsInv0003 and HsInv0201, have the expected non-recombining region lengths. And inversion HsInv0409, also at high frequency, is in the lower tail of the distribution. Therefore, we do not have any evidence of the recombination inhibition extending past the breakpoints. Observed variation could be explained by differences in local recombination levels or mutation age, none of which we are controlling for.

The other alternative, an increased recombination at the flanking regions created by relocation of recombination events inhibited within the inversion

region, does not have support either. Although five inversions have shorter non-recombining regions than would be expected, including some population with a value within the lower 1% of their corresponding SNP distributions, all of them are mediated by NAHR mechanisms and the most likely explanation is recurrence. Only NAHR inversions HsInv0030 and HsInv0040 have non-recombining regions within the values frequently seen in SNPs at the same frequencies.

## 3.5 Inversion history and dynamics

The inhibition of recombination between orientations offers separated genealogies for inverted and ancestral chromosome groups, from which to infer the past history. The amount of sequence differentiation between orientations should increase with time, making it easier to measure inversion age. However, recurrence obscures the picture by transferring a possibly differentiated haplotypes into the other group and allowing the creation of intermediate haplotypes by recombination with the other chromosomes in the same orientation. In this section we investigate the history of individual inversions by estimating their age from divergence levels and analysing the relationship between haplotypes and inversion orientations.

### 3.5.1 Inversion age

A commonly used method to estimate the age of inversion polymorphisms consists in measuring the excess of pairwise sequence differences between orientations, i.e the net divergence, in regions near breakpoints with suppressed recombination (Hasson and Eanes 1996). The expected differences between two sequences in a free-recombination setting at the time of origin is unknown and it is usually estimated from the observed differences between chromosomes with the ancestral orientation. Here, we used the nucleotide diversity in the ancestral orientation or in the derived allele if it has higher diversity (as happens in high-frequency inversions,see Figure 3.14), that was found to be more accurate (see Methods).

#### 3.5.1.1 Simulations

First, in order to measure the error associated to the estimator, we analysed previously simulated inversions with known age and applied the same strategy to estimate their age. Figure 3.20 shows the difference between the estimated and the simulated age for different frequency and age categories. Positive values mean age overestimates and negative, underestimates. We can see that while for low-frequency recent inversions age can be underestimated, the statistic tends to overestimate the age at higher frequency and older inversions, typically by an amount of 250,000-500,000 years. Additionally, variance is very high and for inversions younger than one million years, estimated age can easily double the real age. Finally, inversions were simulated in a constant-sized panmictic population, and any complex demographic history is likely to increase the error. Therefore, all results from this method need to be taken with caution.

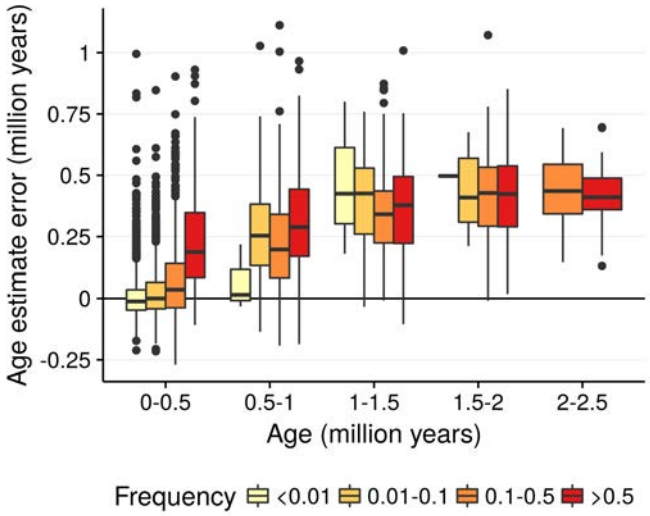


**Figure 3.20: Error in age estimate**. Performance of the used estimator in simulated inversions of different frequencies and ages. Generation time is assumed to be 25 years.

#### 3.5.1.2 Real inversion estimates

Next, we applied the same strategy to the real inversions. To do so, both pairwise differences between orientations and within orientation were estimated using 1000GP phase 3 sequences for all individuals, when fixed SNPs allowed unambiguous inversion phasing, and only for inversion homozygous individuals otherwise. Two low-frequency inversions, HsInv0061 and HsInv0790, do not have homozygous individuals with the alternative orientation and were dropped from the analysis. We considered the inverted region and the extended non-recombining region when defined. A first age estimate was obtained by dividing the net divergence between orientations by twice a constant substitution rate of $1 \times 10^{-9}$ changes per bp per year (Figure 3.21 left panel). Then, in order to control for local differences in substitution rates, we obtained local estimates from the divergence with chimpanzee and gorilla genomes, considering 6 and 8 million years as split times with humans (Figure 3.21 right panel).

NAHR show moderate absolute divergences, with all age estimates except one below 500,000 years. Inversion HsInv0040 has the highest estimate of 860,319 to 940,263 years (considering either the constant substitution rate or the chimpanzee and gorilla-based ones). This is the only NAHR inversion with fixed SNPs from the 21 analysed. The low values of most NAHR inversions are likely a consequence of recurrence that limits divergence between orientations, resulting in most estimated ages being too low to be feasible from their geographical distribution. Inversions polymorphic in both African and non-African populations, as it is the case of all 21 NAHR inversions analysed, should be generally older than the population split time. However, assuming a constant substitution rate, 13 of them have estimated ages below 50,000 years (lower bound for the out-of-Africa event).

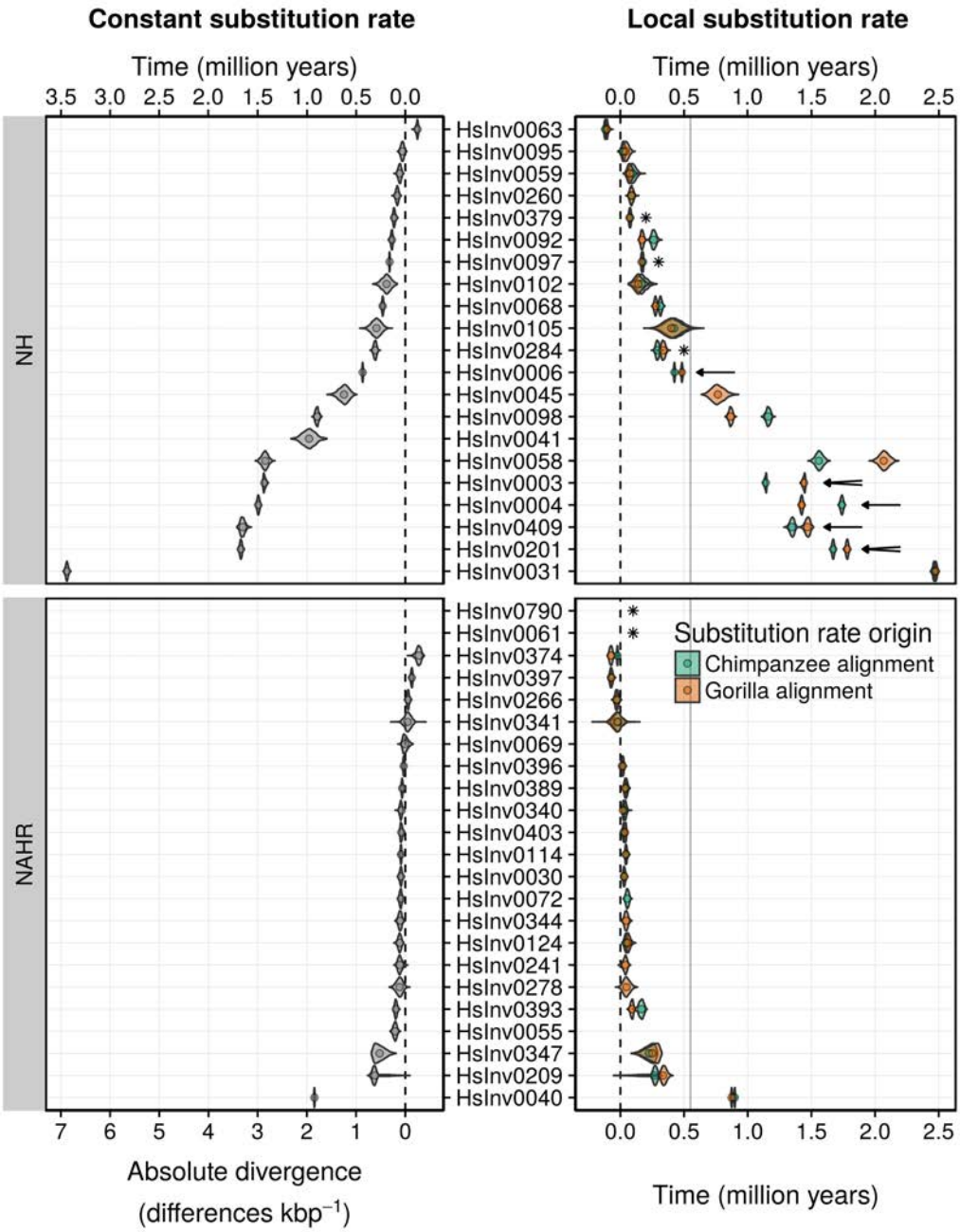


**Figure 3.21: Inversion age estimates**. Bootstrapped distributions of age estimates. Left panel shows absolute divergence between orientations converted into years using a constant substitution rate of $1 \times 10^{-9}$ changes per bp per year. Right panel shows the conversion into years using a local substitution rate estimated from local divergence with chimpanzee and gorilla, assuming split times at 6 and 8 million years ago. Dashed lines indicate present and grey line the lower bound for split time between Neanderthal-Denisova and modern humans. Asterisks: inversions polymorphic only in African populations or in non-African populations. Simple arrows: inversions with derived allele in Neanderthal genome. Double arrows: inversions with derived allele in Neanderthal and Denisova genomes.

In contrast, NH inversions show a wide range of absolute divergences, with six inversions estimated to have appeared more than one million years ago. With some exceptions, most notably inversion HsInv0063 that has a negative estimate, ages of most NH inversions are consistent with their distributions. For instance, the derived orientation of five inversions can be found in Denisova or Neanderthal sequenced genomes (HsInv0003, HsInv0004, HsIvn0006, HsInv0201 and HsInv0409, represented with arrows in Figure 3.21 from Giner-Delgado et al. (in prep.)), and all of them except inversion HsInv0006 have ages older than 550,000-750,000 years (estimated range of split time between Neanderthal-Denisova and modern humans). Estimated age for HsInv0006 ranges between 407,795 and 495,470 years, that despite being more recent than the split time, it is very close to the lower bound. Note that in this case it could not be the result of an introgression, because the derived allele is found at high frequency in Africa. Thus the age is likely to be underestimated. In addition, NH inversions that are restricted to either African (HsInv0284) or non-African populations (HsInv0097 and HsInv0379) have a relatively recent estimated age, also consistent with their limited distribution. Only inversion HsInv0095 has an estimate of between 22,582 and 41,258 years, too recent for their presence inside and outside Africa, but close.

### 3.5.2 History reconstruction from haplotypes

We next wanted to visualize and understand inversion haplotype diversity and distribution within orientations and populations. Evolutionary relationships for large regions with recombination are difficult to reconstruct and require the use of flexible representations, such as reticulated networks, capable to accommodate past recombination events. However, in haplotype networks each sequence is reduced to a node or edge, making it difficult to understand at the same time haplotype relationships and spatial distribution of alleles along the sequence. Here, we opted for a simpler way of representing the similarities between haplotypes in a hierarchical clustering, combined with the visualization of the population where each haplotype is found as well as the orientation if unambiguously known (haplotypes in homozygotes for the inversion or inversions with fixed SNPs). We analysed non-singleton accessible SNPs in the inverted region (plus the non-recombining region when defined). All inversions had enough variants except HsInv0041, that was excluded from the analysis.

Figure 3.22 (plus figure A.4 in Appendix A representing all analysed inversions) show the haplotypes in a central panel with an annotated dendrogram with the clustering at the left and the similarity matrix used for the clustering at the right. Haplotypes are represented as rows of allele combinations. Alleles in black indicate a derived mutation and in white the ancestral allele. Those SNPs without a defined ancestral allele are represented as light

grey for the allele in the HG19 reference genome and in dark grey for the alternative allele. Annotations in the tips of the dendrogram represent the orientation where the haplotype is found as arrows (rightwards means the orientation found in genome HG18) and the populations of the individuals that carry them in the same color code as usual (blues Africa, greens East Asia, reds Europe and orange South Asia).

#### 3.5.2.1 General trends

We found that in all NH inversions, haplotypes carried by chromosomes with the derived orientation cluster together. In five inversions (HsInv0003, HsInv0004, HsInv0031, HsInv0058 and HsInv0201), the two main clusters divide the two orientations (see example A in Figure 3.22), consistent with old inversions that have had time to diverge. Inversion HsInv0409 has a similar pattern, except that one of the haplotypes in the derived cluster can also be found in one homozygote individual with the ancestral orientation (in this case the reference genome has the derived orientation), probably indicating that either the inversion genotype or the haplotype of that individual are wrong. Alternatively, it could also be a past event of gene conversion between chromosomes with different orientations. The option of it being an ancestral haplotype shared between orientations seems unlikely, given that they show high levels of differentiation and there is only one individual with the unexpected haplotype. The remaining 14 inversions cluster all together at higher levels (closer to the dendrogram leaves, example B in Figure 3.22). In some cases, other haplotypes from the ancestral orientation are also in the derived cluster or even some haplotypes are shared between orientations. In this case, it could be that inverted haplotypes have not had time to differentiate and are still similar to the haplotype carried by the original inverted chromosome.

In contrast, there are only three NAHR inversions with clear clustering of the haplotypes in the derived or less frequent orientation (HsInv0030, HsInv0040 and HsInv0072). HsInv0040 is the more clear case with two main clusters diving the orientations, while HsInv0030 and HsInv0072 clusters are in higher levels (closer to the dendrogram leaves). The derived orientation in HsInv0790 is only found in heterozygotes and, since it does not have any fixed SNP within 20 kbp from the breakpoints, haplotypes from heterozygous individuals have not been assigned to any orientation. Nevertheless, a cluster of haplotypes only found in heterozygotes can be clearly observed, indicating that most likely they correspond to the chromosomes with the derived orientation. Similarly, inversion HsInv0061 is only present in 11 individuals in heteroygosis, and again there is no homozygote with the derived orientation (in this case the orientation in the reference genome). Although all haplotypes are present in chromosomes with the ancestral orientation,

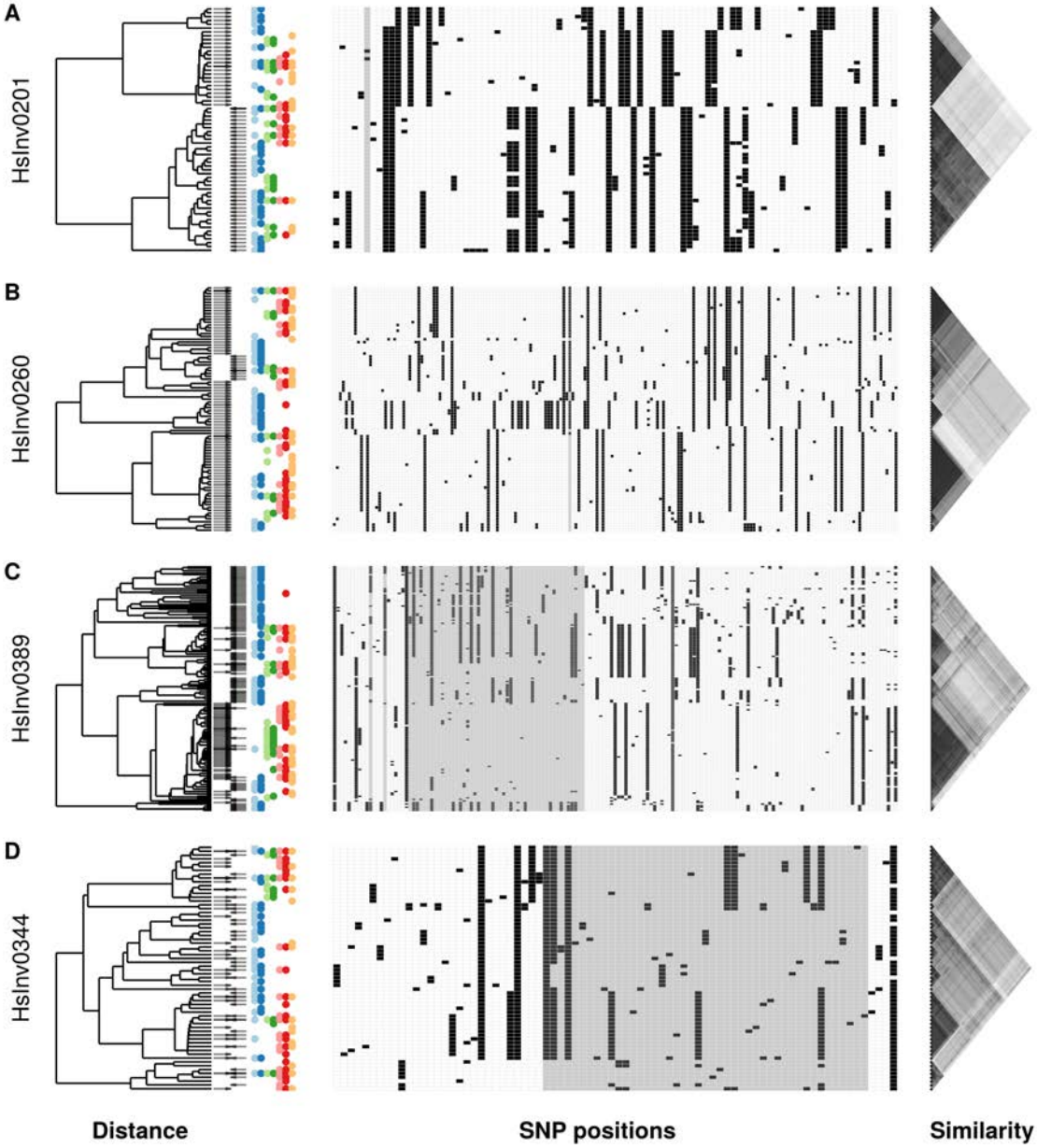


**Figure 3.22: Inversion haplotype examples**. Haplotypes of four inversion regions are shown in the central panels, ordered according to the hierarchical clustering (dendrogram at the left), that is based on distances between haplotypes (similarity matrix at the right). Examples are: A. Old inversion with differentiated haplotypes from a single event; B. More recent single inversion event, with inversion haplotypes similar to other ancestral haplotypes; C. Moderately recurrent inversion; D. Inversion with pervasive recurrence. SNP alleles are coded as follows: white = ancestral, black = derived, light grey = allele in HG19 genome and dark grey = alternative allele (when ancestral allele is unknown). Arrows at the leaves of the dendrogram represent the orientation of the chromosomes with that haplotype (rightward means orientation in reference genome HG18). Populations of the individuals with each haplotype are indicated as dots of different colours (blues, African; greens, East Asia; reds, Europe; orange, South Asia).

the 11 heterozygous individuals have the same frequent haplotype in at least one of their chromosomes, meaning that all chromosomes with the derived orientation could have the same haplotype. The other 19 NAHR inversions have at least two different haplotypes with minor orientation, suggesting multiple independent inversion events (i.e. recurrence). In some cases different inversion events can be guessed, as in the case of inversion HsInv0389 (example C in Figure 3.22), where one main event at high frequency can be seen at the middle of the dendrogram and there are at least three minor events more affecting a reduced number of haplotypes. In other cases, many haplotypes in all positions of the dendrogram can be found in both orientations, as in HsInv0344 (example D in Figure 3.22). However, there are three main factors that make it difficult to quantify the number of independent inversion events: 1) recombination is not inhibited between chromosomes in the same orientation, so some haplotypes could be combinations of previous haplotypes; 2) gene conversion (or double crossovers) could transfer haplotype parts between orientations, specially in long inversions; and 3) SNP genotypes may have increased error rates from mapping and imputation that can create additional artefactual haplotypes.

#### 3.5.2.2 Specific cases

Inversion HsInv0832 is located in non-pseudoautosomal area of chromosome Y and therefore there is no possible recombination between different chromosomes confounding the independent inversion events. That simplifies the interpretation and restricts the noise to genotype errors. Thus, it is a good model to test the strategy. Figure 3.23 shows the five inversion events that can be identified. Most haplotypes in the analysed individuals are in the same orientation as reference genome HG18 (and HG19). In African populations the alternative orientation is the most frequent, although at least inversion two events may be necessary (African population and events are represented in blue in Figure 3.23). Additionally, there could three more inversion events one found in a JPT individual (dark green) and two more in GIH individuals (orange). These results coincide with those of a manual analysis based on the known phylogeny of chromosome Y haplotypes, demonstrating the validity of the used strategy (Giner-Delgado et al. in prep.).

The haplotype analyses also allowed us to examine in depth the possible genotype and phase errors. We had seen from the linked variant classification that the 12.7-kb inversion HsInv0063 has the shared variant rs74405082 between ancestral and derived chromosomes. In order to determine if it could be an event of gene conversion or double crossover, we located the carriers in the haplotype representation. Two JPT individuals have the SNP variant in the same haplotype (green in Figure 3.24), and a third individual from LWK population has the same allele in a different haplotype (blue in Figure

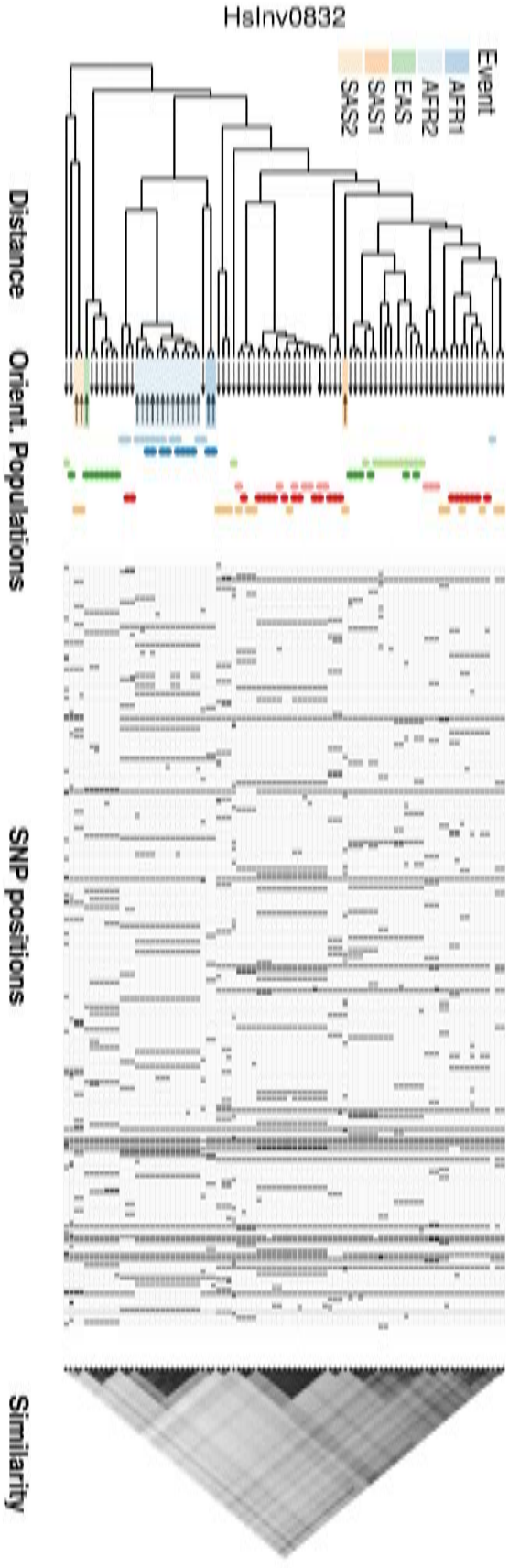


**Figure 3.23: Recurrence in inversion HsInv0832**. Five possible independent inversion events are highlighted in the dendrogram. See Figure 3.22 for details on the representation of each panel.

3.24). The differences between the two haplotypes surrounding the shared SNP imply that if it was the result of a gene conversion event or double crossover, it would have to extend less than 650 bp or we would have had extra shared SNPs in the same chromosomes. The proximity to the 3' inversion breakpoint (only at 577 bp) makes a recombination event less likely. An alternative explanation would be that they are two independent mutations at the same location, perhaps supported by the fact that the carrier haplotypes are in distant populations. In summary, several options could explain the pattern, including gene conversion.

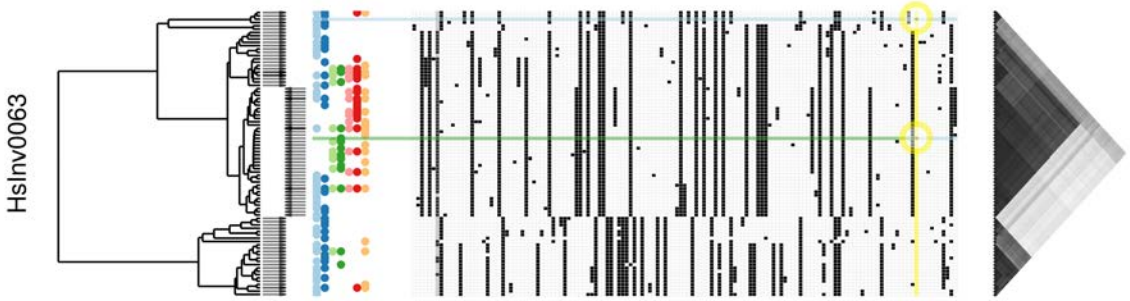


**Figure 3.24: Shared SNP in inversion HsInv0063**. The shared SNP rs74405082 is highlighted in yellow. Carrier haplotypes indicated in blue (one LWK individual) and green (two JPT individuals). See Figure 3.22 for details on the representation of each panel.

## 3.6 Inversions under selection

In order to detect inversions that could have a beneficial effect and could deserve further functional characterization, we integrated complementary signals consistent with past and ongoing positive or balancing selection. Since most inversions in the data set are present in multiple continents, we considered two main selective scenarios. First, long-term balancing selection, where an inversion would be maintained at intermediate frequencies in all or most populations. Second, local positive selection on standing variation, that would increase the frequency of the beneficial allele in a certain environment. Other possibilities, such as hard sweeps or adaptive introgressions, seem unlikely. Hard sweeps would be fixed or nearly fixed in one or several related populations and absent in the rest, and the only population- or continent-specific inversions are at low frequencies. Adaptive introgression would require the absence of the inverted allele in populations without contact with the possible donor archaic hominin, which does not fit all the inversions present worldwide.

Specifically, we defined the signatures suggestive of long-term balancing selection scenario as: (i) intermediate frequencies in all populations; (ii) old origin; (iii) possibly present in other archaic hominins; and (iv) excess of variants at intermediate frequency. And the signatures of positive selection as: (i) high frequency of the selected allele in a population or related populations; (ii) high population differentiation, with lower frequencies in other populations; (iii) relatively recent inversion origin; (iv) low local nucleotide diversity with excess of low frequency variants. The different characteristics analysed in previous sections are summarized in Figure 3.25.

### 3.6.1 Signatures of balancing selection

Several inversions show patterns consistent with long-term balancing selection, although no formal test has been applied that takes into account inversion peculiarities. Therefore, neutrality cannot be ruled out. The three inversions estimated to be the oldest (HsInv0031, HsInv0201, and HsInv0058) have also intermediate frequencies and low population differentiation. Tajima's D values are relatively high (more than 1.5 in several populations), consistent with the possibility that they are under balancing selection. However, we have seen from simulations that old neutral inversions at intermediate frequencies are expected to give high values of the statistic. Among them, only inversion HsInv0201 has a clear functional effect, with the derived orientation associated to a deletion removing a coding exon of *SPINK14* gene. It is interesting that the derived allele of inversion HsInv0201 is also present in the genomes of Neanderthal and Denisova. Additionally,

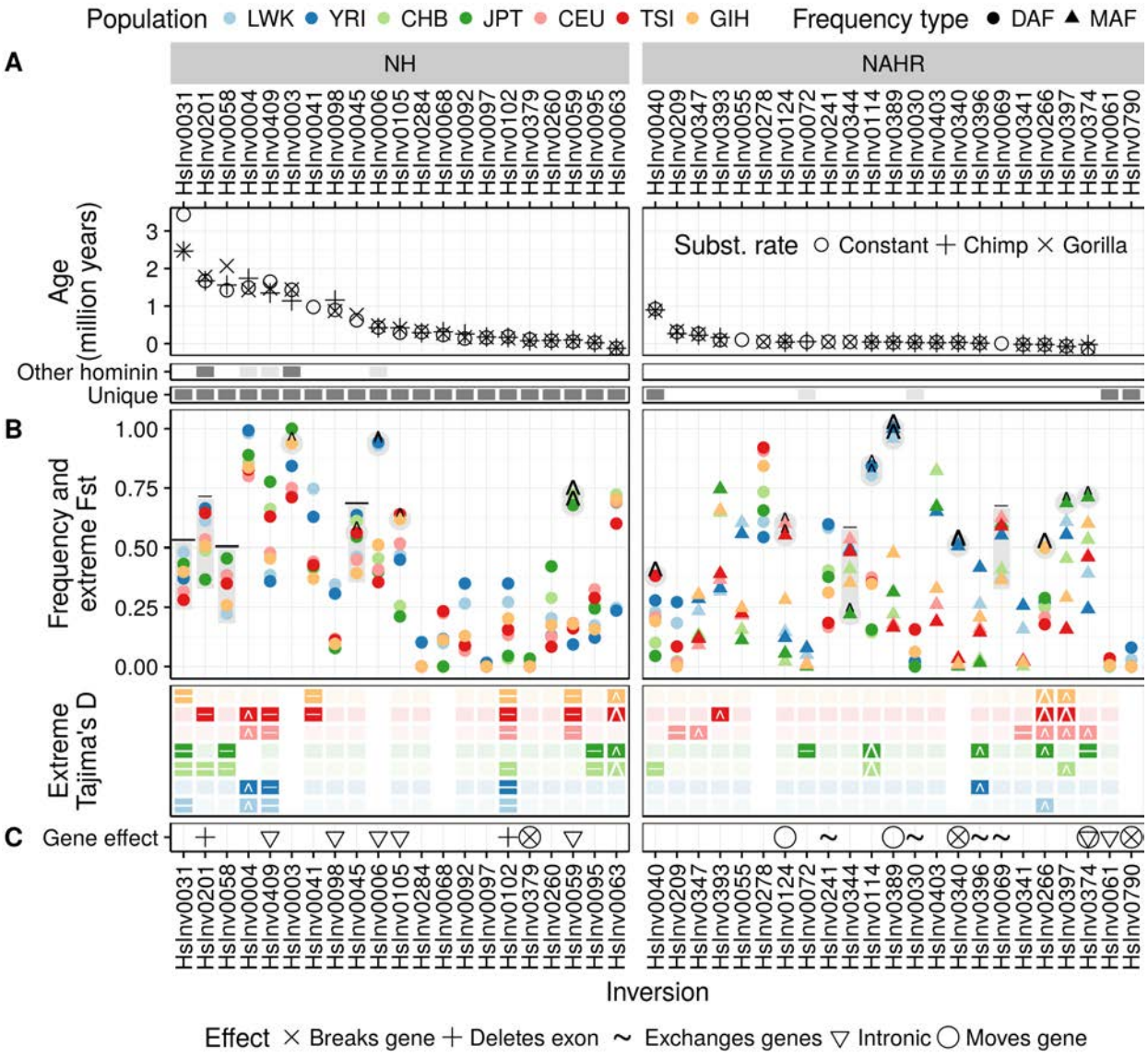


**Figure 3.25: Evolutionary history, selective signatures and functional effect of the 45 inversions**. Panels from top to bottom as follows. **A. Evolutionary history**. A.1 Age estimates from Figure 3.21. A.2 Presence in other hominin genomes. Dark grey: derived orientation is found in Denisova and Neanderthal genomes; light grey: only in Neanderthal; white: absent or unknown. A.3 Unique inversions. Dark grey: all inverted chromosomes are likely to have a unique origin; light grey: they could have a unique origin; white: they show patterns consistent with recurrence. **B. Selective signatures**. B.1 Derived (circle) or minor (triangle) allele frequencies and inversions with unusual $F_{ST}$ in grey, indicating likely population(s) driving the signal. Signs $\wedge$ indicate high population differentiation and sign – low population differentiation, consistent with positive and balancing selection, respectively, as indicated in Tables 3.3 and 3.4. Sign size indicates strength of signature. **B.2** Extreme Tajima's D. Highlighted inversions in Tajima's D panel have values over |1.5| (small sign) and |2| (large sign). Signs indicate positive or balancing selection like in the $F_{ST}$ panel. **C. Functional effect**. Effect on genes as detailed in Figure 3.1.

inversions HsInv0409 and HsInv0041 have also relatively high Tajima's D values. As before, that could be signals of neutral old inversion polymorphisms (∼1 million years) that increase population structure and nucleotide diversity. In this case, population frequencies span a wider range, so they do not seem to be maintained at certain frequency by strong balancing selection. Three more inversions (HsInv045, HsInv0069 and HsInv0344) have low-differentiation and intermediate frequencies, but without high Tajima's D values. Interestingly, inversion HsInv0102 has high Tajima's D values, while showing relatively low frequencies and usual population differentiation levels. In that case, if balancing selection is acting on the region, it is probably maintaining another polymorphism.

### 3.6.2 Signatures of positive selection

In the case of positive selection, there are more clear candidates. Among the NH inversions, HsInv0006 and HsInv0059 stand out, both located in introns of protein-coding genes *DSTYK* and *GABRR1*. Inversion HsInv0006 is nearly fixed in Africa, while populations in other continents have frequencies around 50%. Frequency of inversion HsInv0059 is much higher in East Asian populations than in the other populations. Despite having a frequency near 75%, the haplotypic diversity is very limited, with only three haplotypes in the cluster analysis (Figure A.4 in Appendix A). Interestingly, in South Asian and European populations Tajima's D has relatively high values, probably because of the differentiated haplotypes of the chromosomes with the ancestral orientation. Inversion HsInv0004 shows low Tajima's D values and has high frequencies in all populations. However, since the inversion is estimated to be old, the Tajima's D signature could be caused by the reduced variation in the derived orientation (from the absence of recombination rather than from a fast increase in frequency). Finally, inversion HsInv0063 has a less clear but special pattern. It is the only inversion with a negative estimate of the age, probably because the similarity of the inverted haplotypes to the ancestral orientation ones (Figure A.4 in Appendix A), coupled to a high overall diversity of the region. Despite the likely recent origin (although it has to pre-date out-of-Africa expansion), all non-African populations have frequencies over 50%, suggesting a fast increase in frequency.

Several NAHR inversions have also large frequency differences in some populations. However, the interpretation is more complicated when the inversion has appeared multiple times in different haplotypes. For instance, inversions HsInv0389 and HsInv0340 have very different frequencies in African and non-African populations, but it is difficult to know which orientation has suffered the frequency change and is leading the differences, since the ancestral orientation is unknown. In both cases the inversions include genes that are moved, and HsInv0340 additionally disrupts a non-coding RNA. Similarly,

inversion HsInv0114 has higher frequencies in Africa. However, East Asian populations have Tajima's D values more consistent with recent expansion or positive selection, which could indicate that the most frequent orientation in East Asia was recently at much lower frequencies. Low tajima's D are also observed in inversion HsInv0397 for the populations with apparently lower frequencies, which could have an equivalent interpretation. Finally, two more inversions stand out because of their high population differentiation. Inversion HsInv0266 is at high frequency at GIH, although it has low Tajima's D at most population, meaning that frequencies may have increased in other populations too. Inversion HsInv0124 has higher frequency in Europe. Interestingly, this inversion moves *IFITM1* gene, and two more members of the same gene family are located close to the breakpoints.

# Chapter 4

# Discussion

The subtle but powerful effect of inversions in natural populations has fascinated evolutionary biologists for a century. However, in the current era of genomics and big data, inversions escape the grasp of current technologies and remain largely overlooked. During the last few years, the InvFEST Project has committed to address our lack of understanding about human polymorphic inversions by validating, invalidating and genotyping with targeted methods a big proportion of those predicted in the human genome. One of the most useful and complete data sets for population and evolutionary genetics has been generated by genotyping 45 common inversions in seven populations of diverse ancestry. The 550 samples chosen belong to the International HapMap Project (The International HapMap 3 Consortium 2010), and most of them are also studied in the 1000 Genomes Project (1000GP) (The 1000 Genomes Project Consortium 2015). The result is a combination of the high-quality inversion data with low-coverage sequencing data for up to 434 samples, plus other functional information for many of them.

This thesis takes advantage of the population data generated to investigate the basic evolutionary properties of human inversions, largely inaccessible until now. It represents the first detailed analysis of selective forces acting on genome-wide human polymorphic inversions and their impact in the human genome, and also offers a characterization of individual inversions for future functional candidate studies. The work confirms the fundamental differences between inversions created by different molecular mechanisms, which has implications for the design of strategies to detect new inversions and study their association to human diseases and traits. It also describes the footprints of the restricted recombination between orientations and its effect on sequence variation in the two inversion types. Finally, inversions showing patterns consistent with selection are highlighted for further characterization.

## 4.1 Human polymorphic inversions

The 45 inversions studied here allow us to learn about inversion polymorphisms in the human genome. But to what extent do they represent all inversions in the human genome? Most of the inversions here are common, with only two inversions below a frequency of 1% in the population sample. In addition, more than two thirds are smaller than 10 kbp, and only two are over 100 kbp. Finally, just above half of the inversions mediated by NAHR and the rest by other replication and repaired-based mechanisms. How are other inversions in the human genome?

### 4.1.1 Inversion frequency

We know the population frequency of only a small fraction of the inversions described, and they tend to have high frequencies. The 17q21.31 inversion (HsInv0573) is known to be segregating at different frequencies across the world, and up to 34% in South Europe (Alves et al. 2015). Frequency of 8p23.1 inversion (HsInv0501) ranges from nearly absent in America to nearly fixed in Africa (Salm et al. 2012). Inversion in 16p11.2 (HsInv0786) also has high frequencies, up to 49% in North Europe (González et al. 2014). Similar to the well-studied cases, inversions in our data set are common, with 41 out of 45 segregating in different continents. And even the least frequent inversions are found in five individuals in heterozygosis (HsInv0097 and HsInv0379), a global frequency of 0.5%.

While high frequencies could be the norm if all inversions were under some mutation-drift equilibrium, as a consequence of high recurrence rates, this seems unlikely at least for non-recurrent inversions. It is more probable that the high frequencies of the inversions studied in a targeted manner result from the strong ascertainment bias, and that, as other types of variation, most inversions in the genome are expected to be at low frequencies. The final phase of the 1000GP included 786 inversion predictions, and more than a third (292) was found only in one chromosome of the 5,008 surveyed (Figure 4.1). Similar numbers are found in other types of variants, such as deletions (Figure 4.1) (Sudmant et al. 2015). However, note that 54% of the 229 inversions characterized in the 1000GP are classified as inverted duplications (Sudmant et al. 2015), that would be considered false inversions in the InvFEST database (Martínez-Fundichely et al. 2014). Therefore, although real frequency distribution is unknown, the data set here is likely to represent a minority of high-frequency inversions, with a potential increased impact on population local nucleotide diversity and recombination than lower frequency inversions.

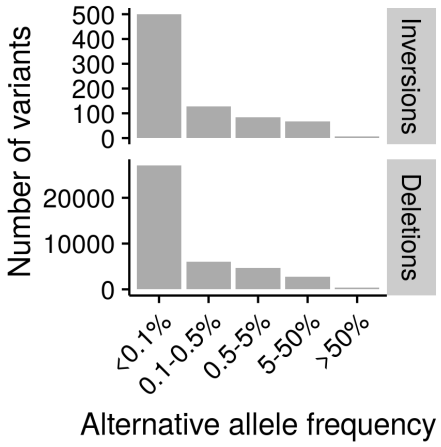


**Figure 4.1: Frequency distribution of the 1000GP inversions and deletions.** Global frequencies in the 2504 individuals from the phase 3 of the 1000GP.

### 4.1.2 Inversion size

Our knowledge about the size of human inversions is strongly conditioned by the limitations of the detection methods. Initial microscopic methods only detected megabase-long inversions. Paired-end mapping strategies can also have lower sensitivity to detect short inversions, as seen in the data used here. Long polymorphic inversions are known to be segregating in humans (e.g. the 4.5-Mbp 8p23.1 or 835-kbp inversion 17q21.31), and in the InvFEST database there are two inversions longer than 5 Mbp estimated to be at more than 1% of frequency (Martínez-Fundichely et al. 2014).

Nevertheless, the current information suggests that small inversions are more abundant than long ones. In InvFEST database, most of the validated or predicted inversions are less than 100 kbp in size (541/616, Figure 4.2 A) (Martínez-Fundichely et al. 2014). The 1000GP only studied predicted inversions within the range of 250 bp and 50 kbp (Sudmant et al. 2015), presumably to reduce false discovery rates. In any case, the size distribution within their studied range shows a clear enrichment of short inversions (Figure 4.2 C). If we focus on inversions smaller than 10 kbp in InvFEST database the pattern is similar (Figure 4.2 B). Mechanistically, it seems natural that inversions involving breakpoints at close distance are more frequently created than those inverting long stretches of genome and this is what is expected if breakpoints occur at random (Cáceres, Barbadilla, and Ruiz 1997). The inversions in the data set are all relatively short, between 83 bp and 415 kbp, although only two are over 100 kbp. Thus, they are representing the lower end of the distribution, which is probably the most abundant.

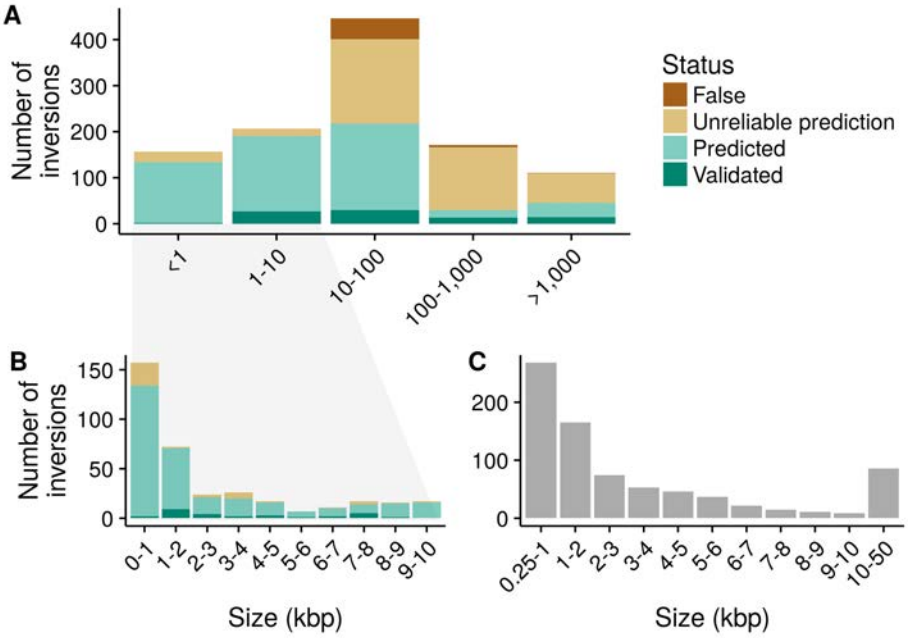


**Figure 4.2: Inversion sizes in InvFEST and the 1000GP**. A. Size of the 1,092 inversions in InvFEST. B. Detail of the size distribution of InvFEST inversions below 10 kbp. C. Size of the 786 inversions in the 1000GP phase 3. Inversion predictions with sizes smaller than 250 bp and larger than 50 kbp were filtered out in the 1000GP workflow.

### 4.1.3 Mechanisms of formation

Inversions are created by different mechanisms with different sequence signatures. While errors in replication or DNA repair can leave clean breakpoints, NAHR uses inverted repeats as a substrate for inversion generation. As a consequence, all inversions mediated by NAHR are surrounded by repetitive sequence. In particular, over half (24/45) of the inversions in the data set are likely created by NAHR, and that coincides with the proportion estimated in some studies. For example, Pang et al. (2013) analysed the mechanism of formation of 117 predicted inversions and found that 55% of them where likely created by NAHR. It was also suggested that NAHR may be more prevalent in long inversions. Therefore, the studied inversions would represent well inversion diversity.

However, the performance of detection methods based on sequence mapping is strongly reduced by ambiguous mapping (Lucas Lledó and Cáceres 2013). There are no possible signatures of split reads when the repeats are longer than the read length. In addition, only read pairs from sequence templates longer than the repeat size can span the repetitive region at the breakpoints and still detect the inversion. Thus, the real number of inversions created by NAHR could be much larger. The comparison between the inversions in the

study and those analysed by the 1000GP offers an illustration of the extent of the limitations. The 1000GP only detected 14 inversions out of the 45 studied here. From the remaining 31, six fall outside the size range considered by the 1000GP (although inversion HsInv0006 was detected, despite being $<$ 250 bp). The other 25 were missed, and an 80% of the them have inverted repeats at the breakpoints.

A clearer picture can emerge from inversions detected by Strand-seq technology, that does not rely on traditional mapping signatures to detect inversions. Instead, a single DNA strand can be sequenced at low-coverage and mapped to a reference genome, and changes in the reference strand where the reads are mapped to indicate the presence of an inversion. In the recent study were the technique was applied to detect inversions, the authors found that 77% of inversion predictions where either flanked by palindromic or non-palindromic segmental duplications or by assembly gaps (Sanders et al. 2016). The breakpoint architecture of the remaining 23% did not contain segmental duplications, but we do not know if they could have some other repetitive sequence (for instance, 8 of the 24 NAHR inversions in our data set have other forms of inverted repeats). Because of the low-coverage, the technology has higher power to detect long inversions than shorter ones. In any case, it confirms that NAHR inversions are more abundant than previously thought.

In summary, our data set is enriched in high frequencies, representing mostly short inversions (that are probably more abundant than long ones) and could have a higher NH:NAHR ratio than found in the genome. Therefore, we are representing an important fraction of all inversions, but findings may not apply to inversions at lower frequencies or longer ones.

## 4.2 Inversion effect on fertility

Inversions have the potential to decrease fertility in heterozygote carriers by generating unbalanced chromosomes in the first meiotic division. However, unbalanced chromosomes can only result from crossovers within the inverted segment, and perhaps that never happens in short inversions in the first place. Recombination could be physically impossible for inversions that are too short to create a loop to pair or that do not contain recombination hotspots or synaptic initiation points (Torgasheva and Borodin 2010). So one of our questions was, are polymorphic inversions in the human genome impacting fertility? And specifically, are the inversions in the data set impacting fertility?

In order to detect systematic selective pressures we examined the frequency distribution of inversions. The initial hypothesis was that if inversions are

deleterious, their frequencies will be lower than frequencies of neutral mutations under the same ascertainment process. The comparison with the frequency distribution of SNPs should allow us to detect differences in the underlying distribution of fitness effects. Therefore, we sampled neutral SNPs from 1000GP, simulated the ascertainment model and compared the frequency distributions.

The problem was that the comparison turned out to be more difficult than planned. Our first strategies only considered the enrichment in high frequencies due to a small and ancestrally diverse detection panel. Also, we performed forward-in-time simulations with demography and also sampled real genome variation present in a small panel. However, in both cases the expected frequencies were significantly lower than those observed in inversions. It was then when we realized that the inversion detection method was more sensitive to high frequency variants, so detected variants would be even more biased towards high frequencies. The final model is described in section 3.1.2.1 and successfully explains most of the enrichment in high frequency observed in NH inversions. We finally opted for using only data from real genome variation to implicitly correct for demographic effect on frequencies (instead of explicitly simulating it).

Frequencies of NH inversions can be completely explained by the ascertainment model, without needing to include any selective factor, and that means that we have no evidence of systematic negative selection from the frequency distribution analysis. Nevertheless, it is important to take into account that the strong ascertainment bias may affect the power to detect differences in the underlying distribution of fitness effects. Also, as seen above, we only compare common inversions with relatively small sizes, for instance, long inversions at lower frequencies are not represented, which might give different results.

On the other hand, the frequencies of NAHR inversions are still higher than those obtained from SNPs under the ascertainment simulation. However, recurrence is a more likely explanation than pervasive positive selection favouring inversions at high frequencies. Even more, these inversions could be more deleterious than shorter NH inversions and be only kept at high frequencies by recurrence, in a mutation-selection equilibrium. In that case, and bearing in mind that NAHR inversions seem to be more common in the genome, the accumulated effect on fertility of heterozygous inversions in an individual could be significant. Unfortunately, we cannot estimate this just from the frequency distribution.

A negative correlation between frequency and length could also be indicative of negative selection against inversion size (that increase the probability of recombination within the inverted region). Indeed, long inversions in the data set are at relatively low frequencies: HsInv0379 (415 kbp) 2-3% in East

Asia and HsInv0790 (202 kbp) at 3-8% in Africa. In addition, in the detailed study of inversion HsInv0379, simulations suggested that it was likely under negative selection (Puig et al. 2015a). A parallel analysis suggests that when controlling for gene content, frequency and genetic length of inversions are negatively correlated (Giner-Delgado et al. in prep.). However, the ascertainment bias against low frequency small inversions was not taken into consideration and it could explain part of the correlation observed. It is also unclear if the two longest inversions in the data set, both of which actually break genes, could be leading the correlation, with only inversions over a certain size (say ∼100 kbp) being under negative selection because of the creation of unbalanced chromosomes in meiosis.

In summary, the deleterious effect of recombination within inversions in heterozygotes is not clear in the frequency data. Nevertheless, we do not rule out a subtle cumulative impact fertility of all heterozygous inversions carried by an individual, that should be tested in an independent sample. Currently, studies are under way to use available fertility data on big cohorts with SNP genotype data to try to determine the effect of inversions and obtain a more clear answer to that question.

## 4.3 Recombination inhibition

It is clear that inversions inhibit recombination between alleles in opposite orientations, either physically or indirectly from the formation of unbalanced chromosomes. This alteration of the recombination can impact on the neighbouring variation at a population level. We have studied the recombination inhibition and the consequences using computer simulations and directly on the inversion data set. In order to have a reference of expected recombination levels, we have also compared the measures from the inversion regions with other regions of the genome. What have we learnt?

### 4.3.1 Limitations of low-coverage sequencing data

The ideal data to study the recombination patterns in inversion regions would be high-quality and phase-resolved genotypes for all polymorphisms in the region (including the inversion) in a large sample of individuals. Here we are using reliable inversion genotypes and low-coverage sequence data for 434 individuals. While it is a remarkable data set, there are several sources of error that we need to control for.

The 1000GP sequences are probabilistically genotyped and phased in a complex pipeline that has been optimized for the short-read data (The 1000

Genomes Project Consortium 2015). However, the cost-effective low-coverage strategy implies higher genotype and phase error rates than other technologies. According to the authors, the average phasing error is every 1,062 kbp, and some times it represents just a flip error (just one variant with alleles in opposite haplotypes) (The 1000 Genomes Project Consortium 2015). They also report that when trusting the most likely genotype (genotype hard coding), the errors are higher in heterozygote calls than in homozygote calls. In biallelic SNPs, genotype concordance with SNP arrays is 98.8% for REF/REF, 90.0% for REF/ALT and 99.0% for ALT/ALT (The 1000 Genomes Project Consortium 2015). Additionally, it is known that the error rates are not uniform across the genome.

As already mentioned, the 1000GP provide genome accessibility masks that define those regions where the short-read technology can be used with confidence, which define as accessible the 76.9% (strict mask) and the 95.9% (pilot mask) of the non-N fraction, respectively (The 1000 Genomes Project Consortium 2015). Unfortunately, inversions are located in regions of difficult access to short-reads. Even when excluding the inverted repeats and indels at the breakpoints, the average accessible region inside the inversion is 50.3% for the strict mask, clearly lower than the genome average. For the pilot mask the difference is smaller, an average of 93.5% of the inverted region is accessible, although this proportion is highly variable between inversions (shown in Figure A.3 in Appendix A). Also, a misaligned read extending through the breakpoint could generate a false variant at either side of the breakpoint. Therefore, we have to take into consideration the possible error sources in order to reach conclusions about the recombination and sequence variation in inversion regions.

During the different analyses, care was taken to minimize the impact of the data errors on the results. From the beginning we considered that the inverted repeats and indels associated to the breakpoints could have a high genotype error rate. Thus, we removed the inverted repeats to avoid false variants created from the mapping of reads from the other copy. In addition, we removed the deletion regions because SNPs called and imputed there have the wrong haploidy for individuals with the deletion (two alleles are always called although there is only one copy of the region). At the same time, that also removes a big part of the sequence flanking the breakpoints, reducing errors caused by read misalignments.

The nucleotide diversity was first assessed with a sliding window approach using the R package `PopGenome` (Pfeifer et al. 2014), and we chose to include initially all positions and filter windows with less than 80% of accessible positions according to the pilot mask. Although practical to implement, the approach includes noise and misses useful information of windows that fall under the threshold. Therefore, for the other sequence analyses we decided to finely select the regions included, by limiting the measures to the acces-

sible regions only. We also adopted a two-step approach, first using less information but more reliable, and later using as much information as possible. In the first step we only used the genotype information, without relying on the phase, and only for variants accessible according to the strict mask. Later we also included phase data and lowered the threshold to the pilot accessibility mask in order to include information for as many inversions as possible. This careful analysis made sure that we could extract the more reliable conclusions possible from the available data.

### 4.3.2 Genetic flux measure

If homologue chromosomes in a heterozygous individual create a loop in the inversion region, viable recombination can take place only as gene conversion or double crossover. Several factors have been suggested to control the loop formation (Torgasheva and Borodin 2010), and size is an important one. The detection of genetic exchange between orientations (or genetic flux) suggests that homologues synapse and are able to recombine. While all inversions irrespective of their mutational dynamics are subjected to the same recombination processes, the recombination footprints are easier to detect in inversions with a unique origin. In recurrent inversions, similar haplotypes are commonly found in both orientations, which makes almost impossible the identification of inter-orientation recombination events. And even if a haplotype only seems to have one orientation, we cannot be certain that the haplotype exists in the opposite orientation in the unsampled population. Therefore, evidences of recombination between orientations were analysed only in NH inversions, all with a unique origin.

The distribution of linked variants suggest that recombination between orientations is inhibited throughout the length of the NH inversions in the data set. And if recombination does take place, it has to be at very low rates. In some cases, we have identified a few shared variants between orientations, but nearly all of them are located in the regions tagged as non-accessible by the strict mask of the 1000GP. The only NH inversion with a shared variant unlikely to be a genotype error is the 12.7-kbp inversion HsInv0063. It is unclear if this variant comes from genetic flux between orientations, they are independent mutations or sequencing error in multiple reads. The shared variant, rs74405082, is located at 577 bp from the 3' breakpoint and only present in three individuals. On one hand, its proximity to the breakpoint suggests that homologous paring and recombination during synapsis may be difficult, if not impossible. But on the other, if a gene conversion event had taken place in the shared SNP region, the tract could be as long as 647 bp without creating any additional shared SNP. Gene conversion tracts in humans have been estimated to have a wide range of lengths, from few base-pairs to kilobases (Williams et al. 2015), so the possible lengths here would

be within normal values.

The similarity between nucleotide diversity levels in NH inversions and those predicted by InvertFREGENE simulations (that do not consider recombination between orientations) also suggests that genetic flux between orientations is not an important factor in the local variation patterns. Small rates of genetic flux between orientations would be expected to increase nucleotide diversity within derived inverted chromosomes (Navarro, Barbadilla, and Ruiz 2000; Guerrero, Rousset, and Kirkpatrick 2012). Here we observe a strongly reduced nucleotide diversity in inverted chromosomes, which is consistent with variation only arising from new mutations. The increase in the total nucleotide diversity of the region and the divergence between orientations point to the same direction too.

The apparent lack of genetic flux may be caused by an absence of synapsis, generating a bubble (asynapsis) or chromosome pairing without homology (heterosynapsis), physically inhibiting the initiation of recombination. Alternatively, the probability of recombination could be very low in short inversions, meaning that recombination may never happen even without an inversion. However, that is probably not the case in most inversions, because we observe the consequences of recombination inhibition in the increase of total nucleotide diversity. For instance, inversion HsInv0041 has the highest increase in total nucleotide diversity with respect to the flanking regions, and it is very short (107 bp), frequent (global frequency 50%) and with an estimated age of $\sim$1 million years.

The absent (or nearly absent) genetic flux found here contrasts with the patterns in longer inversions in *Drosophila* species, where a complete inhibition of recombination is normally detected only within a short distance from the breakpoints (ad shown in *D. melanogaster* for example in Corbett-Detig and Hartl (2012)). While we can assume that low levels of genetic flux may be the norm in short human inversions, we can not extrapolate the observations to longer inversions in the human genome. Indeed, gene conversion or double crossovers have been reported in inversion 17q21.31 (HsInv0573) (Zody et al. 2008; Steinberg et al. 2012). And it has also been shown that recombination between orientations does take place in some cytogenetically visible inversions (Anton et al. 2005).

### 4.3.3 Recombination past the breakpoints

Inversions do affect recombination within the inverted region, but their impact on the neighbouring sequence is unclear. In theory, inversions could either reduce recombination or increase it. A reduction could be caused, for instance by creating physical stress on the synapsing chromosomes in asynap-

sis or creating a loop. The extension of the recombination inhibition seems to happen in some species where the divergence associated to the limited recombination extends as much as 2 Mbp away (Machado, Haselkorn, and Noor 2007; McGaugh and Noor 2012). The opposite scenario, an increase of recombination outside of the inversion, is also possible. If total number of recombination events are regulated in chromosome or cell, recombination that can not take place within the inversion could relocate to the flanking regions. The extreme situation of an effect on other chromosomes is known as interchromosomal effect and it has been described in some cytogenetically visible inversions (Anton et al. 2005).

The results obtained here do not show strong support for either effect. However, the trend is towards a moderate extension of the recombination to the flanking regions. We compared several measures of recombination in other regions in the genome using SNPs at the same frequency as the inversions. Again, recurrence in NAHR inversions confound the pattern, so the recombination levels can be interpreted easier in NH inversion. The proportion of NH inversions with nearby variants in complete linkage disequilibrium (tag variants) is slightly higher than the proportion found in genome-wide SNPs, after removing those within the inverted region. The proportion of fixed SNPs in the flanking regions and the distance until the first exchanged variant (non-recombining region) are also higher than the average found in the same-frequency SNPs. Still, the difference falls within the normal differences expected in the relatively small number of inversions tested.

It is important to note that we did not control for recombination rates, only for variant frequency. Therefore, the higher linkage disequilibrium with the flanking regions could be a consequence of inversions being preferentially located in regions of low recombination. This may be an important factor if inversions in the size range tested have a small but deleterious effect from the creation of unbalanced chromosomes in meiosis, that favours those in low-recombining regions.

Another consideration is the possible presence of false tag variants created by misalignments of reads across the breakpoints. In the analyses, we are excluding the regions with deletions, insertions or microhomology at the breakpoints. Thus, we are removing an important fraction of the regions susceptible to misalignments. Nevertheless, the problem could still exist for inversions with clean breakpoints or with insertions in the alternative alleles. We inspected manually some of the fixed SNPs close to the breakpoints (e.g. rs557593764 at only four positions before the breakpoint of inversion HsInv0097), and we did not find evidences of misalignment. What was clearly visible was the start of the inversion breakpoint as an abrupt end of several read alignments at the first discordant position. In any case, this should affect only the variants closer to the breakpoints and its impact in the measures used should be minimal.

### 4.3.4 Additional resources

All our analyses of recombination patterns in inversion regions are based on 1000GP phase 3 data, and mostly on the hard-coded genotypes (analysing genotypes as certain values, instead of probabilities). Although valuable information can be obtained, the low coverage data means that genotype and phase errors can miss variation in repetitive regions or confound individual recombination events. Several alternative approaches could complement the results obtained here and get further insight into recombination and altered diversity levels in inversion regions.

First, the 1000GP data could be re-analysed using our information about the inversion breakpoints and sequence in the alternative orientation. Reads in the region (and possibly also those unmapped) could be re-mapped to the two alternative references to identify more accurately variants around the breakpoints. The strategy would be similar to the *reference-assisted reassembly* used by Corbett-Detig and Hartl (2012) or the breakpoint library-based inversion genotyping used by Breakseq software (Lam et al. 2009). Also, instead of using the SNP and indel genotypes, we could use genotype likelihoods with tools such as ANGSD (Korneliussen, Albrechtsen, and Nielsen 2014).

Second, deep sequencing and long-read sequencing are available for at least eight individuals genotyped in this study. The 1000GP also sequenced at high coverage (30X) some individuals used here (The 1000 Genomes Project Consortium 2015). Among them, a CEU trio (NA12891, NA12892 and NA12878), a YRI trio (NA19238, NA19239 and NA19240), a TSI female (NA20502), and a GIH male (NA20845). Additionally, the YRI trio has been analysed by a diverse array of technologies, including long reads and strand-seq within The Human Genome Structural Variation Consortium, for a study that has not yet been published (`http://www.internationalgenome.org/data-portal/data-collection/structural-variation`). The offspring of the CEU trio has also been analysed by the Genome in a Bottle Consortium using different sequencing technologies (Zook et al. 2014) and the entire trio is included in the pedigree analysed in the Illumina Platinum study, that obtained a high-coverage phased data (Eberle et al. 2017). Finally, new data generated from the 1000GP is going to be integrated by the International Genome Sample Resource (Clarke et al. 2017). Thus, all these available data from the same individuals genotyped experimentally here could be used to refine the classification of sequence variants present in inversion haplotypes and to study in greater detail single recombination events.

Finally, tag variants can be used to infer inversion genotypes in other well-characterized populations or families. It would be of special interest to use data sets where current meiotic recombination has been directly measured.

Linkage disequilibrium patterns can be used to measure recombination (Auton and McVean 2007), but one is limited to the historical recombination, as an average of the footprints of all past recombinations in the ancestors. But in inversions, the estimated recombination is an average of the free recombination that has taken place through time between chromosomes of the same orientation and the inhibited recombination between chromosomes in opposite orientations. Thus, it is likely to underestimate the real recombination rate. Therefore, in order to get a good estimate of the recombination rate within chromosomes in the same orientation, we would have to measure it directly in a homozygous individual. Promising techniques are being developed that allow for detailed individual maps of recombination or initiation of recombination (Pratto et al. 2014; Ottolini et al. 2015), which are likely to allow this type of analyses in the near future.

## 4.4 Recurrence

Most NAHR inversions show signs of recurrence in the human populations analysed. In particular, diverse evidences point towards a widespread recurrence of inversions flanked by inverted repeats. First, their frequency distribution shows an excess of high frequency and cosmopolitan inversions. Second, linkage disequilibrium levels with neighbouring variants are much lower than those in SNPs at the same frequency. Third, most of the times there is virtually no divergence between orientations and none of the two orientations show a strongly reduced nucleotide diversity as found in NH inversions. Finally, the two orientations are mixed in the haplotype clustering with many shared variants and haplotypes are frequently shared between orientations. There are a few exceptions and inversions with insufficient data, but overall the pattern is clear and contrasts strongly with that of NH mediated inversions.

Two of the exceptions are inversions HsInv0040 and HsInv0790, which have patterns completely consistent with a unique origin of all inverted chromosomes. There are tag variants, an absence of shared variants between orientations and the haplotypes in opposite orientations are clearly differentiated. Inversion HsInv0061 is at low frequency in non-African populations and it is also consistent with a unique origin. The 11 inversion heterozygotes have at least one chromosome with the most frequent haplotype in the inverted region (that is the same as the reference genome). Therefore, all inverted chromosomes are likely to have the same sequence, which has not diverged from the sequences in the ancestral orientation. The unique origin of inversion HsInv0061 cannot be ruled out, given that it does not have shared variants in the accessible sequence. Finally, in inversions HsInv0030 and HsInv0072, despite having shared variants between orientations, all homozy-

gotes for the minor orientation have similar haplotypes and a unique origin can not be ruled out. All the remaining 19 NAHR inversions show clear signs of recurrence.

The results here confirm and expand those found in Aguado et al. (2014), where 14 NAHR inversions were studied in a smaller sample of the CEU population. In that study four inversions were found to be recurrent, four inversions had non-conclusive data but seemed to point to recurrence, four were considered compatible with a unique origin and two did not have enough data. Now we have re-analysed 13 of the 14 inversions there (inversion HsInv0286 is not included in the 45-inversion data set), and in all cases we have found clear signatures of recurrence. That includes HsInv0114, HsInv0347 and HsInv0396, that were considered unique in the CEU sample. In Aguado et al. (2014), those three inversions had SNPs in complete linkage ($r^2 = 1$, HsInv0114 and HsInv0347) or ($r^2 \geq 0.9$, HsInv0396) high disequilibrium with the inversion. However, since the study was only assessing the association in CEU individuals, some inversion recurrence events in other populations or at low frequency in CEU were missed. In the new analysis, none of the three inversions had global perfect tag SNPs, but we still found CEU-specific perfect tag variants for HsInv0114 and HsInv0347, and variants in high linkage disequilibrium for HsInv0396 (Table B.4 in Appendix B), explaining the previous results. Likewise, it could be possible that recurrence events are discovered for inversions with an apparent unique origin (such as HsInv0040) when new populations and more individuals are studied.

The fact that 14 out of the 24 inversions are polymorphic in chimpanzee or gorilla also points towards recurrence as the norm for NAHR inversions. While shared polymorphisms between humans and other apes exist, they are only a minority and are associated with balancing selection maintaining them (Leffler et al. 2013). Therefore, the vast majority of polymorphic inversions present in different species probably represent genomic regions that have been inverting recurrently throughout millions of years. Indeed, inversion HsInv0389 has been shown to be recurrent in mammals, where the inverted duplications remain highly identical through homogenizing gene conversion between copies (Cáceres et al. 2007). It would be interesting to check if others of the studied inversions show a similar pattern of long range recurrence.

### 4.4.1 Recurrence determinants

Recurrence is prevalent in NAHR inversions, but independent inversion events seem to appear at different rates in different inversions. We find inversions with a single event (e.g. HsInv0040 and HsInv0790), others with a main inversion event at high frequency and a few secondary ones (e.g. HsInv0114, HsInv0124 or HsInv0389), and inversions with all kinds of haplotypes found

in either orientation (e.g. HsInv0241 or HsInv0344). What makes a sequence invert at higher or lower rates?

Inverted repeat size and identity are clear candidates to play a role in recurrence, increasing the opportunities for recombination to happen. In addition, the distance between the inverted repeats could be an important factor. It has been noted that inversion and repeat sizes show a positive correlation in humans (Sanders et al. 2016; Shao et al. 2017). That is, distant repeats may need to be longer in order to recombine and create an inversion. Sanders et al. (2016) also noted that the distribution of inversions was uneven across the genome, and suggested that inversion generation could be inhibited in some chromosomes were they do not observe inversions, such as chromosome 13 and 18. While we do have inversions in chromosome 13 (HsInv0340 and HsInv0341) –but not in chromosome 18–, their observations could be related to different inverted repeat abundance between chromosomes. Finally, three-dimensional genome organization together with proximity to recombination hotspots and hotspot strength are expected to strongly influence NAHR events, as it has been described for structural variants in general (Mills et al. 2011; Escaramís, Docampo, and Rabionet 2015).

Inversions studied here support the notion that inversion and repeat sizes are important factors of recurrence. The two inversions where all the chromosomes come from a single event have especially small repeats (HsInv0040) or are especially long (HsInv0790) (Table 4.1). In the other extreme, we find highly-recurrent inversions, such as HsInv0241 and HsInv0344. In these cases, despite having intermediate repeat and inversion sizes, repeats are large in proportion to the inversion size. Therefore, the size ratio between inversion and inverted repeats seems the most important factor of the ones analysed here. There is no clear trend regarding the average identity between repeats (Table 4.1), which is perhaps explained by the non-uniform identity across the repeat length, that usually have long central regions completely identical in all cases.

Other inversions seem to follow the same trends. For instance, inversion HsInv0069 has the highest repeat size/inversion size ratio and the highest fraction of shared variants inside. And inversions HsInv0030 and HsInv0061 have low ratio, and consistently, they are candidates to have a unique origin. In order to properly measure the relative importance of each factor, we would need a comparable measure of recurrence, that we do not currently have. Independent events are difficult to identify, since recombination between a newly inverted haplotype and a previously inverted haplotype can create new combinations. Also, our ability to differentiate haplotypes, and thus separate events, depends on the diversity in the region. New inversion events do not leave any associated signature in the sequence either, that would allow to identify them. As alternatives to directly counting the mutation events, some proxy measure could be used instead. Number of shared haplotypes, shared

**Table 4.1: Inversion architecture and recurrence.**

| Inversion | Chr | Inversion size (kbp) | Repeat size (kbp) | Identity (%) | Size ratio |
|---|---|---|---|---|---|
| *Unique origin* | | | | | |
| HsInv0040 | chr2 | 3.59 | 0.76 | 99.9 | 0.21 |
| HsInv0790 | chr17 | 201.54 | 24.29 | 98.9 | 0.12 |
| *Recurrent* | | | | | |
| HsInv0114 | chr9 | 11.93 | 2.73 | 99.9 | 0.23 |
| HsInv0124 | chr11 | 6.11 | 3.78 | 97.1 | 0.62 |
| HsInv0389 | chrX | 37.61 | 11.36 | 99.2 | 0.30 |
| *Highly recurrent* | | | | | |
| HsInv0241 | chr2 | 9.85 | 13.06 | 98.5 | 1.33 |
| HsInv0344 | chr14 | 7.39 | 7.23 | 99.7 | 0.98 |

variants, or maximum $r^2$ with neighbouring variants are necessarily reflecting recurrence and are correlated. However, random genealogies and the size of the inversion (and of the region accessible to sequencing technologies) do affect those measures as well, making them difficult to compare between inversions.

On top of the other confounding factors, composition of the studied sample could affect recurrence estimates if recurrence rate is not constant across populations. Large differences between population frequencies of some inversions could be caused by different recurrence rate. The most clear case is inversion HsInv0340, that seems to have many inversion events in African populations (Figure A.4 in Appendix A) and has frequencies around 50% there, whereas in the other populations most individuals carry the same orientation. We hypothesise that differences in recombination hotspots between populations (Pratto et al. 2014) or inverted repeat identity and composition could drive the recurrence rate variation.

Inversion HsInv0832 is a special case that allows us to determine individual inversion events, thanks to the absence of recombination in the region of the chromosome Y where it is located. Using variation around the inversion we are able to identify at least five inversion events. The same number is obtained when using chromosome Y haplogroup information published elsewhere for a large fraction of the studied individuals (Giner-Delgado et al. in prep.). Knowing the time to the most recent common ancestor of the sample (∼80,000 years ago, knowing that we have represented the haplogroups B2-T (Poznik et al. 2016)), one can estimate a global recurrence rate of $\sim 6.25 \times 10^{-5}$ inversions per year (one inversion every 16,000 years).

Similar estimates for other inversions are not yet available, but a compre-

hensive overview with comparable recurrence measures in different inversions should help to clarify the mechanisms underlying inversion recurrence. In turn, the knowledge could be extrapolated to gain insight into the risk of diseases caused by recurrent inversions (such as some cases of hemophilia A (Bagnall et al. 2002)) or associated secondary pathogenic rearrangements (reviewed in Puig et al. (2015b)).

## 4.5 Functional effects and selection

An inversion can impact the genomic region where it occurs in multiple ways, from a direct effect disrupting or creating new elements at the breakpoints, to the protection of an optimized haplotype from recombination. Median inversion size in the data set is 4.1 kbp and the median human gene annotated by RefSeq spans 23.6 kbp. Therefore, the recombination-reducing mechanisms, may not play an important role in short inversions. Instead, it is more likely that potentially functional small inversions have a direct impact on the breakpoint sequence or modify the function of some element by changing its position and orientation. However, not all inversions are going to have functional consequences and the frequency distribution of NH inversions suggests that an important fraction could be neutral.

### 4.5.1 Functional candidates

Determining the real functional effect of any variant, and inversions in particular, is not trivial. Some of the inversions are clear candidates to have an impact on the organism, given that they change the sequence of known functional elements. Inversion HsInv0379 is the most clear case and has been studied in depth in Puig et al. (2015a). In this case the inversion separates the promoter and the first coding exon of a zinc finger gene from the rest. The consequences are not only a strong reduction in expression of the affected gene, but also mild changes on expression of other genes, probably regulated by the disrupted transcription factor (Puig et al. 2015a). In addition, the moved exon creates a new fusion transcript with repetitive sequences located in the other end of the inversion (Puig et al. 2015a). There are other candidates with clear effects still to characterize in detail. For example, inversions HsInv0340 and HsInv0790 break a long intergenic non-protein coding RNA and an expressed pseudogene. The derived allele of inversion HsInv0201 is accompanied by a deletion of a coding exon of protein coding gene *SPINK14*. And inversion HsInv0102 inverts a non-coding exon of the gene encoding small G-protein *RHOH*.

Other inversions may affect regulatory elements and indirectly impact gene

expression too. Several inversions created by NAHR have highly identical genes embedded into the repeats (such as HsInv0030 or HsInv0241). Since they are very similar, in some cases the gene sequence remains the same after the inversion but regulatory elements may be exchanged. Inversions located within gene introns are also potential modifiers of transcription and splicing. Finally, genes within long inversions that get inverted can be relocated in a different regulatory environment with altered three-dimensional contacts.

In order to find less evident potential effects, inversion genotypes can be statistically associated to genome-wide gene expression or to other phenotypic traits in large cohorts. In that regard, some of the studied samples are also part of Geuvadis project (Lappalainen et al. 2013), and the direct association between inversion genotype and gene expression has been assessed by other members of the group. Notably, six inversions were found to have some effect on neighbouring or distant genes (Giner-Delgado et al. in prep.). Additionally, the inversion tag variants (Table B.4 in Appendix B) can also be used to screen their associations in other populations without direct inversion genotypes. In the context of the 45-inversion project, inversion tag SNPs have been screened in expression quantitative trait loci (eQTLs) described in the GTEX project (Lonsdale et al. 2013) and genome-wide association studies (GWAS) hits from repositories such a the GWAS Catalog (MacArthur et al. 2017), finding again those functional candidates from Geuvadis with tag SNPs and adding 11 more inversions with some effect (Giner-Delgado et al. in prep.).

However, now we know that most inversions are likely to be missed in SNP-based GWAS. When we checked the ability to tag the inversions with SNPs in commonly used arrays, we found that in the best array nearly half of the inversions did not have any SNP at linkage disequilibrium $r^2 \geqslant 0.8$. And even then, only seven inversions would be perfectly tagged. The consequence of the low representation, especially important in recurrent inversions, is that any genetic contribution to the studied traits has been missed. Perhaps the inclusion of inversion genotypes in association studies could partly help to explain the missing heritability of some traits (Eichler et al. 2010). It can have similar consequences for eQLT studies as well. The association of inversion genotypes with gene expression changes is unlikely to be captured by neighbouring SNPs in recurrent inversions. In other words, we may be missing a large fraction of important associations.

### 4.5.2 Signatures of selection

Another indirect strategy to find functional candidates is to detect the signatures left by natural selection. Since selection acts on phenotypes, footprints on the sequence reveal functional elements and beneficial variants. This

thesis has contributed to this second approach.

Since inversions studied here are mostly at high frequency, we have focused on two modes of selection that could apply to our polymorphisms: positive selection on standing variation and long-term balancing selection. In order to identify inversions that could be evolving or have evolved under the considered scenarios, we have explored three signatures: unusual frequency differentiation measured in terms of $F_{ST}$ compared to that observed in genome-wide SNPs, skewness of nucleotide allele frequency distribution measured by Tajima's D statistic, and the speed of inversion frequency changes by the comparison with its age.

However, inversion characteristics have to be taken into account to interpret the different signatures. We have seen with simulations that inversion's special recombination can affect the values of Tajima's D in inversions with a unique origin: old inversions at intermediate frequencies can create balancing selection signatures and inversions at high frequency can create signatures of positive selection. Therefore, Tajima's D signatures have to be taken with caution and interpreted together with the age and frequency.

With those limitations in mind and combining different signatures, we identified four NH inversions that could be under balancing selection and three under positive selection. The balancing-selection candidates are: inversions HsInv0031, HsInv0201 and HsInv0058, that have old origins, intermediate frequencies and little differentiation; and inversion HsInv0102, that has high Tajima's D despite being at relatively low frequencies. Tajima's D values for the three old inversions do not add much information, because it is what it would be expected from the accumulated divergence. In inversion HsInv0102 the target of selection could be something else, given that there are two differentiated haplotypes in the ancestral orientation. We propose that if balancing selection is maintaining diversity in a region, perhaps the inversion and its inhibition of recombination is being favoured to protect the haplotype diversity. The NH candidates to be under positive selection in some population are: inversions HsInv0006, HsInv0059, with high population differentiation; and inversion HsInv0063, frequent and young with low Tajima's D. Inversion HsInv0006 seems to be positively selected in African populations, HsInv0059 in East Asia and HsInv0063 in non-African populations.

Signal interpretation in recurrent inversions present even more challenges, given that each orientation represents an amalgam of independent mutations. Thus, differences in frequency between population may not reflect differences in selective pressures on the inversion but rather differences in recurrence rate (that could be itself under selection). Also, the age cannot be estimated as a single absolute measure, and even if we were able to identify independent inversion events, we could not estimate the age based on the divergence between orientations.

Low population differentiation and frequencies around 50% of inversions HsInv0069 and HsInv0344, which could be considered signatures of balancing selection, are probably consequence of high recurrence rates of these two inversions, that perhaps keep them at an equilibrium frequency. Inversion HsInv0069 is especially curious, with population frequencies forming two clusters above and below 50%. The unusual thing is that pairs of populations of the same super-population fall in different clusters, with unusually high $F_{ST}$, as if geographical proximity did not count and instead all were in an equilibrium frequency of the population minor allele at ∼40-45%. Five inversions could be regarded as candidates to be under positive selection because of their high differentiation in one population: HsInv0114, HsInv0340 and HsInv0389 in Africa; HsInv0124 in Europe; and HsInv0266 in South Asia. However, in some cases such as HsInv0340, where African populations seem to have many independent inversion events, differences could also reflect variation in recurrence activity rather than selection. Patterns in inversion HsInv0114 are also difficult to interpret because, while population differentiation would point to selection in Africa, Tajima's D values indicate that the other orientation could be the one selected. Inversions that do not seem very recurrent can be interpreted more safely. This includes HsInv0124, where a main inversion event is at high frequency in Europeans, inversion HsInv0266, where most inverted chromosomes carry the same haplotype that may have increased in frequency fast; and perhaps HsInv0389, although more independent inversion events are visible and it is difficult to tell if frequency has increased in African populations or in non-African populations (since the last main ancestral orientation in humans is unknown).

In summary, inversions HsInv0031, Hsnv0058 and HsInv0201 are candidates to be under long-term balancing selection, inversion HsInv0102 may be in a region under balancing selection, and HsInv0006, HsInv0059, HsInv0063, HsInv0124, HsInv0266 and HsInv0389 are candidates to be under positive selection in some population. These ten inversions would deserve further characterization, and six of them have a candidate gene to be involved in the selected effect. Notably, eight out of the ten candidates are among the 17 inversions with some gene expression change, eQTL or GWAS hit found independently by other members of the group (Giner-Delgado et al. in prep.), what would not be expected by chance if we were just detecting random noise (10,000 permutation test, $P = 0.001$).

### 4.5.3 Neutrality tests for inversions

The study of polymorphic inversions in the human genome is still in its infancy and there is a lot of room to develop and adapt neutrality tests to inversion characteristics. For instance, the possibility of measuring the age of the mutation offers a direct mechanism to estimate the selection coefficient

of the derived allele by comparing it to the expected frequency in a diffusion model (e.g. Wiehe and Stephan (1993)). And here we have estimated the age just from the divergence between orientations, but intra-allelic variation of the inverted chromosomes can also be used to estimate the time to the most recent common ancestor of all derived chromosomes, that is a lower-bound estimate of the age (see for instance Rozas et al. (1999) or Corbett-Detig and Hartl (2012)).

With the development of an efficient simulation software capable to represent complex demographies (and possibly selection), we would be able to use measures as divergence, inverted chromosome nucleotide diversity and frequency to estimate the age (and the selection coefficient) using an approach such as the approximate bayesian computation (Sunnåker et al. 2013). We still would have the important uncertainty of the local mutation rate and other model parameters, but the power to explore different scenarios to the observed data would increase greatly.

One class of neutrality tests not used here are those related to the linkage disequilibrium, such as iHS (Voight et al. 2006). The reasoning that prevented us using it was that the inversion effect on linkage disequilibrium patterns could confound the signal. However, if we assume that for some inversions the recombination is unaffected past the breakpoints, we could still use the haplotype length of the flanking regions normally, and it would be something to explore in the future.

Finally, special frameworks for recurrent inversions are needed. If we assume that inversions in general, and NAHR inversions in particular, may have a deleterious effect on heterozygotes, the inversions with high recurrence rate could be modelled in a mutation-selection equilibrium, where selection acts against heterozygotes and symmetrical recurrence pushed inversion frequency to the 50%. Explicit models of recurrent inversions would also help to understand the complex interplay of the inhibition of recombination and multiple inversion events. Therefore, the data and the work presented here open an interesting research area in the study of the evolutionary impact of inversions in the human genome.

# Chapter 5

# Conclusions

The conclusions of this work are the following:

1. The frequency distribution of the studied inversions is that expected for neutral variants when controlling for detection biases, which indicates that they are not subjected to strong negative selection. However, inversions generated by non-allelic homologous recombination (NAHR) between inverted repeats show an enrichment of inversions at intermediate frequency.

2. Overall, inversions also show the expected levels of population differentiation, although there are several inversions with unusually high frequency differences between populations.

3. Inversions generated by NAHR show high levels of recurrence and most of them (19/24) have originated multiple times during human evolution. In contrast, all inversions generated by non-homologous mechanisms (NH) have single origins.

4. Recurrence strongly reduces linkage disequilibrium (LD) between NAHR inversions and neighbouring variants, which limits the use of tag SNPs to infer inversion genotypes.

5. Single-origin inversions have more tag variants than SNPs at the same frequency, as a consequence of recombination inhibition between orientations. No genetic flux between orientations is detected in inverted regions, suggesting that recombination is completely inhibited in heterozygotes.

6. Nucleotide diversity is strongly affected in genomic regions with single-origin inversions, as predicted by computer simulations. Older inversions tend to increase total nucleotide diversity, while younger ones at very high frequency could have the opposite effect.

7. There are no clear alterations of the LD in flanking genomic regions of single-origin inversions, which suggests that recombination outside inverted regions is not increased and inhibition could just extend a few kilobases from the breakpoint, if any.

8. The ages of single-origin inversion have been estimated from sequence divergence between orientations, with nine of them having appeared more than 500,000 years ago and the oldest around 2.5 million years ago.

9. Ten inversions are candidates to be under positive or balancing selection and deserve further characterization, and over half are located within gene regions.

# Bibliography


Abyzov, A., S. Li, D. R. Kim, M. Mohiyuddin, A. M. Stütz, et al. (2015). "Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms." *Nature communications* 6, p. 7256.

Aguado, C., M. Gayà-Vidal, S. Villatoro, M. Oliva, D. Izquierdo, 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.

Alves, J. M., A. C. Lima, I. A. Pais, N. Amir, R. Celestino, et al. (2015). "Reassessing the Evolutionary History of the 17q21 Inversion Polymorphism." *Genome biology and evolution* 7.12, pp. 3239–3248.

Andolfatto, P., F. Depaulis, and A. Navarro (2001). "Inversion polymorphisms and nucleotide variability in Drosophila". *Genetical Research* 77.1, pp. 1–8.

Andolfatto, P., J. D. Wall, and M. Kreitman (1999). "Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of Drosophila melanogaster". *Genetics* 153.3, pp. 1297–1311.

Andrés, A. M., M. J. Hubisz, A. Indap, D. G. Torgerson, J. D. Degenhardt, et al. (2009). "Targets of balancing selection in the human genome". *Molecular biology and evolution* 26.12, pp. 2755–2764.

Anton, E., J. Blanco, J. Egozcue, and F. Vidal (2005). "Sperm studies in heterozygote inversion carriers: A review". *Cytogenetic and Genome Research* 111.3-4, pp. 297–304.

Antonacci, F., J. M. Kidd, T. Marques-Bonet, M. Ventura, P. Siswara, et al. (2009). "Characterization of six human disease-associated inversion polymorphisms". *Human Molecular Genetics* 18.14, pp. 2555–2566.

Auton, A. and G. McVean (2007). "Recombination rate estimation in the presence of hotspots". *Genome Research* 17.8, pp. 1219–27.

Ayala, D., P. Acevedo, M. Pombi, I. Dia, D. Boccolini, et al. (2017). "Chromosome inversions and ecological plasticity in the main African malaria mosquitoes". *Evolution* 71.3, pp. 686–701.

Bachtrog, D. (2013). "Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration". *Nature Reviews Genetics* 14.2, pp. 113–124.

Bagnall, R. D., N. Waseem, P. M. Green, and F. Giannelli (2002). "Recurrent inversion breaking intron 1 of the factor VIII gene is a frequent cause of severe hemophilia A". *Blood* 99.1, pp. 168–174.

Bank, C., G. B. Ewing, A. Ferrer-admettla, M. Foll, and J. D. Jensen (2014). "Thinking too positive? Revisiting current methods of population genetic selection inference". *Trends in Genetics* 30.12, pp. 540–546.

Baudat, F., Y. Imai, and B. de Massy (2013). "Meiotic recombination in mammals: localization and regulation." *Nature reviews. Genetics* 14.11, pp. 794–806.

Beaumont, M. A. and D. J. Balding (2004). "Identifying adaptive genetic divergence among populations from genome scans". *Molecular Ecology* 13.4, pp. 969–980.

Berg, J. J. and G. Coop (2014). "A population genetic signal of polygenic adaptation". *PLoS Genetics* 10.8, e1004412.

Boyle, E. A., Y. I. Li, and J. K. Pritchard (2017). "An expanded view of complex traits: from polygenic to omnigenic". *Cell* 169.7, pp. 1177–1186.

Brandvain, Y. and S. I. Wright (2016). "The Limits of Natural Selection in a Nonequilibrium World". *Trends in Genetics* 32.4, pp. 201–210.

Cáceres, A. and J. R. González (2015). "Following the footprints of polymorphic inversions on SNP data: from detection to association tests". *Nucleic acids research* 43.8, e53.

Cáceres, A., S. S. Sindi, B. J. Raphael, M. Cáceres, and J. R. González (2012). "Identification of polymorphic inversions from genotypes." *BMC bioinformatics* 13, p. 28.

Cáceres, M., A. Barbadilla, and A. Ruiz (1997). "Inversion length and breakpoint distribution in the Drosophila buzzatii species complex: Is inversion length a selected trait?" *Evolution* 51.4, pp. 1149–1155.

Cáceres, M., J. C. McDowell, J. Gupta, S. Brooks, G. G. Bouffard, et al. (2007). "A recurrent inversion on the eutherian X chromosome". *Proceedings of the National Academy of Sciences of the United States of America* 104.47, pp. 18571–18576.

Cáceres, M., S. Villatoro, and C. Aguado (2015). *Inverse multiplex ligation-dependent probe amplification (iMLPA), an in vitro method of genotyping multiple inversions*. US Patent App. 14/647,718.

Cáceres, M. (2010). *INVFEST - Evolutionary and functional analysis of polymorphic inversions in the human genome*. URL: `http://cordis.europa.eu/project/rcn/93873_en.html` (visited on 08/01/2017).

Cao, H., H. Wu, R. Luo, S. Huang, Y. Sun, et al. (2015). "De novo assembly of a haplotype-resolved human genome". *Nature Biotechnology* 33.6, pp. 617–622.

Carr, D. (1962). "Chromosomal anomalies with special reference to Klinefelter's syndrome." *Transactions of the American Association of Genito-Urinary Surgeons* 54, pp. 9–14.

Carvalho, C. M. B. and J. R. Lupski (2016). "Mechanisms underlying structural variant formation in genomic disorders." *Nature reviews. Genetics* 17.4, pp. 224–238.

Casillas, S. and A. Barbadilla (2017). "Molecular population genetics". *Genetics* 205.3, pp. 1003–1035.

Chaisson, M. J. P., J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, et al. (2015). "Resolving the complexity of the human genome using single-molecule sequencing". *Nature* 517.7536, pp. 608–611.

Cheng, E. Y., Y. J. Chen, C. M. Disteche, and S. M. Gartler (1999). "Analysis of a paracentric inversion in human oocytes: Nonhomologous pairing in pachytene". *Human Genetics* 105.3, pp. 191–196.

Chiang, C., A. J. Scott, J. R. Davis, E. K. Tsang, X. Li, et al. (2016). "The impact of structural variation on human gene expression". *Nature Genetics* 49.5, pp. 692–699.

Clarke, L., S. Fairley, X. Zheng-Bradley, I. Streeter, E. Perry, et al. (2017). "The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data". *Nucleic Acids Research* 45.D1, pp. D854–D859.

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, et al. (2010). "Origins and functional impact of copy number variation in the human genome". *Nature* 464.7289, pp. 704–712.

Corbett-Detig, R. B. and D. L. Hartl (2012). "Population genomics of inversion polymorphisms in Drosophila melanogaster". *PLoS Genetics* 8.12, e1003056.

Coyne, J. and H. Orr (2004). *Speciation*. Sunderland, MA: Sinauer Associates.

Crawford, J. E. and B. P. Lazzaro (2012). "Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data". *Frontiers in Genetics* 3, pp. 1–13.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, et al. (2011). "The variant call format and VCFtools". *Bioinformatics* 27.15, pp. 2156–2158.

Davydov, E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, et al. (2010). "Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP ++". *PLoS Computational Biology* 6.12, e1001025.

De Filippo, C., F. M. Key, S. Ghirotto, A. Benazzo, J. R. Meneu, et al. (2016). "Recent Selection Changes in Human Genes under Long-Term Balancing Selection". *Molecular Biology and Evolution* 33.6, pp. 1435–1447.

de Vries, A. and B. D. Ripley (2016). *ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'*. R package version 0.1-20. URL: `https://CRAN.R-project.org/package=ggdendro`.

DeGiorgio, M., K. E. Lohmueller, and R. Nielsen (2014). "A model-based approach for identifying signatures of ancient balancing selection in genetic data." *PLoS genetics* 10.8, e1004561.

Dobzhansky, T. (1970). *Genetics of the Evolutionary Process*. Columbia University Press.

Duret, L. and P. F. Arndt (2008). "The impact of recombination on nucleotide substitutions in the human genome". *PLoS Genetics* 4.5.

Eberle, M. A., E. Fritzilas, P. Krusche, M. Källberg, B. L. Moore, et al. (2017). "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". *Genome Research* 27.1, pp. 157–164.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." *Nature reviews. Genetics* 11.6, pp. 446–50.

Escaramís, G., E. Docampo, and R. Rabionet (2015). "A decade of structural variants: Description, history and methods to detect structural variation". *Briefings in Functional Genomics* 14.5, pp. 305–314.

Eslami Rasekh, M., G. Chiatante, M. Miroballo, J. Tang, M. Ventura, et al. (2017). "Discovery of large genomic inversions using long range information". *BMC Genomics* 18.1, p. 65.

Falconer, E., M. Hills, U. Naumann, S. S. Poon, E. A. Chavez, et al. (2012). "DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution". *Nature methods* 9.11, pp. 1107–1112.

Fan, S., M. E. B. Hansen, Y. Lo, and S. A. Tishkoff (2016). "Going global by adapting local: A review of recent human adaptation." *Science* 354.6308, pp. 54–59.

Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal, and B. Servin (2013). "Detecting signatures of selection through haplotype differentiation among hierarchically structured populations". *Genetics* 193.3, pp. 929–941.

Ferfouri, F., P. Clement, D. M. Gomes, M. Minz, E. Amar, et al. (2009). "Is classic pericentric inversion of chromosome 2 inv(2)(p11q13) associated with an increased risk of unbalanced chromosomes?" *Fertility and Sterility* 92.4, 1497.e1–4.

Ferretti, L., A. Klassmann, E. Raineri, T. Wiehe, S. E. Ramos-Onsins, et al. (2017). "The expected neutral frequency spectrum of linked sites". *bioRxiv*. DOI: 10.1101/100123.

Feuk, L., J. R. MacDonald, T. Tang, A. R. Carson, M. Li, et al. (2005). "Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies". *PLoS Genetics* 1.4, pp. 489–498.

Flores, M., L. Morales, C. Gonzaga-Jauregui, R. Domínguez-Vidaña, C. Zepeda, et al. (2007). "Recurrent DNA inversion rearrangements in the human genome". *Proceedings of the National Academy of Sciences of the United States of America* 104.15, pp. 6099–6106.

Frichot, E., S. D. Schoville, G. Bouchard, and O. François (2013). "Testing for associations between loci and environmental gradients using latent factor mixed models". *Molecular Biology and Evolution* 30.7, pp. 1687–1699.

Gardner, R. M., G. R. Sutherland, and L. G. Shaffer (2011). *Chromosome abnormalities and genetic counseling*. 61. OUP USA.

Gazave, E., L. Ma, D. Chang, A. Coventry, F. Gao, et al. (2014). "Neutral genomic regions refine models of recent rapid human population growth." *Proceedings of the National Academy of Sciences of the United States of America* 111.2, pp. 757–762.

Gibbs, R. A., J. Rogers, M. G. Katze, R. Bumgarner, G. M. Weinstock, et al. (2007). "Evolutionary and biomedical insights from the rhesus macaque genome". *Science* 316.5822, pp. 222–234.

Giner-Delgado, C., S. Villatoro, J. Lerga-Jaso, M. Gayà-Vidal, M. Oliva, et al. (in prep.). "Functional and evolutionary impact of polymorphic inversions in the human genome".

González, J. R., A. Cáceres, T. Esko, I. Cuscó, M. Puig, et al. (2014). "A Common 16p11.2 Inversion Underlies the Joint Susceptibility to Asthma and Obesity". *The American Journal of Human Genetics*.

Gordon, D., J. Huddleston, M. J. Chaisson, C. M. Hill, Z. N. Kronenberg, et al. (2016). "Long-read sequence assembly of the gorilla genome". *Science* 352.6281, aae0344.

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, et al. (2011). "Demographic history and rare allele sharing among human populations." *Proceedings of the National Academy of Sciences of the United States of America* 108.29, pp. 11983–11988.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, et al. (2010). "A Draft Sequence of the Neandertal Genome". *Science* 328.5979, pp. 710–722.

Grossman, S. R., I. Shlyakhter, I. Shylakhter, E. K. Karlsson, E. H. Byrne, et al. (2010). "A composite of multiple signals distinguishes causal variants in regions of positive selection." *Science* 327.5967, pp. 883–886.

Gu, W., F. Zhang, and J. R. Lupski (2008). "Mechanisms for human genomic rearrangements". *PathoGenetics* 1.1, p. 4.

Guerrero, R. F., F. Rousset, and M. Kirkpatrick (2012). "Coalescent patterns for chromosomal inversions in divergent populations". *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1587, pp. 430–438.

Günther, T. and G. Coop (2013). "Robust identification of local adaptation from allele frequencies". *Genetics* 195.1, pp. 205–220.

Gurdasani, D., T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, et al. (2015). "The African Genome Variation Project shapes medical genetics in Africa." *Nature* 517.7534, pp. 327–332.

Haasl, R. J. and B. A. Payseur (2016). "Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication". *Molecular Ecology* 25.1, pp. 5–23.

Harewood, L., K. Kishore, M. D. Eldridge, S. Wingett, D. Pearson, et al. (2017). "Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours". *Genome Biology* 18.1, p. 125.

Harris, R. S. (2007). "Improved pairwise alignment of genomic DNA". PhD thesis. The Pennsylvania State University.

Hasson, E. and W. F. Eanes (1996). "Contrasting histories of three gene regions associated with In(3L)Payne of Drosophila melanogaster". *Genetics* 144.4, pp. 1565–1575.

Hehir-Kwa, J. Y., T. Marschall, W. P. Kloosterman, L. C. Francioli, J. A. Baaijens, et al. (2016). "A high-quality human reference panel reveals the complexity and distribution of genomic structural variants". *Nature Communications* 7, p. 12989.

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, et al. (2011). "Classic selective sweeps were rare in recent human evolution." *Science* 331.6019, pp. 920–924.

Herrero, J., M. Muffato, K. Beal, S. Fitzgerald, L. Gordon, et al. (2016). "Ensembl comparative genomics resources". *Database* 2016.

Hoffmann, A. A. and L. H. Rieseberg (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, pp. 21–42.

Holsinger, K. E. and B. S. Weir (2009). "Genetics in geographically structured populations: defining, estimating and interpreting F(ST)." *Nature reviews. Genetics* 9.10, pp. 639–650. DOI: 10.1038/nrg2611.

Huddleston, J., M. J. Chaisson, K. M. Steinberg, W. Warren, K. Hoekzema, et al. (2017). "Discovery and genotyping of structural variation from long-read haploid genome sequence data". *Genome Research* 27.5, pp. 677–685.

Huddleston, J. and E. E. Eichler (2016). "An incomplete understanding of human genetic variation". *Genetics* 202.4, pp. 1251–1254.

Huerta-Sánchez, E., X. Jin, Asan, Z. Bianba, B. M. Peter, et al. (2014). "Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA". *Nature* 512.7513, pp. 194–197.

Hughes, A. L. and M. Nei (1988). "Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection". *Nature* 335.6186, pp. 167–170.

Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, et al. (2004). "Detection of large-scale variation in the human genome". *Nature Genetics* 36.9, pp. 949–951.

Iskow, R. C., O. Gokcumen, and C. Lee (2012). "Exploring the role of copy number variants in human adaptation". *Trends in Genetics* 28.6, pp. 245–257.

Jong, S. de, I. Chepelev, E. Janson, E. Strengman, L. H. van den Berg, et al. (2012). "Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner". *BMC Genomics* 13.1, p. 458.

Kapun, M., D. K. Fabian, J. Goudet, and T. Flatt (2016). "Genomic Evidence for Adaptive Inversion Clines in Drosophila melanogaster". *Molecular Biology and Evolution* 33.5, pp. 1317–1336.

Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, et al. (2004). "The UCSC Table Browser data retrieval tool". *Nucleic acids research* 32.Database issue, pp. D493–D496.

Kent, W. J. (2002). "BLAT - The BLAST-like alignment tool". *Genome Research* 12.4, pp. 656–664.

Kent, W. J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler (2003). "Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes". *Proceedings of the National Academy of Sciences* 100.20, pp. 11484–11489.

Key, F. M., Q. Fu, F. Romagné, M. Lachmann, and A. M. Andrés (2016). "Human adaptation and population differentiation in the light of ancient genomes". *Nature Communications* 7, p. 10775.

Key, F. M., J. C. Teixeira, C. de Filippo, and A. M. Andrés (2014). "Advantageous diversity maintained by balancing selection in humans". *Current Opinion in Genetics and Development* 29, pp. 45–51.

Kidd, J. M., G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, et al. (2008). "Mapping and sequencing of structural variation from eight human genomes". *Nature* 453.7191, pp. 56–64.

Kim, Y. and W. Stephan (2002). "Detecting a local signature of genetic hitchhiking along a recombining chromosome". *Genetics* 160.2, pp. 765–777.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences". *Journal of Molecular Evolution* 16.2, pp. 111–120.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

Kirkpatrick, M. (2010). "How and why chromosome inversions evolve". *PLoS Biology* 8.9, e1000501.

Kirkpatrick, M. and B. Barrett (2015). "Chromosome inversions, adaptive cassettes and the evolution of species' ranges". *Molecular Ecology* 24.9, pp. 2046–2055.

Kirkpatrick, M. and A. Kern (2012). "Where's the money? inversions, genes, and the hunt for genomic targets of selection". *Genetics* 190.4, pp. 1153–1155.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, et al. (2007). "Paired-end mapping reveals extensive structural variation in the human genome". *Science* 318.5849, pp. 420–426.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen (2014). "ANGSD: Analysis of next generation sequencing data". *BMC Bioinformatics* 15.1, p. 356.

Krimbas, C. B. and J. R. Powell (1992). *Drosophila inversion polymorphism*. Boca Raton: CRC Press, p. 560.

Krumsiek, J., R. Arnold, and T. Rattei (2007). "Gepard: A rapid and sensitive tool for creating dotplots on genome scale". *Bioinformatics* 23.8, pp. 1026–1028.

Lahn, B. T. and D. C. Page (1999). "Four Evolutionary Strata on the Human X Chromosome". *Science* 286.5441, pp. 964–967.

Lam, H. Y. K., X. J. Mu, A. M. Stütz, A. Tanzer, P. D. Cayting, et al. (2009). "Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library". *Nature Biotechnology* 28.1, pp. 47–55.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, et al. (2001). "Initial sequencing and analysis of the human genome". *Nature* 409.6822, pp. 860–921.

Lappalainen, T., M. Sammeth, M. R. Friedländer, P. A. C. 't Hoen, J. Monlong, et al. (2013). "Transcriptome and genome sequencing uncovers functional variation in humans". *Nature* 501.7468, pp. 506–511.

Leffler, E. M., Z. Gao, S. Pfeifer, L. Ségurel, A. Auton, et al. (2013). "Multiple instances of ancient balancing selection shared between humans and chimpanzees". *Science* 340.6127, pp. 1578–1582.

Levy, S., G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, et al. (2007). "The diploid genome sequence of an individual human". *PLoS Biology* 5.10, pp. 2113–2144.

Lewontin, R. C. and J. Krakauer (1973). "Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms". *Genetics* 74.1, pp. 175–195.

Li, H. (2011). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". *Bioinformatics* 27.21, pp. 2987–2993.

Locke, D. P., L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, et al. (2011). "Comparative and demographic analysis of orang-utan genomes". *Nature* 469.7331, pp. 529–533.

Lonsdale, J., J. Thomas, M. Salvatore, R. Phillips, E. Lo, et al. (2013). "The Genotype-Tissue Expression (GTEx) project." *Nature genetics* 45.6, pp. 580–5.

Lowry, D. B. and J. H. Willis (2010). "A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation". *PLoS Biology* 8.9.

Lucas Lledó, J. I. and M. Cáceres (2013). "On the Power and the Systematic Biases of the Detection of Chromosomal Inversions by Paired-End Genome Sequencing". *PLoS ONE* 8.4, e61292.

Lucas-Lledó, J. I., D. Vicente-Salvador, C. Aguado, and M. Cáceres (2014). "Population genetic analysis of bi-allelic structural variants from low-coverage sequence data with an expectation-maximization algorithm". *BMC Bioinformatics* 15.1, p. 163.

MacArthur, J., E. Bowler, M. Cerezo, L. Gil, P. Hall, et al. (2017). "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". *Nucleic Acids Research* 45.D1, pp. D896–D901.

MacDonald, J. R., R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer (2014). "The Database of Genomic Variants: a curated collection of structural variation in the human genome". *Nucleic Acids Research* 42.Database issue, pp. D986–D992.

Machado, C. A., T. S. Haselkorn, and M. A. F. Noor (2007). "Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in Drosophila pseudoobscura and D. persimilis". *Genetics* 175.3, pp. 1289–1306.

Machiela, M. J. and S. J. Chanock (2015). "LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants". *Bioinformatics* 31.21, pp. 3555–3557.

Mallick, S., H. Li, M. Lipson, I. Mathieson, M. Gymrek, et al. (2016). "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations". *Nature* 538.7624, pp. 201–206.

Martínez-Fundichely, A., M. Oliva, D. Vicente-Salvador, C. Aguado, D. Izquierdo, et al. (in prep.). "Accurate characterization/prediction of inversions in the human genome from paired-end mapping data with the GRIAL algorithm".

Martínez-Fundichely, A., S. Casillas, R. Egea, M. Ràmia, A. Barbadilla, et al. (2014). "InvFEST, a database integrating information of polymorphic inversions in the human genome". *Nucleic Acids Research* 42.D1, pp. D1027–D1032.

Mathieson, I., I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, et al. (2015). "Genome-wide patterns of selection in 230 ancient Eurasians". *Nature* 528.7583, pp. 499–503.

Maynar Smith, J. and J. Haigh (1974). "The hitch-hiking effect of a favourable gene". *Genetical research* 23.1, pp. 23–35.

McGaugh, S. E. and M. A. F. Noor (2012). "Genomic impacts of chromosomal inversions in parapatric Drosophila species". *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1587, pp. 422–429.

McVey, M. and S. E. Lee (2008). "MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings". *Trends in Genetics* 24.11, pp. 529–538.

Meyer, M., M. Kircher, M. .-.-T. Gansauge, H. Li, F. Racimo, et al. (2012). "A high-coverage genome sequence from an archaic Denisovan individual". *Science* 338.6104, pp. 222–226.

Mikkelsen, T. S., L. W. Hillier, E. E. Eichler, M. C. Zody, D. B. Jaffe, et al. (2005). "Initial sequence of the chimpanzee genome and comparison with the human genome". *Nature* 437.7055, pp. 69–87.

Mills, R. E., C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, et al. (2006). "An initial map of insertion and deletion (INDEL) variation in the human genome". *Genome Research* 16.9, pp. 1182–1190.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen, et al. (2011). "Mapping copy number variation by population scale genome sequencing". *Nature* 470.7332, pp. 59–65.

Morel, F., B. Laudier, F. Guérif, M. L. Couet, D. Royère, et al. (2007). "Meiotic segregation analysis in spermatozoa of pericentric inversion carriers using fluorescence in-situ hybridization". *Human Reproduction* 22.1, pp. 136–141.

Morin, S. J., J. Eccles, A. Iturriaga, and R. S. Zimmerman (2017). "Translocations, inversions and other chromosome rearrangements". *Fertility and Sterility* 107.1, pp. 19–26.

Moses, M. J., P. A. Poorman, T. H. Roderick, and M. T. Davisson (1982). "Synaptonemal Complex Analysis of Mouse Chromosomal Rearrangements. IV. Synapsis and Synaptic Adjustment in Two Paracentric Inversions". *Chromosoma* 64, pp. 457–474.

Navarro, A., A. Barbadilla, and A. Ruiz (2000). "Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in drosophila". *Genetics* 155.2, pp. 685–698.

Navarro, A. and A. Ruiz (1997). "On the fertility effects of pericentric inversions". *Genetics* 147.2, pp. 931–933.

Navarro, A., E. Betran, A. Barbadilla, and A. Ruiz (1997). "Recombination and Gene Flux Caused by Gene Conversion and Crossing Over in Inversion Heterokaryotypes". *Genetics* 146.2, pp. 695–709.

Nei, M. and W.-. H. Li (1979). "Mathematical model for studying genetic variation in terms of restriction endonucleases." *Proceedings of the National Academy of Sciences of the United States of America* 76.10, pp. 5269–5273.

Nielsen, R. (2004). "Population genetic analysis of ascertained SNP data." *Human genomics* 1.3, pp. 218–224.

Nielsen, R. (2005). "Molecular signatures of natural selection". *Annual Review of Genetics* 39, pp. 197–218.

Nielsen, R., J. M. Akey, M. Jakobsson, J. K. Pritchard, S. Tishkoff, et al. (2017). "Tracing the peopling of the world through genomics". *Nature* 541.7637, pp. 302–310.

Ohta, T. (1972). "Population Size and Rate of Evolution". *Journal of Molecular Evolution* 1.4, pp. 305–314.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, et al. (2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". *Nucleic acids research* 44.D1, pp. D733–D745.

O'Reilly, P. F., L. J. M. Coin, and C. J. Hoggart (2010). "invertFREGENE: Software for simulating inversions in population genetic data". *Bioinformatics* 26.6, pp. 838–840.

Otto, S. P. and T. Lenormand (2002). "Resolving the Paradox of Sex and Recombination". *Nature Reviews Genetics* 3.4, pp. 252–261.

Ottolini, C. S., L. J. Newnham, A. Capalbo, S. A. Natesan, H. A. Joshi, et al. (2015). "Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates". *Nature Genetics* 47.7, pp. 727–735.

Pang, A. W. C., O. Migita, J. R. Macdonald, L. Feuk, and S. W. Scherer (2013). "Mechanisms of formation of structural variation in a fully sequenced human genome". *Human Mutation* 34.2, pp. 345–354.

Pasvol, G., D. J. Weatherall, and R. J. Wilson (1978). "Cellular mechanism for the protective effect of haemoglobin S against P. falciparum malaria". *Nature* 274.5672, pp. 701–703.

Peischl, S., E. Koch, R. F. Guerrero, and M. Kirkpatrick (2013). "A sequential coalescent algorithm for chromosomal inversions". *Heredity* 111.3, pp. 200–9.

Perry, G. H., N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, et al. (2007). "Diet and the evolution of human amylase gene copy number variation." *Nature genetics* 39.10, pp. 1256–1260.

Pfeifer, B., U. Wittelsburger, S. E. Ramos-Onsins, and M. J. Lercher (2014). "PopGenome: An efficient swiss army knife for population genomic analyses in R". *Molecular Biology and Evolution* 31.7, pp. 1929–1936.

Poznik, G. D., Y. Xue, F. L. Mendez, T. F. Willems, A. Massaia, et al. (2016). "Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences". *Nature Genetics* 48.6, pp. 593–599.

Pratto, F., K. Brick, P. Khil, F. Smagulova, G. V. Petukhova, et al. (2014). "Recombination initiation maps of individual human genomes". *Science (New York, N.Y.)* 346.6211, p. 1256442.

Prüfer, K., K. Munch, I. Hellmann, K. Akagi, J. R. Miller, et al. (2012). "The bonobo genome compared with the chimpanzee and human genomes". *Nature* 486.7404, pp. 527–531.

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman, et al. (2014). "The complete genome sequence of a Neanderthal from the Altai Mountains." *Nature* 505.7481, pp. 43–49.

Puig, M., D. Castellano, L. Pantano, C. Giner-Delgado, D. Izquierdo, et al. (2015a). "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., S. Casillas, S. Villatoro, and M. Cáceres (2015b). "Human inversions and their functional consequences". *Briefings in Functional Genomics* 14.5, pp. 369–379.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, et al. (2007). "PLINK: A tool set for whole-genome association and population-based linkage analyses". *American Journal of Human Genetics* 81.3, pp. 559–575.

Pybus, M., P. Luisi, G. M. Dall'Olio, M. Uzkudun, H. Laayouni, et al. (2015). "Hierarchical boosting: A machine-learning framework to detect and clas-

sify hard selective sweeps in human populations". *Bioinformatics* 31.24, pp. 3946–3952.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Racimo, F., S. Sankararaman, R. Nielsen, and E. Huerta-Sánchez (2015). "Evidence for archaic adaptive introgression in humans." *Nature reviews. Genetics* 16, pp. 359–371.

Radke, D. W. and C. Lee (2015). "Adaptive potential of genomic structural variation in human and mammalian evolution." *Briefings in functional genomics* 5.14, pp. 358–368.

Rogers, J. and R. A. Gibbs (2014). "Comparative primate genomics: emerging patterns of genome content and dynamics". *Nature Reviews Genetics* 15.5, pp. 347–359.

Rogers, R. L. (2015). "Chromosomal rearrangements as barriers to genetic homogenization between archaic and modern humans". *Molecular biology and evolution* 32.12, pp. 3064–3078.

Rozas, J., C. Segarra, G. Ribó, and M. Aguadé (1999). "Molecular population genetics of the rp49 gene region in different chromosomal inversions of Drosophila subobscura". *Genetics* 151.1, pp. 189–202.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, et al. (2002). "Detecting recent positive selection in the human genome from haplotype structure". *Nature* 419.6909, pp. 832–837.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, et al. (2007). "Genome-wide detection and characterization of positive selection in human populations." *Nature* 449.7164, pp. 913–918.

Salm, M. P. A., S. D. Horswell, C. E. Hutchison, H. E. Speedy, X. Yang, et al. (2012). "The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism." *Genome research* 22.6, pp. 1144–53.

Sanders, A. D., M. Hills, D. Porubský, V. Guryev, E. Falconer, et al. (2016). "Characterizing polymorphic inversions in human genomes by single-cell sequencing". *Genome Research* 26.11, pp. 1575–1587.

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, et al. (2012). "Insights into hominid evolution from the gorilla genome sequence." *Nature* 483.7388, pp. 169–175.

Schouten, J. (2003). *Multiplex ligatable probe amplification*. US Patent App. 10/218,567.

Schrider, D. R. and A. D. Kern (2016). "S/HIC: Robust identification of soft and hard sweeps using machine learning". *PLoS Genetics* 12.3, e1005928.

Schrider, D. R. and A. D. Kern (2017). "Soft sweeps are the dominant mode of adaptation in the human genome". *Molecular Biology and Evolution* 34.8, pp. 1863–1877.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, et al. (2004). "Large-scale copy number polymorphism in the human genome". *Science* 305.5683, pp. 525–528.

Shao, H., D. Ganesamoorthy, T. Duarte, M. D. Cao, C. Hoggart, et al. (2017). "npInv: accurate detection and genotyping of inversions mediated by non-allelic homologous recombination using long read sub-alignment". *bioRxiv*, p. 178103.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro (1995). "Properties of statistical tests of neutrality for DNA polymorphism data". *Genetics* 141.1, pp. 413–429.

Small, K., J. Iber, and S. T. Warren (1997). "Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats". *Nature Genetics* 16.1, pp. 96–99.

Stefansson, H., A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, et al. (2005). "A common inversion under selection in Europeans". *Nature Genetics* 37.2, pp. 129–137.

Steinberg, K. M., F. Antonacci, P. H. Sudmant, J. M. Kidd, C. D. Campbell, et al. (2012). "Structural diversity and African origin of the 17q21.31 inversion polymorphism". *Nature Genetics* 44.8, pp. 872–880.

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, pp. 555–558.

Sturtevant, A. H. (1921). "A case of rearrangement of genes in Drosophila". *Proceedings of the National Academy of Sciences of the United States of America* 7.8, p. 235.

Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton, et al. (2015). "UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age". *PLoS Medicine* 12.3, e1001779.

Sudmant, P. H., T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, et al. (2015). "An integrated map of structural variation in 2,504 human genomes". *Nature* 526.7571, pp. 75–81.

Sunnåker, M., A. G. Busetto, E. Numminen, J. Corander, M. Foll, et al. (2013). "Approximate Bayesian computation." *PLoS computational biology* 9.1, e1002803.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism". *Genetics* 123.3, pp. 585–595.

Teague, B., M. S. Waterman, S. Goldstein, K. Potamousis, S. Zhou, 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, pp. 10848–10853.

The 1000 Genomes Project Consortium (2010). "A map of human genome variation from population-scale sequencing". *Nature* 467.7319, pp. 1061–1073.

The 1000 Genomes Project Consortium (2012). "An integrated map of genetic variation from 1,092 human genomes". *Nature* 491.7422, pp. 56–65.

The 1000 Genomes Project Consortium (2015). "A global reference for human genetic variation". *Nature* 526.7571, pp. 68–74.

The Genome of the Netherlands Consortium (2014). "Whole-genome sequence variation, population structure and demographic history of the Dutch population". *Nature Genetics* 46.8, pp. 818–825.

The International HapMap 3 Consortium (2010). "Integrating common and rare genetic variation in diverse human populations". *Nature* 467.7311, pp. 52–58.

The International HapMap Consortium (2005). "A haplotype map of the human genome." *Nature* 437.7063, pp. 1299–320.

The UK10K Consortium (2015). "The UK10K project identifies rare variants in health and disease". *Nature* 526.7571, pp. 82–90.

Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov (2013). "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". *Briefings in bioinformatics* 14.2, pp. 178–192.

Torgasheva, A. A. and P. M. Borodin (2010). "Synapsis and recombination in inversion heterozygotes." *Biochemical Society transactions* 38.6, pp. 1676–1680.

Torgasheva, A. A., N. B. Rubtsov, and P. M. Borodin (2013). "Recombination and synaptic adjustment in oocytes of mice heterozygous for a large paracentric inversion". *Chromosome Research* 21.1, pp. 37–48.

Turchin, M. C., C. W. K. Chiang, C. D. Palmer, S. Sankararaman, D. Reich, et al. (2012). "Evidence of widespread selection on standing variation in Europe at height-associated SNPs." *Nature genetics* 44.9, pp. 1015–1019.

Tuzun, E., A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison, et al. (2005). "Fine-scale structural variation of the human genome". *Nature Genetics* 37.7, pp. 727–732.

Vicente-Salvador, D., M. Puig, M. Gayà-Vidal, S. Pacheco, C. Giner-Delgado, et al. (2016). "Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution". *Human Molecular Genetics* 26.3, pp. 567–581.

Vitti, J. J., S. R. Grossman, and P. C. Sabeti (2013). "Detecting natural selection in genomic data." *Annual review of genetics* 47, pp. 97–120.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard (2006). "A map of recent positive selection in the human genome." *PLoS biology* 4.3, e72.

Wall, J. D. and D. Yoshihara Caldeira Brandt (2016). "Archaic admixture in human history". *Current Opinion in Genetics and Development* 41, pp. 93–97.

Weckselblatt, B. and M. K. Rudd (2015). "Human Structural Variation: Mechanisms of Chromosome Rearrangements". *Trends in Genetics* 31.10, pp. 587–599.

Weir, B. S. and C. C. Cockerham (1984). "Estimating F-Statistics for the Analysis of Population Structure". *Evolution* 38.6, p. 1358.

White, M. J. D. (1978). *Modes of evolution*. San Francisco, CA: W.H. Freeman.

Wiehe, T. H. and W. Stephan (1993). "Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster." *Molecular biology and evolution* 10.4, pp. 842–854.

Williams, A. L., G. Genovese, T. Dyer, N. Altemose, K. Truax, et al. (2015). "Non-crossover gene conversions show strong GC bias and unexpected clustering in humans." *eLife* 4, e04637.

Wong, L. P., R. T. H. Ong, W. T. Poh, X. Liu, P. Chen, et al. (2013). "Deep whole-genome sequencing of 100 southeast Asian malays". *American Journal of Human Genetics* 92.1, pp. 52–66.

Yates, A., W. Akanni, M. R. Amode, D. Barrell, K. Billis, et al. (2016). "Ensembl 2016". *Nucleic Acids Research* 44.D1, pp. D710–D716.

Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. Cuo, et al. (2010). "Sequencing of 50 human exomes reveals adaptation to high altitude". *Science* 329.5987, pp. 75–78.

Zody, M. C., Z. Jiang, H. C. Fung, F. Antonacci, L. W. Hillier, et al. (2008). "Evolutionary toggling of the MAPT 17q21.31 inversion region". *Nature Genetics* 40.9, pp. 1076–1083.

Zook, J. M., B. Chapman, J. Wang, D. Mittelman, O. Hofmann, et al. (2014). "Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls". *Nature Biotechnology* 32.3, pp. 246–251.

# Appendix A

# Supplementary figures

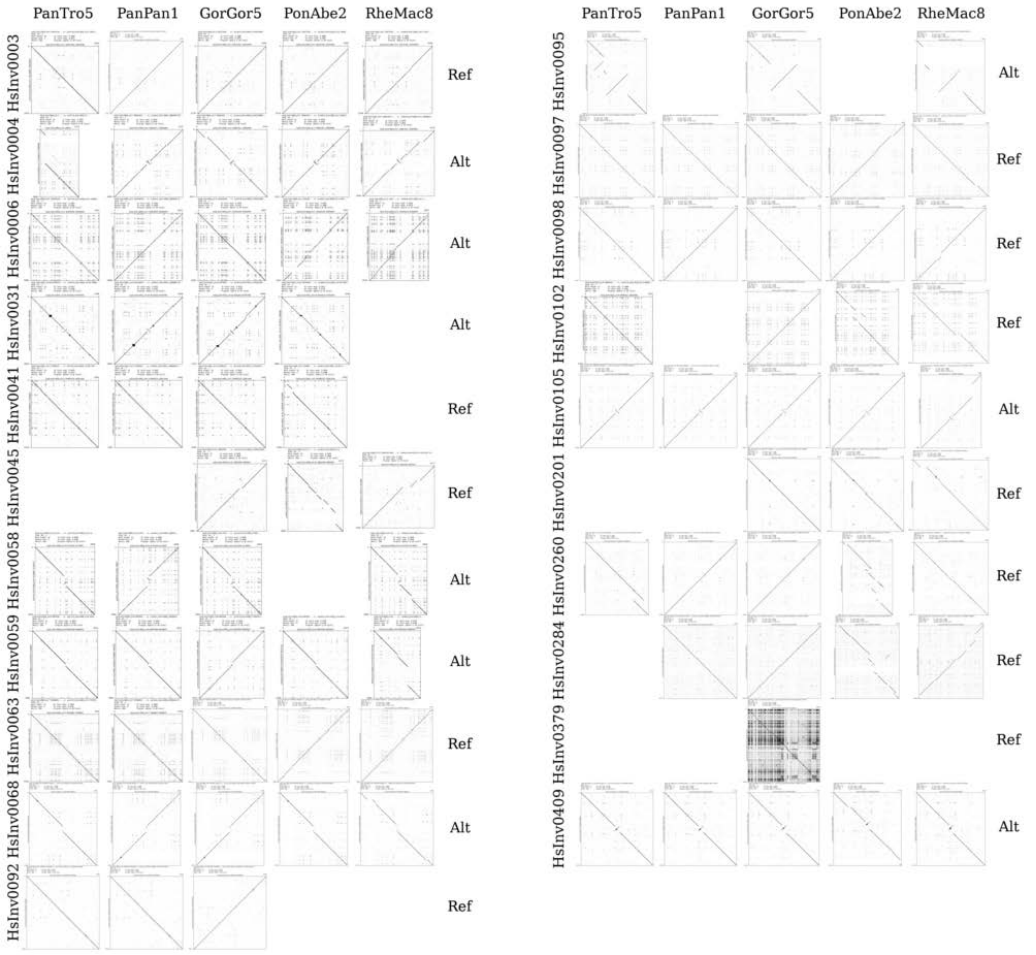


**Figure A.1: Alignments of NH inversions with non-human primate assemblies**. Alignment of each inversion region plus flanking 10 kbp in HG18 genome against the most recent assembly of chimpanzee, bonobo, gorilla, orangutan and rhesus macaque. The orientation found in the tested assemblies is annotated on the right, for all inversions it is consistent across species.

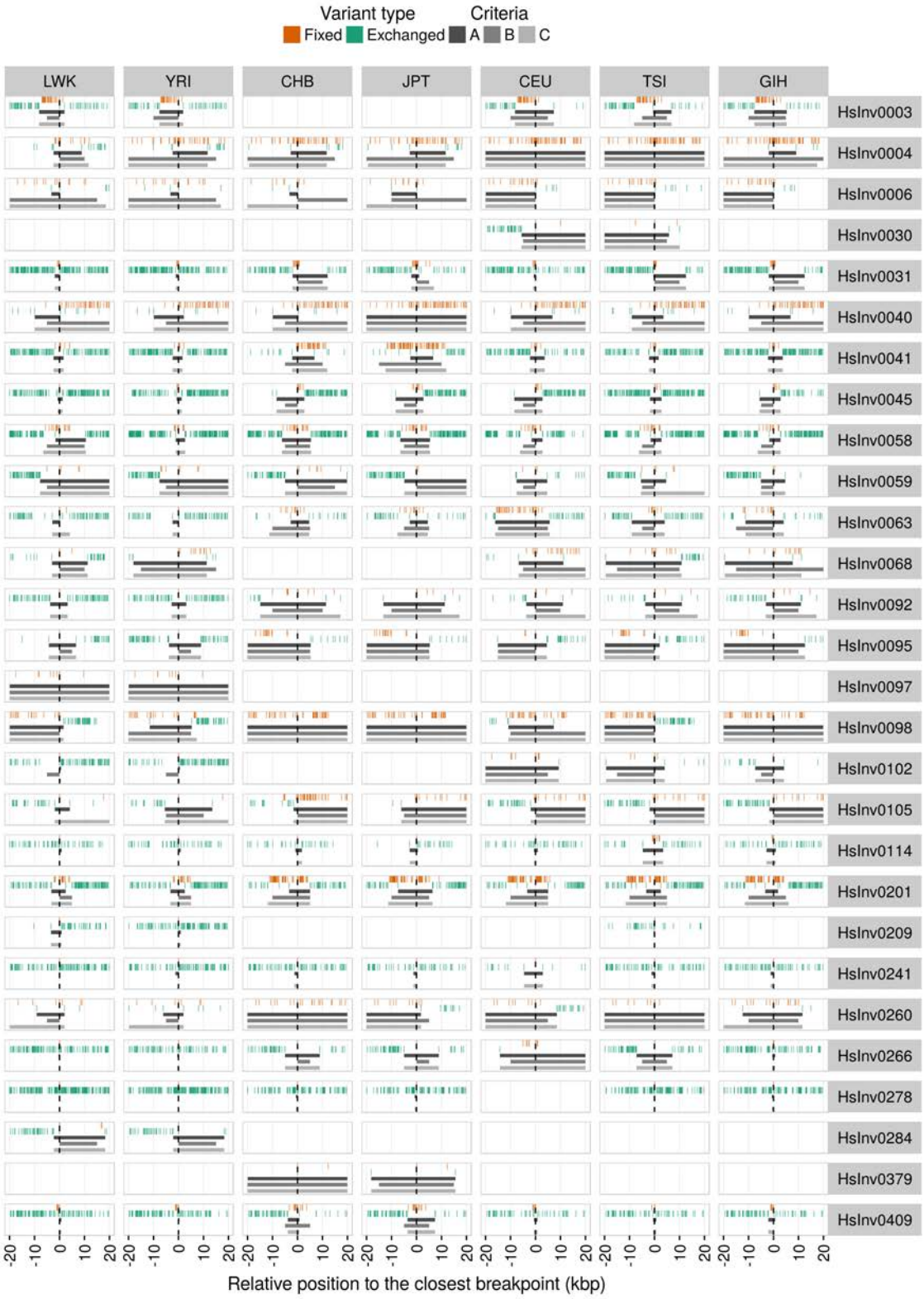


**Figure A.2: Criteria to define the extension of inversion-linked region**. The region defined by each of the three tested criteria are shown as ranges below the distribution of informative variants. Criterion C was chosen for subsequent analyses.

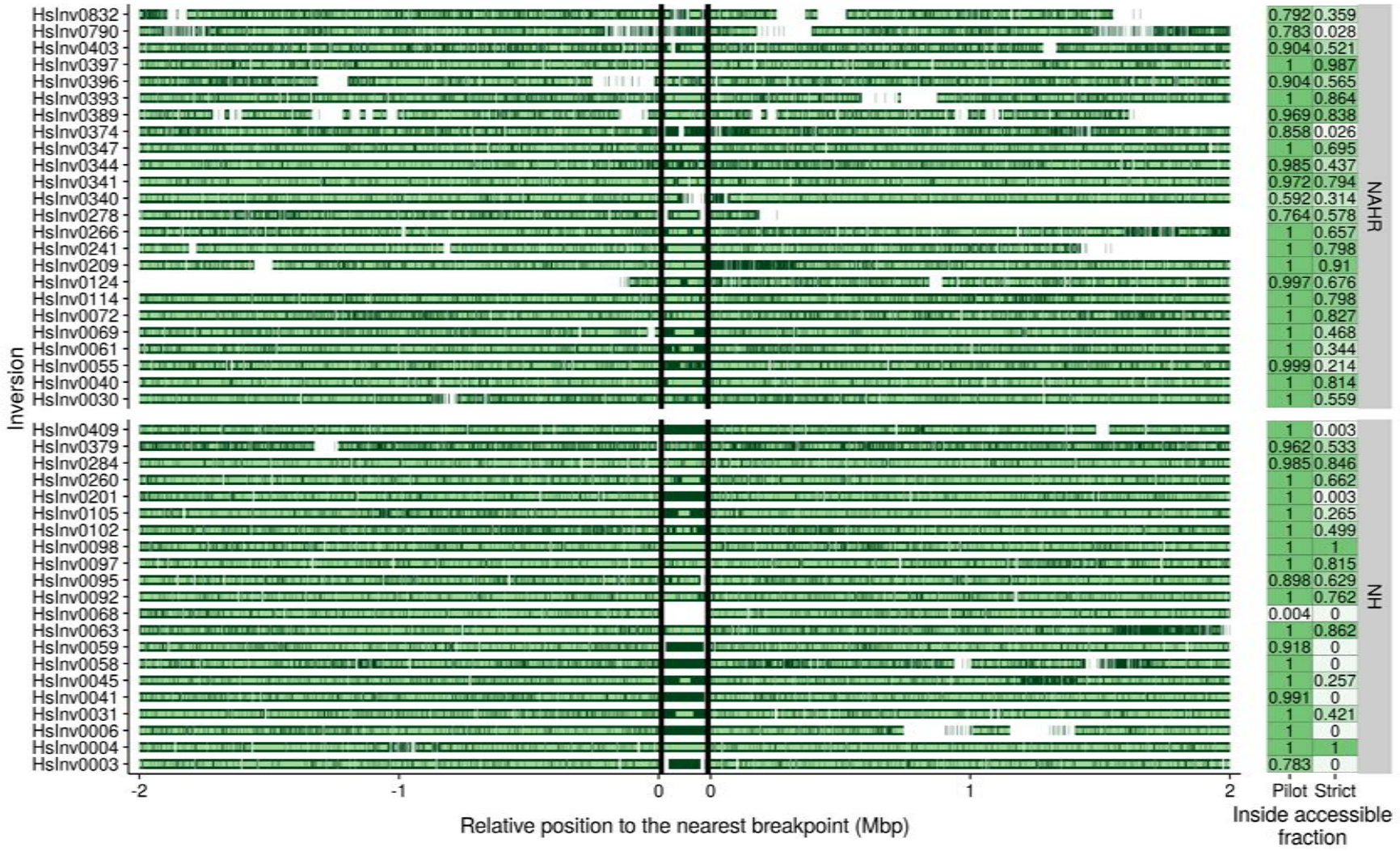


**Figure A.3: Inversion regions accessible to 1000GP technologies**. Accessible areas are shown in rectangles as defined by pilot (dark green) and strict (light green) accessibility masks. Breakpoint regions have been omitted since they are excluded from most sequence analyses. The inside region of each inversion has been scaled to the same width for representation purposes.

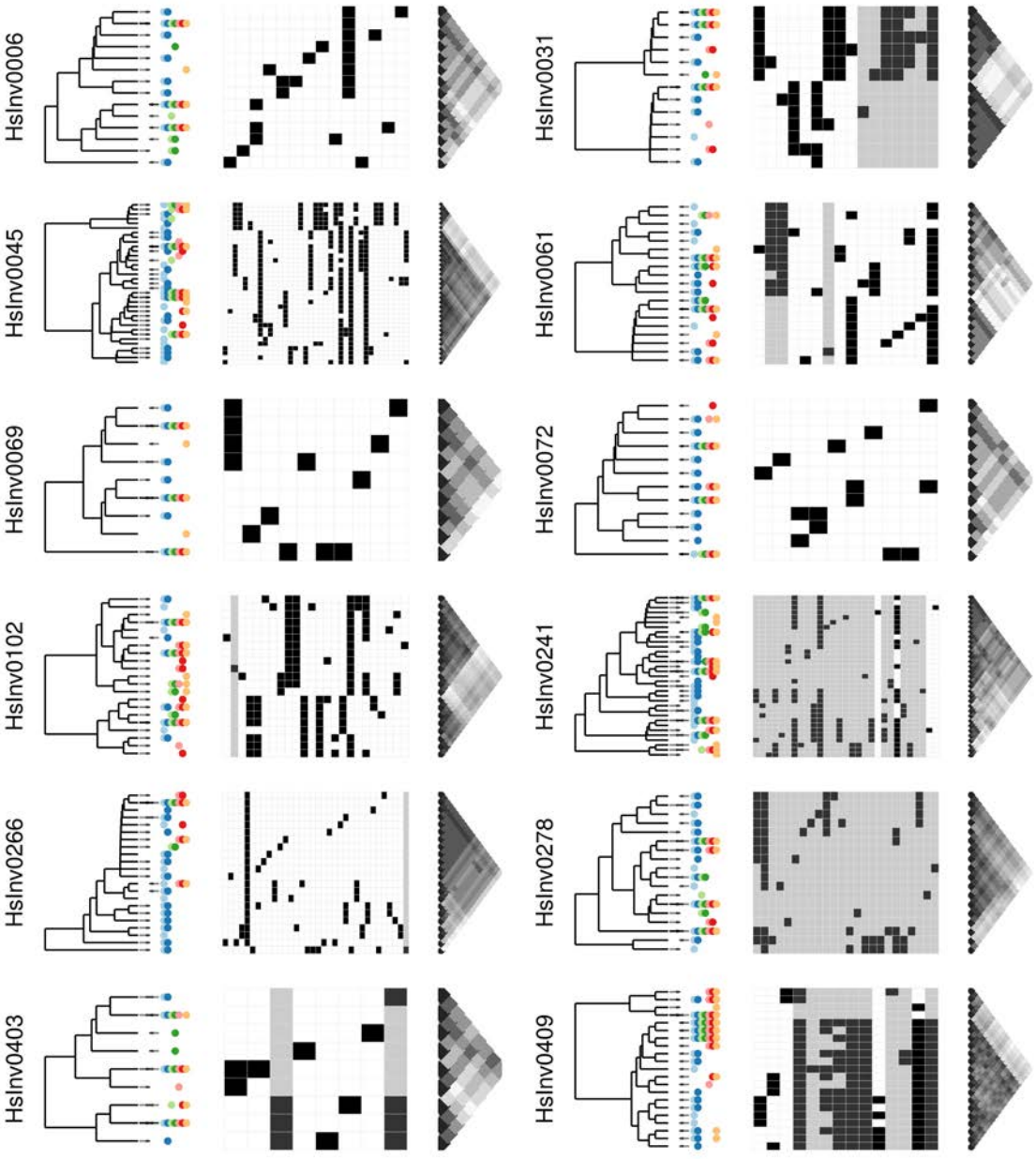


**Figure A.4: Haplotype alignment and clustering**. Haplotype visualization for all inversions. Arrows represent orientation of the chromosomes with that haplotype: pointing right means orientation in reference genome HG18 and pointing left means alternative orientation (for ancestral orientation see Figure 3.2). Populations of the individuals with each haplotype are indicated as dots (blues indicate African; greens, East Asia; reds, Europe; orange, South Asia). Color code for SNP alleles in haplotype alignment: white/black are ancestral/derived alleles, light/dark grey are reference HG19/alternative alleles (when ancestral unknown). In this page, inversions with short haplotypes.

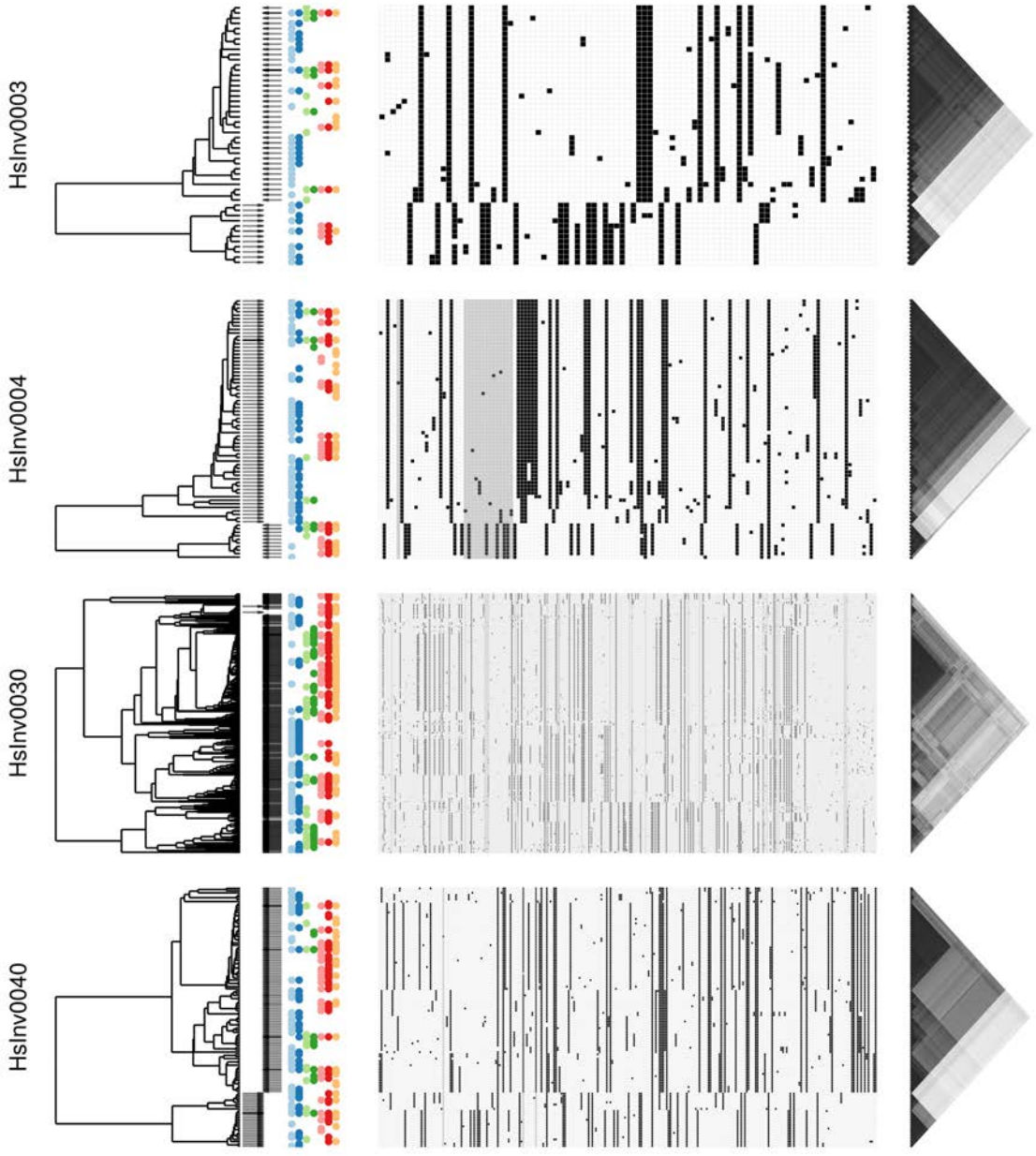


**Figure A.4 continued**. Inversions with medium-sized haplotypes.

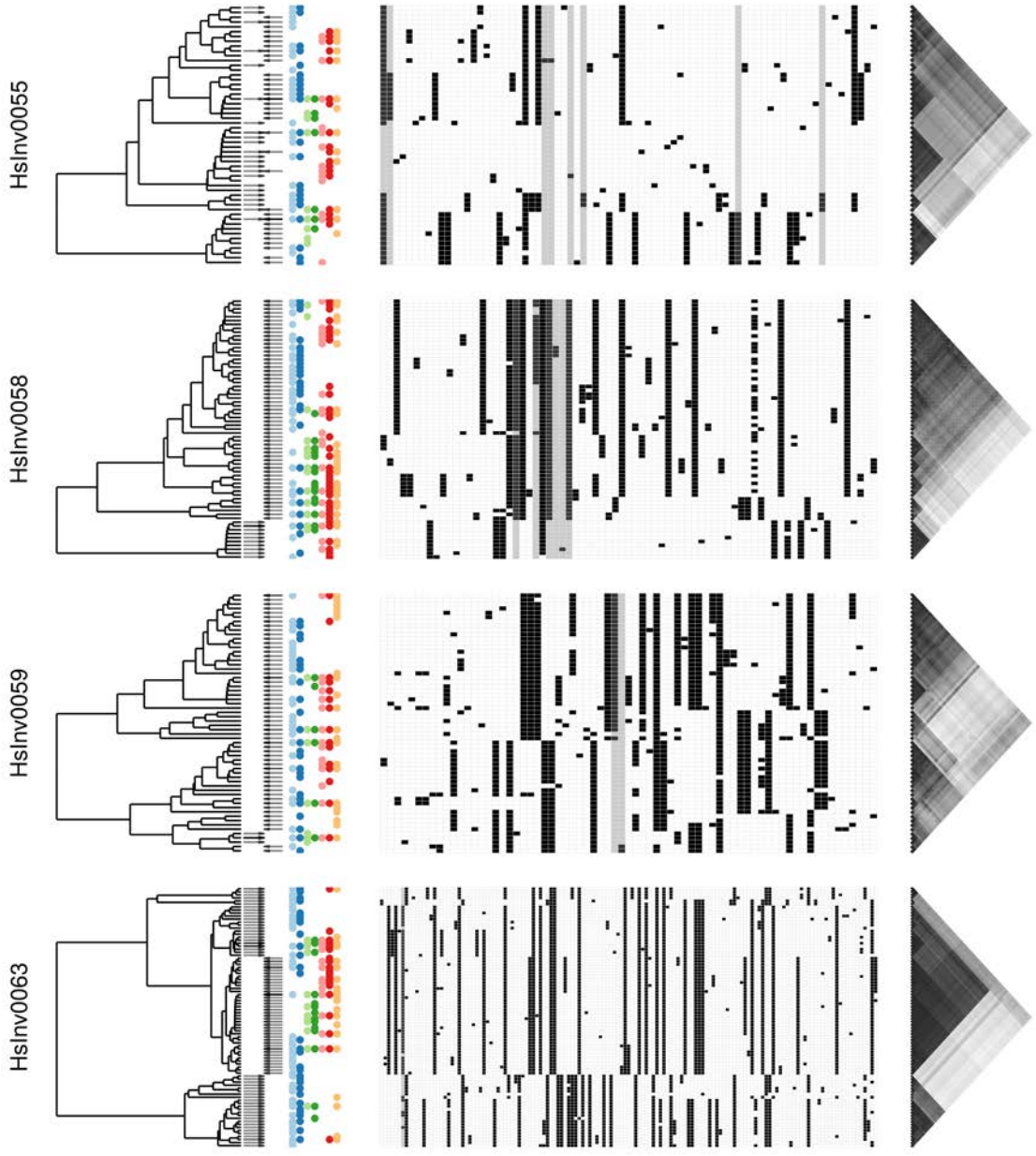


**Figure A.4 continued**. Inversions with medium-sized haplotypes.

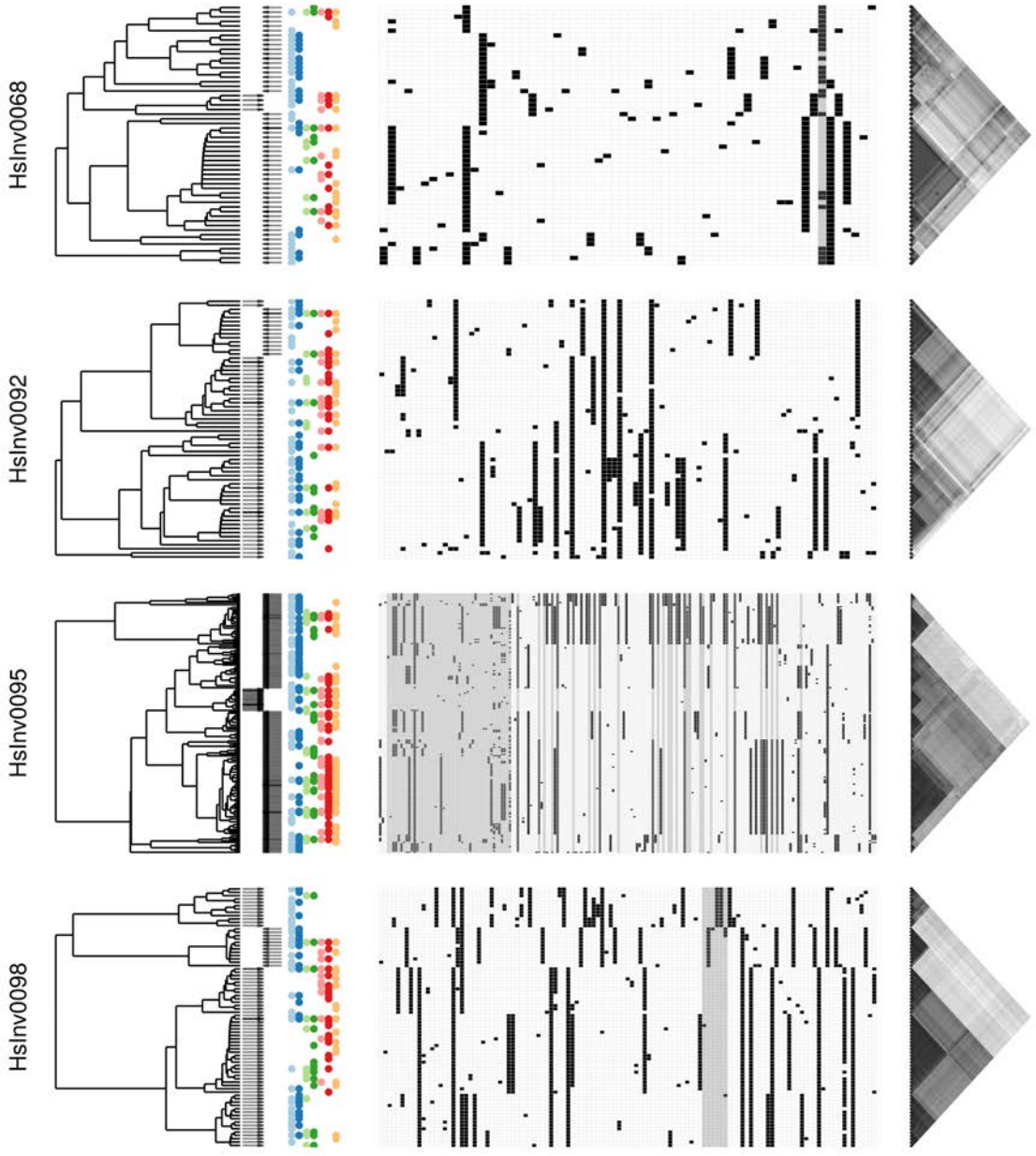


**Figure A.4 continued**. Inversions with medium-sized haplotypes.

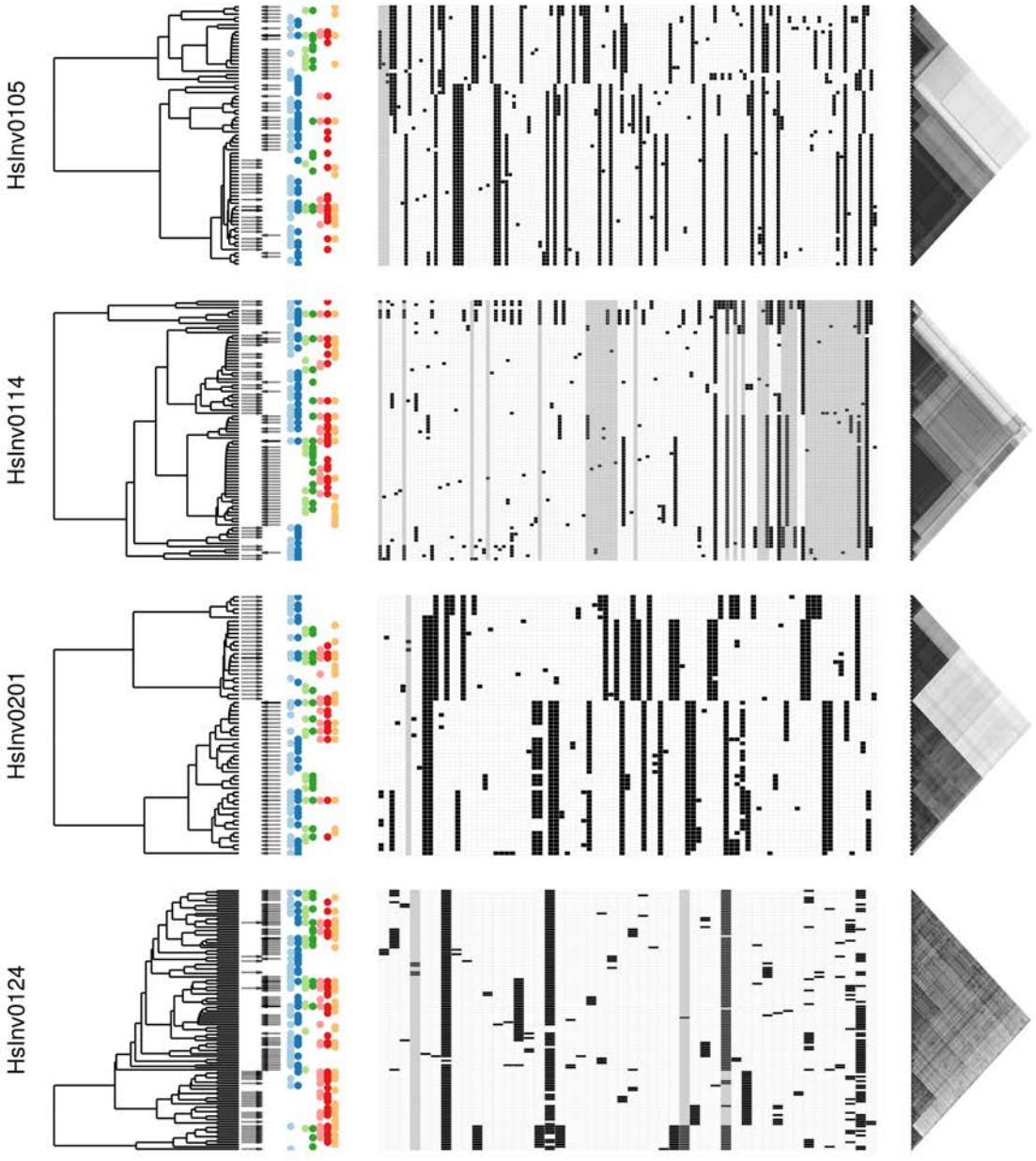


**Figure A.4 continued**. Inversions with medium-sized haplotypes.

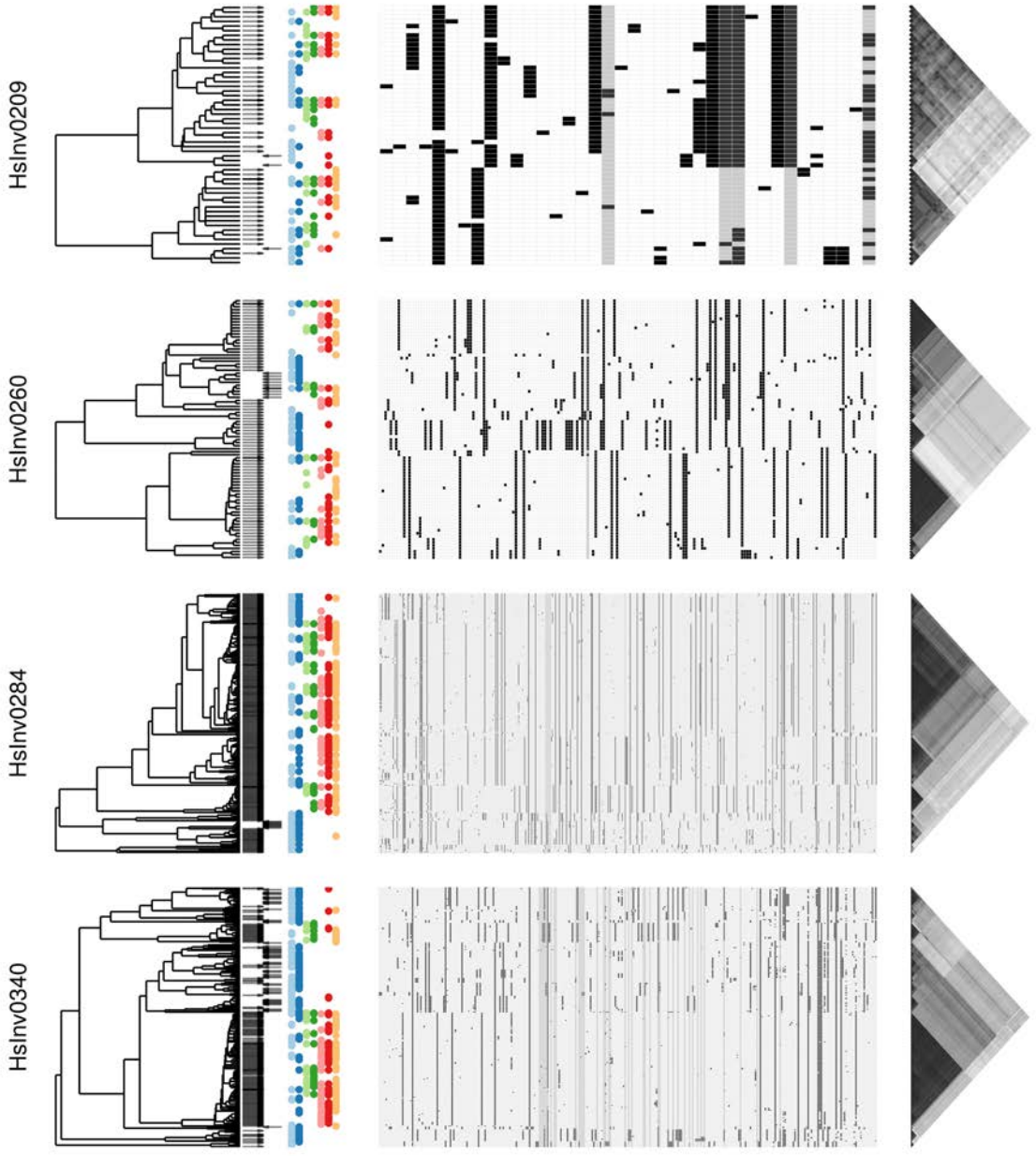


**Figure A.4 continued**. Inversions with medium-sized haplotypes.

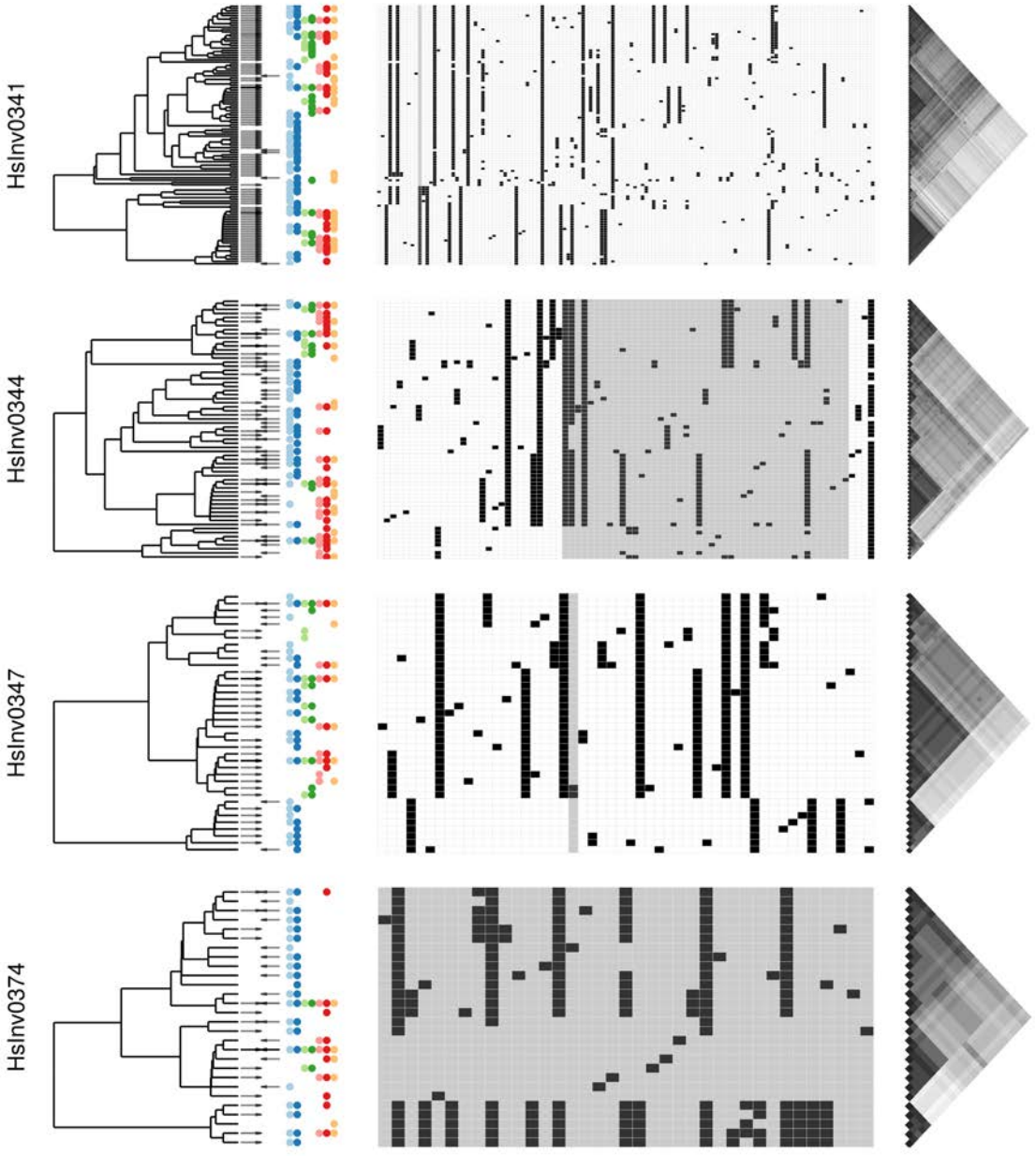


**Figure A.4 continued**

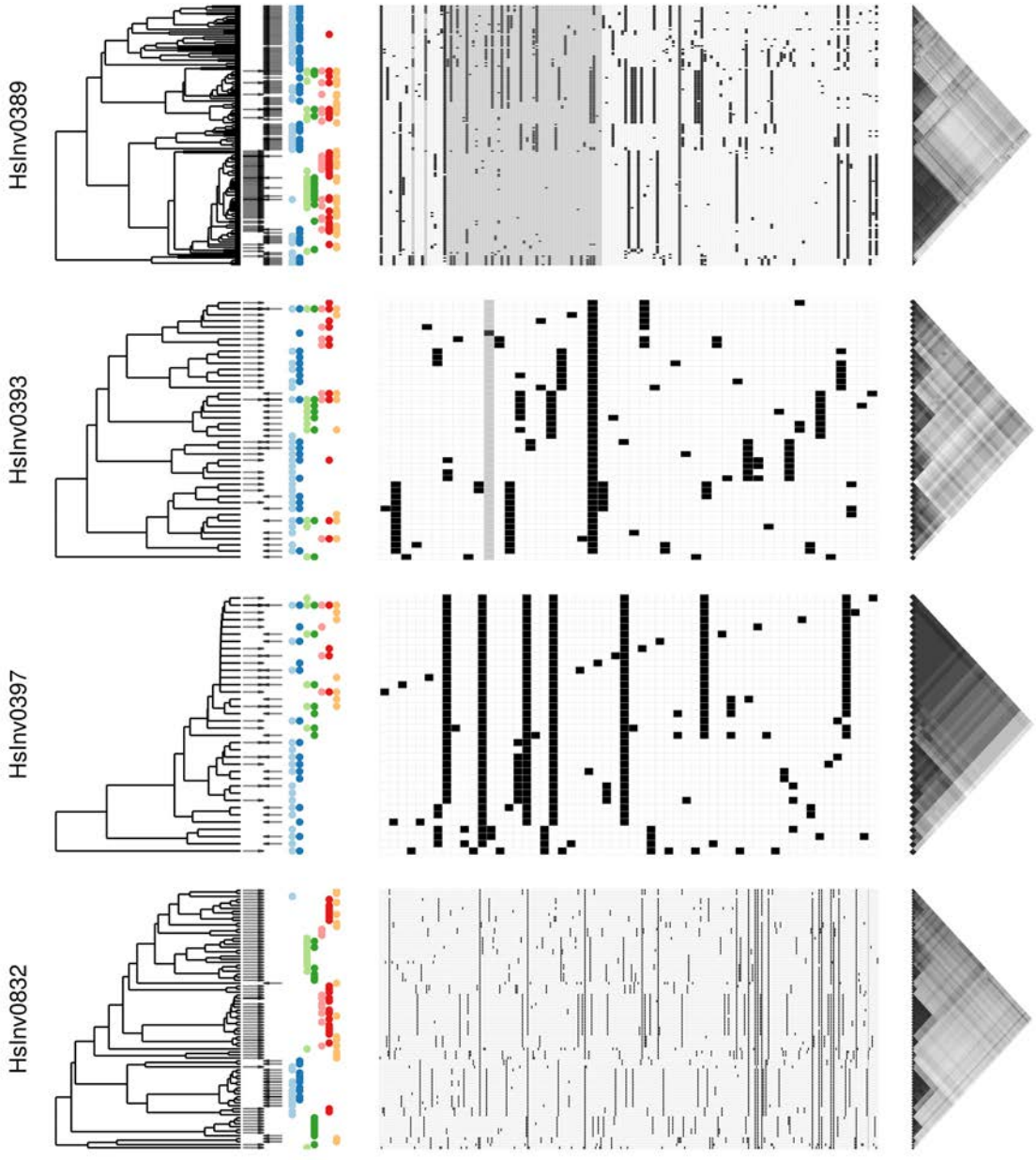


**Figure A.4 continued**. Inversions with medium-sized haplotypes.

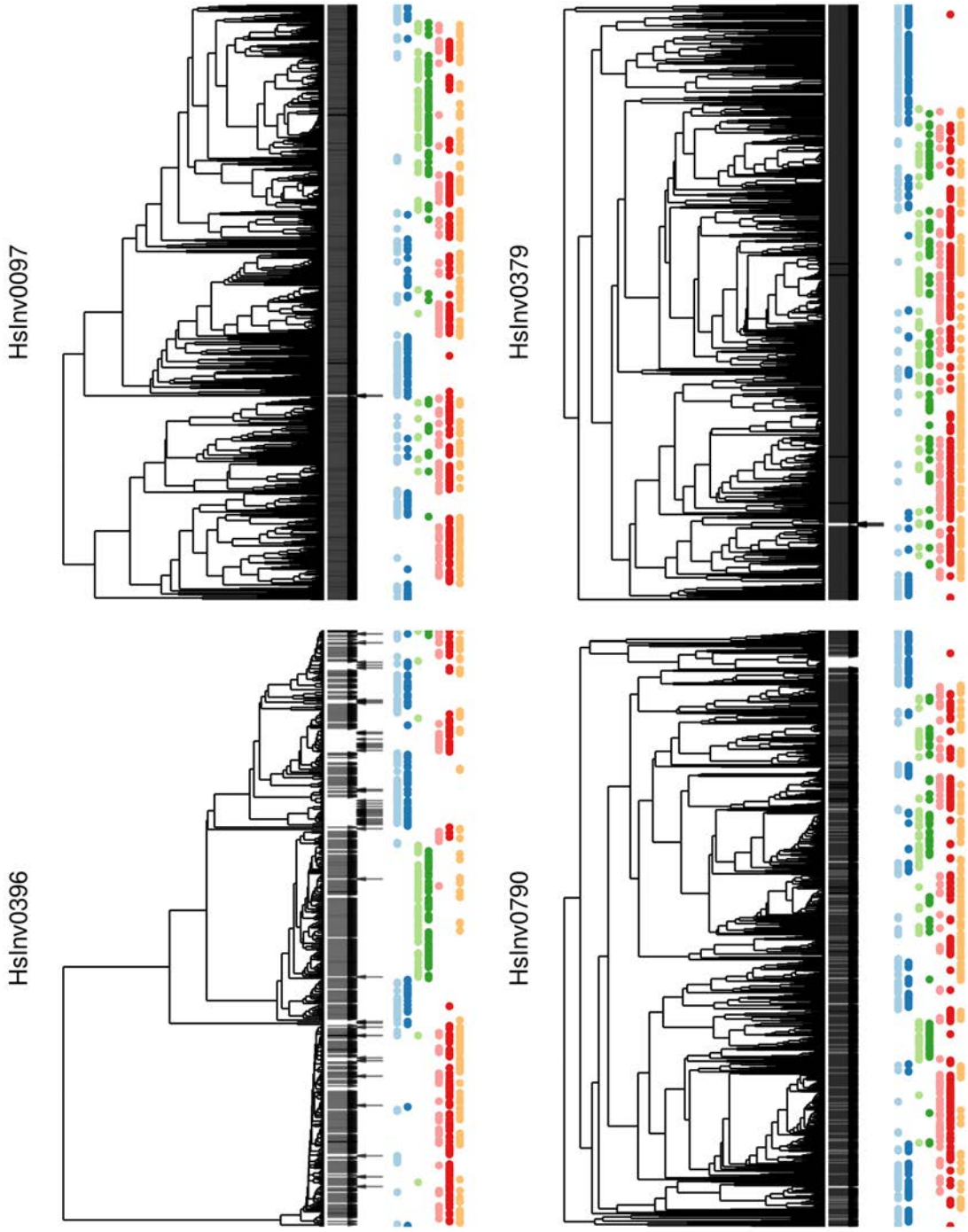


**Figure A.4 continued**. Inversions with long haplotypes, here only showing the annotated dendrograms.

# Appendix B

# Supplementary tables

**Table B.1: Breakpoint annotation of NH inversions**. Coordinates in the reference genome HG18.

| ID | Source [d] | Chr | Breakpoint 1 | Breakpoint 2 | Inner size (bp) | Indel at breakpoint 1 | Indel size and type | Indel at breakpoint 2 | Indel size and type |
|---|---|---|---|---|---|---|---|---|---|
| HsInv0003 | 5 | chr1 | 185731452-185733101 | 185733351-185733355 | 249 | 185731452-185733101 | 1650-bp Alt Del | 185733351-185733352 | 2-bp Alt Del |
| HsInv0004 | 5, 2 | chr1 | 196023412-196023413 | 196024608-196024609 | 1194 | 196023411-196023411 | 1-bp Ref Ins | - | - |
| HsInv0006 | 5 | chr1 | 203445230-203445294 | 203445378-203445457 | 83 | 203445254-203445294 | 41-bp Ref Dup | 203445378-203445416 | 39-bp Ref Dup |
| HsInv0031 [a] | 5, 2, Others | chr16 | 83746215-83746259 | 83747296-83747305 | 1036 | - | - | 83747302-83747303 | 35-bp Ref Del |
| HsInv0041 | 5 | chr2 | 225001193-225001224 | 225001332-225002195 | 107 | 225001196-225001224 | 29-bp Alt Del | 225001332-225002195 | 864-bp Alt Del |
| HsInv0045 [b] | 5 | chr21 | 26942440-26942558 | 26943499-26943617 | 940 | - | - | 26943504-26943597 | 94-bp Alt Del |
| HsInv0058 | 5 | chr6 | 31117199-31117201 | 31118075-31118075 | 873 | 31117201-31117202 | 2188-bp Ref Del | 31118075-31118076 | 630-bp Ref Del |
| HsInv0059 | 5 | chr6 | 89980347-89980355 | 89980662-89980670 | 306 | 89980354-89980355 | 2-bp Ref Dup | 89980662-89980668; 89980670-89980671 | 7-bp Ref Dup; 618-bp Ref Del |
| HsInv0063 | 5, 2, Others | chr7 | 70058904-70064121 | 70076816-70076823 | 12694 | 70058905-70058906; 70058906-70064121 | 6 bp Alt Dup; 5216-bp Alt Del | - | - |
| HsInv0068 | 5 | chr9 | 76087959-76087960 | 76088210-76088213 | 249 | 76087960-76087961 | 1231-bp Ref Del | 76088209-76088210 | 325-bp Ref Del |
| HsInv0092 | 6 | chr6 | 130889879-130889879 | 130893988-130893988 | 4108 | - | - | - | - |
| HsInv0095 | 6, 2 | chr4 | 89066188-89066188 | 89077724-89077723 | 11535 | - | - | 89077723-89077724 | 2-bp Ref Del |
| HsInv0097 | 6 | chr3 | 109202512-109202535 | 109223744-109224821 | 21208 | 109202511-109202512; 109202512-109202535 | 88-bp Alt Dup; 24-bp Alt Del | 109223744-109224821; 109224821-109224822 | 1078-bp Alt Del; 1-bp Alt Ins |
| HsInv0098 [c] | 6 | chr3 | 41337147-41337165 | 41338116-41338809 | 950 | 41337146-41337147; 41337147-41337165 | 12-bp Alt Dup; 19-bp Alt Del | 41338116-41338790 | 675-bp Alt Del |
| HsInv0102 | 6, 1 | chr4 | 39911422-39911423 | 39913454-39913455 | 2030 | - | - | - | - |
| HsInv0105 | 6 | chr7 | 40845901-40845900 | 40846995-40846994 | 1094 | 40845900-40845901 | 2-bp Ref Del | - | - |
| HsInv0201 | 6, 3 | chr5 | 147533233-147534432 | 147534809-147534971 | 376 | 147533233-147534432 | 1200-bp Alt Del | 147534809-147534971 | 163-bp Alt Del |
| HsInv0260 | 6 | chr2 | 184277421-184277442 | 184279274-184280256 | 1831 | 184277423-184277424; 184277424-184277437 | 980-bp Alt Dup; 14-bp Alt Del | - | - |
| HsInv0284 | 6 | chr6 | 142229155-142229154 | 142261543-142261544 | 32388 | 142229154-142229155 | 2-bp Alt Dup | - | - |
| HsInv0379 | 4 | chr19 | 21621974-21621973 | 22037101-22037100 | 415127 | - | - | - | - |
| HsInv0409 | 6 | chrX | 6147047-6147049 | 6148391-6148392 | 1341 | 6147049-6147050 | 8-bp Ref Del | 6148385-6148390; 6148390-6148391 | 6-bp Ref Dup; 357-bp Ref Del |

[a] With low identity inverted repeats. IR1: chr16:83746171-83746455; IR2: chr16:83747100-83747348; IR size: 285/249; %IR: 83,2%; Repeat type: AluSq2/AluSx1

[b] With low identity inverted repeats. IR1: chr21:26942303-26942558; IR2: chr21:26943499-26943755; IR size: 256/257; %IR:85,6% ; Repeat type: AluSx1

[c] With low identity inverted repeats. IR1: chr3:41337143-41337439; IR2: chr3:41338517-41338813; IR size: 297/297; %IR: 86,2%; Repeat type: AluSz/AluSz6

[d] Annotation source. 1- Korbel et al. (2007) Science, 318:420-426. 2- Pang et al. (2013) Human Mutation, 34(2): 345-354. 3- Lucas-Lledó et al. (2014) BMC Bioinformatics, 15(1):163. 4- Puig et al. (2015) Plos Genetics, 11(10):e1005495. 5- Vicente-Salvador et al. (2016) Human Molecular Genetics, 26(3): 567-581. 6- Martínez-Fundichely et al. (in prep.)

**Table B.2: Breakpoint annotation of NAHR inversions**. Coordinates in the reference genome HG18.

| ID | Source [a] | Chr | Breakpoint 1 | Breakpoint 2 | Inner size (bp) | Inverted repeat at breakpoint 1 | Inverted repeat at breakpoint 2 | IR size | %IR identity | Repeat type |
|---|---|---|---|---|---|---|---|---|---|---|
| HsInv0030 | 3 + 1 + Others | chr16 | 73797041-73797599 | 73814160-73814718 | 16560 | 73796066-73797613 | 73814148-73815693 | 1548/1546 | 99.6% | SD |
| HsInv0040 | 3 | chr2 | 138720715-138721469 | 138725059-138725814 | 3589 | 138720715-138721469 | 138725059-138725814 | 755/756 | 99.9% | US+TEs [b] |
| HsInv0055 | 3 | chr5 | 63797584-63802465 | 63808718-63813599 | 6252 | 63796505-63802503 | 63808673-63814674 | 5999/6002 | 96.6% | L1PA7/L1PA3 |
| HsInv0061 | 3 | chr6 | 107275246-107275899 | 107277574-107278229 | 1674 | 107275246-107275899 | 107277574-107278229 | 654/656 | 99.7% | US+TEs [c] |
| HsInv0069 | 3 | chr9 | 114907364-114913787 | 114914988-114921411 | 1200 | 114906753-114913787 | 114914988-114922014 | 7035/7027 | 98.6% | SD |
| HsInv0072 | 3 | chrX | 45432027-45433183 | 45435670-45436826 | 2486 | 45431993-45433440 | 45435413-45436856 | 1448/1444 | 98.4% | L1PA13 |
| HsInv0114 | 2 | chr9 | 125778473-125781206 | 125793139-125795872 | 11932 | 125778473-125781206 | 125793139-125795872 | 2734/2734 | 99.9% | SD |
| HsInv0124 | 2 | chr11 | 300225-301836 | 307945-309555 | 6108 | 297676-301836 | 307945-311340 | 4161/3396 | 97.1% | SD |
| HsInv0209 | 2 | chr11 | 70953485-70960513 | 70965419-70972446 | 4905 | 70953485-70960513 | 70965419-70972446 | 7029/7028 | 93.7% | SD |
| HsInv0241 | 2 | chr2 | 241264234-241270541 | 241280391-241286847 | 9849 | 241260597-241273879 | 241277057-241289892 | 13283/12836 | 98.5% | SD |
| HsInv0266 | 4 | chr4 | 189106589-189108085 | 189113168-189114664 | 5082 | 189106410-189108334 | 189112920-189114842 | 1925/1923 | 98.5% | L1PA5 |
| HsInv0278 | 2 | chr5 | 180455356-180458186 | 180460457-180463248 | 2270 | 180455356-180458186 | 180460457-180463248 | 2831/2792 | 97.9% | SD |
| HsInv0340 | 2 | chr13 | 63188985-63193242 | 63237334-63241591 | 44091 | 63188926-63199511 | 63231065-63241650 | 10586/10586 | 99.9% | SD |
| HsInv0341 | 2 | chr13 | 79291903-79294413 | 79312733-79315243 | 18319 | 79289177-79295346 | 79311807-79317965 | 6170/6159 | 98.4% | L1PA3 |
| HsInv0344 | 2 | chr14 | 34079802-34086813 | 34094204-34101227 | 7390 | 34079608-34086828 | 34094188-34101421 | 7221/7234 | 99.7% | SD |
| HsInv0347 | 2 | chr14 | 60141001-60142582 | 60148138-60149719 | 5555 | 60141001-60142592 | 60148128-60149719 | 1592/1592 | 99.9% | THE1C |
| HsInv0374 | 4 | chr17 | 25967879-25973016 | 25976743-25981879 | 3726 | 25967273-25973025 | 25976734-25982494 | 5753/5761 | 99.5% | SD |
| HsInv0389 | 2 | chrX | 153219602-153228808 | 153266422-153275629 | 37613 | 153217456-153228808 | 153266422-153277781 | 11353/11360 | 99.2% | SD |
| HsInv0393 | 2 | chrX | 100739178-100743833 | 100753236-100757891 | 9402 | 100739110-100743833 | 100753236-100757955 | 4724/4720 | 99.7% | SD |
| HsInv0396 | 2 | chrX | 72132652-72141803 | 72214354-72223499 | 72550 | 72132629-72142124 | 72214044-72223522 | 9496/9479 | 99.5% | SD |
| HsInv0397 | 2 | chrX | 105396369-105408238 | 105417867-105429736 | 9628 | 105376433-105408592 | 105417499-105432595 | 32160/15097 | 93.8% | SD |
| HsInv0403 | 2 | chrX | 75278106-75282269 | 75285318-75289479 | 3048 | 75278068-75282398 | 75285195-75289517 | 4331/4323 | 99.4% | US+TEs [d] |
| HsInv0790 | 2 | chr17 | 18442024-18466271 | 18667812-18692134 | 201540 | 18442024-18466271 | 18667812-18692134 | 24248/24323 | 98.9% | SD |
| HsInv0832 | 2 | chrY | 16496132-16504854 | 16517493-16526218 | 12638 | 16496122-16504854 | 16517493-16526227 | 8733/8735 | 99.9% | SD |

[a] Annotation source. 1- Pang et al. (2013) Human Mutation, 34(2): 345-354. 2- Aguado et al. (2014) Plos Genetics, 10(3):e1004208. 3- Vicente-Salvador et al. (2016) Human Molecular Genetics, 26(3): 567-581. 4- Martínez-Fundichely et al. (in prep.)

[b] Unique sequence + several TEs (MER3, L1ME3A, Tigger1)

[c] Unique sequence + several TEs (ALuJR, L2a and MIR3)

[d] Unique sequence + several TEs (L2/L2a, L4, MamGyp, MER41C and MER4A)

**Table B.3: Inversion frequencies**.

| Inversion | Ancestral orientation | Frequency of alternative orientation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | GLB | LWK | YRI | CHB | JPT | CEU | TSI | GIH |
| HsInv0003 | Reference | 0.833 | 0.722 | 0.843 | 1.000 | 1.000 | 0.750 | 0.711 | 0.938 |
| HsInv0004 | Alternative | 0.116 | 0.019 | 0.007 | 0.144 | 0.111 | 0.200 | 0.172 | 0.163 |
| HsInv0006 | Alternative | 0.413 | 0.069 | 0.057 | 0.544 | 0.600 | 0.592 | 0.644 | 0.489 |
| HsInv0030 | Alternative | 0.934 | 0.988 | 0.979 | 1.000 | 1.000 | 0.842 | 0.844 | 0.938 |
| HsInv0031 | Alternative | 0.617 | 0.519 | 0.629 | 0.567 | 0.568 | 0.683 | 0.719 | 0.601 |
| HsInv0040 | Alternative | 0.775 | 0.772 | 0.721 | 0.900 | 0.956 | 0.792 | 0.618 | 0.809 |
| HsInv0041 | Reference | 0.500 | 0.747 | 0.629 | 0.411 | 0.422 | 0.442 | 0.428 | 0.371 |
| HsInv0045 | Reference | 0.514 | 0.463 | 0.636 | 0.611 | 0.544 | 0.450 | 0.561 | 0.393 |
| HsInv0055 | - | 0.676 | 0.395 | 0.443 | 0.844 | 0.889 | 0.783 | 0.775 | 0.753 |
| HsInv0058 | Alternative | 0.660 | 0.778 | 0.616 | 0.544 | 0.544 | 0.617 | 0.650 | 0.742 |
| HsInv0059 | Alternative | 0.753 | 0.907 | 0.907 | 0.278 | 0.322 | 0.817 | 0.839 | 0.820 |
| HsInv0061 | Alternative | 0.987 | 1.000 | 1.000 | 1.000 | 0.978 | 0.975 | 0.966 | 0.994 |
| HsInv0063 | Reference | 0.538 | 0.247 | 0.236 | 0.722 | 0.689 | 0.692 | 0.600 | 0.702 |
| HsInv0068 | Alternative | 0.874 | 0.901 | 0.884 | 1.000 | 1.000 | 0.775 | 0.767 | 0.888 |
| HsInv0069 | Recurrent | 0.495 | 0.375 | 0.551 | 0.409 | 0.600 | 0.625 | 0.589 | 0.365 |
| HsInv0072 | Recurrent | 0.976 | 0.950 | 0.922 | 1.000 | 1.000 | 0.989 | 0.992 | 0.993 |
| HsInv0092 | Reference | 0.160 | 0.265 | 0.350 | 0.078 | 0.089 | 0.067 | 0.089 | 0.129 |
| HsInv0095 | Alternative | 0.782 | 0.827 | 0.879 | 0.744 | 0.756 | 0.675 | 0.711 | 0.843 |
| HsInv0097 | Reference | 0.005 | 0.019 | 0.014 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| HsInv0098 | Reference | 0.171 | 0.346 | 0.307 | 0.078 | 0.078 | 0.117 | 0.111 | 0.096 |
| HsInv0102 | Reference | 0.188 | 0.272 | 0.350 | 0.033 | 0.044 | 0.133 | 0.156 | 0.202 |
| HsInv0105 | Alternative | 0.512 | 0.531 | 0.550 | 0.744 | 0.789 | 0.483 | 0.361 | 0.382 |
| HsInv0114 | Alternative | 0.537 | 0.196 | 0.157 | 0.856 | 0.844 | 0.625 | 0.646 | 0.652 |
| HsInv0124 | Recurrent | 0.720 | 0.854 | 0.879 | 0.978 | 0.944 | 0.400 | 0.449 | 0.719 |
| HsInv0201 | Reference | 0.561 | 0.611 | 0.664 | 0.489 | 0.367 | 0.533 | 0.644 | 0.506 |
| HsInv0209 | Reference | 0.091 | 0.184 | 0.271 | 0.022 | 0.011 | 0.017 | 0.084 | 0.000 |
| HsInv0241 | Recurrent | 0.372 | 0.582 | 0.597 | 0.405 | 0.378 | 0.167 | 0.184 | 0.312 |
| HsInv0260 | Reference | 0.183 | 0.204 | 0.129 | 0.289 | 0.422 | 0.125 | 0.083 | 0.174 |
| HsInv0266 | Reference | 0.288 | 0.269 | 0.271 | 0.250 | 0.289 | 0.208 | 0.178 | 0.500 |
| HsInv0278 | Alternative | 0.246 | 0.392 | 0.457 | 0.267 | 0.344 | 0.093 | 0.080 | 0.157 |
| HsInv0284 | Reference | 0.032 | 0.105 | 0.101 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| HsInv0340 | Recurrent | 0.167 | 0.513 | 0.507 | 0.000 | 0.000 | 0.008 | 0.034 | 0.011 |
| HsInv0341 | Recurrent | 0.077 | 0.158 | 0.257 | 0.000 | 0.023 | 0.025 | 0.022 | 0.017 |
| HsInv0344 | Recurrent | 0.436 | 0.487 | 0.493 | 0.411 | 0.221 | 0.534 | 0.483 | 0.354 |
| HsInv0347 | Recurrent | 0.195 | 0.234 | 0.286 | 0.133 | 0.122 | 0.092 | 0.118 | 0.303 |
| HsInv0374 | - | 0.475 | 0.392 | 0.243 | 0.533 | 0.711 | 0.467 | 0.461 | 0.601 |
| HsInv0379 | Reference | 0.005 | 0.000 | 0.000 | 0.022 | 0.033 | 0.000 | 0.000 | 0.000 |
| HsInv0389 | Recurrent | 0.498 | 0.958 | 1.000 | 0.221 | 0.313 | 0.178 | 0.165 | 0.478 |
| HsInv0393 | Recurrent | 0.474 | 0.317 | 0.330 | 0.647 | 0.746 | 0.367 | 0.391 | 0.657 |
| HsInv0396 | Recurrent | 0.195 | 0.263 | 0.417 | 0.029 | 0.015 | 0.159 | 0.144 | 0.209 |
| HsInv0397 | Recurrent | 0.393 | 0.608 | 0.553 | 0.456 | 0.687 | 0.156 | 0.158 | 0.291 |
| HsInv0403 | Recurrent | 0.475 | 0.650 | 0.650 | 0.824 | 0.672 | 0.261 | 0.189 | 0.328 |
| HsInv0409 | Alternative | 0.485 | 0.615 | 0.641 | 0.338 | 0.224 | 0.522 | 0.370 | 0.545 |
| HsInv0790 | Reference | 0.018 | 0.029 | 0.080 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| HsInv0832 | Recurrent | 0.336 | 0.816 | 1.000 | 0.000 | 0.043 | 0.000 | 0.000 | 0.250 |

**Table B.4: Complete list of inversion tag variants**.

Table in electronic format only.
List of all biallelic SNPs and small indels associated with each inversion with an $r^2$ value equal or higher than 0.8 in the 434 individuals analysed or restricted to specific populations or super-populations. Variant positions are in HG19. The SNP or indel allele associated with the alternative orientation is annotated in the fifth column. Remaining columns have the $r^2$ values globally (GLB) and in each superpopulation and population. $< 0.8$ indicates that the variant is not tagging the inversion at that level. NA indicates that the inversion is absent or fixed in that population or superpopulation.