U. PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

U. PORTO

FMUP FACULDADE DE MEDICINA
UNIVERSIDADE DO PORTO

U. PORTO

INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

Understanding the impact of *chromosomal inversions* on the evolution of the human genome

Tese de Candidatura ao grau de Doutor em *Biologia Básica e Aplicada* submetida ao
Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.
Porto, Portugal.

Porto
Junho, 2014

João Miguel Fernandes Alves
jalves@ipatimup.pt

João Miguel Fernandes Alves

Understanding the impact of **chromosomal inversions** on the evolution of the human genome

Tese de Candidatura ao grau de Doutor em
Biologia Básica e Aplicada submetida ao
Instituto de Ciências Biomédicas Abel Salazar
da Universidade do Porto.
Porto, Portugal.

Supervision: Professor Doutor António Amorim dos Santos
Professor Catedrático @ FCUP
IP — Grupo Genética das Populações @ IPATIMUP
Faculdade de Ciências da Universidade do Porto, Porto. PORTUGAL

Co-Supervision: Doutor Lounès Chikhi
Chargé de Recherche 1ère Classe - CNRS
IP — Genética da Conservação e das Populações @ IGC
Université Paul Sabatier, Toulouse. FRANCE

Co-Supervision: Professor Doutor Peter Heutink
IP — German Center for Neurodegenerative Diseases @ DZNE
German Center for Neurodegenerative Diseases, Tübingen. GERMANY

Research work coordinated by:

25 ANOS
A DESCOBRIR O CANCRO
ipatimup

FUNDAÇÃO CALOUSTE GULBENKIAN
Instituto Gulbenkian de Ciência

Apoio financeiro da Fundação para a Ciência e Tecnologia e do FSE
no âmbito do Quadro Comunitário de Apoio
SFRH/BD/51004/2010

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Apoio financeiro da European Molecular Biology Organization
ASFT 520-2013

EMBO

Aos meus pais,

Abstract

The significance of genomic rearrangements (structural modifications that may involve loss or gain of genetic material) in evolution and their consequences in human health have been long recognized, in particular when involving large, cytogenetically detectable changes. However, their importance in genetic research has been overshadowed for decades in favor of smaller mutational changes of a different nature more amenable to be studied at the population level. Recently, advances in genome sequencing have revealed that one subtype of these rearrangements — chromosomal inversions — is far more common than previously admitted. At this stage, a very large number of inverted rearrangements have been detected and validated in the human genome, but such finding is difficult to reconcile with the classical interpretation of inversions as a common mechanism causing subfertility or even reproductive isolation and ultimately leading to speciation. Moreover, despite the improved molecular and computational methods allowing an exponential increase in the discovery of chromosomal inversions, these rearrangements remain poorly characterized at the population level, and our understanding of its evolutionary impact is still largely limited.

Therefore, in this thesis, we make use of dense population genetic data to explore the evolutionary processes shaping two polymorphic inversions on the human genome that have been subject to rigorous scientific scrutiny over the last years — *8p23-inv* and *17q21.31-inv*. The results presented here suggest that the genetic barrier created by these rearrangements facilitate the accumulation of sequence divergence over time, leading to significant changes in the recombination landscape of the affected regions (chapter II). Moreover, by targeting a large number of present-day populations, we showed that the spread of these rearrangements can result from complex past demographic changes (and distinct dispersal trajectories), without the need of invoking selection (chapter III).

Altogether, the studies presented in this thesis highlight the role of these rearrangements as drivers of genome evolution (even at an intraspecific level) while providing relevant insights into the processes shaping the frequency and distribution of inversion polymorphisms in human populations.

Resumo

O significado evolutivo de rearranjos genómicos (modificações estruturais podendo ou não envolver perda ou ganho de material genético) e as suas consequências na saúde humana são há muito reconhecidos, principalmente em casos envolvendo alterações maiores e citogeneticamente detetáveis. Contudo, a sua importância em investigação decresceu durante décadas em prol de mudanças menores, mais simples e passíveis de ser estudadas ao nível populacional. Recentemente, os avanços em sequenciação genómica revelaram que um subtipo de rearranjo genómico — inversões cromossómicas — é mais comum do que previamente admitido. Hoje em dia, um número muito grande de rearranjos invertidos foram já detetados e validados no genoma humano. No entanto, esta descoberta é difícil de reconciliar com a interpretação clássica de inversões como mecanismos passíveis de causar subfertilidade ou mesmo isolamento reprodutivo, e em última instância especiação. Além disso, apesar dos avanços moleculares e computacionais que permitiram um aumento exponencial na descoberta de inversões cromossómicas, estes rearranjos continuam insuficientemente caracterizados ao nível populacional, e o nosso conhecimento acerca dos seus possíveis impactos evolutivos é ainda escasso.

Assim, nesta tese, procuramos usar dados genéticos de várias populações para explorar os processos evolutivos de duas inversões cromossómicas no genoma humano, que têm sido estudadas em detalhe ao longo dos últimos anos — *8p23-inv* e *17q21-inv*. Os resultados apresentados nesta tese sugerem que a barreira genética criada por este tipo de rearranjo facilita a acumulação de diferenças entre sequências com orientação oposta, levando a alterações significativas do padrão de recombinação das regiões afetadas (capítulo II). Adicionalmente, através do uso de um grande número de amostras de várias populações humanas, mostramos que a propagação destes rearranjos pode resultar de fenómenos demográficos complexos (com trajetórias de dispersão distintas), sem ser necessário invocar forças seletivas (capítulo III).

No geral, os estudos apresentados nesta tese demostram o importante papel destes rearranjos na evolução genómica (mesmo atuando a um nível intraespecífico), ao mesmo tempo que revelam detalhes extremamente importantes acerca dos processos que contribuiram para a distribuição e frequência de inversões cromossómicas, em populações humanas, observadas nos dias de hoje.

Acknowledgments

There are **many** people to whom I owe thanks.

I first became interested in population genetics during my first year at the IGC, under the supervision of Lounès Chikhi. It is, therefore, only fair to dedicate the first few lines of this acknowledgement letter to him. It is shamefully cheesy, but I have recently realized that Lounès is now a daily influence on my work. He is an incredibly smart and talented researcher (he is into Ben Stiller, though) and over the years he hugely shaped the way I think about science. I am grateful for his mentorship, continuous support, and friendship.

Most of my PhD research time was spent working with my supervisor, António Amorim, to whom I owe my profound gratitude for the support and inspiration. He was always there for me and I am extremely thankful for the many hours of advice over these years.

The success and final outcome of this project would not have been possible without the help and encouragment provided by Alexandra Lopes. I was fortunate enough to have her support and guidance.

I would also like to thank Peter Heutink who kindly agreed to co-supervise my work, for his valuable advice and support. I take this opportunity to extend my thanks to all members of the GABBA PhD Program, especially to those that continuously try to maintain such a wonderful environment for young students like me.

Thanks to all members (*past and present*) of the PCG group at the IGC for the friendship and the MANY hours of insightful discussion during these last few years. In particular, I would like to thank Bárbara Parreira for her *kind* introduction to script writting; Jordi Salmona, Rasmus Heller and Reeta Sharma for the help with NGS data analysis; A special word to Isa Pais for the ENORMOUS support during lab work. Thanks to Hugo Viaud and Maxime Verger for help modifying and developing new simulation and statistical tools. Thanks to all members of the GEPO group at IPATIMUP, (specially my office buddies) for the relaxed and friendly environment.

Thanks to IPATIMUP and the IGC for providing the platform for my work. A special word to the IGC Bioinformatics Unit (Daniel Sobral, Paulo Almeida and Renato Alves) for their willingness to help me solve problems. I would also like to thank João Costa and Isabel Marques for providing great technical support during my experimental work at the IGC. Thanks to all collaborators, including David Comas, Nadir Amir, Leonor Gusmão, Cíntia Alves, Lara Henriques, Patrícia Castro, and Ricardo Celestino.

I also need to thank Maria João Leão, for her constant support (not only, but especially) during my first year at the IGC under an Optimus Alive fellowship.

I am extremely grateful to my **family**. In particular, I thank my parents (!), my brother and sister, for their encouragement and incredible support. Finally, I would like to thank Lígia, whose companionship has been so important.

Contents

List of Figures

Chapter II

Chapter III

General Discussion

Appendix A

Appendix B

List of Tables

Chapter II

Chapter III

Appendix A

General Introduction

When chromosomal inversions were first documented by Alfred Sturtevant in the beginning of the 20th century (1921), a previously unknown instability of eukaryotic genomes structure was exposed. During the (highly productive) decades that followed Sturtevant discovery, groundbreaking studies carried out by Dobzhansky and others (*e.g.* Da Cunha & Dobzhansky, 1954) in *Drosophila* flies provided the first glimpse of the potential evolutionary value of these (and other types of) complex rearrangements, as individuals within a species, as well as distinct species, were shown to "*differ in gene arrangement*" (Dobzhansky, 1938). Given the ability of chromosomal inversions to suppress recombination between alternative structural types (*i.e.* recombination in hetero-karyotypes would result in unbalanced abortive gametes), new ideas on chromosome evolution quickly began to emerge, placing inversions as candidate players in adaptive phenomena and species divergence.

However, with the birth of molecular genetics, which took place in the 1970s, the interest in genome reorganization and structural variants decreased in favor of simpler nucleotide changes (*e.g.* SNPs). Only recently, with the development of modern molecular biology methods to variants detection, did the study of structural rearrangements re-gain momentum and it is now becoming increasingly clear that the molecular and evolutionary effects of inversions are far more complex than previously predicted. Indeed, recent comparative analysis at multiple levels have suggested that inversions are a relatively common source of genetic variation. In humans, for instance, more than 1000 inversions have been identified and validated, including all autosomes. However, the impact of most rearrangements remains either unclear or unidentified, and even with the improved molecular and statistical methods allowing an exponential increase of chromosomal inversions discovery, there is still much debate concerning the processes driving the spread of these rearrangements in the human lineage.

In this thesis, we address some of the fundamental issues regarding the role of inversion rearrangements in evolution, with a special focus on the human genome. By incorporating several complementary strategies in cytogenetics, population genetics, and genomics, we explore the molecular and population dynamics of two common inversion polymorphisms segregating in human populations. Furthermore, by applying inferential statistical methods we try to quantify the relative contribution of ancient and recent factors, including selection and drift, in shaping the frequency, distribution and genetic diversity of these inversions in present-day populations.

In **chapter I** we review the latest theoretical and empirical work dedicated to chromosomal inversions in the human genome, either as disease-associated variants or as segregating polymorphisms in human populations. We discuss the recent advances made in the structural and genetic characterization of inversion polymorphisms - highlighting the major drawbacks in the current strategies as well as important issues that have as yet received little attention. Moreover, we explore some of the evolutionary and demographic scenarios that have been invoked to explain the presence, maintenance and apparent rise in frequency of the previously mentioned chromosomal inversions in different human populations. Finally, we outline the main aims of this work, emphasizing the extent to which our findings will help clarify the role of these drastic rearrangements in the context of genome evolution.

In **chapter II**, we evaluate the fine-scale effects of chromosomal inversions in recombination by applying a method to indirectly estimate recombination rates from population genotype data. Using publicly available genotype information from the largest known inversion polymorphism segregating in the human genome, we show that the distribution of genetic recombination is largely heterogeneous between inversion types, further highlighting the role of chromosomal inversions as evolutionary significant elements acting at intraspecific level.

In **chapter III**, we reassess the distribution of one polymorphic inversion in the human genome that became the focus of intense research in the last decade due to its high degree of complexity, both in terms of genetic diversity and structural plasticity — ***17q21 inv***. Here, we will draw the evolutionary history of this inversion by refining the diversity patterns of the inversion-associated haplotype (*i.e.* H2) in several meta-population groups that have been overlooked in previous studies. Moreover, using state-of-the-art methods to haplotype reconstruction, we provide some insights concerning the processes shaping the evolution of this inversion polymorphism in the human genome.

Chapter IV will be dedicated to the main conclusions of this thesis. Here, we discuss the core findings of this work and suggest new paths for future research, including some of our *ongoing* work.

CHAPTER I

The study of chromosomal inversions in retrospect

João M Alves[1,2,3] , Alexandra M Lopes[2], Lounès Chikhi[3,4] & António Amorim[2,5]

[1]Doctoral Program in Areas of Basic and Applied Biology (*GABBA*), University of Porto, Portugal; [2]IPATIMUP - Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal; [3]Instituto Gulbenkian de Ciência (IGC), Oeiras, Portugal; [4] CNRS (Centre National de la Recherche Scientifique), Université Paul Sabatier, Ecole Nationale de Formation Agronomique, Unité Mixte de Recherche 5174 EDB (Laboratoire Évolution & Diversité Biologique), F-31062 Toulouse, France; [5]Faculdade de Ciências da Universidade do Porto, Porto, Portugal.

Review Article:
”On the Structural Plasticity of the Human Genome: Chromosomal Inversions Revisited”
Current Genomics, 2012, 13, 623-632

Introduction

Over the last years, a growing number of geneticists and evolutionary biologists are shifting their attention from single nucleotide polymorphisms (SNPs) towards bigger and more complex alterations in the architecture of eukaryotic genomes thus going back to some of the oldest genetic markers (*e.g.* Dobzhansky (1951), Kirkpatrick (2010)). With the aid of novel and powerful molecular biology techniques (*e.g.* highthroughput sequencing platforms, array-Comparative Genomic Hybridization and SNP microarrays (see Alkan *et al.* (2011) for a review), the study of the structural plasticity of the genome has gained momentum. Indeed, we are currently witnessing major advances in the field of molecular and computational genomics with increasingly high quality whole-genome data accumulating for several species and fast improvements in computational and statistical tools that allow the extraction of reliable information from these sources.

This has led to the discovery, validation and characterization of a whole set of different types of structural variants (SVs) and it is now evident that genomic variation is far more complex than previously thought (Alkan *et al.* (2009)). SVs can be defined as a wide variety of balanced and unbalanced genomic rearrangements of different sizes. They range from Copy Number Variants (CNVs) such as insertions, deletions, and duplications, all being unbalanced, to chromosomal inversions (balanced) and translocations (unbalanced or balanced). Biomedical and clinically oriented research became particularly focused in genomic imbalances, and architectural changes, with genome-wide association studies (GWAS) regularly highlighting the involvement of SVs in several genomic disorders (Craddock *et al.* (2010); Gonzalez *et al.* (2005); Fanciulli *et al.* (2007)). At present, much attention is being directed to the identification of the mechanisms and processes involved in their formation, however uncertainty remains regarding the contribution of these heteromorphisms to phenotype differences between individuals, since most variants described have been found in healthy individuals (Feuk (2007); Bailey & Eichler (2006); Conrad *et al.* (2010)).

Here, we consider a particular subtype of rearrangement — chromosomal inversions — that has been increasingly recognized as a relatively common source of variation contrary to early predictions from classical cytogenetics (Kirkpatrick (2010)).

Inversions alter the orientation of a specific genomic sequence and, for decades, they have been interpreted as a potential mechanical cause of subfertility (and ultimately reproductive isolation) since cross-over events (*i.e.* recombination) between inverted and non-inverted segments could result in unbalanced, and generally abortive, gametes (Hoffmann & Rieseberg (2008), Kirkpatrick (2010)). From an evolutionary point of view, inversions became recognized as privileged systems to study major processes (*e.g.* selection) (Navarro *et al.* (2000)) under the generalist idea often held but not always well defined that they could protect chromosomal regions from gene flow, and therefore act as an initial step towards genomic divergence (Rieseberg (2001)). Indeed, studies in chromosomal evolution have repeatedly attributed important evolutionary roles to these structural rearrangements, with several lines of evidence suggesting their involvement in phenotypic variability (Joron *et al.* (2011); Fragata *et al.* (2009)), adaptive divergence within species (Lowry *et al.* (2010), Ayala *et al.* (2010)), and in the origin and evolution of sex chromosomes in mammals (Kirkpatrick (2010)).

In humans, however, the role of inversions in disease or genome evolution remains unclear (Feuk (2007), Antonacci *et al.* (2009)). At this stage, more than 1000 inversions have been deposited in the Database of Genomic Variants (Iafrate *et al.* (2004)), involving all 22 autosomes, but the fact that only two inversion polymorphisms have been fully characterized at the population level (Stefansson *et al.* (2005), Zody *et al.* (2008), Donnelly *et al.* (2010), Salm *et al.* (2012)) clearly illustrates the necessity of studying inversion polymorphisms at a larger scale(Antonacci *et al.* (2009)).

On the detection of balanced structural variants

The detection of inversions was traditionally limited to large-scale microscopically visible rearrangements via karyotype analysis using classical G-banding techniques (Wilson *et al.* (1970), de la Chapelle *et al.* (1974), O'Neill *et al.* (2004)). With the implementation of improved comparative genomic strategies, both at population and species level, an extraordinary amount of previously unknown inversions were identified in recent years (Hara *et al.* (2011), Feuk *et al.* (2005)). While most experimental techniques (*e.g.* FISH, PGFE, Fusion-PCR) remain laborious and target-based (Turner *et al.* (2006)), where one can only test the presence of a predicted inversion in a specific genomic location, new computational approaches have been recently introduced to identify or predict the location of inversions, from SNP array data and next-generation sequencing (NGS) data, at a genome-wide level (Bansal *et al.* (2007), Sindi *et al.* (2010), Tuzun *et al.* (2005), Kidd *et al.* (2008)).

For example, Bansal *et al.* (2007) developed a statistical method to detect large polymorphic chromosomal segments ($>$ 200 Kb) that are inverted in the majority of the chromosomes in a population, with respect to the human reference sequence and applied it to HapMap data. Even with limited statistical power to detect polymorphisms at frequencies lower than 0.25 (with respect to the human reference), a list of 176 candidate inversions was generated using this model, which overlapped with several previously known inversion polymorphisms. However, since the model uses patterns of strong, long range linkage disequilibrium (LD) to access putative sites of inversion rearrangements, some predicted inversions might be artifacts and may just represent regions of high LD due to low recombination or recent selective sweeps, as noted by the authors.

More recently, Sindi *et al.* (2010) applied a probabilistic model, using differences in haplotype block structure, to identify inversion polymorphisms from this type of data. In opposition to Bansal *et al.* (2007), their method was able to predict inversion frequencies and detect inversions that are the minor allele in the population (*i.e.* where most individuals had the reference ”non-inverted” haplotype). Furthermore, they generated a set of 355 putative inversion polymorphisms using SNP data from 4 populations (CEU, YRI, CHB+JPT), overlapping with several inversion polymorphisms that have already been validated by others, or for which direct evidence exists (Alkan *et al.* (2011)). While it was possible to identify known inversion polymorphisms in both studies, hence validating the methods used, there are still several limitations that need to be considered when predicting inversion rearrangements from SNP-haplotype data. The proposed computational models rely on the assumption that (i) SNP haplotypes can be used as a proxy of the inversion status and (ii) strong LD is expected in regions harboring inversion rearrangements. As a consequence, only ancient inversions which have accumulated divergent mutations are likely to be captured. Another issue is that both models implicitly assume a single origin but multiple independent events might have given rise to the presence of a given inversion in different haplotypic backgrounds. Indeed, in the attempt to characterize 6 human disease- associated inversion polymorphisms, Antonacci *et al.* (2009) showed, with the exception of one inversion (*i.e.* 17q21.31), no remarkable correlation between SNP-based haplotypes and the inversion structure. The authors concluded that each of these inversions may have occurred multiple times in the human lineage, on different haplotype backgrounds, providing evidence of recurrence. Similar results were later observed for the 17q21.31 inversion, in a cytogenetically-based study (Rao *et al.* (2010)), where some individuals homozygous for the SNP-defined haplotype, previously thought to be completely associated with the inversion (*i.e.* H2), were in fact heterokaryotic (*i.e.* inversion heterozygotes). In the latter case, it shows that the inverted haplotype is sometimes oriented like the non-inverted haplotype. In summary, identifying inversions by means of high density SNP data is promising but far from a trivial task, and in those cases where an inversion has arisen independently on at least 2 distinct haplotype backgrounds, genotyping methods based on SNP data are prone to artifacts (*e.g.* false negatives).

Alternatively, sequence-based computational approaches have been recently introduced to detect SVs (including inversions) making use of specific sequence data signatures (see Alkan *et al.* (2011)). Among others, paired-end mapping (PEM) algorithms are showing promising results in genome wide detection of inversion rearrangements as they are able to assess the orientation of paired-end reads, therefore allowing the identification of discordant mapping to a reference genome (Korbel *et al.* (2007), Medvedev *et al.* (2009)). A series of recent publications (Tuzun *et al.* (2005), Kidd *et al.* (2008), Levy *et al.* (2007)) have applied this new method to identify structural variants in the human genome, and while 56 inversions were found using a single individual (Tuzun *et al.* (2005), Kidd *et al.* (2008) analyzing 8 genomes, identified a total of 217 inversions (but see Feuk (2007) for a more comprehensive review).

Considerable technological improvements have boosted our ability to assay inversion variants in the human genome. NGS is becoming a routinely used tool in many biological fields (Nothnagel *et al.* (2011), Baird *et al.* (2008), Davey *et al.* (2011), Garvin *et al.* (2010), Hohenlohe *et al.* (2010)), and has already contributed (and is still contributing) to a better understanding of the architecture of the human genome. Nevertheless, such technologies still represent a challenge to present-day research (Nothnagel *et al.* (2011), Alkan *et al.* (2010)). For instance, inversion breakpoints are generally enriched in runs of duplicated segments of DNA (*e.g.* segmental duplications (SDs)), which greatly limits the ability to unambiguously map breakpoint regions (Feuk *et al.* (2005)). Also, upon discovery, independent validation methods are still required to confirm the orientation of a specific chromosome segment.

Ultimately, validation studies that simultaneously take into account the limitations of the computational and molecular tools and experimental procedures are crucially needed to estimate the error rates of SNP- or NGS- inferred inversion rearrangements. Indeed, a recent review (Alkan *et al.* (2010)) explored the main limitations of the current approaches to discovering structural variants, highlighting the importance of designing algorithms that incorporate multiple methodologies to improve power, robustness, sensitivity and specificity.

Impact of inversions on genome evolution

Molecular Effects of Inversions

As balanced rearrangements, inversions do not involve quantitative alteration in the content of cellular DNA (at least no significant change in theory), but the reorganization of a genomic segment induces an alteration of the original genetic background which may have several repercussions. Although much uncertainty remains regarding the direct effects of inversions at the molecular level (*e.g.* gene expression patterns), it has been shown that some inversions can result in major phenotypic alterations. For instance, the split of the mammalian *Hoxd* gene cluster into two independent pieces, using an experimental technique (STRING) that induced an inversion rearrangement (Spitz *et al.* (2005)), was responsible for the loss of expression of *Hoxd* genes during limb development. One likely explanation to this observation is that the artificial repositioning of the genes within the inverted region, relatively to flanking regulatory elements, led to changes in patterns of gene activity (Feuk *et al.* (2005), Hoffmann & Rieseberg (2008)). Inversions exert some of their effects indirectly, by imposing new regimens of molecular evolution on the DNA sequences encompassed by them. This is due to a reduction, or even suppression of recombination within these segments in *heterokaryotypes*. As subtle as it may seem, such effect can have drastic consequences since, by acting as a genetic barrier, an inversion may ”freeze” an alternative allelic/haplotypic sequence in a population (Hey (2003)). Indeed, ever since their first identification in the 1920s (Sturtevant (1921)), inversions have been particularly investigated for their putative role in population divergence and speciation phenomena (Rieseberg (2001), Kirkpatrick & Barton (2006), Navarro & Barton (2003a), Noor *et al.* (2001), Faria *et al.* (2011)). While classic models (*e.g.* hybrid dysfunction model of speciation such as the Bateson-Dobzhansky-Muller (Orr (1996)) often rely on the idea of fertility cost to hybrids, overlooking the mechanisms by which rearrangements become established in the first place, new inversion-based speciation models (Rieseberg (2001), Kirkpatrick & Barton (2006), Navarro & Barton (2003a), Noor *et al.* (2001)) have been proposed in recent years invoking the suppression of recombination as a major process for genetic diversification and speciation. Recombination is regarded as one of the major evolutionary processes since it is responsible for the genetic shuffling and introduction of new allelic combinations, upon which selection can act (Faria *et al.* (2011)). Once an inversion arises in a population, recombination in that region becomes suppressed between chromosomes with different orientations (with the exception of double crossovers within large inverted regions). Virtually all ”suppressed-recombination” models explicitly suggest that such rearrangements provide a window of opportunity for the accumulation of differences between the two chromosomal configurations that could culminate in the evolution of reproductive isolation (Rieseberg (2001), Kirkpatrick & Barton (2006), Navarro & Barton (2003a), Noor *et al.* (2001), Faria *et al.* (2011)). At present, observations supporting these new models have been reported for several species, including birds (Huynh *et al.* (2011)), mammals (Bardhan & Sharma (2000)), insects (Joron *et al.* (2011)) and plants (Lowry *et al.* (2010)). In primates much controversy has been building up in the last years over the contribution of suppressed recombination to the divergence of ancestral populations of humans and chimpanzees and, in spite of many efforts, accelerated evolution in rearranged *versus* collinear chromosomes between the two species has not been definitely proven (Navarro & Barton (2003b), Lu *et al.* (2003), Zang *et al.* (2004)). However, since the scope of this thesis falls exclusively on human polymorphicinversions, we will not explore further the role of inversions on speciation, but instead we will focus on the possible mechanisms and processes by which inversions rise in frequency and may become established in populations.

From Genomic Novelties to Established Polymorphisms

The spread of these rearrangements can result from a combination of several factors, largely influenced by populations demography, ecology and evolutionary history. It has been argued (Ayala *et al.* (2010), Kirkpatrick & Barton (2006)) that an inversion could rise in frequency because it brings together locally adapted genes that become ”protected” from introgression, due to a local reduction in recombination. According to this scenario, the selective advantage is not directly related to the new chromosomal structure but to its favorable genetic (*i.e.* haplotypic) composition (Spirito (1998)). As a consequence, the distribution of such inversions may display clines related with local adaptation (Ayala *et al.* (2010)). Non-ecological processes, such as meiotic drive (*i.e.* a process in which an allele is over-transmitted in gametes during meiosis), might also influence the frequency and distribution of an inversion polymorphism by distorting its segregation (Kirkpatrick (2010)). However, while this is theoretically possible such processes do not appear to be general features in establishing inversions in human populations, since most rearrangements seem to segregate normally. As any other type of mutation, inversions are affected by evolutionary forces. On this basis, random genetic drift, selection and gene flow (*i.e.* migration) can play major roles in shaping their distribution and frequencies across populations. For example,Spirito *et al.* (1993), using a multi-deme model of local extinction and recolonization, observed that even underdominant inversions could, by chance, persist or rise to fixation in populations. However, the authors noted that this scenario is only achieved in cases of small effective population size, where drift causes the maintenance or rise in frequency of the rearrangement albeit the systematic pressure of selection. In contrast, if the rearrangement offers a selective advantage to the carriers, its fixation is more likely, due to the expected advantage of the inversion homozygotes (Faria *et al.* (2011), Spirito (1998), Spirito *et al.* (1993)). In humans, numerous inversion variants of different sizes segregate in populations (Feuk (2007), Bansal *et al.* (2007), Hoffmann & Rieseberg (2008)). Although the vast majority falls within the 10 to 100kb size interval, there are several inversion polymorphisms with sizes greater than 1Mb in length (Feuk (2007)). Such findings are not necessarily surprising as, in theory, the impact of an inversion is primarily related with its breakpoints location (Feuk (2007)) and if no gene is disrupted, even large inversions may be neutral and, thus, spread within and between populations through stochastic processes. However, in the absence of a robust high-throughput method to genotype balanced rearrangements, much uncertainty remains regarding the incidence of inversions in humans, how they are distributed throughout populations and their frequency as polymorphic variants.

Human Polymorphic Inversions

Aside from a small number of examples that come from indirect studies focusing on human diseases (Antonacci *et al.* (2009), Bugge *et al.* (2000), Shaw & Lupski (2004)), only a couple of inversions have been extensively characterized at the population level (Stefansson *et al.* (2005), Zody *et al.* (2008), Donnelly *et al.* (2010), Salm *et al.* (2012)). Namely, (i) the 8p23.1 inversion that spans a 4.5 Mb region and is considered the largest polymorphic inversion known in the human genome (Salm *et al.* (2012)), and (ii) the smaller but still very large 900 Kb inversion at 17q21.31 which attains relatively high frequencies in several European populations.

The 8p23.1 Inversion (*8p23.1-inv*)

Initial studies (Antonacci *et al.* (2009), Bosch *et al.* (2009)) have made clear that this particular segment presents a very complex genomic architecture mainly due to the two large blocks of segmental duplications (SDs) it contains. Although considered a neutral polymorphism (Salm *et al.* (2012)), it has been repeatedly argued (Hollox *et al.* (2008)) that, due to the presence of these highly identical structures, subsequent rearrangements via non allelic homologous recombination (NAHR i.e. a mechanism of illegitimate recombination between sequences of high identity) can cause syndromic phenotypes (*e.g.* microdeletion syndromes) in the offspring of heterozygous mothers. However, the exact molecular mechanisms leading to disease phenotypes remain to be elucidated (but see below). Another important aspect of the 8p23.1-inv is the number of genes encompassed. The region contains at least 50 genes (Bosch *et al.* (2009)), among which the BLK - *B lymphocyte kinase* - gene that has been associated with systemic lupus erythematosus (SLE), rheumatoid arthritis (RA) and other autoimmune diseases (Simpfendorfer *et al.* (2012)). Interestingly, it has been suggested that the risk alleles are specific to the non-inverted configuration (Salm *et al.* (2012)). In order to characterize its worldwide distribution, Salm et al. have recently applied an innovative approach to diploid SNP-genotype data (Salm *et al.* (2012)). Taking into consideration the limitations of most SNP-based tagging methods to identify inversions, as we noted above, the authors have designed a new and powerful multidimensional scaling (MDS) algorithm called PFIDO (Phase-Free Inversion Detection Operator) to efficiently categorize almost 2000 individuals from 56 populations by inversion status. According to their results, this inversion polymorphism displays a worldwide clinal distribution with frequencies reaching 79% in a Mozabite sample (Algeria), 63% in an Italian sample and 25% in a ”Manchu” sample (North-East Asia), which, the authors claimed would be consistent with demographic models of early human expansions out of Africa. However, since no single SNP was perfectly correlated with the inversion status, the 8p23.1 inversion may not act as an absolute recombination barrier and low levels of gene flow between inverted haplotypes may have occurred throughout its evolution. This is not necessarily surprising given the size of the inversion, which may allow for some double cross-over events.

Based on these results, the authors concluded that the *8p23-inv* appears to have evolved neutrally (or under very weak selective pressure) in humans. Moreover, given the correlation between the genetic substructure and the inversion status, they suggested that recurrent events were also infrequent across this region in the *Homo* lineage.

The 17q21.31 Inversion (*17q21.31-inv*)

Another relatively common inversion polymorphism that became the focus of intense research in the last years is located at 17q21.31. In contrast to the *8p23-inv*, early studies suggested (Stefansson *et al.* (2005)) that the 900 kb inversion polymorphism is undergoing selection in Europeans. After analyzing more than 29,000 Icelandic individuals,Stefansson *et al.* (2005) observed that females carrying either one or two copies had more children, and, applying coalescent simulations, concluded that positive selection is likely acting on the rearrangement. More recently,Zody *et al.* (2008) analyzed the evolutionary history of the same inverted region, using data from several non-human primates. According to their results, this particular segment was prone to multiple recurrent events throughout primate evolution, which contributed to the complex duplicated architecture of the region. Moreover, they highlighted the emergence of directly oriented blocks of segmental duplications (SDs) in the human H2 haplotype (inversionassociated haplotype). SDs can act as substrates of nonallelic homologous recombination (NAHR) that can result in microdeletions and microduplications events, often associated with disease (Zody *et al.* (2008), Flores *et al.* (2007), Gu *et al.* (2008)). On this basis,Zody *et al.* (2008) proposed that, due to the negative selection against the H2 haplotype, the H1 ”chromosome” rose to high frequencies in humans. However, the high frequency of the H2 chromosome in some European populations (between 5% and 35%) was explained by founder effects during the peopling of Europe following the Out-of-Africa human colonization of the continent. Similar demographic interpretations were subsequently given byDonnelly *et al.* (2010) after analyzing a more detailed global distribution of the 17q21.31 haplotypes, using SNPs and short tandem repeats (STRs) polymorphisms. They found low frequencies of the H2 haplotype in most of the 63 non-European populations. Based on these observations, their model favored a complete fixation of the H1 haplotype followed by a *de novo* occurrence in the *Homo* line, hence explaining its patchy distribution.Donnelly *et al.* (2010) also concluded that the Neolithic transition, rather than the first out of Africa wave, might be responsible for its present-day distribution in Europe. Interestingly, two new and independent studies have focused on the duplicated architecture of the 17q21.31 region to further investigate its evolutionary trajectory (Steinberg *et al.* (2012), Boettger *et al.* (2012)). Using NGS data from more than 800 individuals and applying a strategy that combined BAC-based assemblies, read depth-base copy number estimates, BAC pool sequencing and FISH,Steinberg *et al.* (2012) have identified distinct copy number polymorphisms (CNPs), including a short (CNP155) and long duplication (CNP205) exclusively associated with the H2 and H1 haplotypes, respectively. On the basis of these architectural differences, the authors were able to define four main structural haplotypes classified according to the inversion status and copy-number status. Furthermore, the frequency of the *17q21.31-inv* in the African continent was reassessed by surveying a large collection of new population samples from different sources (*e.g.* 1000Genomes). Remarkably, it was reported that the different inversion associated haplotypes (namely H2 and H2D) were segregating at fairly high frequencies (*e.g.* 7% in Maasai population) in several African ancestry groups, in opposition to earlier observations (Donnelly *et al.* (2010)). In light of these new results,Steinberg *et al.* (2012) proposed a new model where an ancestral H2 haplotype arose in central or eastern Africa and spread to southern regions before the emergence of anatomically modern humans. Approximately 2.3 Million years ago the region (re-)inverted back to the direct orientation and the resulting genomic configuration (H1) spread throughout the *Homo* lineage becoming the predominant haplotype. The authors also note that the complex duplicated architecture of extant haplotypes (H2D and H1D) represents younger evolutionary events, as the duplications in the two major clades (H1 and H2) have occurred independently. Another important conclusion from this study was finding that only one haplotype (H2D) predisposes to the syndromic 17q21.31 microdeletion, via NAHR. This configuration is characterized by the presence of directly oriented homologous SDs flanking the disease-critical region and it is associated with a duplication of the KANSL1 locus. Intriguingly, this chromosomal variant appears to be enriched in some European populations, with frequencies reaching 25%, and with virtually no genetic variation between carriers.

Similar conclusions were reached in a parallel study by Boettger *et al.* (2012), where two duplications of the KANSL1 locus, one in each genomic background (H1 and H2), have also been reported. According to the authors, these architectural changes lead to a similar alteration at the molecular level creating a new transcript of the *KANSL* gene which may have an impact on female fertility, as demonstrated in a *Drosophila* mutant (Yu *et al.* (2010)), strengthening the initial idea of selection (Stefansson *et al.* (2005)). In summary, the (i) contradictory hypotheses raised to explain the high genetic divergence observed between the inverted and non-inverted configuration in modern humans, and (ii) the conflicting scenarios proposed to explain the expansion of inversion-carrying haplotype across populations, highlight two very important features of genetic data. First, complex spatial phenomena (*e.g.* human demographic expansions, contractions, and admixture events) can produce selection-like signatures in the genome (Klopfstein *et al.* (2005), Currat *et al.* (2006)). And secondly, species-specific characteristics, such as migration rates, population size, etc., are crucial when modeling genetic data. It is well known that human populations have gone through massive changes in size and distribution in the past, including expansions, bottlenecks, and admixture events, which resulted in distinct genetic diversity patterns among populations.

However, quantifying the contribution of past events to the genetic pool of present-day populations remains a difficult task (Goldstein & Chikhi (2002), Chikhi (2009)) in which new modeling approaches are needed. Due to the complexity of the 17q21.31 region, the evolutionary history of this inversion remains a debated issue (Steinberg *et al.* (2012)). Although one cannot rule out the possibility of selection (nor a possible contribution of the *Homo neanderthalensis* (Hardy *et al.* (2005))), it is quite likely that different demographic histories could produce the same patterns of variation with or without selection. Identifying the scenarios that best explain these patterns is a challenge that may be overcome with some recent advances in population genetics inference (Goldstein & Chikhi (2002), Beaumont & Rannala (2004)).

Simulation and Inferential Tools

One important question is whether there is an appropriate statistical framework which would allow us to choose among a set of currently proposed scenarios the most appropriate. Recent advances in population genetics modeling suggest that it may be possible thanks to improved simulation programs and to Approximate Bayesian Computation (ABC), which may provide part of the answer. In a few words, the ABC framework relies on the use of very large numbers of simulations under one or several models. The observed (or real) genetic data are summarized by several summary statistics such as the number of alleles or the expected heterozygosity. The simulated data are also summarized and compared to the observed data. The scenarios or parameter values that produce simulated data that are closest to the observed data are then considered to be the most likely (Beaumont *et al.* (2002), Beaumont (2010) for a review). The ABC methodology relies on the ability to simulate genetic data very efficiently and rapidly, which was made possible thanks to the development of the coalescent theory (Hudson (1990)). In the last ten years the ABC framework has gained momentum and has been widely applied. It is the focus of intense research (Wegmann *et al.* (2009), Blum (2010), Sousa *et al.* (2009, 2011)) which suggests that it is a very flexible approach to model choice and parameter estimation. In the case of genomic data and inversions, one of the main constraints is the limitation in terms of simulating tools. While simulating large numbers of loci under the coalescent is relatively straightforward (Beaumont (2010)), even at a genome-wide scale (Carvajal-Rodríguez (2010)), the simulation of inversions has unfortunately received little attention with few exceptions (O'Reilly *et al.* (2010)).

To our knowledge, invertFREGENE (O'Reilly *et al.* (2010)) is the first (and probably the only) software allowing the introduction of a single inversion polymorphism of specific length into a population. The authors ingeniously modified a version of a previously published software (Chadeau-Hyam *et al.* (2008)) to incorporate the possibility of modeling neutral inversion rearrangements under a finite sites mutation model. The invertFREGENE software provides the possibility of simulating very large inversions, and to account for complex demographic scenarios to study the fate of inversions. Several features like the incorporation of population substructure, instantaneous expansions and contractions, are also allowed. However, there are several limitations which make it difficult for statistical inference. Indeed, invertFREGENE allows the simulation of inversions by specifying a ”target” frequency (for instance the observed frequency today) but, since the number of simulations that actually took place in order to reach this target frequency is not kept, it is difficult to identify the parameter values most likely to produce the observed data. In other words, each run only gives the output for one successful inversion that reached the given target frequency. However, given that the code is freely available it should be possible to modify it so as to circumvent this limitation. By using its core simulation engine, one could in principle develop an ABC approach that would allow us to identify models of recent human evolution with and without selection that best explain the current distribution of inversions in human populations. Recent simulation work by Li & Jakobsson (2012) has for instance shown that the use of between several hundreds and a couple of thousands of SNPs, provides major improvements in the estimation of parameters. They did not explore the issue of model choice but other studies have done it with smaller number of loci (Sousa *et al.* (2011), Fagundes *et al.* (2007)). For instance, Fagundes *et al.* (2007) were able to identify which model of human evolution was best supported using only 50 independent DNA sequences. With the arrival of genomic data, one could potentially determine how different regions of the genome are best explained by models with or without selection. Inverted regions could easily be typed for hundreds of SNPs and their demographic history compared to that of other regions. However, the general ABC framework has its limits. For instance, using forward-in-time simulators, such as invertFREGENE, could prove computationally very demanding. The ABC framework might however be modified to account for these computational challenges. One way could be to use fewer simulations and/or fewer scenarios. This would still allow us to identify the most likely among several scenarios. For instance, Rasteiro et al. (2012) used a complex spatial framework to identify among 45 different scenarios which ones were most likely to explain genetic data from European populations. They focused on genetic diversity and differentiation in Y chromosome and mtDNA data and tried to determine mgration patterns among hunter-gatherers and farmers during the Neolithic transition. Thus, despite these limitations, ABC is currently one of the most flexible and powerful approaches to explore the properties of genomic data, including inversions.

Overlooked Issues and Future Perspectives

Inversion Hotspots

From an evolutionary perspective, the presence of almost identical duplicated sequences in inversion breakpoints is also intriguing. Consider, for instance, the whole-genome comparative study by Murphy *et al.* (2005) where the genome organization of 8 mammalian species was analyzed in order to identify patterns of chromosome evolution. Using homologous synteny blocks (HSBs) they have identified several regions of chromosome breakage that apparently have been reused throughout evolution (*i.e.* independent breaks occurring at the same chromosomal sites). Interestingly, the authors have also observed that most of primate-specific breaks involve inversions that have been generated via NAHR between duplicated HSBs. Further support was later provided by Caceres *et al.* (2007) who identified another example of long-term breakpoint reuse throughout mammalian evolution in a genomic segment containing a polymorphic inversion on the human X chromosome. By sequence comparison between 28 placental mammals, the authors have suggested that at least 10 independent recurrent events must be considered to accommodate the present-day genomic structures observed in different species. In addition, recurrent events within multiple primate lineages have also been proposed for the 17q21.31 region (Zody *et al.* (2008)). Overall, these results appear to suggest that some genomic locations might exhibit greater rearrangement activity than others. One interesting possibility is that some regions represent conserved inversion hotspots that could have been maintained due to important functional or regulatory properties associated with the duplications (Caceres *et al.* (2007)). Indeed, after analyzing a specific class of duplicated structures, defined as inverted repeats (IRs), Warburton *et al.* (2004) hypothesized that their maintenance during primate evolution could be linked to important regulatory mechanisms controlling deleterious gene expression on sex-chromosomes. In conclusion, future work is still needed in order to determine the distribution of these apparently non-randomly distributed break sites, as studies analyzing at depth the population genetics of inversions are scarce in the literature.

Inversions and Recombination Rate

Many authors have also overlooked the effect of chromosomal inversions on the overall recombination rate, despite the vital role of crossing over during meiosis for proper chromosome segregation (Fledel-Alon *et al.* (2011), Stevison *et al.* (2011)). In humans, as in many other organisms (Petes (2001), Jensen-Seaman *et al.* (2004)), recombination is affected by several genomic features, such as location (*e.g.* lower rates near centromeres and higher near telomeres), and gene density (but see Coop & Przeworski (2007) for a more detailed review). Interestingly, it has also been shown that most recombination events (approximately 80%) are concentrated in small genomic regions of 1-2 kb, known as recombination hotspots (Goldstein & Weale (2001), Goldstein & Chikhi (2002), Myers *et al.* (2005)). The PRDM9 gene was recently described (Baudat *et al.* (2010), Berg *et al.* (2010)) as a major regulator of human recombination hotspots, with allelic variants of this gene influencing the differential usage of recombination hotspots. However, one might hypothesize that if an inversion happens to encompass an active hotspot, recombination will likely become inhibited in that particular region, disturbing the overall recombination rate by possibly de-localizing crossing-over events to different locations. For instance, it has been argued that, in *Drosophila* species, inversions significantly increase the recombination rate throughout the rest of the genome Stevison *et al.* (2011). Interestingly, it has also been consistently reported that polymorphisms on the H2 (inverted) haplotype in 17q21.31 are associated with an increase of the genome-wide recombination rate in heterozygous females (Stefansson *et al.* (2005), Chowdhury *et al.* (2009)). Have these inversions trapped specific variants of more active recombination hotspot determinants? That is an intriguing possibility; however, the recombination machinery might be extremely different between these species, since no recombination hotspots were ever reported in *Drosophila* (Coop & Przeworski (2007)). On an evolutionary time-scale, inversions may lead to new stabilizing points of the map of recombination events within the affected chromosome, as has been recurrently observed in the establishment of dimorphic sex chromosomes of mammals and other distantly related vertebrate taxa (Lahn & Page (1999), Skaletsky *et al.* (2003), Ross *et al.* (2005)) as well as in plants (Matsunaga (2006)). In fact, incipient heteromorphic sex chromosomes (Y and Z chromosomes) often differentiate via the accumulation of inversion rearrangements that prevent recombination over increasingly large regions with their homologues. Nevertheless, recombination and successful disjunction are maintained and therefore the recombination machinery may be more labile than would be expect *a priori*. Moreover, since current estimates suggest that approximately 25,000 putative hotspots exist in the human genome (Myers *et al.* (2005)) understanding how inversion rearrangements might affect or contribute to differential hotspot usage will be a challenging task.

Conclusions

Given the increased interest on chromosomal rearrangements, scientists are now beginning to recognize inversions as important players shaping genetic variation. Over the last decade, fundamental questions began to emerge focusing on their molecular properties (Flores *et al.* (2007), Gu *et al.* (2008)), on the mechanisms responsible for their origin (Lee *et al.* (2007)), on their evolutionary significance (Kirkpatrick (2010), Hey (2003), Brown & O'Neill (2010)) and on their role in speciation (Rieseberg (2001), Sturtevant (1921), Hey (2003), Navarro & Barton (2003a), Noor *et al.* (2001), Faria *et al.* (2011)). In humans, extensive sequencing efforts have revealed a somewhat surprising abundance of inversions segregating as polymorphisms (Feuk *et al.* (2005)). This observation is in sharp contrast with previous expectations that suggested a direct impact of inversions on fertility (Brown & O'Neill (2010)). However, as seen above, it is evident that such impact might be influenced by a combination of multiple processes (Nosil & Feder (2011), Guerrero *et al.* (2012)).

As genomic information continues to accumulate in publicly available databases, new *in silico* approaches combined with evidence of human demographic history based on archaeological and linguistic theories - might prove useful when exploring the role of inversion polymorphisms as evolutionary significant elements. Nevertheless, genetic data should be used with extreme caution as different plausible scenarios might fit the observed patterns of present day diversity (Currat *et al.* (2006)). In our opinion, due to its flexibility, robustness and efficiency, ABC strategies should be considered in future studies, as these approaches allow us to quantify the relative contribution of ancient and recent factors, including selection, in shaping the genetic structure of present-day populations. Even if ABC modeling only represents an approximation, it surely constitutes a promising statistical inferential framework to reconstruct important aspects of the evolutionary history of populations.

In conclusion, with the work presented in this thesis, we expect to contribute to the general understanding of the evolutionary implications of inversion rearrangements in the human genome, by exploring the potential role of inversions in evolution from different perspectives (chapter II and chapter III). Moreover, it should be mentioned that we are currently performing a series of *simulations* using a modified version of invertFREGENE (forward-in-time simulator) software (O'Reilly *et al.* (2010)). Although the work has not yet been finalised, as we need to expand our simulation framework to account for other possible scenarios, we have found interesting patterns when the simulated data evolves under neutral processes. In the final chapter of this thesis (chapter IV), we will discuss our preliminary data in greater detail and highlight how these new strategies may allow us to further understand the impact of this specific type of structural variability in evolution.

References

Alkan, C., *et al.* 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*, **41(10)**, 1061–67.

Alkan, C., *et al.* 2010. Limitations of next-generation genome sequence assembly. *Nature methods*, **8**, 61–65.

Alkan, C., *et al.* 2011. Genome structural variation discovery and genotyping. *Nature reviews genetics*, **12(5)**, 363–76.

Alves, J.M., *et al.* 2012. On the Structural Plasticity of the Human Genome: Chromosomal Inversions Revisited. *Current Genomics*, **13(8)**, 623–632.

Antonacci, F., *et al.* 2009. characterization of six human disease-associated inversion polymorphisms. *human molecular genetics*, **18(14)**, 2555–2566.

Auton, A., & McVean, G.A. 2007. Recombination rate estimation in the presence of hotspots. *Genome Research*, **18(14)**, 2555–2566.

Ayala, D., *et al.* 2010. chromosomal inversions, natural selection and adaptation in the malaria vector anopheles funestus. *molecular biology and evolution*, **28(1)**, 745–758.

Bailey, J.A., & Eichler, E.E. 2006. primate segmental duplications: crucibles of evolution, diversity, and disease. *nature reviews*, **7**, 552–564.

Baird, N.A., *et al.* 2008. rapid snp discovery and genetic mapping using sequenced rad markers. *plos one*, **3**, 110–115.

Bansal, V., *et al.* 2007. evidence for large inversion polymorphisms in the human genome from hapmap data. *genome research*, **17(2)**, 219–30.

Bardhan, A., & Sharma, T. 2000. meiosis and speciation: a study in a speciating mus terricolor complex. *journal of genetics*, **79**, 105–111.

Baudat, F., *et al.* 2010. prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. *science*, **237**, 836–840.

Beaumont, M.A. 2010. approximate bayesian computation in evolution and ecology. *annual review of ecology, evolution, and systematics*, **41(1)**, 379–406.

Beaumont, M.A., *et al.* 2002. approximate bayesian computation in population genetics. *journal of genetics*, **162**, 2025–2035.

Berg, I.L., *et al.* 2010. prdm9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *nature genetics*, **42**, 859–863.

Blum, M.B. 2010. approximate bayesian computation: a nonparametric perspective. *journal of the american statistical association*, **105(491)**, 1178–1187.

Boettger, L.M., *et al.* 2012. structural haplotypes and recent evolution of the human 17q21.31 region. *nature genetics*, **44(8)**, 881–885.

Bosch, N., *et al.* 2009. nucleotide, cytogenetic and expression impact of the human chromosome 8p23.1 inversion polymorphism. *plos one*, **4(12)**, e8269.

Botigué, LR., *et al.* 2013. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proceedings of the National Academy of Sciences*, **110(29)**, 11791–11796.

Brown, G.M., *et al.* 1998. Genetic analysis of meiotic recombination in humans by use of sperm typing: reduced recombination within a heterozygous paracentric inversion of chromosome 9q32-q34.3. *American Journal of Human Genetics*, **62**, 1484–1492.

Brown, J.F., & O'Neill, R.J. 2010. chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *annual review of genomics and human genetics*, 291–316.

Browning, SR., & Browning, BL. 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics*, **81**, 1084–1097.

Bugge, M., *et al.* 2000. disease associated balanced chromosome rearrangements: a resource for large scale genotype-phenotype delineation in man. *journal of medical genetics*, **37**, 858–865.

Caceres, M., *et al.* 2007. a recurrent inversion on the eutherian x chromosome. *proceedings of the national academy of sciences*, **104(47)**, 18571–18576.

Campbell, CD., *et al.* 2011. Population-genetic properties of differentiated human copy-number polymorphisms. *American Journal of Human Genetics*, **88(3)**, 317–332.

Carvajal-Rodríguez, A. 2010. simulation of genes and genomes forward in time. *current genomics*, **11**, 58–61.

Chadeau-Hyam, M., *et al.* 2008. fregene: simulation of realistic sequence-level data in populations and ascertained samples. *bmc bioinformatics*, **9**, 364.

Cheung, V.G., *et al.* 2007. Polymorphic variation in human meiotic recombination. *American Journal of Human Genetics*, **80(3)**, 526–530.

Chikhi, L. 2009. update to chikhi et al.'s ”clinal variation in the nuclear dna of europeans” (1998): genetic data and storytelling-from archaeogenetics to astrologenetics-. *human biology*, **81(5-6)**, 639–643.

Chowdhury, R., *et al.* 2009. genetic analysis of variation in human meiotic recombination. *plos genetics*, **5**, e1000648.

Clark, A.G., *et al.* 2010. Contrasting methods of quantifying fine structure of human recombination. *Annual Reviews in of Genomics and Human Genetics*, **11**, 45–64.

Conrad, D.F., *et al.* 2010. origins and functional impact of copy number variation in the human genome. *nature*, **464**, 704–712.

Conrad, D.F., & Hurles, M.E. 2007. Polymorphic variation in human meiotic recombination. *American Journal of Human Genetics*, **80(3)**, 526–530.

Consortium, International HapMap. 2003. The International HapMap Project. *Nature*, **426(6968)**, 789–796.

Consortium, The 1000 Genomes Project. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Coop, G., & Przeworski, M. 2007. an evolutionary view of human recombination. *nature reviews genetics*, **8(1)**, 23–34.

Craddock, N., *et al.* 2010. genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls. *nature*, **464(7289)**, 713–20.

Currat, M., *et al.* 2006. comment on ”ongoing adaptive evolution of aspm, a brain size determinant in homo sapiens” and ”microcephalin, a gene regulating brain size, continues to evolve adaptively in humans”. *science*, **313**, 172.

Davey, J.W., *et al.* 2011. genome-wide genetic marker discovery and genotyping using next-generation sequencing. *nature review genetics*, **12**, 499.

de la Chapelle, A., *et al.* 1974. pericentric inversions of human chromosomes 9 and 10. *american journal of human genetics*, **26**, 746–766.

Delaneau, O., *et al.* 2013. Haplotype estimation using sequence reads. *American Journal of Human Genetics*, **93(4)**, 787–796.

Deng, Y., & Tsao, B.P. 2010. Genetic susceptibility to systemic lupus erythematosus in the genomic era. *Nature Reviews Rheumatology*, **6**, 683–692.

Depristo, M., *et al.* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Dobzhansky, T. 1951. *genetics and the origin of species.* oxford: columbia university press: new york.

Donnelly, M.P., *et al.* 2010. the distribution and most recent common ancestor of the 17q21 inversion in humans. *american journal of human genetics*, **86(2)**, 161–71.

Fagundes, N.J.R., *et al.* 2007. statistical evaluation of alternative models of human evolution. *proceedings of the national academy of sciences*, **104(17)**, 614–619.

Fanciulli, M., *et al.* 2007. fcgr3b copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *nature genetics*, 721–723.

Faria, R., *et al.* 2011. role of natural selection in chromosomal speciation. *In: encyclopedia of life sciences*. chichester, uk: John Wiley & Sons, Ltd.

Faria, R., & Navarro, A. 2010. Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends in Ecology and Evolution*, **25**, 660–669.

Farré, M., *et al.* 2013. Recombination Rates and Genomic Shuffling in Human and Chimpanzee - A New Twist in the Chromosomal Speciation Theory. *Molecular Biology and Evolution*, **30(4)**, 853–864.

Fearnhead, P., *et al.* 2004. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics*, **167(4)**, 2067–2081.

Feuk, L. 2007. inversion variants in the human genome: role in disease and genome architecture. *genome medicine*, **2(2)**, 11.

Feuk, L., *et al.* 2005. discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee dna sequence assemblies. *plos genetics*, **1(4)**, e56.

Fledel-Alon, A., *et al.* 2011. variation in human recombination rates and its genetic determinants. *plos one*, **6(6)**, e20321.

Flores, M., *et al.* 2007. recurrent dna inversion rearrangements in the human genome. *proceedings of the national academy of sciences*, **104(15)**, 6099–106.

Fragata, I., *et al.* 2009. contrasting patterns of phenotypic variation linked to chromosomal inversions in native and colonizing populations in drosophila subobscura. *journal of evolutionary biology*, **23(1)**, 112–123.

Garvin, M.R., *et al.* 2010. application of single nucleotide polymorphisms to non-model species: a technical review. *molecular ecology resources*, **10(6)**, 915–934.

Goldstein, D.B., & Chikhi, L. 2002. human migrations and population structure : what we know and why it matters. *annual review of genomics and human genetics*, **3**, 129–152.

Goldstein, D.B., & Weale, M.E. 2001. population genomics: linkage disequilibrium holds the key. *current biology*, **11**, 576–579.

Gonzalez, E., *et al.* 2005. the influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility. *science*, **307**, 1434–40.

Goudet, J. 2005. Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184–186.

Gu, W., *et al.* 2008. mechanisms for human genomic rearrangements. *pathogenetics*, **1(4)**.

Guerrero, R.F., *et al.* 2012. coalescent patterns for chromosomal inversions in divergent populations. *philosophical transactions of the royal society b: biological sciences*, **367(1587)**, 430–438.

Hara, Y., *et al.* 2011. abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. *bmc evolutionary biology*, **11(1)**, 308.

Hardy, J., *et al.* 2005. evidence suggesting that homo neanderthalensis contributed the h2 mapt haplotype to homo sapiens. *biochemical society transactions*, **33(4)**, 582–585.

Henn, BM., *et al.* 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genetics*, **8(1)**, e1002397.

Hey, J. 2003. speciation and inversions: chimps and humans. *bioessays*, 825–828.

Hoffmann, A.A., & Rieseberg, L.H. 2008. revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation-. *annual review of ecology, evolution, and systematics*, **39**, 21–42.

Hohenlohe, P.A., *et al.* 2010. population genomics of parallel adaptation in three-spine stickleback using sequenced rad tags. *plos genetics*, **6(2)**, e1000862.

Hollox, E.J., *et al.* 2008. defensins and the dynamic genome: what we can learn from structural variation at human chromosome band 8p23.1. *genome research*, **18**, 1686–1697.

Hubert, R., *et al.* 1994. High resolution localization of recombination hot-spots using sperm typing. *Nature Genetics*, **7**, 420–434.

Hudson, R. 1990. gene genealogies and the coalescent process. *In: oxford surveys in evolutionary biology*. oxford, new york: oxford university press: oxford.

Huynh, L.Y., *et al.* 2011. chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (zonotrichia albicollis). *heredity*, **106**, 537–546.

Iafrate, A.J., *et al.* 2004. Detection of large-scale variation in the human genome. *Nature Genetics*, **36(9)**, 949–51.

Jensen-Seaman, M.I., *et al.* 2004. comparative recombination rates in the rat, mouse, and human genomes. *genome research*, **14**, 528–538.

Jobling, M.A, Hurles, M.E., & Tyler-Smith, C. 2004. *Human Evolutionary Genetics: Origins, Peoples and Disease*. New York: Garland Science: New York.

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Joron, M., *et al.* 2011. chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *nature*, **477**, 203–206.

Keinan, A., & Reich, D. 2010. Human population differentiation is strongly correlated with local recombination rate. *PLoS Genetics*, **6(3)**, e1000886.

Kidd, J.M., *et al.* 2008. mapping and sequencing of structural variation from eight human genomes. *nature*, **453(7191)**, 56–64.

Kirkpatrick, M. 2010. how and why chromosome inversions evolve. *plos biology*, **8**, 9.

Kirkpatrick, M., & Barton, N. 2006. chromosome inversions, local adaptation and speciation. *genetics*, **173(1)**, 419–434.

Klopfstein, S., *et al.* 2005. the fate of mutations surfing on the wave of a range expansion. *molecular biology and evolution*, **23**, 482–490.

Korbel, J.O., *et al.* 2007. paired-end mapping reveals extensive structural variation in the human genome. *science*, **318**, 420–426.

Laayouni, H., *et al.* 2011. Similarity in recombination rate estimates highly correlates with genetic differentiation in humans. *PLoS One*, **6(3)**, e17913.

Lahn, B.T., & Page, D.C. 1999. four evolutionary strata on the human x chromosome. *science*, **286(5441)**, 964–967.

Lee, J.A., *et al.* 2007. a dna replication mechanism for generating non-recurrent rearrangements associated with genomic disorders. *cell*, **131**, 1235–1247.

Levy, S., *et al.* 2007. the diploid genome sequence of an individual human. *plos biology*, **5**, e254.

Li, H., & Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **Epub**.

Li, J., *et al.* 2006. A new method for detecting human recombination hotspots and its applications to the HapMap ENCODE data. *American Journal of Human Genetics*,

79(4), 628–639.

Li, N., & Stephens, M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165(4)**, 2213–2233.

Li, S., & Jakobsson, M. 2012. estimating demographic parameters from large scale population genomic data using approximate bayesian computation. *bmc genetics*, **13**, 22.

Lichten, M., & Goldman, A.S. 1995. Meiotic recombination hotspots. *Annual Review of Genetics*, **29**, 423–444.

Lopes, AM., *et al.* 2013. Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene DMRT1. *PLoS Genetics*, **9(3)**, e1003349.

Lowry, D.B., *et al.* 2010. a widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *plos biology*, **8(9)**, e1000500.

Lu, J., *et al.* 2003. comment on ”chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes”. *science*, **302**, 988.

MacDonald, J.R., *et al.* 2013. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*, **42(1)**, 986–992.

Matsunaga, S. 2006. sex chromosome-linked genes in plants. *genes & genetic systems*, **81**, 219–226.

McVean, G.A., *et al.* 2004. The fine-scale structure of recombination rate variation in the human genome. *Science*, **304(5670)**, 581–584.

Medvedev, p., *et al.* 2009. computational methods for discovering structural variation with next-generation sequencing. *nature methods*, **6**, 13–20.

Murphy, W.J., *et al.* 2005. dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *science*, **309**, 613–617.

Myers, S., *et al.* 2005. a fine-scale map of recombination rates and hotspots across the human genome. *science*, **310**, 321–324.

Myers, S., *et al.* 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, **40**, 1124–1129.

Navarro, A., & Barton, N. 2003a. accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *evolution*, **57(3)**, 447–59.

Navarro, A., & Barton, N. 2003b. chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes. *science*, **300(5617)**, 321–324.

Navarro, A., *et al.* 1997. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics*, **146**, 695–709.

Navarro, A., *et al.* 2000. effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in drosophila. *genetics*, **155(2)**, 685–698.

Noor, M.A.F., *et al.* 2001. chromosomal inversions and the reproductive isolation of species. *proceedings of the national academy of sciences*, **98**, 12084–12088.

Nosil, P., & Feder, J.L. 2011. genomic divergence during speciation: causes and consequences. *philosophical transactions of the royal society b: biological sciences*, **367(1587)**, 332–342.

Nothnagel, M., *et al.* 2011. technology-specific error signatures in the 1000 genomes project data. *human genetics*, 505–516.

O'Neill, R.J., *et al.* 2004. centromere dynamics and chromosome evolution in marsupials. *the journal of heredity*, **95(5)**, 375–381.

O'Reilly, P.F., *et al.* 2010. invertfregene: software for simulating inversions in population genetic data. *bioinformatics*, **26(6)**, 838–840.

Orr, A. 1996. Dobzhansky, bateson, and the genetics of speciation. *genetics*, **144**, 1331–1335.

Parvanov, E.D., *et al.* 2010. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science*, **327(5967)**, 835.

Petes, T.D. 2001. meiotic recombination hot spots and cold spots. *nature review genetics*, **2**, 360–369.

Pickrell, J.K., *et al.* 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, **19**, 826–837.

Purandare, S.M., & Patel, P.I. 1997. Recombination hot spots and human disease. *Genome Research*, **7(8)**, 773–786.

Purcell, S., *et al.* 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, **81(3)**, 559–575.

Rao, P.N., *et al.* 2010. recurrent inversion events at 17q21.31 microdeletion locus are linked to the mapt h2 haplotype. *cytogenetic and genome research*, **90095**, 275–279.

Richards, M., *et al.* 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. *American Journal of Human Genetics*, **67(5)**, 12511276.

Rieseberg, L.H. 2001. chromosomal rearrangements and speciation. *trends in ecology and evolution*, **16(7)**, 351–358.

Ross, M.T., *et al.* 2005. the dna sequence of the human x chromosome. *nature*, **434**, 325–337.

Salm, M.P.A., *et al.* 2012. he origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *genome research*, **22(6)**, 1144–1153.

Serre, D., *et al.* 2005. Large-scale recombination rate patterns are conserved among human populations. *Genome Research*, **15(11)**, 1547–1552.

Shaw, C.J., & Lupski, J.R. 2004. implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *human molecular genetics*, **13(1)**, 57–64.

Simpfendorfer, K.R., *et al.* 2012. the autoimmunity-associated blk haplotype exhibits cis-regulatory effects on mrna and protein expression that are prominently observed in b cells early development. *human molecular genetics*, **21(17)**, 3918–3925.

Sindi, S.S., *et al.* 2010. identification and frequency estimation of inversion polymorphisms from haplotype data. *journal of computational biology*, **17(3)**, 517–531.

Skaletsky, H., *et al.* 2003. the male-specific region of the human y chromosome is a mosaic of discrete sequence classes. *nature*, **423(6942)**, 825–37.

Sousa, V., *et al.* 2009. approximate bayesian computation without summary statistics: the case of admixture. *journal of genetics*, **181**, 187–197.

Sousa, V., *et al.* 2011. population divergence with or without admixture: selecting models using an abc approach. *heredity*, **108**, 521–530.

Spirito, F. 1998. *endless forms: species and speciation*. oxford, new york: oxford university press: oxford.

Spirito, F., *et al.* 1993. the establishment of underdominant chromosomal rearrangements in multi-deme systems with local extinction and colonization. *theoretical population biology*, **44**, 80–94.

Spitz, F., *et al.* 2005. inversion induced disruption of the hoxd cluster leads to the partition of regulatory landscapes. *nature genetics*, **37**, 889–893.

Stefansson, H., *et al.* 2005. a common inversion under selection in europeans. *nature genetics*, **37(2)**, 129–37.

Steinberg, K.M., *et al.* 2012. structural diversity and african origin of the 17q21.31 inversion polymorphism. *nature genetics*, **44(8)**, 872–880.

Stephens, M., *et al.* 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Stevison, L.S., *et al.* 2011. effects of inversions on within- and between-species recombination and divergence. *genome biology and evolution*, **3**, 830–841.

Sturtevant, A.H. 1921. a case of rearrangement of genes in drosophila. *proceedings of the national academy of sciences*, **7**, 235–237.

Sudmant, PH., *et al.* 2010. Diversity of human copy number variation and multicopy genes. *Science*, **330(6004)**, 641–646.

Tamura, K., *et al.* 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution. *Molecular Biology and Evolution*, **28**, 2731–2739.

Turner, D.J., *et al.* 2006. assaying chromosomal inversions by single-molecule haplotyping. *nature methods*, **3**, 439–445.

Tuzun, E., *et al.* 2005. fine-scale structural variation of the human genome. *nature genetics*, **37**, 727–732.

Warburton, P.E., *et al.* 2004. inverted repeat structure of the human genome: the x-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *genome research*, **14**, 1861–1869.

Wegmann, D., *et al.* 2009. efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *journal of genetics*, **182**, 1207–1218.

Wilson, M.G., *et al.* 1970. inherited pericentric inversion of chromosome nb. 4. *american journal of human genetics*, **22**, 679–690.

Yu, L., *et al.* 2010. e(nos)/cg4699 required for nanos function in the female germ line of drosophila. *genesis*, **48**, 161–170.

Zang, J., *et al.* 2004. testing the chromosomal speciation hypothesis for humans and chimpanzees. *genome research*, **14**, 845–851.

Zilhão, J. 2001. Radiocarbon evidence for maritime pioneer colonization at the origins of farming in west Mediterranean Europe. *Proceedings of the National Academy of Sciences*, **98(24)**, 1418014185.

Zody, M.C., *et al.* 2008. evolutionary toggling of the mapt 17q21.31 inversion region. *nature genetics*, **40(9)**, 1076–83.

CHAPTER II

Inversion polymorphism determines local recombination heterogeneity across human population

João M Alves[1,2,3], Lounès Chikhi[3,4], António Amorim[2,5] & Alexandra M Lopes[2]

[1]Doctoral Program in Areas of Basic and Applied Biology (*GABBA*), University of Porto, Portugal; [2]IPATIMUP - Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal; [3]Instituto Gulbenkian de Ciência (IGC), Oeiras, Portugal; [4] CNRS (Centre National de la Recherche Scientifique), Université Paul Sabatier, Ecole Nationale de Formation Agronomique, Unité Mixte de Recherche 5174 EDB (Laboratoire Évolution & Diversité Biologique), F-31062 Toulouse, France; [5]Faculdade de Ciências da Universidade do Porto, Porto, Portugal.

Original Researh Article:
”The *8p23-inv* polymorphism determines local recombination heterogeneity across human populations”
Genome, Biology & Evolution, 2014
(**doi:** 10.1093/gbe/evu064)

Abstract

For decades, chromosomal inversions have been regarded as fascinating evolutionary elements as they are expected to suppress recombination between chromosomes with opposite orientations, leading to the accumulation of genetic differences between the two configurations over time. Here, making use of publicly available population genotype data for the largest polymorphic inversion in the human genome (*8p23-inv*), we assessed whether this inhibitory effect of inversion rearrangements led to significant differences in the recombination landscape of two homologous DNA segments, with opposite orientation. Our analysis revealed that the accumulation of genetic differentiation is positively correlated with the variation in recombination profiles. The observed recombination dissimilarity between inversion types is consistent across all populations analyzed, and surpasses the effects of geographic structure, suggesting that both structures (orientations) have been evolving independently over an extended period of time, despite being subjected to the very same demographic history. Aside this mainly independent evolution, we also identified a short segment (350 kb, less than 10% of the whole inversion) in the central region of the inversion where the genetic divergence between the two structural haplotypes is diminished. While it is difficult to demonstrate it, this could be due to gene flow (possibly via double-crossing over events), which is consistent with the higher recombination rates surrounding this segment. This study demonstrates for the first time that chromosomal inversions influence the recombination landscape at a fine-scale, and highlights the role of these rearrangements as drivers of genome evolution.

Introduction

Genetic recombination is one of the key evolutionary processes affecting variation throughout the genome. This process, generally mediated by homology, involves the exchange of genetic information between two homologous chromosomes (or between different albeit homologous regions of the same chromosome) (Faria & Navarro (2010)), potentially disrupting the relationship between alleles at those loci and ensuring new allelic combinations. Traditionally, recombination has been estimated using pedigree-based or sperm-typing methods, by directly counting the products of meiosis (Hubert *et al.* (1994), Brown *et al.* (1998)).

Given that such techniques are impracticable at a population level (Clark *et al.* (2010)), recent years have witnessed the rise and improvement of statistical inferential approaches to indirectly detect recombination events at a genome wide scale, from population genetic data (Li & Stephens (2003), McVean *et al.* (2004), Auton & McVean (2007)). In general, these methods rely on the assumption that Linkage Disequilibrium LD (*i.e.* non-random association of alleles) is significantly reduced in regions that are exposed to recombination (Clark *et al.* (2010)). Studies applying these alternative methods (Fearnhead *et al.* (2004), Li *et al.* (2006)) validated the empirical evidence from the late 1990s (Lichten & Goldman (1995), Purandare & Patel (1997)) that suggested an uneven distribution of recombination along the genome. In other words, recombination appears to be clustered in specific genomic regions, now known as recombination hotspots. Such finding encouraged the emergence of fine-scale comparative analysis at multiple levels (Jensen-Seaman *et al.* (2004), Serre *et al.* (2005), Cheung *et al.* (2007)) and it has become increasingly clear that, even though the global recombination landscape is largely conserved among humans, local recombination patterns are significantly heterogeneous between different present-day populations (Serre *et al.* (2005), Cheung *et al.* (2007), Keinan & Reich (2010), Laayouni *et al.* (2011), Fledel-Alon *et al.* (2011)).

Interestingly, a particular type of chromosomal rearrangement inversions became subject of intense research in the last few years due to their negative effects on recombination (Hoffmann & Rieseberg (2008), Kirkpatrick (2010), Alves *et al.* (2012)). Inversions are known to suppress recombination between differently oriented chromosomal segments, and it has been suggested that such rearrangements may play an important role shaping species divergence and evolution (Kirkpatrick & Barton (2006), Ayala *et al.* (2010)). At present more than 1000 inversions have been identified and validated in the human genome (Iafrate *et al.* (2004)), but only a few have been studied at a population scale (Stefansson *et al.* (2005), Antonacci *et al.* (2009), Donnelly *et al.* (2010), Steinberg *et al.* (2012), Salm *et al.* (2012)).

In the following study we focus on the human 8p23 region. This region harbors the largest polymorphic inversion known in the human genome (Antonacci *et al.* (2009)) A 4Mb long paracentric inversion that shows a strong clinal distribution in human populations, with frequencies varying between 80% (in Africa), 50% (in Europe) and 20% (in Asia) (Salm *et al.* (2012)). Even though the 8p23 region harbors several candidate loci for natural selection (Pickrell *et al.* (2009)), and genes related to autoimmune disorders (Deng & Tsao (2010)), a model of neutral evolution shaped mostly by demographic factors has been suggested to explain its current distribution (Salm *et al.* (2012)).

Here, we used this inversion polymorphism to study the evolution of recombination in a novel way. Taking advantage of the fact that this inversion is frequent in several human populations, our first aim is to quantify the distribution of recombination along the 4MB genomic segment and to determine whether the recombination landscape has evolved differently in the two chromosomal orientations. While recombination is expected to be suppressed (or extremely rare) between heterokaryotypes (*i.e.* individuals heterozygous for the orientation), chromosomes with the same orientation should still be able to recombine freely across the region (Conrad & Hurles (2007)). Indeed, chromosomal segments with opposing orientations may be seen as two different ”sub-populations” subjected to the same demographic history while independently accumulating mutations and recombination events, leading to increasing divergence over time. By comparing the recombination patterns of inverted and non-inverted chromosomes, we thus expect to gain insight on the evolution of recombination following a drastic chromosome rearrangement.

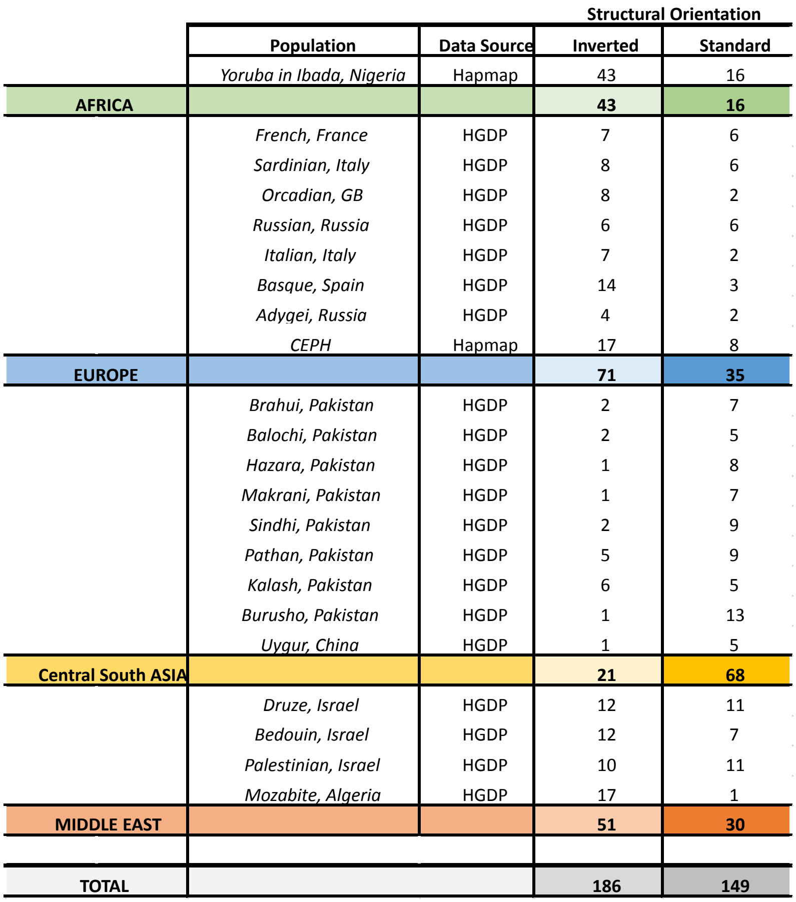
Materials & Methods

Genotype Data, Inference of Inversion Status & Population sets

Genotype data were obtained from the Stanford Human Genome Diversity Project (HGDP) website and subsequently stored as a single raw file using the *Plink* software (Purcell *et al.* (2007)). Individuals were grouped according to continental origin, as in Salm *et al.* (2012), and 4 distinct groups were thus defined, Sub-Saharan Africa, Europe, Middle East, and Central South Asia. Altogether 1,447 SNPs were identified for the whole data set (685 individuals) with an average spacing of 3.1 kb. Note that the geographical groupings above are only used as practical units devised to achieve sufficient sample sizes. The *PFIDO* (Phase-Free Inversion Detection Operator) *R* package (Salm *et al.* (2012)) was then used to infer the orientation of the 8p23 region. This package uses a database of genotypes for which the inversion profile is known and a statistical approach to then assign new multi-locus genotypes to one of the three possible inversion statuses (*i.e.* two different homokaryotypes and one heterokaryotype). This step was independently applied to the 4 meta-population groups since, in each region, different SNPs may be statistically associated to the inversion status. Moreover, since no single SNP can be used as proxy of the inversion status (*i.e.* no inversion marker has yet been identified), PFIDO was applied following the package recommendations on the entire SNP set.

Due to the low coverage of the HGDP SNP panel, we were unable to accurately predict the 8p23 orientation in the Sub-Saharan individuals with PFIDO. Given that the International Hapmap Project (Consortium (2003)) comprises a larger density marker panel ($>$ 4000 SNPs encompassing the region), we retrieved the available genotype data for the YRI population (Yoruba in Ibadan, Nigeria) from the project Phase II (release 23) and applied the same procedure as above. We were thus able to infer the 8p23 orientation for all YRI individuals, and used them as our African group for the remaining of the study. To minimize any bias related to the source of the data for the African sample, we additionally obtained genotype information and inferred the 8p23 orientation in individuals from the Hapmap CEU population (Utah residents with Northern and Western European ancestry from the CEPH collection). These samples were merged to the HGDP European set once we confirmed that no bias could be identified (see below).

Once the orientation was determined for the different groups, each was again split according to the inversion status. As our working set was mostly composed of unphased genotype data, heterokaryotypes were excluded from the analysis to avoid inaccurate recombination rate estimates. A list of the number of individuals used in this study is shown in **Table I**.



			Structural Orientation	
	Population	Data Source	Inverted	Standard
	Yoruba in Ibada, Nigeria	Hapmap	43	16
AFRICA			**43**	**16**
	French, France	HGDP	7	6
	Sardinian, Italy	HGDP	8	6
	Orcadian, GB	HGDP	8	2
	Russian, Russia	HGDP	6	6
	Italian, Italy	HGDP	7	2
	Basque, Spain	HGDP	14	3
	Adygei, Russia	HGDP	4	2
	CEPH	Hapmap	17	8
EUROPE			**71**	**35**
	Brahui, Pakistan	HGDP	2	7
	Balochi, Pakistan	HGDP	2	5
	Hazara, Pakistan	HGDP	1	8
	Makrani, Pakistan	HGDP	1	7
	Sindhi, Pakistan	HGDP	2	9
	Pathan, Pakistan	HGDP	5	9
	Kalash, Pakistan	HGDP	6	5
	Burusho, Pakistan	HGDP	1	13
	Uygur, China	HGDP	1	5
Central South ASIA			**21**	**68**
	Druze, Israel	HGDP	12	11
	Bedouin, Israel	HGDP	12	7
	Palestinian, Israel	HGDP	10	11
	Mozabite, Algeria	HGDP	17	1
MIDDLE EAST			**51**	**30**
TOTAL			**186**	**149**

Table I: **Number of individuals and corresponding inversion status by geographical origin.**

Also, only SNPs identified within the HGDP data were considered for subsequent analysis, thus minimizing missing data. Finally, a Principal Component Analysis (PCA) was conducted prior to the estimation of recombination rates to examine the consistency of the data. As **Figure 1** shows, all individuals clustered according to the inversion status and continental origin regardless of the data set used.

Recombination rates Estimation

Estimates of recombination rate were obtained using the rhomap program distributed within the *LDHat* package (v2.2) (Auton & McVean (2007)). *LDHat* uses a composite-likelihood scheme, where population-scaled recombination rates are estimated between each pair of consecutive SNPs. Independent runs of rhomap were carried out for all geographical- and orientation- specific groups for a total of 10,000,000 iterations with a burn-in of 100,000 iterations. Samples were taken every 5,000 iterations after the burn-in, with block and hotspot penalties set to zero. Given that (i) *LDHat* ignores non-polymorphic positions, and (ii) the HGDP panel is composed of SNPs that are not globally segregating as polymorphisms (*i.e.* some SNPs are monomorphic in certain populations), the comparison of the results for different populations requires that intervals be defined which will then be comparable. To do this we adopted a similar approach to McVean *et al.* (2004) and the local recombination rates were ”averaged” by summing all estimated values over non-overlapping segments of 20kb. This approach has the advantage of allowing a direct comparison of recombination estimates, while maintaining a good resolution of the recombination landscape.

Recombination dissimilarity & Genetic differentiation

To determine whether the different structural haplotypes had similar recombination profiles we compared the sets of *rho* values estimated in the previous section. Spearman rank correlation coefficients were obtained between each pair of structural haplotype (Inverted *versus* Standard) within each geographical group using a 500 kb sliding-window approach (in other words: 25 *rho* values were used to compute one correlation coefficient). These coefficients were then transformed into dissimilarity values by subtracting them from 1, as in Laayouni *et al.* (2011). In parallel, the same 500kb blocks were used to estimate the differentiation between the two structural orientations. F_{st} values were computed using the *Hierfstat R* package (Goudet (2005)). Finally the obtained dissimilarity measures were compared to the corresponding F_{st} estimates. All

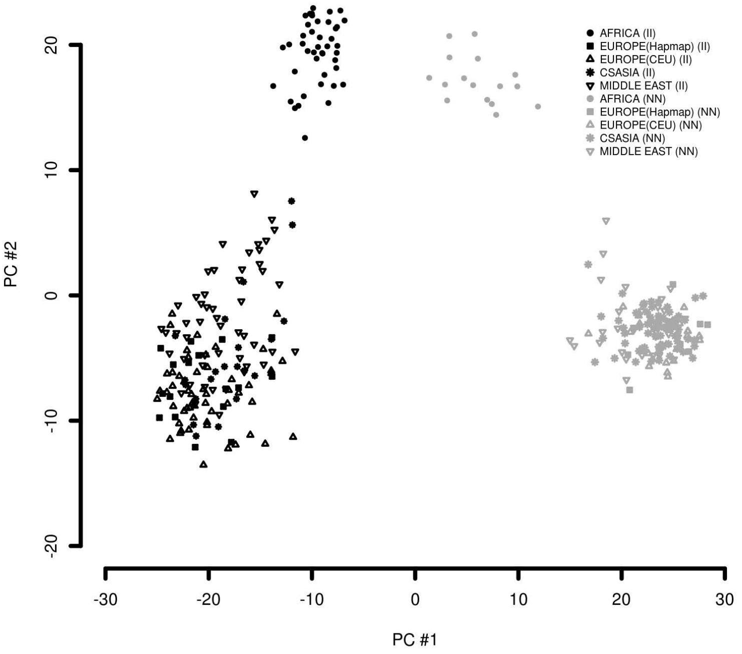


Figure 1: **Global genetic stratification at the 8p23 region** - *Principal Component Analysis* (PCA) performed on HGDP (n = 251) and HapMap (n = 84) population genotype data. A total of 1,447 SNPs were used. Each dot corresponds to one individual, with distinct symbols representing geographical- and orientation- specific groups. The first principal component (*i.e.* horizontal axis) illustrates the strong genetic differentiation between the two main haplotypes (Inverted / Standard)

statistical analyses were performed using the *R* software. To further evaluate the variation in the distribution of recombination within the 8p23 region, we applied the same method and performed pairwise comparisons between geographical groups within each structural haplotype.

Results

Recombination patterns along the 8p23 region

Recombination patterns along the 8p23 region Population-scaled recombination rates (4Ne*r*/Kb) were inferred for a total of 335 individuals from eight distinct groups (according to inversion-status and geographical origin) using the 1,447 SNPs identified. The cumulative plot of the proportion of recombination occurring in a given fraction of the sequence (**Figure 2**) shows that there is an uneven distribution of recombination across the interval. While such distribution is expected based on previous genome-wide recombination maps (McVean *et al.* (2004), Clark *et al.* (2010)), it is interesting to note differences between the populations analyzed (see below).

Figure 3 shows the recombination profile of each group for the 8p23 region. While there is a good overall agreement in the large-scale patterns of recombination (*i.e.* the

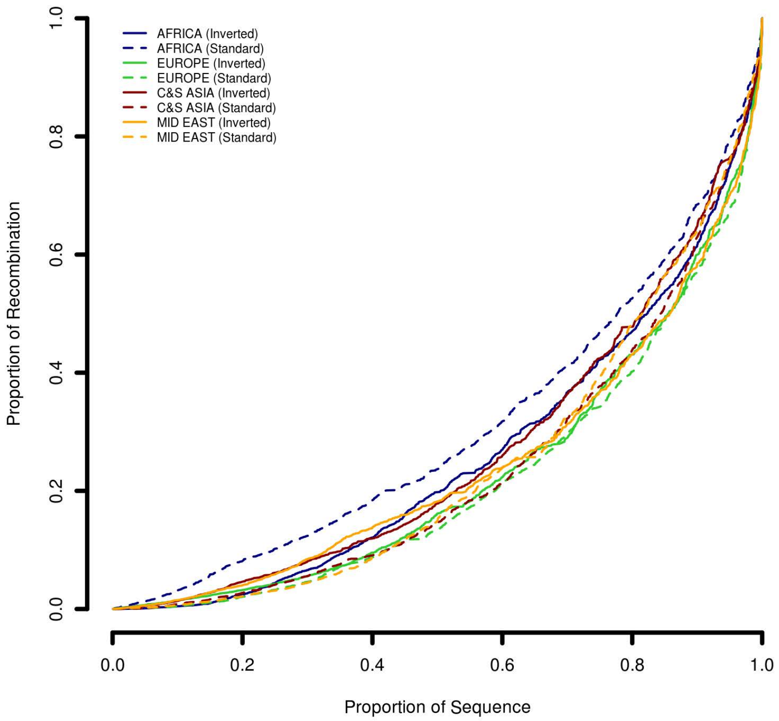


Figure 2: **Cumulative recombination patterns at the 8p23 region** Uneven distribution of recombination within the 8p23 region: The cumulative plot illustrates the proportion of recombination occurring in a given portion of the sequence. Recombination estimates have been sorted in increasing order of intensity.[1]

strongest peaks are shared across all analyzed groups), significant differences in local recombination estimates are also observable, suggesting that we have sufficient power to detect fine-scale variation inside the region. **Table II** shows the mean recombination rate across all SNPs for each group. Significant differences in recombination rates between the groups were confirmed by a repeated measures ANOVA test (p-value $<$ 0.00001). Interestingly, it appears that a significant part and perhaps most of the variation in the recombination landscape is associated with the chromosomal rearrangement. Indeed, a much stronger concordance can be observed between the profiles of individuals sharing the same chromosomal configuration (*i.e.* orientation) than between individuals sharing the same continental origin but having different orientations (**Figure 3**). For instance, we can identify a peak around 9.5 Mbp that is shared between all ”standard individuals” but absent or much weaker in the inverted chromosomes. Another similar example can be found around 11.0 Mbp, where a relatively strong peak shared between all non-African inverted chromosomes is substantially weaker in the standard chromosomes.

[1] **Abbreviations:** C&S ASIA, Central and South Asia; MID EAST, Middle East.

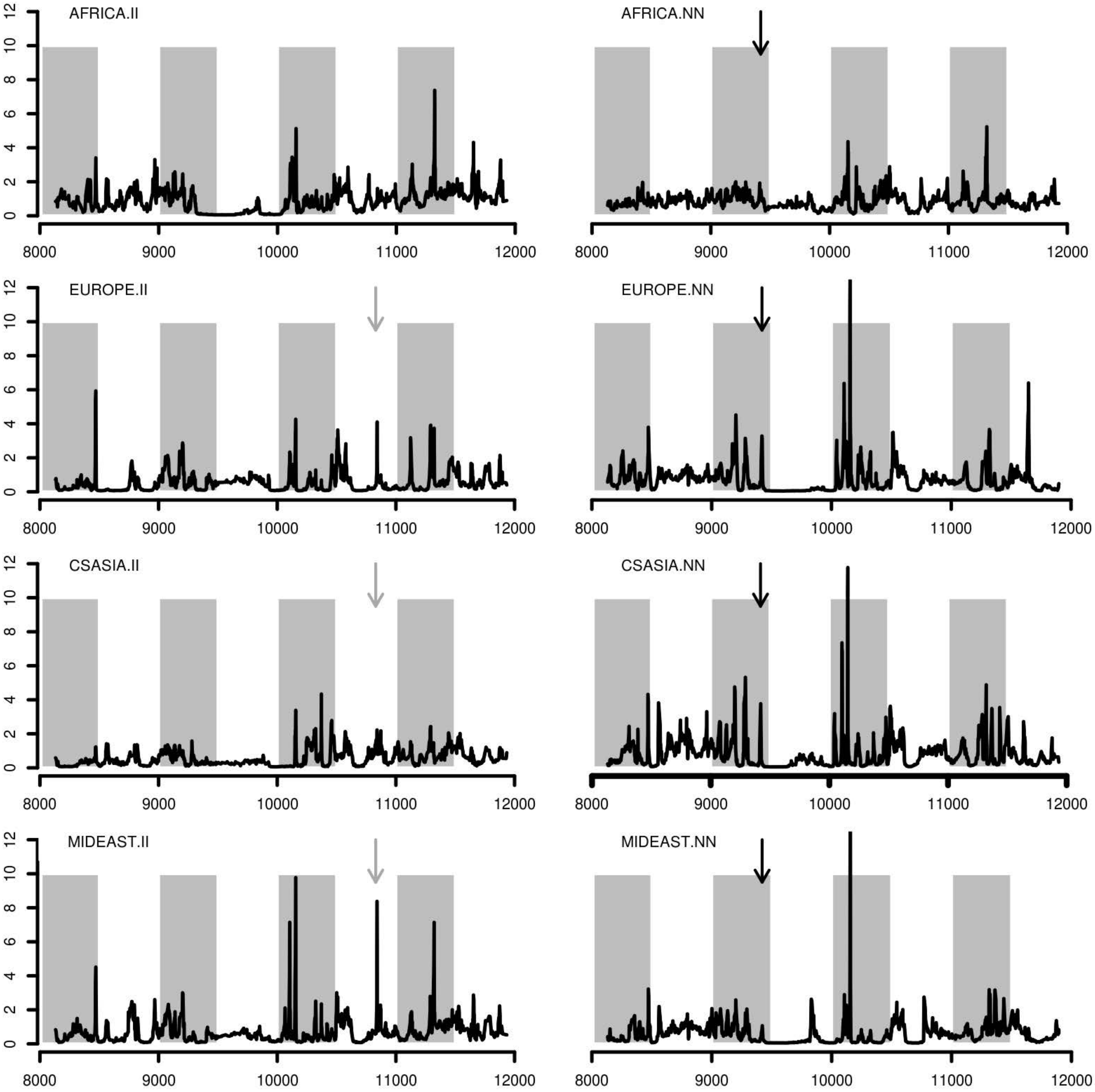
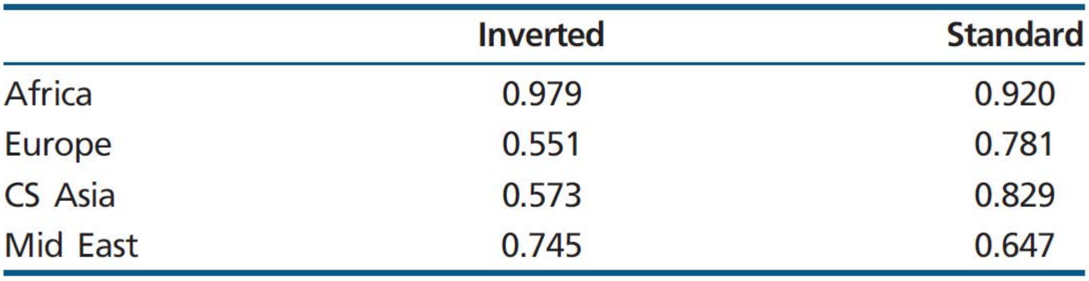


Figure 3: **Distribution of Recombination rates** Recombination estimates for each (geographical and orientation specific) group obtained from LdHat. The left and right panel show the results for the Inverted (II) and Non-inverted (NN) subtypes, respectively. Grey bins of 500Kb are displayed to ease comparisons between the different groups along the region. Drawn arrows denote recombination events that appear to be unique (or much more frequent) in one of the two structural haplotypes

”II”, Inverted orientation; **”NN”**, Standard orientation



	Inverted	Standard
Africa	0.979	0.920
Europe	0.551	0.781
CS Asia	0.573	0.829
Mid East	0.745	0.647

Table II: **Mean recombination rate across the 8p23 for the different groups.**

Given that LD-based recombination estimates are influenced by the allele frequencies (hereafter, AFs) of the used markers, and that the ability to reliably resolve recombination events may become progressively weaker for SNPs showing low minor AFs (MAF) (Auton & McVean (2007), Laayouni *et al.* (2011)), we next placed the estimated recombination rates for each SNP in five bins ordered according to increasing MAF. Note that each group was treated independently, since the AFs varied across populations and inversion status and, therefore, the same SNP will not necessarily fall in the same MAF bin for every group. We then performed a repeated measures analysis of variance with the recombination estimates as the dependent variable and our results showed that, indeed, significantly lower recombination rates were found for SNPs with lower MAF (p-value < 0.005). However, the effect disappears once only SNPs with MAF > 0.1 are considered. A new repeated measures ANOVA excluding the local recombination estimates for SNPs with MAF < 0.1 was applied and the differences in recombination rates between the groups remained highly significant (p-value < 0.00001).

Influence of the inversion rearrangement on recombination patterns

Considering that recent genome-wide studies (Keinan & Reich (2010), Laayouni *et al.* (2011)) have argued that the amount of recombination variation might be positively correlated with the genetic distance found between populations, we next asked whether there was a relationship between the recombination dissimilarity and the genetic distance (as measured by F_{st}) between the structural pairs for each geographical group (**Figure 4**). A statistically significantly positive association was found (r^2= 0.27, p-value < 0.0015) indicating that the genetic divergence between the haplotypes is correlated with the observed dissimilarity in the recombination patterns across the region. Although these results were obtained from recombination rates estimated at SNPs showing MAF $>$ 0.1, a similar, but less robust, significant positive association is also detected when the global set is used (**Supp. Figure 1**; r^2 = 0.11, p-value $<$ 0.05).

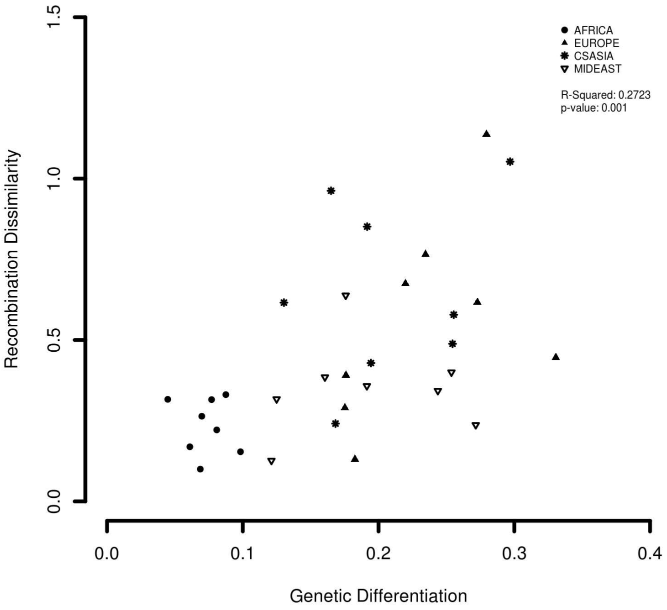


Figure 4: **Relationship between dissimilarity in recombination patterns and genetic differentiation between inversion types**
The figure illustrates the relationship between the dissimilarity observed in the patterns of recombination and the F_{st} values between the two major haplotypes for SNPs with MAF $>$ 0.1 (p-value $<$ 0.0015). The different symbols represent population specific comparisons. Also, each plotted value represents the relationship observed in a genomic window of 500 Kb. In total, 8 non-overlapping windows per population are shown. Comparisons between each chromosomal form were independently performed for each population.

The same method was then applied to test whether population differences within each major haplotype could also account for some of the heterogeneity in the estimated patterns of recombination. Both sets (*i.e.* Inverted and Standard) were analyzed independently and pairwise comparisons were performed between population pairs. The results are shown in **Figure 5**. Here, a much less clear relationship was found suggesting that perhaps the limited degree of divergence, for this genomic region, between the populations under study may be insufficient to produce clear departures between the estimated recombination patterns. Indeed, only for Standard chromosomes is the recombination dissimilarity positively associated with the genetic differentiation between populations (r^2=0.3246, p-value $<$ 0.0001) and this effect is mainly driven by the differences between African and non-African chromosomes (**Figure 6)**.

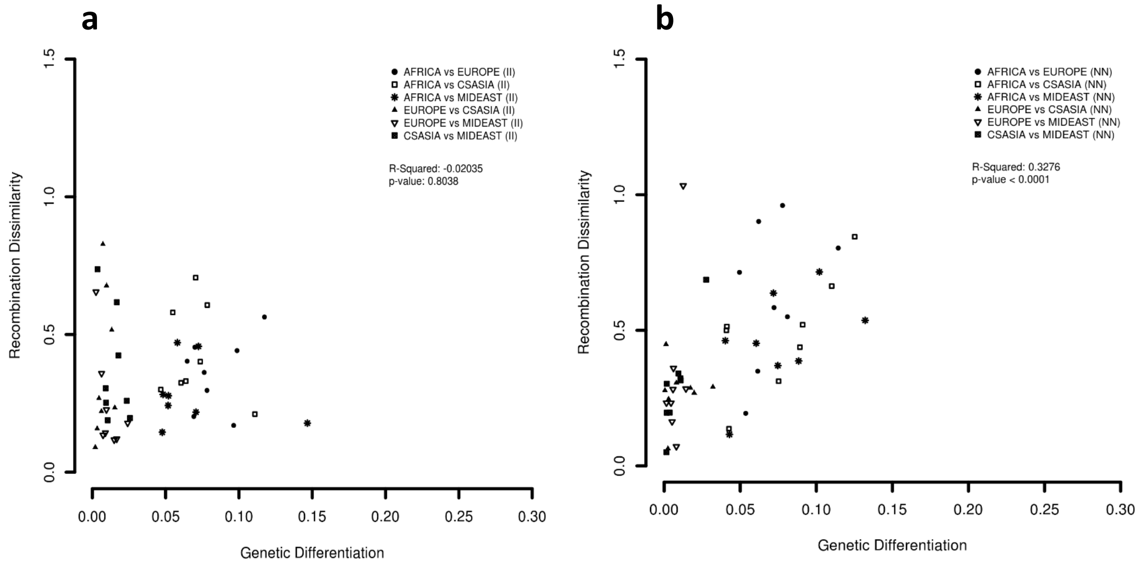


Figure 5: **Relationship between dissimilarity in recombination patterns and genetic differentiation between geographical regions** - Dissimilarity in recombination rate and F_{st} values based on 6 pairwise comparisons between all geographical regions within the **(a)** Inverted (II) haplotype and **(b)** Standard (NN) haplotype.

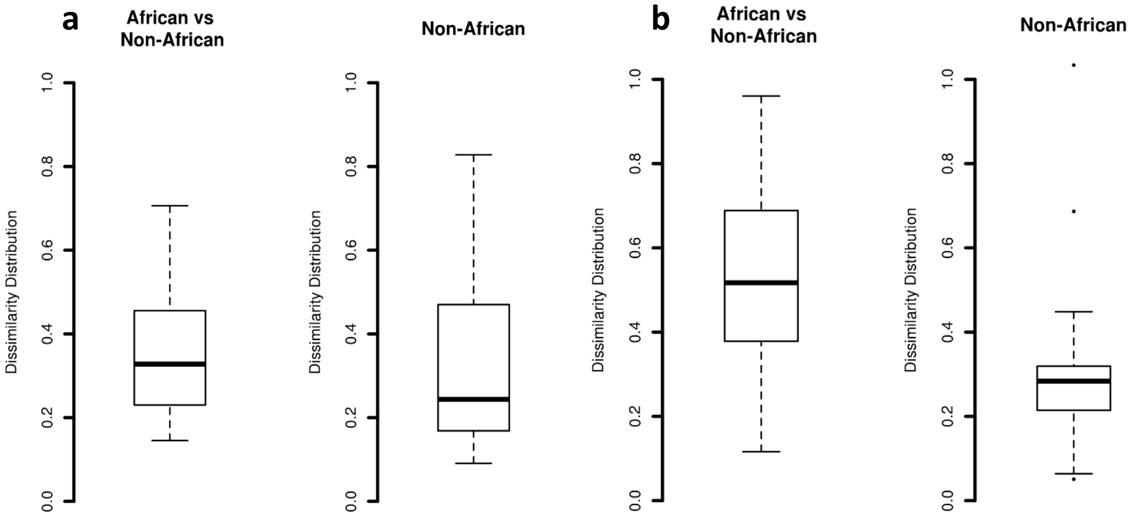


Figure 6: **Asymmetry in recombination dissimilarity scores between African and Non-African groups** - Boxplots displaying the asymmetry in the distribution of recombination dissimilarities between African *versus* Non-African groups for the **(a)** Inverted and **(b)** Standard haplotypes. For each figure, the "African *versus* Non-African" boxplot represents the distribution of dissimilarity scores observed when comparing the African set *versus* all other groups, while the "Non-African" boxplot represents the distribution of dissimilarity scores observed between all Non-African sets.

Robustness of the association maintaining sample size across the different groups

So far it has been argued that LD-based estimates of population-scaled recombination rate are generally robust to sample size change (Serre *et al.* (2005), and references therein) However, given the relatively large disparity in group size in our analysis (minimum of 16 and maximum of 71 samples), we investigated whether the results obtained in the previous section persisted in a scenario where all geographical- and orientation-specific groups had equal sample size. Even with a sample size of 16 individuals, the inferred recombination profiles of each group were very similar to the ones estimated for the original set, despite slight differences in the intensity and magnitude of the strongest recombination peaks (**Figure 7**). In addition, the local recombination heterogeneity was still significantly associated with genetic divergence (r^2= 0.15, p-value $<$ 0.015) (**Figure 8**).

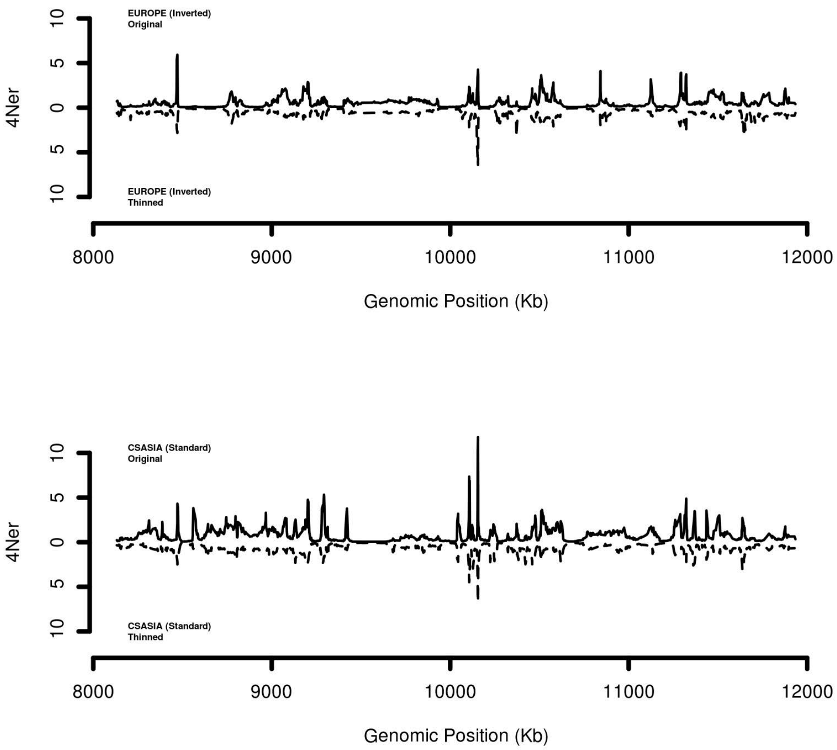


Figure 7: **Validation of the relationship between recombination dissimilarity and genetic distance using a smaller data-set** - Mirror plots of recombination rate obtained for different sample size. Top plot illustrate the recombination estimates for the "European (Inverted)" set (solid line original data; dashed line thinned data (n=16)). Bottom plot illustrate the recombination estimates for the "Central South Asian (Standard)" set (solid line original data; dashed line thinned data (n=16))

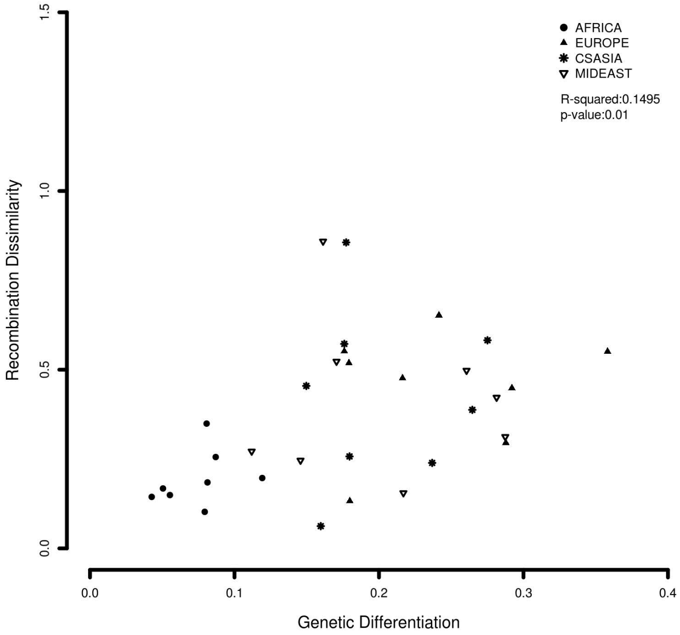


Figure 8: **Relationship between the dissimilarities observed in the patterns of recombination and genetic differentiation (F_{st} values) between the two major haplotypes for data sets of equal size** - Aside from the ”African (Standard)” group, we randomly picked 16 individuals from each group and *re-measured* the recombination dissimilarity and F_{st} values between the inverted and the standard haplotype (p-value < 0.015).

Gene-flow within the 8p23 region

Although theory predicts that recombination should be prohibited between inverted regions (Hoffmann & Rieseberg (2008), Kirkpatrick (2010), Faria & Navarro (2010), Alves *et al.* (2012)), it has been proposed that limited gene flow may have occurred between the two major haplotypes at 8p23. Using inferential methods to ancestral sequence reconstruction, Salm *et al.* (2012) found individual genomes bearing interspersed runs of distinct ancestry (*i.e.* ”Inverted”-ancestry and ”Standard”-ancestry), and concluded that double-recombination events have, to some extent, homogenized the genetic diversity of the region. We, therefore, examined whether similar signals could be identified in our data. Interestingly, a 350-kb segment encompassing the center of the inversion showed significantly lower levels of diversity ($\Pi = 5.2\ 10^{-5}$) (**Figure 9**) that overlapped with a region of deflated F_{st} ($F_{st} = 0.11$), when compared to the average diversity over the whole interval ($\Pi = 12.5\ 10^{-5}$; $F_{st} = 0.17$). Moreover, this segment is flanked by regions with signals of higher recombination activity (*i.e.* putative hotspots), that are shared between the two chromosomal forms. In order to explore this effect in greater detail we performed a principal component analysis with SNPs located within the portion showing lower divergence between inverted and standard haplotypes, and for comparison, in two flanking regions (5’ and 3’). When SNPs located in the central segment of the 8p23 inversion were analyzed, no clear segregation of inverted and standard chromosomes was observed. In contrast, a much cleaner structured environment with only a slight overlap was obtained when including SNPs within each of the flanking regions (**Figure 10**).

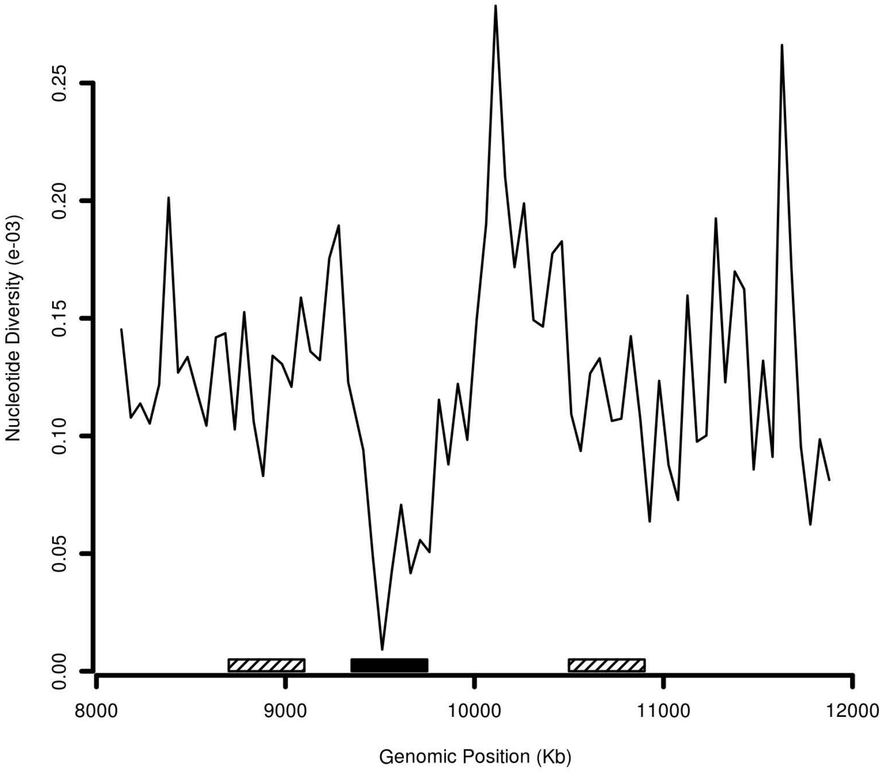


Figure 9: **Evidence of gene flow within the 8p23 regions** - Nucleotide diversity across the 8p23 region. The central filled rectangle highlights the region of reduced diversity; dashed rectangles represent the flanking regions (5 and 3) randomly picked for comparison (see text)

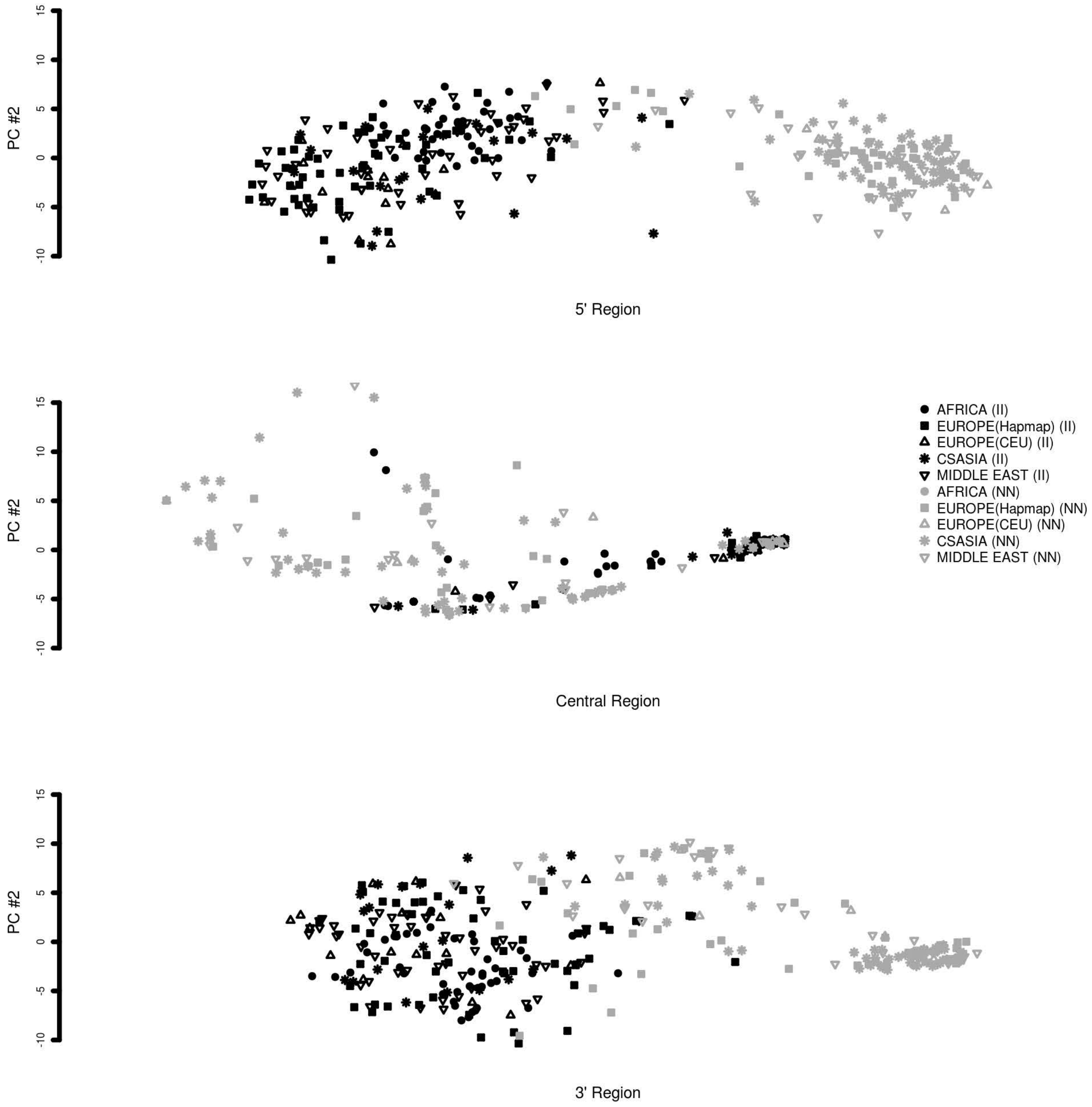


Figure 10: **Evidence of gene flow within the 8p23 regions** - PCA on three non-overlapping regions within the 8p23 (see left plot). The top and bottom plot represent the distribution of individual genotypes for the 5 and 3 regions used for comparison. The center plot shows the distribution of individual genotypes in the region of reduced diversity. Each dot represents one individual, with distinct symbols representing the geographical and orientation specific groups.

Discussion

Inversions have long been regarded as privileged systems to study major evolutionary processes, potentially playing a significant role in species divergence. By preventing gene flow between two different structural types, these rearrangements are thought to allow the accumulation of mutations, representing an initial step towards chromosomal differentiation that may ultimately lead to speciation (see Alves *et al.* (2012) for a detailed review). Here, we took advantage of the co-existence of two groups of structurally distinct chromosomes and assessed how inverted rearrangements influence the evolutionary trajectory of the affected genomic region.

In agreement with what has been described in genome-wide surveys comparing the recombination profiles of different human populations (McVean *et al.* (2004), Clark *et al.* (2010), Keinan & Reich (2010), Laayouni *et al.* (2011), Fledel-Alon *et al.* (2011)), we found evidence supporting a strong correlation between recombination dissimilarities and genetic divergence. Our results indicate that the presence of the rearrangement largely contributed to the accumulation of distinct mutation and recombination events between inversion types, which resulted in extended local recombination heterogeneity within the 8p23 segment.

The genetic differences found between the two major haplotypes (*i.e.* Inverted and Standard) surpassed the differentiation found at the population level (*i.e.* geographical stratification), suggesting that both orientations have been around throughout most human evolutionary history (note the range of F_{st} values in **Figure 3** and **Figure 4**). Indeed, a recent study found that the inversion may have occurred as a single event in the human lineage somewhere around 200-600 kya (Salm *et al.* (2012)), (*i.e.* before modern human emergence) with the inhibition of recombination leading to the formation of two highly divergent haplotype families segregating within populations (Antonacci *et al.* (2009), Salm *et al.* (2012)). The clear differentiation between the two configurations in our PCA further supports the hypothesis of a single very ancient inversion event.

It is, therefore, not unexpected that the rearrangement exerts a stronger effect on the variation of the recombination patterns than population structure. Early migrations of modern humans (*i.e.* Out of Africa) are believed to have started approximately 100kya (Jobling *et al.* (2004)) with complex spatial demographic phenomena (*e.g.* expansions, contractions, admixture events) being particularly responsible for much of the variation identified between present-day human populations. This variation has induced fine-scale differences in recombination patterns between populations, with multiple lines of evidence now suggesting that recombination is a rapidly evolving process partially controlled by the surrounding DNA sequence (Baudat *et al.* (2010), Parvanov *et al.* (2010)). Recent studies using genome wide data have focused on the recombination heterogeneity accumulated at shorter timescales (*i.e.* separation of human populations) (Keinan & Reich (2010), Laayouni *et al.* (2011), Fledel-Alon *et al.* (2011)). Given that the inversion event pre-dates human expansions (Salm *et al.* (2012)), our results are not only consistent with these previous findings but they also extend the analysis into a greater time depth and therefore into a genomic region of increased evolutionary significance.

Despite the overall genetic distinctiveness of the two major haplotypes, we also identified a short region of weaker differentiation at the center of the inversion. While this could be caused by other factors (*e.g.* stochasticity in the mutation process), it supports previous claims of moderate gene flow between inversion-types throughout the evolution of this genomic region (Salm *et al.* (2012)). Indeed, genetic exchange between inverted arrangements may be possible via double cross-over events in inversion loops, with the probability of recombination increasing with physical distance from the inversion breakpoints (Navarro *et al.* (1997), Faria & Navarro (2010), Stevison *et al.* (2011)). Given the size of the *8p23-inv*, it is surely plausible that double cross-over events may have occurred within inversion heterozygotes.

As (i) our primary goal was to evaluate the recombination heterogeneity within the 8p23 segment, and (ii) the SNP density was below optimal for an accurate high resolution inference ($<$ 1/kb), we intentionally avoided to characterize the precise location of recombination hotspots. Hotspots are defined as regions showing enriched recombination rate by several orders of magnitude and have been repeatedly associated with a 13 bp sequence motif that is specifically recognized by PRDM9, a rapidly evolving protein believed to be involved in speciation in mammals (Baudat *et al.* (2010), Parvanov *et al.* (2010)). While the connection between PRDM9 and recombination hotspots is not perfect (Myers *et al.* (2008), Berg *et al.* (2010)), approximately 40% of hotspots in the human genome contain this motif which remains, so far, one of the very few known determinants of meiotic recombination. Interestingly, in a recent comparative study of fixed inverted differences between the genomes of humans and chimpanzees, Farré *et al.* (2013) have proposed that the lower recombination activity observed within inverted segments was linked to a lower density of PRDM9 binding motifs found within these regions, when compared to collinear regions on the same chromosome. In our analysis, and likely due to the short evolutionary time scale that this rearrangement represents, we have found no depletion of PRDM9 motifs within the inverted segment (54.35 motifs/Mb) when compared to collinear regions (44.42 motifs/Mb) in the entire chromosome 8. Nevertheless, it will be interesting to test with deep sequenced data whether the differences in recombination profiles between inverted and standard chromosomes may be partly explained by sequence variants within PRDM9 motifs.

In conclusion, while confirming that recombination is likely suppressed in inverted regions (*i.e.* recombination is almost entirely restricted to chromosomes oriented in the same direction) in global terms, our work showed that fine-scale recombination patterns are evolving differently between chromosomal forms, highlighting the role of inversions as evolutionary significant elements acting at intraspecific level. Also, we provided evidence that this effect is robust to differences in the proportion of inverted to standard chromosomes in a population, since the trend was shared by several geographical regions where the two haplotypes segregate at considerably different frequencies. This work will therefore contribute to a better understanding of recombination heterogeneity at a population level, and reinforce the need to extend these studies to other known inverted regions on the human genome in order to obtain a more comprehensive and meaningful human recombination map. As information on the architectural plasticity of the human genome continues to accumulate (MacDonald *et al.* (2013)), future studies should also consider the implications of these rearrangements in genome-wide selection scans, given that the long-range LD patterns usually manifested within chromosomal inversions may generate signals that could be confounded with selection. As demonstrated here, controlling for inversion-type may help circumvent these limitations. Moreover, our work suggests a new research line devoted to the unveiling of the sequences internal to the inversions that allow for double recombination, and thus overcoming the meiotic problems associated with this rearrangement.

References

Alkan, C., *et al.* 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*, **41(10)**, 1061–67.

Alkan, C., *et al.* 2010. Limitations of next-generation genome sequence assembly. *Nature methods*, **8**, 61–65.

Alkan, C., *et al.* 2011. Genome structural variation discovery and genotyping. *Nature* *reviews genetics*, **12(5)**, 363–76.

Alves, J.M., *et al.* 2012. On the Structural Plasticity of the Human Genome: Chromosomal Inversions Revisited. *Current Genomics*, **13(8)**, 623–632.

Antonacci, F., *et al.* 2009. characterization of six human disease-associated inversion polymorphisms. *human molecular genetics*, **18(14)**, 2555–2566.

Auton, A., & McVean, G.A. 2007. Recombination rate estimation in the presence of hotspots. *Genome Research*, **18(14)**, 2555–2566.

Ayala, D., *et al.* 2010. chromosomal inversions, natural selection and adaptation in the malaria vector anopheles funestus. *molecular biology and evolution*, **28(1)**, 745–758.

Bailey, J.A., & Eichler, E.E. 2006. primate segmental duplications: crucibles of evolution, diversity, and disease. *nature reviews*, **7**, 552–564.

Baird, N.A., *et al.* 2008. rapid snp discovery and genetic mapping using sequenced rad markers. *plos one*, **3**, 110–115.

Bansal, V., *et al.* 2007. evidence for large inversion polymorphisms in the human genome from hapmap data. *genome research*, **17(2)**, 219–30.

Bardhan, A., & Sharma, T. 2000. meiosis and speciation: a study in a speciating mus terricolor complex. *journal of genetics*, **79**, 105–111.

Baudat, F., *et al.* 2010. prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. *science*, **237**, 836–840.

Beaumont, M.A. 2010. approximate bayesian computation in evolution and ecology. *annual review of ecology, evolution, and systematics*, **41(1)**, 379–406.

Beaumont, M.A., *et al.* 2002. approximate bayesian computation in population genetics. *journal of genetics*, **162**, 2025–2035.

Berg, I.L., *et al.* 2010. prdm9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *nature genetics*, **42**, 859–863.

Blum, M.B. 2010. approximate bayesian computation: a nonparametric perspective. *journal of the american statistical association*, **105(491)**, 1178–1187.

Boettger, L.M., *et al.* 2012. structural haplotypes and recent evolution of the human 17q21.31 region. *nature genetics*, **44(8)**, 881–885.

Bosch, N., *et al.* 2009. nucleotide, cytogenetic and expression impact of the human chromosome 8p23.1 inversion polymorphism. *plos one*, **4(12)**, e8269.

Botigué, LR., *et al.* 2013. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proceedings of the National Academy of Sciences*, **110(29)**, 11791–11796.

Brown, G.M., *et al.* 1998. Genetic analysis of meiotic recombination in humans by use of sperm typing: reduced recombination within a heterozygous paracentric inversion of chromosome 9q32-q34.3. *American Journal of Human Genetics*, **62**, 1484–1492.

Brown, J.F., & O'Neill, R.J. 2010. chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *annual review of genomics and human genetics*, 291–316.

Browning, SR., & Browning, BL. 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics*, **81**, 1084–1097.

Bugge, M., *et al.* 2000. disease associated balanced chromosome rearrangements: a resource for large scale genotype-phenotype delineation in man. *journal of medical genetics*, **37**, 858–865.

Caceres, M., *et al.* 2007. a recurrent inversion on the eutherian x chromosome. *proceedings of the national academy of sciences*, **104(47)**, 18571–18576.

Campbell, CD., *et al.* 2011. Population-genetic properties of differentiated human copy-number polymorphisms. *American Journal of Human Genetics*, **88(3)**, 317–332.

Carvajal-Rodríguez, A. 2010. simulation of genes and genomes forward in time. *current genomics*, **11**, 58–61.

Chadeau-Hyam, M., *et al.* 2008. fregene: simulation of realistic sequence-level data in populations and ascertained samples. *bmc bioinformatics*, **9**, 364.

Cheung, V.G., *et al.* 2007. Polymorphic variation in human meiotic recombination. *American Journal of Human Genetics*, **80(3)**, 526–530.

Chikhi, L. 2009. update to chikhi et al.'s "clinal variation in the nuclear dna of europeans" (1998): genetic data and storytelling-from archaeogenetics to astrologenetics-. *human biology*, **81(5-6)**, 639–643.

Chowdhury, R., *et al.* 2009. genetic analysis of variation in human meiotic recombination. *plos genetics*, **5**, e1000648.

Clark, A.G., *et al.* 2010. Contrasting methods of quantifying fine structure of human recombination. *Annual Reviews in of Genomics and Human Genetics*, **11**, 45–64.

Conrad, D.F., *et al.* 2010. origins and functional impact of copy number variation in the human genome. *nature*, **464**, 704–712.

Conrad, D.F., & Hurles, M.E. 2007. Polymorphic variation in human meiotic recombination. *American Journal of Human Genetics*, **80(3)**, 526–530.

Consortium, International HapMap. 2003. The International HapMap Project. *Nature*, **426(6968)**, 789–796.

Consortium, The 1000 Genomes Project. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Coop, G., & Przeworski, M. 2007. an evolutionary view of human recombination. *nature reviews genetics*, **8(1)**, 23–34.

Craddock, N., *et al.* 2010. genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls. *nature*, **464(7289)**, 713–20.

Currat, M., *et al.* 2006. comment on ”ongoing adaptive evolution of aspm, a brain size determinant in homo sapiens” and ”microcephalin, a gene regulating brain size, continues to evolve adaptively in humans”. *science*, **313**, 172.

Davey, J.W., *et al.* 2011. genome-wide genetic marker discovery and genotyping using next-generation sequencing. *nature review genetics*, **12**, 499.

de la Chapelle, A., *et al.* 1974. pericentric inversions of human chromosomes 9 and 10. *american journal of human genetics*, **26**, 746–766.

Delaneau, O., *et al.* 2013. Haplotype estimation using sequence reads. *American Journal of Human Genetics*, **93(4)**, 787–796.

Deng, Y., & Tsao, B.P. 2010. Genetic susceptibility to systemic lupus erythematosus in the genomic era. *Nature Reviews Rheumatology*, **6**, 683–692.

Depristo, M., *et al.* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Dobzhansky, T. 1951. *genetics and the origin of species.* oxford: columbia university press: new york.

Donnelly, M.P., *et al.* 2010. the distribution and most recent common ancestor of the 17q21 inversion in humans. *american journal of human genetics*, **86(2)**, 161–71.

Fagundes, N.J.R., *et al.* 2007. statistical evaluation of alternative models of human evolution. *proceedings of the national academy of sciences*, **104(17)**, 614–619.

Fanciulli, M., *et al.* 2007. fcgr3b copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *nature genetics*, 721–723.

Faria, R., *et al.* 2011. role of natural selection in chromosomal speciation. *In: encyclopedia of life sciences*. chichester, uk: John Wiley & Sons, Ltd.

Faria, R., & Navarro, A. 2010. Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends in Ecology and Evolution*, **25**, 660–669.

Farré, M., *et al.* 2013. Recombination Rates and Genomic Shuffling in Human and Chimpanzee - A New Twist in the Chromosomal Speciation Theory. *Molecular Biology and Evolution*, **30(4)**, 853–864.

Fearnhead, P., *et al.* 2004. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics*, **167(4)**, 2067–2081.

Feuk, L. 2007. inversion variants in the human genome: role in disease and genome architecture. *genome medicine*, **2(2)**, 11.

Feuk, L., *et al.* 2005. discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee dna sequence assemblies. *plos genetics*, **1(4)**, e56.

Fledel-Alon, A., *et al.* 2011. variation in human recombination rates and its genetic determinants. *plos one*, **6(6)**, e20321.

Flores, M., *et al.* 2007. recurrent dna inversion rearrangements in the human genome. *proceedings of the national academy of sciences*, **104(15)**, 6099–106.

Fragata, I., *et al.* 2009. contrasting patterns of phenotypic variation linked to chromosomal inversions in native and colonizing populations in drosophila subobscura. *journal of evolutionary biology*, **23(1)**, 112–123.

Garvin, M.R., *et al.* 2010. application of single nucleotide polymorphisms to non-model species: a technical review. *molecular ecology resources*, **10(6)**, 915–934.

Goldstein, D.B., & Chikhi, L. 2002. human migrations and population structure : what we know and why it matters. *annual review of genomics and human genetics*, **3**, 129–152.

Goldstein, D.B., & Weale, M.E. 2001. population genomics: linkage disequilibrium holds the key. *current biology*, **11**, 576–579.

Gonzalez, E., *et al.* 2005. the influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility. *science*, **307**, 1434–40.

Goudet, J. 2005. Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184–186.

Gu, W., *et al.* 2008. mechanisms for human genomic rearrangements. *pathogenetics*, **1(4)**.

Guerrero, R.F., *et al.* 2012. coalescent patterns for chromosomal inversions in divergent populations. *philosophical transactions of the royal society b: biological sciences*, **367(1587)**, 430–438.

Hara, Y., *et al.* 2011. abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. *bmc evolutionary biology*, **11(1)**, 308.

Hardy, J., *et al.* 2005. evidence suggesting that homo neanderthalensis contributed the h2 mapt haplotype to homo sapiens. *biochemical society transactions*, **33(4)**, 582–585.

Henn, BM., *et al.* 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genetics*, **8(1)**, e1002397.

Hey, J. 2003. speciation and inversions: chimps and humans. *bioessays*, 825–828.

Hoffmann, A.A., & Rieseberg, L.H. 2008. revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation-. *annual review of ecology, evolution, and systematics*, **39**, 21–42.

Hohenlohe, P.A., *et al.* 2010. population genomics of parallel adaptation in three-spine stickleback using sequenced rad tags. *plos genetics*, **6(2)**, e1000862.

Hollox, E.J., *et al.* 2008. defensins and the dynamic genome: what we can learn from structural variation at human chromosome band 8p23.1. *genome research*, **18**, 1686–1697.

Hubert, R., *et al.* 1994. High resolution localization of recombination hot-spots using sperm typing. *Nature Genetics*, **7**, 420–434.

Hudson, R. 1990. gene genealogies and the coalescent process. *In: oxford surveys in evolutionary biology*. oxford, new york: oxford university press: oxford.

Huynh, L.Y., *et al.* 2011. chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (zonotrichia albicollis). *heredity*, **106**, 537–546.

Iafrate, A.J., *et al.* 2004. Detection of large-scale variation in the human genome. *Nature Genetics*, **36(9)**, 949–51.

Jensen-Seaman, M.I., *et al.* 2004. comparative recombination rates in the rat, mouse, and human genomes. *genome research*, **14**, 528–538.

Jobling, M.A, Hurles, M.E., & Tyler-Smith, C. 2004. *Human Evolutionary Genetics: Origins, Peoples and Disease*. New York: Garland Science: New York.

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Joron, M., *et al.* 2011. chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *nature*, **477**, 203–206.

Keinan, A., & Reich, D. 2010. Human population differentiation is strongly correlated with local recombination rate. *PLoS Genetics*, **6(3)**, e1000886.

Kidd, J.M., *et al.* 2008. mapping and sequencing of structural variation from eight human genomes. *nature*, **453(7191)**, 56–64.

Kirkpatrick, M. 2010. how and why chromosome inversions evolve. *plos biology*, **8**, 9.

Kirkpatrick, M., & Barton, N. 2006. chromosome inversions, local adaptation and speciation. *genetics*, **173(1)**, 419–434.

Klopfstein, S., *et al.* 2005. the fate of mutations surfing on the wave of a range expansion. *molecular biology and evolution*, **23**, 482–490.

Korbel, J.O., *et al.* 2007. paired-end mapping reveals extensive structural variation in the human genome. *science*, **318**, 420–426.

Laayouni, H., *et al.* 2011. Similarity in recombination rate estimates highly correlates with genetic differentiation in humans. *PLoS One*, **6(3)**, e17913.

Lahn, B.T., & Page, D.C. 1999. four evolutionary strata on the human x chromosome. *science*, **286(5441)**, 964–967.

Lee, J.A., *et al.* 2007. a dna replication mechanism for generating non-recurrent rearrangements associated with genomic disorders. *cell*, **131**, 1235–1247.

Levy, S., *et al.* 2007. the diploid genome sequence of an individual human. *plos biology*, **5**, e254.

Li, H., & Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **Epub**.

Li, J., *et al.* 2006. A new method for detecting human recombination hotspots and its applications to the HapMap ENCODE data. *American Journal of Human Genetics*, **79(4)**, 628–639.

Li, N., & Stephens, M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165(4)**, 2213–2233.

Li, S., & Jakobsson, M. 2012. estimating demographic parameters from large scale population genomic data using approximate bayesian computation. *bmc genetics*, **13**, 22.

Lichten, M., & Goldman, A.S. 1995. Meiotic recombination hotspots. *Annual Review of Genetics*, **29**, 423–444.

Lopes, AM., *et al.* 2013. Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene DMRT1. *PLoS Genetics*, **9(3)**, e1003349.

Lowry, D.B., *et al.* 2010. a widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *plos biology*, **8(9)**, e1000500.

Lu, J., *et al.* 2003. comment on ”chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes”. *science*, **302**, 988.

MacDonald, J.R., *et al.* 2013. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*, **42(1)**, 986–992.

Matsunaga, S. 2006. sex chromosome-linked genes in plants. *genes & genetic systems*, **81**, 219–226.

McVean, G.A., *et al.* 2004. The fine-scale structure of recombination rate variation in the human genome. *Science*, **304(5670)**, 581–584.

Medvedev, p., *et al.* 2009. computational methods for discovering structural variation with next-generation sequencing. *nature methods*, **6**, 13–20.

Murphy, W.J., *et al.* 2005. dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *science*, **309**, 613–617.

Myers, S., *et al.* 2005. a fine-scale map of recombination rates and hotspots across the human genome. *science*, **310**, 321–324.

Myers, S., *et al.* 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, **40**, 1124–1129.

Navarro, A., & Barton, N. 2003a. accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *evolution*, **57(3)**, 447–59.

Navarro, A., & Barton, N. 2003b. chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes. *science*, **300(5617)**, 321–324.

Navarro, A., *et al.* 1997. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics*, **146**, 695–709.

Navarro, A., *et al.* 2000. effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in drosophila. *genetics*, **155(2)**, 685–698.

Noor, M.A.F., *et al.* 2001. chromosomal inversions and the reproductive isolation of species. *proceedings of the national academy of sciences*, **98**, 12084–12088.

Nosil, P., & Feder, J.L. 2011. genomic divergence during speciation: causes and consequences. *philosophical transactions of the royal society b: biological sciences*, **367(1587)**, 332–342.

Nothnagel, M., *et al.* 2011. technology-specific error signatures in the 1000 genomes project data. *human genetics*, 505–516.

O'Neill, R.J., *et al.* 2004. centromere dynamics and chromosome evolution in marsupials. *the journal of heredity*, **95(5)**, 375–381.

O'Reilly, P.F., *et al.* 2010. invertfregene: software for simulating inversions in population genetic data. *bioinformatics*, **26(6)**, 838–840.

Orr, A. 1996. Dobzhansky, bateson, and the genetics of speciation. *genetics*, **144**, 1331–1335.

Parvanov, E.D., *et al.* 2010. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science*, **327(5967)**, 835.

Petes, T.D. 2001. meiotic recombination hot spots and cold spots. *nature review genetics*, **2**, 360–369.

Pickrell, J.K., *et al.* 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, **19**, 826–837.

Purandare, S.M., & Patel, P.I. 1997. Recombination hot spots and human disease. *Genome Research*, **7(8)**, 773–786.

Purcell, S., *et al.* 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, **81(3)**, 559–575.

Rao, P.N., *et al.* 2010. recurrent inversion events at 17q21.31 microdeletion locus are linked to the mapt h2 haplotype. *cytogenetic and genome research*, **90095**, 275–279.

Richards, M., *et al.* 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. *American Journal of Human Genetics*, **67(5)**, 12511276.

Rieseberg, L.H. 2001. chromosomal rearrangements and speciation. *trends in ecology and evolution*, **16(7)**, 351–358.

Ross, M.T., *et al.* 2005. the dna sequence of the human x chromosome. *nature*, **434**, 325–337.

Salm, M.P.A., *et al.* 2012. he origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *genome research*, **22(6)**, 1144–1153.

Serre, D., *et al.* 2005. Large-scale recombination rate patterns are conserved among human populations. *Genome Research*, **15(11)**, 1547–1552.

Shaw, C.J., & Lupski, J.R. 2004. implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *human molecular genetics*, **13(1)**, 57–64.

Simpfendorfer, K.R., *et al.* 2012. the autoimmunity-associated blk haplotype exhibits cis-regulatory effects on mrna and protein expression that are prominently observed in b cells early development. *human molecular genetics*, **21(17)**, 3918–3925.

Sindi, S.S., *et al.* 2010. identification and frequency estimation of inversion polymorphisms from haplotype data. *journal of computational biology*, **17(3)**, 517–531.

Skaletsky, H., *et al.* 2003. the male-specific region of the human y chromosome is a mosaic of discrete sequence classes. *nature*, **423(6942)**, 825–37.

Sousa, V., *et al.* 2009. approximate bayesian computation without summary statistics: the case of admixture. *journal of genetics*, **181**, 187–197.

Sousa, V., *et al.* 2011. population divergence with or without admixture: selecting models using an abc approach. *heredity*, **108**, 521–530.

Spirito, F. 1998. *endless forms: species and speciation*. oxford, new york: oxford university press: oxford.

Spirito, F., *et al.* 1993. the establishment of underdominant chromosomal rearrangements in multi-deme systems with local extinction and colonization. *theoretical popu-*

lation biology, **44**, 80–94.

Spitz, F., *et al.* 2005. inversion induced disruption of the hoxd cluster leads to the partition of regulatory landscapes. *nature genetics*, **37**, 889–893.

Stefansson, H., *et al.* 2005. a common inversion under selection in europeans. *nature genetics*, **37(2)**, 129–37.

Steinberg, K.M., *et al.* 2012. structural diversity and african origin of the 17q21.31 inversion polymorphism. *nature genetics*, **44(8)**, 872–880.

Stephens, M., *et al.* 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Stevison, L.S., *et al.* 2011. effects of inversions on within- and between-species recombination and divergence. *genome biology and evolution*, **3**, 830–841.

Sturtevant, A.H. 1921. a case of rearrangement of genes in drosophila. *proceedings of the national academy of sciences*, **7**, 235–237.

Sudmant, PH., *et al.* 2010. Diversity of human copy number variation and multicopy genes. *Science*, **330(6004)**, 641–646.

Tamura, K., *et al.* 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution. *Molecular Biology and Evolution*, **28**, 2731–2739.

Turner, D.J., *et al.* 2006. assaying chromosomal inversions by single-molecule haplotyping. *nature methods*, **3**, 439–445.

Tuzun, E., *et al.* 2005. fine-scale structural variation of the human genome. *nature genetics*, **37**, 727–732.

Warburton, P.E., *et al.* 2004. inverted repeat structure of the human genome: the x-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *genome research*, **14**, 1861–1869.

Wegmann, D., *et al.* 2009. efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *journal of genetics*, **182**, 1207–1218.

Wilson, M.G., *et al.* 1970. inherited pericentric inversion of chromosome nb. 4. *american journal of human genetics*, **22**, 679–690.

Yu, L., *et al.* 2010. e(nos)/cg4699 required for nanos function in the female germ line of drosophila. *genesis*, **48**, 161–170.

Zang, J., *et al.* 2004. testing the chromosomal speciation hypothesis for humans and chimpanzees. *genome research*, **14**, 845–851.

Zilhão, J. 2001. Radiocarbon evidence for maritime pioneer colonization at the origins of farming in west Mediterranean Europe. *Proceedings of the National Academy of*

Sciences, **98(24)**, 1418014185.

Zody, M.C., *et al.* 2008. evolutionary toggling of the mapt 17q21.31 inversion region. *nature genetics*, **40(9)**, 1076–83.

CHAPTER III

The evolutionary history of a common polymorphic inversion

João M Alves[1,2,3], Alexandra M Lopes[2], Ana Lima[2], Isa Pais[3], Nadir Amir[4], Ricardo Celestino[2], David Comas[5], Peter Heutink[6], Lounès Chikhi[3,7] & António Amorim[2,8]

[1]Doctoral Program in Areas of Basic and Applied Biology (*GABBA*), University of Porto, Portugal; [2]IPATIMUP - Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal; [3]Instituto Gulbenkian de Ciência (IGC), Oeiras, Portugal; [5]Faculty of Nature and Llife Sciences, Abderrahmane Mira University of Bejaia, Faculty of Nature and Life Sciences, Targa Ouzemmour, 06000 Bejaia, Algeria; [5]Institut de Biologia Evolutiva (Consejo Superior de Investigaciones Científicas-Universitat Pompeu Fabra), Departament de Ciències Experimentals i de la Salut, Human Genome Diversity, Universitat Pompeu Fabra, 08003 Barcelona, Spain; [6]Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands ;[7] CNRS (Centre National de la Recherche Scientifique), Université Paul Sabatier, Ecole Nationale de Formation Agronomique, Unité Mixte de Recherche 5174 EDB (Laboratoire Évolution & Diversit Biologique), F-31062 Toulouse, France; [8]Faculdade de Ciências da Universidade do Porto, Porto, Portugal.

Manuscript in preparation
Running Title:
"Population and evolutionary processes shaping the 17q21 inversion in humans"

Abstract

In contrast to early predictions from classical cytogenetic studies, a particular subtype of balanced rearrangement - chromosomal inversion - has been recently recognized as a common source of genetic variation. A striking example is the widespread polymorphic inversion that lies on chromosome 17q21.31. Due to the lack of recombination, two major haplotype families (H1 and H2) have been derived from its orientation

(Standard and Inverted, respectively). Also, both haplotype hamiles have independently experienced partial duplications of the *KANSL1* gene, and several distinct structural forms that differ in copy-number have already been identified. However, the evolutionary processes driving the spread of this rearrangement in the human lineage remain unclear, since conflicting scenarios have been raised to explain its uneven distribution in modern day populations (*i.e.* selective advantage vs. random demographic effects).

Here, we updated the frequency and distribution of the inversion-associated haplotype lineage (*i.e.* H2) in several meta-population groups that have been overlooked in previous studies. Interestingly, we have found a patchy distribution of the H2 haplotype in African populations, with North Africans displaying a much higher frequency of both H2 subtypes (*i.e.* H2D, with the partial duplication of the *KANSL1* gene, and H2', lacking the duplication) when compared to Sub-Saharan groups. Also, our genetic differentiation estimates revealed that the H2 subtypes found in North Africa are genetically closer to ”Non-African” haplotypes, which is consistent with recent genome-wide studies. Furthermore, our results suggest that the partial duplication of the *KANSL1* gene may be recurrent within the H2 lineage, further highlighting the structural complexity of the region.

Introduction

A widespread polymorphic inversion that lies on chromosome 17q21 became the focus of intense research in the last decade due to its high degree of complexity, both in terms of genetic diversity and structural plasticity. This rearrangement spans nearly 1Mb, involves several genes implicated in complex neurodegenerative disorders (Stefansson *et al.* (2005), Zody *et al.* (2008), Antonacci *et al.* (2009)), and two clearly distinct major haplotype families - H1 and H2 - have been derived from its orientation (Standard and Inverted, respectively) (Stefansson *et al.* (2005), Zody *et al.* (2008), Antonacci *et al.* (2009), Donnelly *et al.* (2010)).

With the recent advances in deep re-sequencing data analysis, new studies focused on the structural variability of the human genome (Sudmant *et al.* (2010), Campbell *et al.* (2011), Lopes *et al.* (2013)) have revealed that the 17q21 segment is also considerably polymorphic at the copy-number (hereon, CN) level, pushing further the already complex architecture of the region. These copy-number changes mainly consist of two distinct duplications overlapping the *KANSL1* gene, and were initially thought to be restricted to European populations (Sudmant *et al.* (2010)). More recently, (Steinberg *et al.*, 2012) updated the frequency and distribution of these CNPs surveying a much larger population panel and, using a complementary strategy between cytogenetics, comparative genomics and population genetics, demonstrated that each of these CN polymorphisms were distinctively associated with the previously reported structural haplotypes. Indeed, while a long 205-kb CN duplication was found to be polymorphic in H1 chromosomes, a distinct, and shorter, 155-kb polymorphic duplication was only detected in the H2 lineage.

From an evolutionary perspective, these copy-number variants are therefore believed to have arisen independently in the two haplotype families. Although it has been argued that the origin of the duplication events pre-dates early human expansions (*e.g.* Out-of-Africa), the ”derived” haplotypes (*i.e.* H1D and H2D), enriched in duplicated copies, are found in much higher frequencies in present-day Non-African populations [Sudmant 2010, Steinberg *et al.* (2012), Boettger *et al.* (2012)]. As a consequence, new ideas regarding the evolutionary history of the region emerged, placing these CNPs as a potential source of selective advantage. This was particularly evident in European populations, considering that the inversion-associated haplotype carrying a duplicated copy of the short CNP - H2D - is present at high frequencies (up to 25%) but displays very low levels of genetic diversity (Steinberg *et al.* (2012); Boettger *et al.* (2012)). Nevertheless, the processes driving the spread of the duplication-specific haplotypes in the human lineage remain unclear, as the observed patterns could also be the result of the complex demographic history of populations, without the need of invoking selection (Zody *et al.* (2008), Donnelly *et al.* (2010), Steinberg *et al.* (2012)).

In the current study, we propose to further explore the evolutionary history of 17q21 region by focusing on the inversion-associated haplotype family (*i.e.* H2). By combining complementary genotype data from multiple sources, we refined the distribution of the H2 sub-haplotypes - H2’ and H2D - in several meta-population groups, including some that have been overlooked in previous studies (*e.g.* North Africa). Considering the high frequencies of the H2D haplotype in Southern European populations, and the recent claims that extensive migrations from North Africa contributed to a high genetic diversity in South Europe (Henn *et al.* (2012), Botigué *et al.* (2013)), it seems plausible that such patterns are the result of considerable uni- (or bi-) directional gene flow between the two continents. In order to increase the resolution of our analysis, we reassessed the diversity patterns within the H2 lineage using a unique panel of H2-specific single nucleotide polymorphisms (SNPs), derived from publicly available re-sequencing data and explicitly selected for this purpose. Altogether, the current work provides a finer picture of the distribution and diversity patterns of the 17q21.31 H2 haplotypes in human geographical groups that are expected to hold the most informative genetic signatures to understand its evolutionary trajectory.

Materials & Methods

Available Datasets and Samples

SNP Genotype Datasets: Genotype data for the majority of the North African samples used in this study, as well as some Southern European samples (*e.g.* Spanish Basque), were retrieved from a recently published database(Henn *et al.* (2012), Botigué *et al.* (2013)). Sardinian SNP data, genotyped using Affymetrix 500K SNP array, were available at PH laboratory. Genotypes of 207 Portuguese individuals, genotyped using Affymetrix 6.0 SNP Array, were retrieved from (Lopes *et al.*, 2013). Publicly available genotypes from a diverse population panel were obtained from the Stanford Human Genome Diversity Project (HGDP).
Sequence Data: Low coverage re-sequencing data were obtained from the 1000 Genomes Project (Phase 1, release v3).
Cell lines: Lymphoblast cell lines from 9 Hapmap individuals were obtained from Coriell Cell repository (http://www.ccr.coriell.org/).
Genomic DNA: DNA was available for a total of 370 individuals from multiple populations.
A list of all individuals used in this study, along with the corresponding inversion- and duplication-status, is shown in Supp. Table 1.

Inference of Inversion- and Duplication-Status

Given that it has been suggested that some H2-chromosomes could carry the non-inverted configuration (Rao *et al.* (2010)), hence suggesting the possibility of recurrence of the inversion in the region, we performed Fluorescent in situ Hybridization (FISH) analysis in a small set of H2 carriers, using the same method described in (Rao *et al.*, 2010) (see above, Supp. Info). Similarly to (Steinberg *et al.*, 2012), we did not find any incongruence between the FISH-determined orientation and the haplotype sequence. Consequently, a set of well-known inversion-markers (Antonacci *et al.* (2009), Donnelly *et al.* (2010)) were used to infer the inversion status for all individual samples (Supp. Info, Supp. Figure 1 and Supp. Table 2). The duplication status of the H2-chromosomes was assessed mainly using two previously reported duplication-markers (Steinberg *et al.* (2012)) (Supp. Info). Nevertheless, whenever re-sequenced data was available (*e.g.* 1000Genomes data), the presence/absence of the duplication was further confirmed using sequence read-depth information, as in Sudmant *et al.* (2010) (Supp. Info, Supp. figure 2 and Supp. Table 3).

Population Sets and Haplotype Frequency estimates

Once the orientation- and duplication-status were confirmed, each individual was then grouped according to geographical origin, and seven major continental groups were thus defined, Sub-Saharan Africa, North Africa, Middle East, South Asia, Asia, South Europe and Southwest Europe. Haplotype frequencies were then estimated using all available information (Table 1 and Supp. Table 1).

H2-Polymorphic SNP Panel

A custom SNP genotyping assay was designed using genomic information derived from previously published deep-sequenced data from (1) an H2 homozygous from Eastern Africa (NA21599) (Steinberg *et al.* (2012)), (2) an H2 homozygous from Central South Asia (NA20890) (Boettger *et al.* (2012)) and (3) low coverage sequences from several European H2 homozygous individuals, available from the 1000Genomes Project (Supp. Table 4). Sequence reads were aligned to the human reference GRCh37 using BWA (Bwa-0.6.2) (Li & Durbin (2010)) and SNP calling was subsequently performed using genome analysis toolkit (GATK) (Depristo *et al.* (2011)) unified genotyper. After quality control filtering, a total of 237 previously unidentified SNPs were found to be polymorphic in at least one individual from the list. Due to the highly repetitive architecture of the 17q21 region, the SNP set was further pruned to avoid variants located within known segmental duplications (n=60). The remaining SNPs (n=177) were then genotyped in all H2 carriers, from whom DNA was available (Supp. Table 1), using SEQUENOM platform. Moreover, 10 inversion-marker SNPs were added to the assay to check for genotype consistency The quality of the obtained genotypes was then verified using Plink software (Purcell *et al.* (2007)), and a final list of 80 SNPs were kept for subsequent analysis (genotype rate: 0.98). It should be mentioned that a high number of the SNPs (n=80) were found to be monomorphic in all individuals. As a consequence, these variants were removed for the remaining of the study as they more likely represent false positives or singletons (Supp. Table 5).

Haplotype Phasing, Genetic Differentiation and Phylogenetic analysis

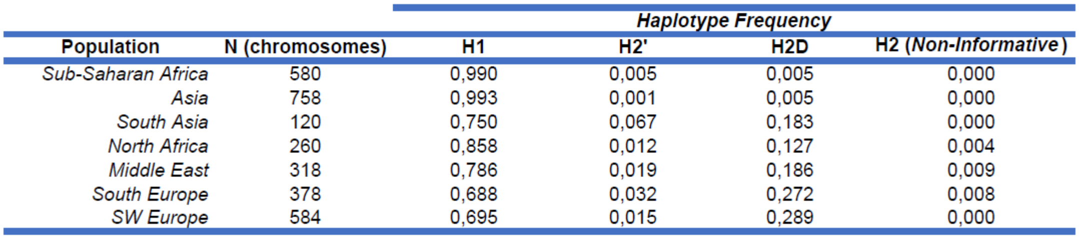
Recent years have witnessed the development of a variety of statistical algorithms to haplotype reconstruction from unrelated genotype data (Stephens *et al.* (2001), Browning & Browning (2007), Delaneau *et al.* (2013)), that generally rely on a preexistent reference list of pre-assembled haplotypes, derived from family-based population sets, to accurately generate phase information. We therefore phased the genotypes from the previous step using SHAPEIT (v2) software (Delaneau *et al.* (2013)) and the 1000Genomes Project data as the reference panel of haplotypes. Following the software recommendations, all genotypes were simultaneously phased for a total of 100 iterations with a burn-in and pruning stage of 25 iterations each. The obtained haplotypes were then stored in VCF format and combined with the ones from the 1000Genomes Project (Phase1.v3). Afterwards, a principal component analysis (PCA) was performed on all haplotypes to check for possible phasing errors (Supp. Figure 3). Chromosomes were finally sorted according to the 3 possible statuses (*i.e.* H1, H2' and H2D) and grouped by continental origin. Nucleotide Diversity estimates were measured for each haplotype using VCFdivstat (v1.0). Genetic differentiation estimates were assessed between all continental pairs using adegenet R package (Jombart (2008)). Finally, a maximum likelihood phylogenetic tree was generated using Mega 5.2 Software (Tamura *et al.* (2011)) for most haplotypes available (see below).

Results

Reassessing the 17q21 Haplotypes Distribution

Using all available SNP genotype data we estimated the frequency of the 17q21 haplotypes in a panel of North African populations, and reassessed the frequency of the inversion-associated haplotypes in several other continental groups by combining our population datasets with publicly available human diversity panels (**Table 1**). Our results indicate that the H2 haplotypes are segregating at a frequency of approximately 15% in North African populations, with the majority of the individuals carrying the H2-

duplicated copy (*i.e.* H2D). Indeed, the H2D was the predominant inversion-associated haplotype in most of the continental groups analyzed, with Southern Europeans populations exhibiting the highest frequency (approximately 30%). The only exception were Sub-Saharan Africans, where the two H2 sub-types segregated at a similar frequencies. Moreover, the highest proportion of the less complex H2 version (*i.e.* H2'), which is considered the ancestral state [Stein/Boet], was found in South Asian individuals with a frequency of 6.7% (Table 1).



		Haplotype Frequency			
Population	**N (chromosomes)**	**H1**	**H2'**	**H2D**	**H2 (*Non-Informative*)**
Sub-Saharan Africa	580	0,990	0,005	0,005	0,000
Asia	758	0,993	0,001	0,005	0,000
South Asia	120	0,750	0,067	0,183	0,000
North Africa	260	0,858	0,012	0,127	0,004
Middle East	318	0,786	0,019	0,186	0,009
South Europe	378	0,688	0,032	0,272	0,008
SW Europe	584	0,695	0,015	0,289	0,000

Table I: **17q21 Haplotypes Frequency in 7 meta-population groups.**[1,2]

17q21 haplotypes Diversity estimates

As noted above, the genetic diversity of the inversion-associated haplotypes was estimated using a unique panel of newly identified H2-specific SNPs (see Methods section). Our results indicate that the duplication-specific haplotype (*i.e.* H2D) is in fact much more diverse than previously predicted (Steinberg *et al.* (2012)), with fairly high levels of nucleotide diversity (**Figure 1**). Interestingly, the H2' displays similar levels of genetic diversity, despite being much less frequent in most continental groups. Furthermore, given that the SNP set was explicitly designed to find polymorphic variants within the H2 lineage, the extremely low levels of diversity found within the H1 haplotype at these positions are not necessarily surprising.

Nucleotide diversity estimates were subsequently calculated for all continental groups within each inversion-associated haplotype (*i.e.* H2' and H2D). Both sub-haplotype

[1] **Abbreviations:** *SW EUROPE* - South West Europe.

[2] **H2(non-informative)**: H2 chromosomes with no allele information for the reported *Duplication markers*

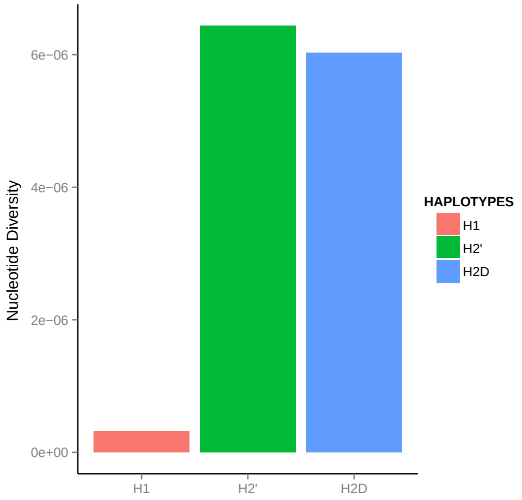


Figure 1: **Nucleotide diversity for each haplotype at 17q21** - Colored barplot represent the average nucleotide diversity for each 17q21 haplotype. Estimates obtained through the analysis of 1602 statistically phased chromomes (H1: 1290; H2D: 270; H2': 42) using the H2-polymorphic SNP panel (see Methods).

sets were analyzed independently and the results are illustrated in **Figure 2**. H2D was found to be more diverse in Southwest and South European populations, compared to all other groups. In contrast, Sub-Saharan African and South Asian populations displayed the least variable H2D chromosomes. The low diversity levels observed in the latter are intriguing, when considering the high frequency of the H2D haplotype in this meta-population group (Table 1).

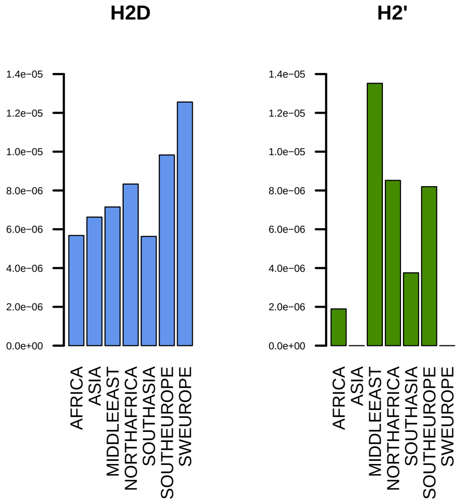


Figure 2: **Nucleotide Diversity of the the 17q21 *inversion-associated* haplotypes *per* Meta-Population Group** Barplots displaying the average nucleotide diversity for each inversion-associated sub-haplotype by meta-population group.

Genetic Differentiation within *inversion-associated* sub-haplotypes

Within the H2’ subtype, the highest nucleotide diversity was found in Middle Eastern populations. Interestingly, Sub-Saharan Africa and South Asia were once more found to have the least diverse chromosomes when compared to all other groups. Nevertheless, it should be noted that (1) this haplotype is virtually absent in Asia and (2) no genomic DNA was available for the H2’ carriers from Southwest Europe (Table 1 and Supp. Table 1). As a consequence, we were unable to estimate the H2’ nucleotide diversity within these continental groups.

Genetic differentiation estimates across meta-populations (assessed through Fst) were next calculated for each H2 sub-haplotype . **Figure 3** shows the genetic distances obtained between all groups for the duplication-specific subtype. Our results indicate that the North African H2D is more closely related to Southern European (Fst=0.04) and Middle Eastern (Fst=0.08) copies than to Sub-Saharan African H2D (Fst=0.22). In addition, the Middle Eastern H2D haplotype appears to be closer to the H2D chromosomes from South Asia, while showing roughly the same genetic distance from the remaining continental groups surveyed. Lastly, the Southern European H2D was found to be more distinct from the Middle Eastern H2D, when compared to the remaining groups. The genetic distances found within the H2’ haplotype are illustrated in **Figure**

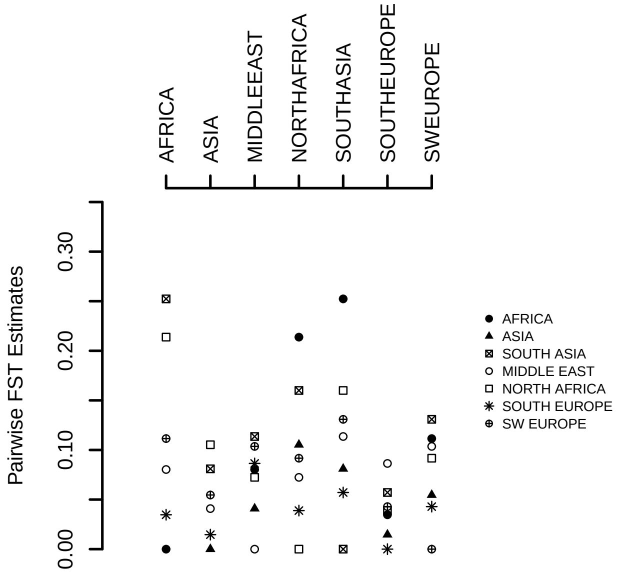


Figure 3: **Genetic Distance estimates between continental groups (within the H2D subtype)** - The figure illustrates the genetic dissimilarity (assessed through pairwise F_{st} estimates) between the distinct meta-population groups for the H2D haplotype. The different symbols represent each meta-population group.

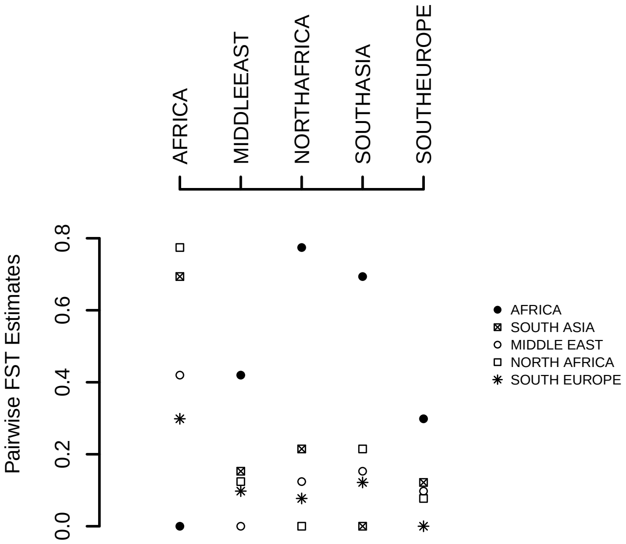


Figure 4: **Genetic Distance estimates between continental groups (within the H2' subtype)** - The figure illustrates the genetic dissimilarity (assessed through pairwise F_{st} estimates) between the distinct meta-population groups for the H2' haplotype. The different symbols represent each meta-population group.

4. Here, our results suggest that the Sub-Saharan African H2' is substantially different from all other H2' chromosomes, as all population-specific H2' chromosomes are found to be genetically closer to each other than to Sub-Saharan Africa. However, given the low number of H2' chromosomes available, these results should be interpreted with caution.

Phylogenetic relationship between the 17q21 haplotypes

It is important to note that so far we have mainly relied on allele frequency information to determine whether the two inversion associated sub-haplotypes display distinct patterns of diversity and differentiation between present-day meta-population groups. While these measures would not be affected by minor intra-haplotypic phasing errors (*e.g.* incorrectly derived haplotypes from an H2D homozygous sample), the same would not apply when considering phylogenetic approaches, where possible errors would likely result in inaccurate evolutionary relationships between chromosomes. In order to circumvent this limitation, we performed a phylogenetic analysis including only heterokaryotype individuals (*i.e.* individuals heterozygous for the orientation) to infer the genetic relationship between haplotypes. Given that one could easily trace phasing errors between the two major haplotype families (*i.e.* H1 vs H2D/H2') (Supp.

Figure 1), this approach has the advantage of avoiding the use of incorrectly phased chromosomes that could be confounded with genetic recombination. The phylogenetic tree of the 17q21 haplotypes derived from heterokaryotypes is thus shown in **Figure 5**. As expected, all chromosomes belonging to the H1 lineage cluster together in a single branch, regardless of the continental origin, clearly separated from all H2 haplotypes.

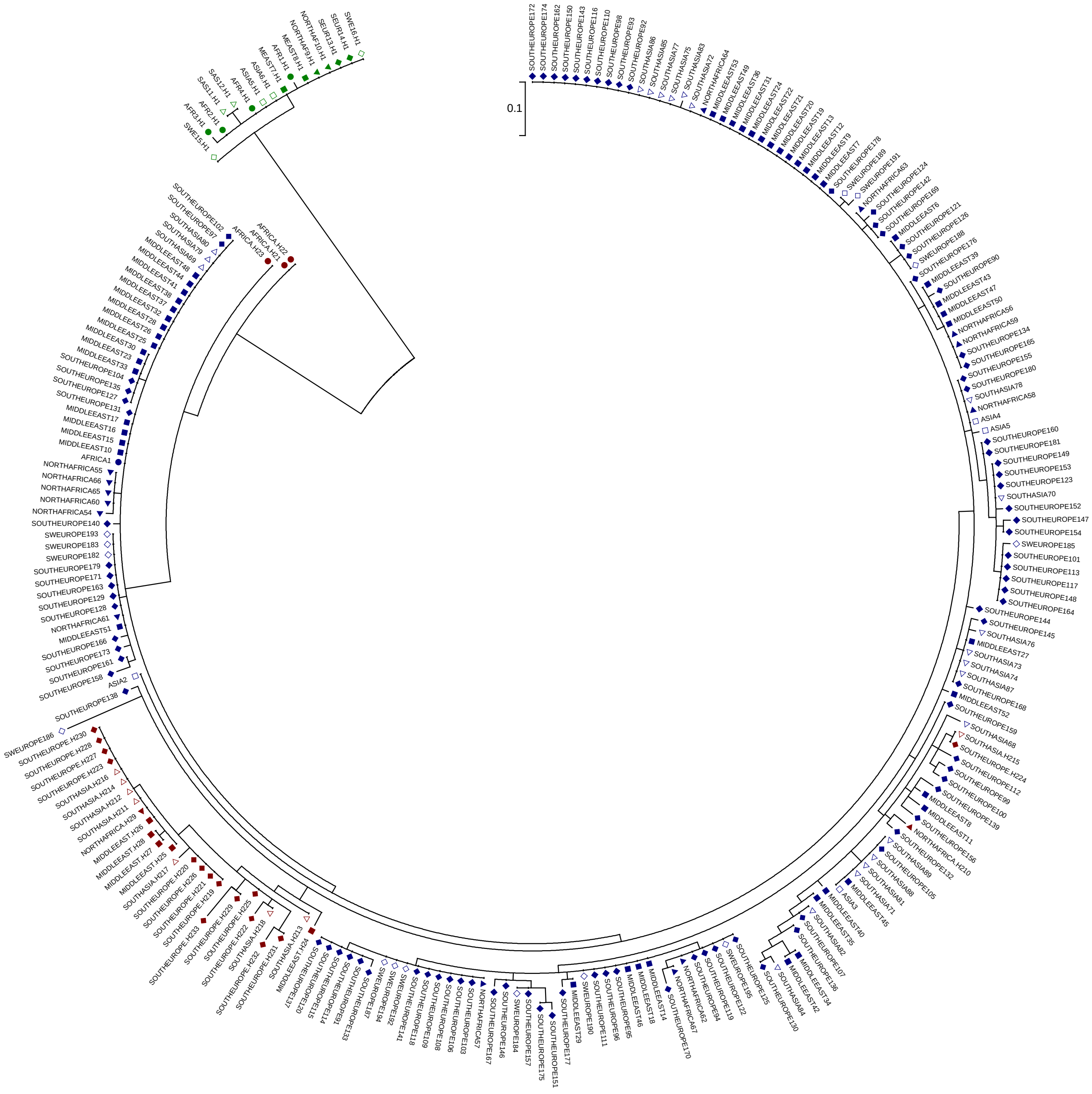


Figure 5: **Phylogenetic Tree of the 17q21 Haplotypes** - Maximum-Likelihood phylogenetic reconstruction of all statistically-derived haplotypes. Distinct colors represent the 3 possible haplotype statuses (H1, H2D, and H2'). Distinct symbols used to represent meta-population groups (see Supp. Figure 4 for higher resolution image).

Furthermore, the African H2' copies appear to represent the ancestral state of the inversion-associated haplotypes, as they were the oldest H2 chromosomes in our dataset (*i.e.* genetically closest to the H1 haplotype). Interestingly, the tree shape also indicates that the remaining H2' chromosomes are genetically distinct from the African copies, as most of the remaining haplotypes cluster together in a predominantly H2D background, hence suggesting the possibility of recurrence throughout human evolution (see Discussion). Finally, it is worth noting that we were unable to identify any signals of population stratification within the H2D lineage.

Discussion

The current study aimed to dissect the complex evolutionary history of the *17q21-inv* by extending previous efforts to uncharacterized populations in order to better understand the processes that have shaped the genetic and architectural diversity of this genomic region. In agreement with recent surveys (Stefansson *et al.* (2005), Zody *et al.* (2008), Donnelly *et al.* (2010), Steinberg *et al.* (2012)), we have found that the inversion-associated haplotype family (*i.e.* H2) is primarily found in Southern European populations where the inversion is segregating at frequencies of approximately 30%. Furthermore, our results demonstrated that both H2 subtypes are also present in North African populations, with most individuals carrying the duplication-specific H2D (frequency of 13%).

Using a unique panel of H2-specific polymorphic variants, we found that the diversity of the inversion-associated haplotypes is considerably higher than initially appreciated (Steinberg *et al.* (2012)). These results are particularly surprising for the H2D sub-type, as previous studies had found essentially no genetic variation within carriers. However, it should be noted that, until now, most estimates were primarily derived from (i) common (non-randomly selected) polymorphisms, where SNP ascertainment bias could be contributing to the lower overall diversity found within the H2 lineage (Stefansson *et al.* (2005), Donnelly *et al.* (2010), Antonacci *et al.* (2009)), and/or (ii) resequenced data from very few individuals (Steinberg *et al.* (2012)).

In addition, we have found that the North African H2 sub-types are genetically much closer to Middle Eastern and South European haplotypes than to the ones present in Sub-Saharan African. While gene flow between North Africa and Europe has long been suspected (Richards *et al.* (2000), Zilhão (2001)), recent genome-wide studies have argued that North African populations exhibit genetic signatures of a very complex demographic history, with contractions and admixture events (with nearby populations) being responsible for its present-day genetic diversity. Indeed, (Henn *et al.*, 2012) have argued that North African populations have initially diverged from Sub-Saharan Africans as part of the early waves of the Out of Africa migrations (approximately 60 Kya). Only afterwards, around 12-40 Kya, the ancestral North Africans started to diverge from the remaining Out of Africa populations as the result of back-to-Africa migrations. The results obtained in this study, are therefore consistent with previous genome-wide findings. Although genetic drift and recent gene flow between North Africa and neighboring populations (*e.g.* South Europe and Middle East) (Botigué *et al.* (2013)) have likely contributed to the observed genetic patterns, the high levels of differentiation found between North African and Sub-Saharan African populations, as well as the intermediate levels of differentiation observed between North Africa and the Middle East and South Europe (**Figure 3** and **Figure 4**), appear to suggest that the inversion-associated haplotypes found in North African populations are not the result of very recent population contacts. However, future work using a larger panel of individuals will be needed to confirm these results.

Also, our phylogenetic analysis revealed that the duplication event within the *KANSL1* gene may not represent a unique event, at least in the H2 background, as previously suggested (Steinberg *et al.* (2012), Boettger *et al.* (2012)). Although our data support a Sub-Saharan African origin for the duplication-specific haplotype (**Figure 5**), we have found that all Non-Sub-Saharan African H2' chromosomes seem to have derived from H2D copies, hence suggesting the possibility of recurrence. While it may be argued that recombination between both inversion-associated haplotypes could have generated similar patterns, the fact that most H2' chromosomes cluster together in a single branch appear to support one of two possible scenarios: either (i) the duplication has recently reverted back to the original configuration (H2') in modern humans, or (ii) the previously identified duplication markers are not fixed between the alternative H2 configurations. Consequently, further work using high coverage re-sequencing data from H2 homokaryotypes will be needed to address this question in greater detail.

In conclusion, while confirming the presence of the H2 haplotype family in North African populations, hence increasing the known distribution of the inversion associated haplotypes to previously uncharacterized human groups, our work showed that the 17q21 H2-family is more diverse than initially appreciated. Furthermore, even though chromosomal inversions represent genomic elements that should be more exposed to genetic drift due to the lack of recombination between differently oriented chromosomes (Alves *et al.* (2012)), our analyses suggest that the North African H2 haplotypes are genetically closer to the haplotypes found in other Non-African populations, further strengthening the results obtained at a genome-wide level. Nevertheless, it should be emphasized that the origin and evolution of the duplication-specific haplotype H2D deserves further consideration, as our phylogenetic analysis appears to indicate that the duplication event may be reversible/recurrent. The results obtained in this study will therefore contribute to the ongoing efforts aimed at understanding the evolutionary processes shaping the 17q21 region. Although the possibility of selection cannot be ruled out, our work showed that the present-day distribution and the genetic signatures within the inverted-associated haplotypes may also be the result of past complex demographic events.

References

Alkan, C., *et al.* 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*, **41(10)**, 1061–67.

Alkan, C., *et al.* 2010. Limitations of next-generation genome sequence assembly. *Nature methods*, **8**, 61–65.

Alkan, C., *et al.* 2011. Genome structural variation discovery and genotyping. *Nature reviews genetics*, **12(5)**, 363–76.

Alves, J.M., *et al.* 2012. On the Structural Plasticity of the Human Genome: Chromosomal Inversions Revisited. *Current Genomics*, **13(8)**, 623–632.

Antonacci, F., *et al.* 2009. characterization of six human disease-associated inversion polymorphisms. *human molecular genetics*, **18(14)**, 2555–2566.

Auton, A., & McVean, G.A. 2007. Recombination rate estimation in the presence of hotspots. *Genome Research*, **18(14)**, 2555–2566.

Ayala, D., *et al.* 2010. chromosomal inversions, natural selection and adaptation in the malaria vector anopheles funestus. *molecular biology and evolution*, **28(1)**, 745–758.

Bailey, J.A., & Eichler, E.E. 2006. primate segmental duplications: crucibles of evolution, diversity, and disease. *nature reviews*, **7**, 552–564.

Baird, N.A., *et al.* 2008. rapid snp discovery and genetic mapping using sequenced rad markers. *plos one*, **3**, 110–115.

Bansal, V., *et al.* 2007. evidence for large inversion polymorphisms in the human genome from hapmap data. *genome research*, **17(2)**, 219–30.

Bardhan, A., & Sharma, T. 2000. meiosis and speciation: a study in a speciating mus terricolor complex. *journal of genetics*, **79**, 105–111.

Baudat, F., *et al.* 2010. prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. *science*, **237**, 836–840.

Beaumont, M.A. 2010. approximate bayesian computation in evolution and ecology. *annual review of ecology, evolution, and systematics*, **41(1)**, 379–406.

Beaumont, M.A., *et al.* 2002. approximate bayesian computation in population genetics. *journal of genetics*, **162**, 2025–2035.

Berg, I.L., *et al.* 2010. prdm9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *nature genetics*, **42**, 859–863.

Blum, M.B. 2010. approximate bayesian computation: a nonparametric perspective. *journal of the american statistical association*, **105(491)**, 1178–1187.

Boettger, L.M., *et al.* 2012. structural haplotypes and recent evolution of the human 17q21.31 region. *nature genetics*, **44(8)**, 881–885.

Bosch, N., *et al.* 2009. nucleotide, cytogenetic and expression impact of the human chromosome 8p23.1 inversion polymorphism. *plos one*, **4(12)**, e8269.

Botigué, LR., *et al.* 2013. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proceedings of the National Academy of Sciences*, **110(29)**, 11791–11796.

Brown, G.M., *et al.* 1998. Genetic analysis of meiotic recombination in humans by use of sperm typing: reduced recombination within a heterozygous paracentric inversion of chromosome 9q32-q34.3. *American Journal of Human Genetics*, **62**, 1484–1492.

Brown, J.F., & O'Neill, R.J. 2010. chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *annual review of genomics and human genetics*, 291–316.

Browning, SR., & Browning, BL. 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics*, **81**, 1084–1097.

Bugge, M., *et al.* 2000. disease associated balanced chromosome rearrangements: a resource for large scale genotype-phenotype delineation in man. *journal of medical genetics*, **37**, 858–865.

Caceres, M., *et al.* 2007. a recurrent inversion on the eutherian x chromosome. *proceedings of the national academy of sciences*, **104(47)**, 18571–18576.

Campbell, CD., *et al.* 2011. Population-genetic properties of differentiated human copy-number polymorphisms. *American Journal of Human Genetics*, **88(3)**, 317–332.

Carvajal-Rodríguez, A. 2010. simulation of genes and genomes forward in time. *current genomics*, **11**, 58–61.

Chadeau-Hyam, M., *et al.* 2008. fregene: simulation of realistic sequence-level data in populations and ascertained samples. *bmc bioinformatics*, **9**, 364.

Cheung, V.G., *et al.* 2007. Polymorphic variation in human meiotic recombination. *American Journal of Human Genetics*, **80(3)**, 526–530.

Chikhi, L. 2009. update to chikhi et al.'s ”clinal variation in the nuclear dna of europeans” (1998): genetic data and storytelling-from archaeogenetics to astrologenetics-. *human biology*, **81(5-6)**, 639–643.

Chowdhury, R., *et al.* 2009. genetic analysis of variation in human meiotic recombination. *plos genetics*, **5**, e1000648.

Clark, A.G., *et al.* 2010. Contrasting methods of quantifying fine structure of human recombination. *Annual Reviews in of Genomics and Human Genetics*, **11**, 45–64.

Conrad, D.F., *et al.* 2010. origins and functional impact of copy number variation in the human genome. *nature*, **464**, 704–712.

Conrad, D.F., & Hurles, M.E. 2007. Polymorphic variation in human meiotic recombination. *American Journal of Human Genetics*, **80(3)**, 526–530.

Consortium, International HapMap. 2003. The International HapMap Project. *Nature*, **426(6968)**, 789–796.

Consortium, The 1000 Genomes Project. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Coop, G., & Przeworski, M. 2007. an evolutionary view of human recombination. *nature reviews genetics*, **8(1)**, 23–34.

Craddock, N., *et al.* 2010. genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls. *nature*, **464(7289)**, 713–20.

Currat, M., *et al.* 2006. comment on ”ongoing adaptive evolution of aspm, a brain size determinant in homo sapiens” and ”microcephalin, a gene regulating brain size, continues to evolve adaptively in humans”. *science*, **313**, 172.

Davey, J.W., *et al.* 2011. genome-wide genetic marker discovery and genotyping using next-generation sequencing. *nature review genetics*, **12**, 499.

de la Chapelle, A., *et al.* 1974. pericentric inversions of human chromosomes 9 and 10. *american journal of human genetics*, **26**, 746–766.

Delaneau, O., *et al.* 2013. Haplotype estimation using sequence reads. *American Journal of Human Genetics*, **93(4)**, 787–796.

Deng, Y., & Tsao, B.P. 2010. Genetic susceptibility to systemic lupus erythematosus in the genomic era. *Nature Reviews Rheumatology*, **6**, 683–692.

Depristo, M., *et al.* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Dobzhansky, T. 1951. *genetics and the origin of species.* oxford: columbia university press: new york.

Donnelly, M.P., *et al.* 2010. the distribution and most recent common ancestor of the 17q21 inversion in humans. *american journal of human genetics*, **86(2)**, 161–71.

Fagundes, N.J.R., *et al.* 2007. statistical evaluation of alternative models of human evolution. *proceedings of the national academy of sciences*, **104(17)**, 614–619.

Fanciulli, M., *et al.* 2007. fcgr3b copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *nature genetics*, 721–723.

Faria, R., *et al.* 2011. role of natural selection in chromosomal speciation. *In: encyclopedia of life sciences*. chichester, uk: John Wiley & Sons, Ltd.

Faria, R., & Navarro, A. 2010. Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends in Ecology and Evolution*, **25**, 660–669.

Farré, M., *et al.* 2013. Recombination Rates and Genomic Shuffling in Human and Chimpanzee - A New Twist in the Chromosomal Speciation Theory. *Molecular Biology and Evolution*, **30(4)**, 853–864.

Fearnhead, P., *et al.* 2004. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics*, **167(4)**, 2067–2081.

Feuk, L. 2007. inversion variants in the human genome: role in disease and genome architecture. *genome medicine*, **2(2)**, 11.

Feuk, L., *et al.* 2005. discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee dna sequence assemblies. *plos genetics*, **1(4)**, e56.

Fledel-Alon, A., *et al.* 2011. variation in human recombination rates and its genetic determinants. *plos one*, **6(6)**, e20321.

Flores, M., *et al.* 2007. recurrent dna inversion rearrangements in the human genome. *proceedings of the national academy of sciences*, **104(15)**, 6099–106.

Fragata, I., *et al.* 2009. contrasting patterns of phenotypic variation linked to chromosomal inversions in native and colonizing populations in drosophila subobscura. *journal of evolutionary biology*, **23(1)**, 112–123.

Garvin, M.R., *et al.* 2010. application of single nucleotide polymorphisms to non-model species: a technical review. *molecular ecology resources*, **10(6)**, 915–934.

Goldstein, D.B., & Chikhi, L. 2002. human migrations and population structure : what we know and why it matters. *annual review of genomics and human genetics*, **3**, 129–152.

Goldstein, D.B., & Weale, M.E. 2001. population genomics: linkage disequilibrium holds the key. *current biology*, **11**, 576–579.

Gonzalez, E., *et al.* 2005. the influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility. *science*, **307**, 1434–40.

Goudet, J. 2005. Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184–186.

Gu, W., *et al.* 2008. mechanisms for human genomic rearrangements. *pathogenetics*, **1(4)**.

Guerrero, R.F., *et al.* 2012. coalescent patterns for chromosomal inversions in divergent populations. *philosophical transactions of the royal society b: biological sciences*, **367(1587)**, 430–438.

Hara, Y., *et al.* 2011. abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. *bmc evolutionary biology*, **11(1)**, 308.

Hardy, J., *et al.* 2005. evidence suggesting that homo neanderthalensis contributed the h2 mapt haplotype to homo sapiens. *biochemical society transactions*, **33(4)**, 582–585.

Henn, BM., *et al.* 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genetics*, **8(1)**, e1002397.

Hey, J. 2003. speciation and inversions: chimps and humans. *bioessays*, 825–828.

Hoffmann, A.A., & Rieseberg, L.H. 2008. revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation-. *annual review of ecology, evolution, and systematics*, **39**, 21–42.

Hohenlohe, P.A., *et al.* 2010. population genomics of parallel adaptation in three-spine stickleback using sequenced rad tags. *plos genetics*, **6(2)**, e1000862.

Hollox, E.J., *et al.* 2008. defensins and the dynamic genome: what we can learn from structural variation at human chromosome band 8p23.1. *genome research*, **18**, 1686–1697.

Hubert, R., *et al.* 1994. High resolution localization of recombination hot-spots using sperm typing. *Nature Genetics*, **7**, 420–434.

Hudson, R. 1990. gene genealogies and the coalescent process. *In: oxford surveys in evolutionary biology*. oxford, new york: oxford university press: oxford.

Huynh, L.Y., *et al.* 2011. chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (zonotrichia albicollis). *heredity*,

106, 537–546.

Iafrate, A.J., *et al.* 2004. Detection of large-scale variation in the human genome. *Nature Genetics*, **36(9)**, 949–51.

Jensen-Seaman, M.I., *et al.* 2004. comparative recombination rates in the rat, mouse, and human genomes. *genome research*, **14**, 528–538.

Jobling, M.A, Hurles, M.E., & Tyler-Smith, C. 2004. *Human Evolutionary Genetics: Origins, Peoples and Disease*. New York: Garland Science: New York.

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Joron, M., *et al.* 2011. chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *nature*, **477**, 203–206.

Keinan, A., & Reich, D. 2010. Human population differentiation is strongly correlated with local recombination rate. *PLoS Genetics*, **6(3)**, e1000886.

Kidd, J.M., *et al.* 2008. mapping and sequencing of structural variation from eight human genomes. *nature*, **453(7191)**, 56–64.

Kirkpatrick, M. 2010. how and why chromosome inversions evolve. *plos biology*, **8**, 9.

Kirkpatrick, M., & Barton, N. 2006. chromosome inversions, local adaptation and speciation. *genetics*, **173(1)**, 419–434.

Klopfstein, S., *et al.* 2005. the fate of mutations surfing on the wave of a range expansion. *molecular biology and evolution*, **23**, 482–490.

Korbel, J.O., *et al.* 2007. paired-end mapping reveals extensive structural variation in the human genome. *science*, **318**, 420–426.

Laayouni, H., *et al.* 2011. Similarity in recombination rate estimates highly correlates with genetic differentiation in humans. *PLoS One*, **6(3)**, e17913.

Lahn, B.T., & Page, D.C. 1999. four evolutionary strata on the human x chromosome. *science*, **286(5441)**, 964–967.

Lee, J.A., *et al.* 2007. a dna replication mechanism for generating non-recurrent rearrangements associated with genomic disorders. *cell*, **131**, 1235–1247.

Levy, S., *et al.* 2007. the diploid genome sequence of an individual human. *plos biology*, **5**, e254.

Li, H., & Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **Epub**.

Li, J., *et al.* 2006. A new method for detecting human recombination hotspots and its applications to the HapMap ENCODE data. *American Journal of Human Genetics*, **79(4)**, 628–639.

Li, N., & Stephens, M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165(4)**, 2213–2233.

Li, S., & Jakobsson, M. 2012. estimating demographic parameters from large scale population genomic data using approximate bayesian computation. *bmc genetics*, **13**, 22.

Lichten, M., & Goldman, A.S. 1995. Meiotic recombination hotspots. *Annual Review of Genetics*, **29**, 423–444.

Lopes, AM., *et al.* 2013. Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene DMRT1. *PLoS Genetics*, **9(3)**, e1003349.

Lowry, D.B., *et al.* 2010. a widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *plos biology*, **8(9)**, e1000500.

Lu, J., *et al.* 2003. comment on ”chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes”. *science*, **302**, 988.

MacDonald, J.R., *et al.* 2013. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*, **42(1)**, 986–992.

Matsunaga, S. 2006. sex chromosome-linked genes in plants. *genes & genetic systems*, **81**, 219–226.

McVean, G.A., *et al.* 2004. The fine-scale structure of recombination rate variation in the human genome. *Science*, **304(5670)**, 581–584.

Medvedev, p., *et al.* 2009. computational methods for discovering structural variation with next-generation sequencing. *nature methods*, **6**, 13–20.

Murphy, W.J., *et al.* 2005. dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *science*, **309**, 613–617.

Myers, S., *et al.* 2005. a fine-scale map of recombination rates and hotspots across the human genome. *science*, **310**, 321–324.

Myers, S., *et al.* 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, **40**, 1124–1129.

Navarro, A., & Barton, N. 2003a. accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *evolution*, **57(3)**, 447–59.

Navarro, A., & Barton, N. 2003b. chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes. *science*, **300(5617)**, 321–324.

Navarro, A., *et al.* 1997. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics*, **146**, 695–709.

Navarro, A., *et al.* 2000. effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in drosophila. *genetics*, **155(2)**, 685–698.

Noor, M.A.F., *et al.* 2001. chromosomal inversions and the reproductive isolation of species. *proceedings of the national academy of sciences*, **98**, 12084–12088.

Nosil, P., & Feder, J.L. 2011. genomic divergence during speciation: causes and consequences. *philosophical transactions of the royal society b: biological sciences*, **367(1587)**, 332–342.

Nothnagel, M., *et al.* 2011. technology-specific error signatures in the 1000 genomes project data. *human genetics*, 505–516.

O'Neill, R.J., *et al.* 2004. centromere dynamics and chromosome evolution in marsupials. *the journal of heredity*, **95(5)**, 375–381.

O'Reilly, P.F., *et al.* 2010. invertfregene: software for simulating inversions in population genetic data. *bioinformatics*, **26(6)**, 838–840.

Orr, A. 1996. Dobzhansky, bateson, and the genetics of speciation. *genetics*, **144**, 1331–1335.

Parvanov, E.D., *et al.* 2010. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science*, **327(5967)**, 835.

Petes, T.D. 2001. meiotic recombination hot spots and cold spots. *nature review genetics*, **2**, 360–369.

Pickrell, J.K., *et al.* 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, **19**, 826–837.

Purandare, S.M., & Patel, P.I. 1997. Recombination hot spots and human disease. *Genome Research*, **7(8)**, 773–786.

Purcell, S., *et al.* 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, **81(3)**, 559–575.

Rao, P.N., *et al.* 2010. recurrent inversion events at 17q21.31 microdeletion locus are linked to the mapt h2 haplotype. *cytogenetic and genome research*, **90095**, 275–279.

Richards, M., *et al.* 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. *American Journal of Human Genetics*, **67(5)**, 12511276.

Rieseberg, L.H. 2001. chromosomal rearrangements and speciation. *trends in ecology and evolution*, **16(7)**, 351–358.

Ross, M.T., *et al.* 2005. the dna sequence of the human x chromosome. *nature*, **434**, 325–337.

Salm, M.P.A., *et al.* 2012. he origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *genome research*, **22(6)**, 1144–1153.

Serre, D., *et al.* 2005. Large-scale recombination rate patterns are conserved among human populations. *Genome Research*, **15(11)**, 1547–1552.

Shaw, C.J., & Lupski, J.R. 2004. implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *human molecular genetics*, **13(1)**, 57–64.

Simpfendorfer, K.R., *et al.* 2012. the autoimmunity-associated blk haplotype exhibits cis-regulatory effects on mrna and protein expression that are prominently observed in b cells early development. *human molecular genetics*, **21(17)**, 3918–3925.

Sindi, S.S., *et al.* 2010. identification and frequency estimation of inversion polymorphisms from haplotype data. *journal of computational biology*, **17(3)**, 517–531.

Skaletsky, H., *et al.* 2003. the male-specific region of the human y chromosome is a mosaic of discrete sequence classes. *nature*, **423(6942)**, 825–37.

Sousa, V., *et al.* 2009. approximate bayesian computation without summary statistics: the case of admixture. *journal of genetics*, **181**, 187–197.

Sousa, V., *et al.* 2011. population divergence with or without admixture: selecting models using an abc approach. *heredity*, **108**, 521–530.

Spirito, F. 1998. *endless forms: species and speciation*. oxford, new york: oxford university press: oxford.

Spirito, F., *et al.* 1993. the establishment of underdominant chromosomal rearrangements in multi-deme systems with local extinction and colonization. *theoretical population biology*, **44**, 80–94.

Spitz, F., *et al.* 2005. inversion induced disruption of the hoxd cluster leads to the partition of regulatory landscapes. *nature genetics*, **37**, 889–893.

Stefansson, H., *et al.* 2005. a common inversion under selection in europeans. *nature genetics*, **37(2)**, 129–37.

Steinberg, K.M., *et al.* 2012. structural diversity and african origin of the 17q21.31 inversion polymorphism. *nature genetics*, **44(8)**, 872–880.

Stephens, M., *et al.* 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Stevison, L.S., *et al.* 2011. effects of inversions on within- and between-species recombination and divergence. *genome biology and evolution*, **3**, 830–841.

Sturtevant, A.H. 1921. a case of rearrangement of genes in drosophila. *proceedings of the national academy of sciences*, **7**, 235–237.

Sudmant, PH., *et al.* 2010. Diversity of human copy number variation and multicopy genes. *Science*, **330(6004)**, 641–646.

Tamura, K., *et al.* 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution. *Molecular Biology and Evolution*, **28**, 2731–2739.

Turner, D.J., *et al.* 2006. assaying chromosomal inversions by single-molecule haplotyping. *nature methods*, **3**, 439–445.

Tuzun, E., *et al.* 2005. fine-scale structural variation of the human genome. *nature genetics*, **37**, 727–732.

Warburton, P.E., *et al.* 2004. inverted repeat structure of the human genome: the x-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *genome research*, **14**, 1861–1869.

Wegmann, D., *et al.* 2009. efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *journal of genetics*, **182**, 1207–1218.

Wilson, M.G., *et al.* 1970. inherited pericentric inversion of chromosome nb. 4. *american journal of human genetics*, **22**, 679–690.

Yu, L., *et al.* 2010. e(nos)/cg4699 required for nanos function in the female germ line of drosophila. *genesis*, **48**, 161–170.

Zang, J., *et al.* 2004. testing the chromosomal speciation hypothesis for humans and chimpanzees. *genome research*, **14**, 845–851.

Zilhão, J. 2001. Radiocarbon evidence for maritime pioneer colonization at the origins of farming in west Mediterranean Europe. *Proceedings of the National Academy of Sciences*, **98(24)**, 1418014185.

Zody, M.C., *et al.* 2008. evolutionary toggling of the mapt 17q21.31 inversion region. *nature genetics*, **40(9)**, 1076–83.

CHAPTER IV

General Discussion

Human genomic diversity is currently acknowledged to encompass more dimensions than those revealed from the now classical SNPs. With the increasing availability of high-throughput genomic data from multiple human populations, a wide range of structural variants (SVs) has become evident, including balanced and unbalanced genomic rearrangements of different sizes (*e.g.* translocations, CNVs, inversions, etc.). Interestingly, ever since high-resolution array-Comparative Genomic Hybridization and microarray technologies were introduced, the contribution of unbalanced SVs to phenotypic variation became widely recognized (Conrad *et al.* (2010)). However, perhaps owing to the difficulty of characterizing balanced rearrangements in large populations, the study of chromosomal inversions has lagged behind and the impact of these structural variants in the human genome remain somewhat elusive.

In this thesis, we have therefore attempted to investigate the role of inverted chromosomal rearrangements in the evolution of the human genome, using genomic information from two well characterized widespread polymorphic inversions - **8p23-inv** and **17q21-inv**. During this final chapter, we will discuss the main findings of our work and suggest areas for future research.

Patterns of recombination rate variation within inversion rearrangements

First, in chapter II, we asked whether chromosomal inversions could contribute to the patterns of recombination rate variation within the human genome, by using population genetic data for the largest known inversion in the human genome.

As inversions are, in theory, expected to suppress recombination between differently oriented chromosomal segments (given that single crossovers between heterokaryotypes will generally result in non-viable gametic products), we estimated the distribution of recombination rates within each structural type in a large dataset that comprised individuals from multiple present-day populations. As described in the Results section, we have found that the observed recombination heterogeneity was primarily caused by the presence of the rearrangement as most variation was found between chromosomal configurations rather than within each orientation (*i.e.* between different populations). Such findings suggest that both structural orientations have been evolving independently over an extended period of time, which led to significant differences in the recombination landscape of the 8p23 region. Interestingly, we additionally found a small segment in the central region of the inversion where the genetic divergence between the two structural haplotypes may have been diluted due to gene flow between the two configurations, suggesting occasional double recombination events.

Our study represented (to our knowledge) the first attempt to examine the evolution of recombination within the same species by exploring the occurrence of a type of structural variant that is more often studied in the context of recombination suppression. The originality of our approach is that we relied on this suppressive effect to study whether recombination has evolved independently in two homologous DNA segments, with opposite orientation that have been subjected to the very same demographic history.

While the results presented in chapter II further highlight the role of chromosomal inversions as elements of evolutionary significance, we should also accept that the study had some limitations. For instance, we used a low density SNP panel to estimate how the patterns of recombination were distributed within the 8p23 region. Consequently, we were unable to properly characterize recombination hotspots in the region and to test whether the observed dissimilarities in recombination could be attributable to sequence variants at the *PRDM9* binding motifs.

Another limitation in our study is that we have only considered the effects of the inversion in the recombination landscape of the affected genomic region. However, it has been argued that inversions are capable of modifying crossover rates throughout the rest of the genome (Stevison *et al.* (2011)). While analyzing the effects of inversions in genome-wide patterns of recombination should prove extremely difficult (or even impossible, given the amount of confounding variables that one may encounter (*e.g.* several inversions might operate simultaneously on an individuals' genome)), one possibility is to investigate whether chromosomal inversions alter the patterns of recombination of the entire chromosome. In addition, this study should be extended to other known inversions in the human genome to test whether the size of the rearrangement is somehow correlated with its effects on genetic recombination.

As highly informative sequenced-derived genomic data continues to accumulate in public databases, future studies should be able to address these issues in greater detail.

Population and evolutionary processes shaping the 17q21 inversion in humans

In chapter III, we re-examined the distribution and diversity patterns of one common polymorphic inversion that became the focus of intense research in the last 10 years, due to the highly complex patterns of genetic differentiation both between and within the different structural configurations. Given the controversy surrounding the evolutionary history of the *17q21-inv*, where conflicting explanations have been put forward to explain its present-day distribution, our initial aim was, thus, to explore previously uncharacterized populations to better understand the processes that have shaped the observed genetic diversity in modern humans. The inclusion of North African populations was, in our opinion, particularly relevant as most studies have, so far, been biased towards sub-Saharan groups even though they may not be the most informative (see chapter III).

Altogether, the results obtained in chapter III highlight four important features of the 17q21 region. First, the distribution of the inversion-associated haplotype is wider than previously appreciated, with North African populations exhibiting the inversion associated haplotypes at fairly high frequencies, in contrast to what is found in most Sub-Saharan populations.

Secondly, most studies have sharply underestimated the genetic diversity of the inversion-associated haplotype family by relying on common SNP panels enriched for positions polymorphic in H1 chromosomes (the most frequent configuration). This was particularly evident for the duplication-specific haplotype (H2D) where nucleotide diversity has been underestimated in previous studies.

Third, the spread of this rearrangement in present-day populations is likely the result of the complex demographic history of human groups, as (*) our phylogenetic analysis suggests an African origin of the H2 haplotype, since the oldest H2 chromosomes (more closely related to H1) were found in Sub-Saharan Africa; and (**) our estimates of genetic distances (based on pairwise F_{st} calculations) isolated Sub-Saharan Africans from the remaining meta-population groups, and revealed intermediate levels of differentiation between North African, Middle Eastern and South European populations. Interestingly, these results are consistent across both H2 subtypes and support previous research suggesting that the ancestral North-African population was part of the first migration wave of modern humans that later migrated back to Africa (Henn *et al.* (2012)).

Fourth, and perhaps more importantly, the partial duplication events at the *KANSL1* locus may be recurrent, at least in the H2 background, further highlighting the complex architecture of the region. It has recently been argued (Steinberg *et al.* (2012)) that the duplication-specific haplotype (H2D) is the only structural configuration at 17q21 that predisposes to clinically significant micro-deletions, as a result of NAHR of homologous segmental duplications in direct orientation (see chapter I). If the duplication event is reversible, then one would expect that the phenotypic relevance of the H2D haplotype is currently being over estimated as most studies (including ours) are indirectly inferring the duplication status relying on diagnostic SNPs. Clinically oriented research is therefore crucial to understand how these polymorphic features can contribute to the establishment of complex human disorders. One possibility is to use targeted re-sequencing information from family trios to evaluate the frequency of the partial duplication event.

In the beginning of this thesis and throughout chapter III, we have mentioned that another important issue surrounding the 17q21 locus is whether selection has played a major role in its evolutionary history. To date, however, this question has never been explicitly tested despite the surprisingly large number of studies claiming selection based on indirect evidence. As previously stated, complex demographic changes can generate specific genetic signatures that may often be mistaken for selection. Quantifying the contribution of past events to the present-day genetic diversity is therefore needed to fully understand the processes driving the spread of the 17q21 inversion in human populations. Interestingly, recent advances in genetic modelling allowed the development of powerful statistical inferential tools (*e.g.* approximate Bayesian computation) that are efficiently being used to reconstruct the demographic history of populations, as well as to infer growth rates, times of divergence and other important population genetics parameters. We, thus, started working with the invertFREGENE software - a *forward-in-time* simulator that allows the introduction of a single inversion polymorphism of specified length into a population, with the possibility of accounting for complex demographic scenarios. Using its core engine, we modified some of its original features and have implemented the possibility of estimating important population genetic statistics (*e.g.* nucleotide and haplotype diversity, Tajimas D, etc.) from the simulated data.

So far, we carried out simulations under a limited set of neutral scenarios to determine whether the spread of this inversion could be explained by stochastic processes (such as wave surfing). As an overview, the simulation framework begins with the introduction of an inversion in a panmictic population of constant size. The inversion will then be transmitted to subsequent generations until a target frequency is reached. Afterwards, a set of summary statistics are estimated from the simulated genetic data and the obtained values are finally compared with the ones observed in real data. Due to the computational cost of the simulations, we could not develop a full ABC approach with hundreds of thousands or millions of simulations. However, we tried to identify a wide array of models and tested a total of 64 scenarios. We used a parameter-based inference model where three important parameters were allowed to vary, namely Effective population size (Ne), mutation rate, and recombination rate. We then tried to identify which combinations of parameter values best explained the present-day distribution of genetic diversity and variation. The results of our (admittedly limited) simulations (n=1600, *i.e.* 25 independent replicates for each parameter combination) are illustrated in **Figure 1** (and Supplementary Figure 1 - Appendix B).

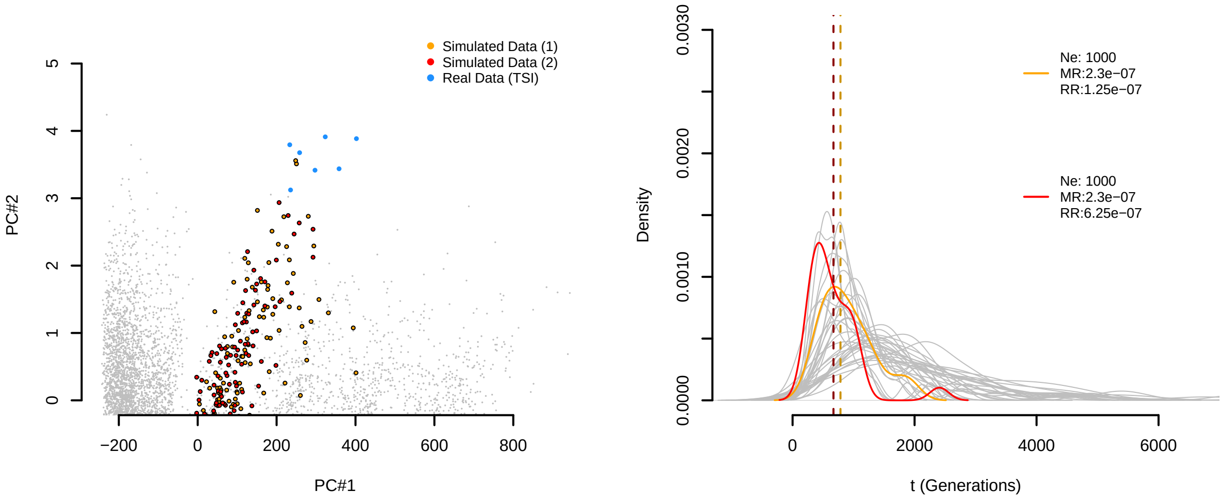


Figure 1: **Single polymorphic inversion simulated in multiple neutral scenarios** - Left plot illustrates a Principal Component Analysis (PCA) between the genetic patterns generated from our simulations (grey dots) with the values obtained for the *17q21-inv* (blue dots). Each dot corresponds to a genomic window of 100Kb (see Appendix B - Supp. Information). Red and orange dots highlight two distinct simulated scenarios that partly overlap with the "real" data. Right figure illustrates the kernel density estimates (for each scenario) of the distribution of the age of the inversion when the target frequency is reached. Red and orange lines correspond to the scenarios highlighted in the left figure. Dashed vertical lines represent the average number of generations for each highlighted scenarios (Red: 690 generations; Orange: 785 generations)

Interestingly, we found that under a neutral model of evolution without any population structure, admixture or any population size change some of the simulated scenarios produce genetic patterns that are similar or partly overlap with the ones found in present-day data. While these results should not (read ”cannot”) be interpreted as conclusive, it suggests that, before invoking selection, genetic patterns like the ones observed in the 17q21 region may need to be interpreted by using neutral models where genetic drift and mutation play major roles. Given that natural populations are much more complex than the ones simulated in our model, new and more realistic scenarios should be incorporated to the simulation framework to account for important demographic events (*e.g.* bottlenecks, expansions, admixture, etc.).

Our results should not be interpreted as a rejection of selection either. Rather we are convinced that scenarios with selection should also be explored. For instance it would also be interesting to test whether similar results could be obtained in a scenario where the inversion is under positive selection, as initially suggested by Stefansson *et al.* (2005).

Another potential issue with our current simulation strategy is the fact that each simulation will stop as soon as the target frequency of the inversion is reached. However, the frequency of inversions, like any other type of mutation, may randomly fluctuate over time, as a result of genetic drift. We have therefore modified the invertFREGENE code to account for new ”stop” conditions (*e.g.* age of the simulated inversion) and we will soon begin to explore this issue in greater depth.

In conclusion, the work presented throughout this thesis suggests that balanced structural changes have had important consequences in human evolution. Given the abundance of chromosomal inversions in polymorphic proportions, additional data is now needed to understand (1) how these rearrangements evolve, (2) which forces and mechanisms are actively maintaining them, and (3) how they contribute to the genetic make-up of human populations.

At this stage, alternative and more flexible strategies should be considered to address some of these issues. As seen during this last chapter, incorporating simulation-based methods may lead to interesting results. In parallel, comparative studies should also analyze the molecular effects of inversions and test whether such effects are correlated with different properties of the rearrangement (*e.g.* size, physical location, etc.).

APPENDIX

Appendix A:
The evolutionary history of a common polymorphic inversion

Supplementary Information

Cytogenetic validation of 17q21 Inversion by Fluorescent *in situ* Hybridization

Individual lymphoblastoid cell lines were obtained for a total of 9 Hapmap individuals from the Coriell Cell Repository (Camden, NJ USA) (Supp. Table 2). DUAL-color FISH was performed, using 3 BAC clones (RP11-403G3, RP11-256F16, and RP11-80L9) directly labeled by nick-translation with biotin, and dioxigenin and hybridized to interphase nuclei, as described in (Rao *et al.* (2010)). The cells were then stained with DAPI, and digital images were obtained using a Leica DMRXA2 fluorescence microscope equipped with a CCD camera and appropriate filters. Patterns of fluorescence signals were analyzed for a minimum of 50 interphase nuclei per individual, in order to statistically determine the orientation of the 17q21 segment.

The results are illustrated in Supp. Figure 1. A green-red-green (GRG) signal pattern corresponds to the direct orientation, while a green-green-red (GGR) signal pattern corresponds to the inverted orientation. In contrast to Rao *et al.* (2010), a perfect correlation between the inversion status and the SNP-defined haplotypes was observed for all individual cell lines. Indeed, the H1 haplotype was found to be associated with the direct orientation while the H2 haplotype was always associated with the inverted configuration.

Consequently, haplotype-informative SNPs (Antonacci *et al.* (2009), Donnelly *et al.* (2010)) were used to classify the remaining individual samples according to inversion status (Supp. Table 1).

Estimation of Copy Number Polymorphisms within the inversion-associated haplotypes

As previously suggested (Sudmant *et al.* (2010)), sequence read-depth information may be used to reliably detect genomic regions that differ in copy-number. We therefore assessed the presence/absence of the short *KANSL1* gene duplication (*i.e.* CNP155) by analyzing the sequence reads from 13 H2 homozygous individuals available from the 1000Genomes Project webpage (1000Genomes Consortium (2012)). Copy-number differences were estimated by counting the number of individual reads that mapped to the region of interest.

The results are shown in Supp. Figure 2. Interestingly, even with low-coverage resequenced data, we were able to classify all individuals according to its copy-number content, as individuals homozygous for the CNP155 H2D show, on average, 4x greater coverage for the genomic region where the duplication is located (chromosome 17: 44,210,855-44,294,624 (assembly GRCh37)) when compared to individuals homozygous for the less complex configuration (*i.e.* H2).

The previously reported duplication-maker SNPs (rs199448, rs199533) were found to be in perfect correlation with the copy-number content (Supp. Table 3). These diagnostic SNPs were then used to classify the remaining individuals according to the duplication status (Supp. Table 1).

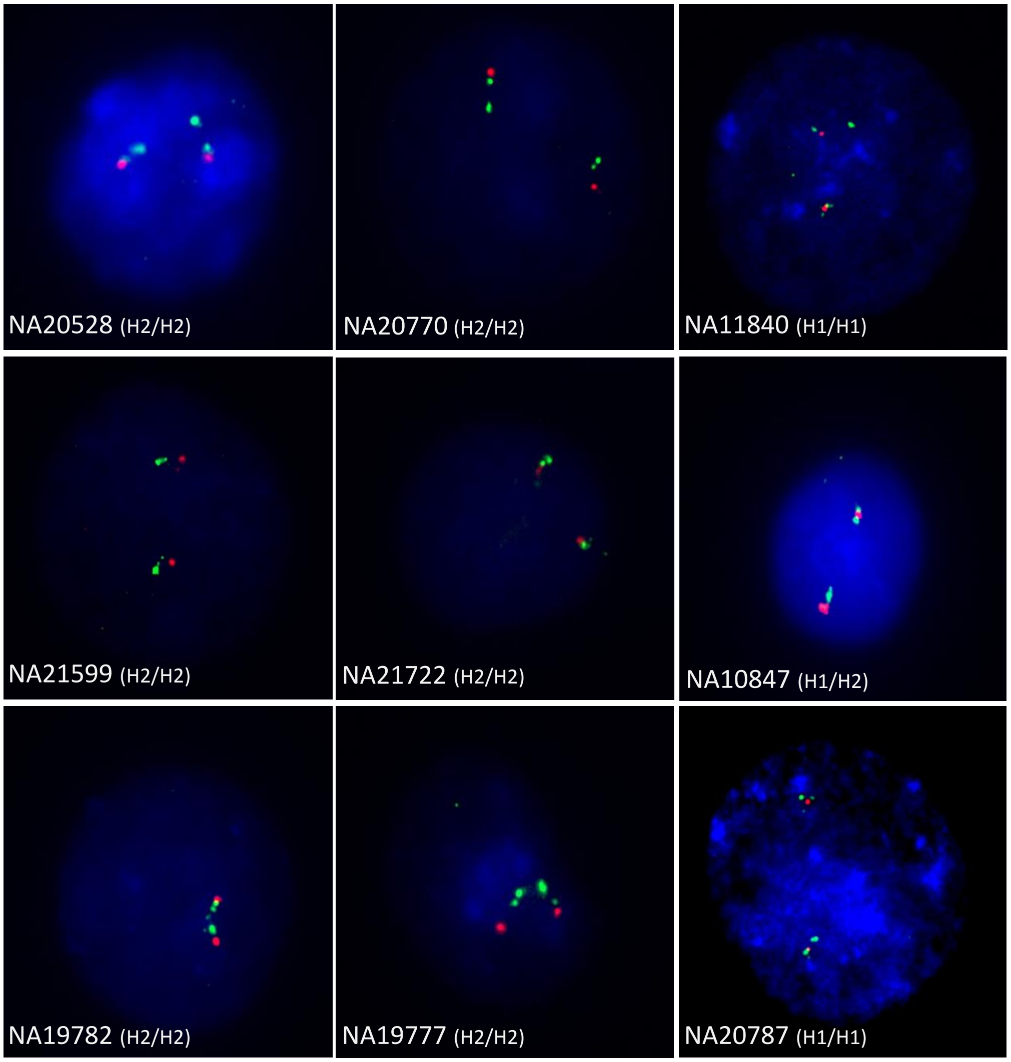


Figure 1: **Cytogenetic validation of 17q21 inversion** - Fluorescent-signal patterns of the 17q21 segment for 9 Hapmap individuals from distinct continental groups. *Green-Red-Green* indicates standard orientation; *Green-Green-Red* indicates inverted orientation. SNP derived haplotypes indicated in each individual digital image. A minimum of 50 interphase nuclei were analyzed for each individual cell line to statistically determine the orientation of the segment.

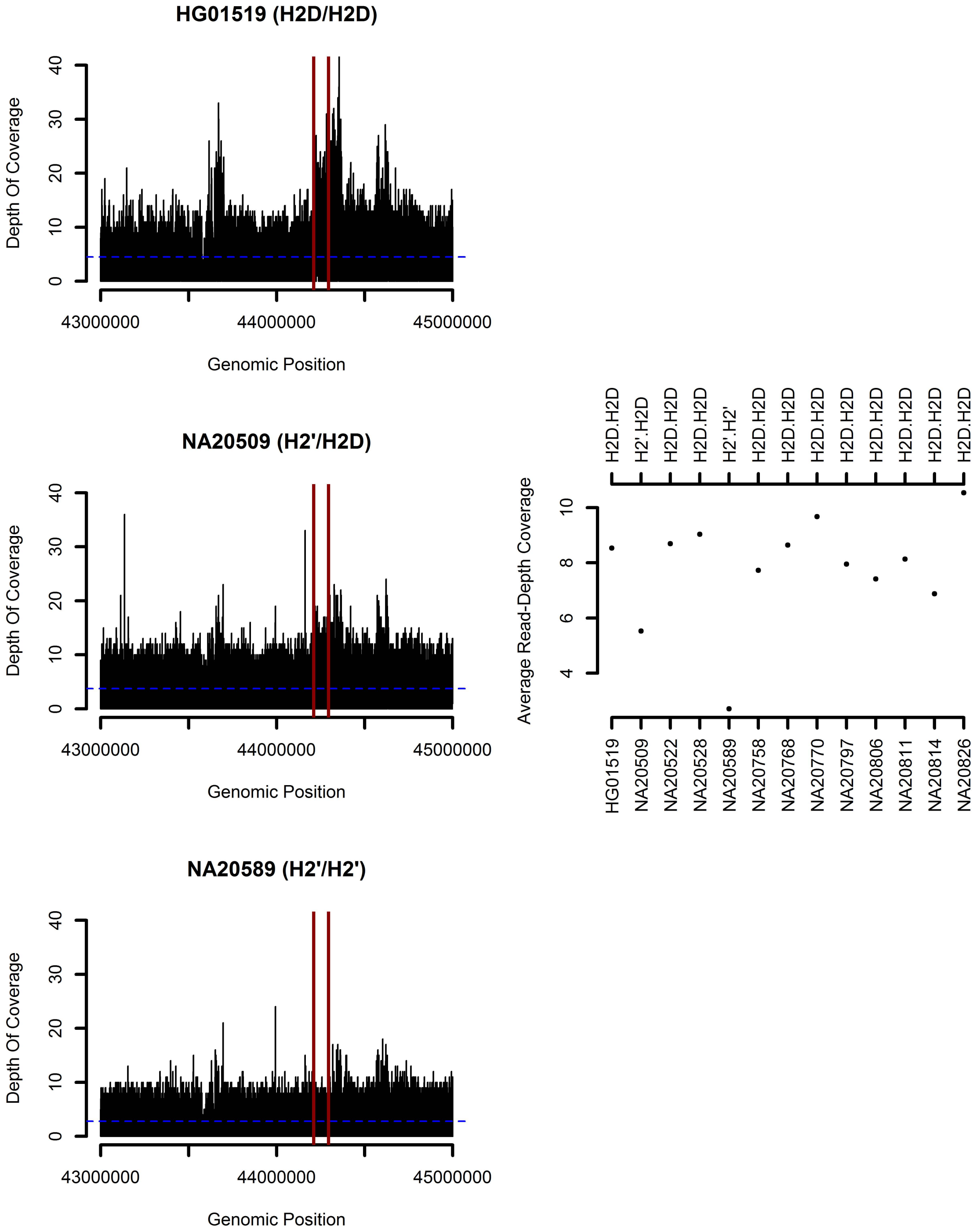


Figure 2: **Copy Number Estimates within the *KANSL1* gene** - Estimates of Copy-number polymorphisms for the inversion-associated haplotypes derived from read-depth coverage analysis. Left-sided Barplots represent 3 individuals from the 1000 Genomes Project with distinct duplication-based genotype - H2D/H2D; H2’/H2D; H2’/H2’. Solid Red delimiters used to specify the duplication genomic location (see above). Right plot illustrate the average read-depth coverage and corresponding duplication-based genotype of 13 individuals from the 1000 Genomes.

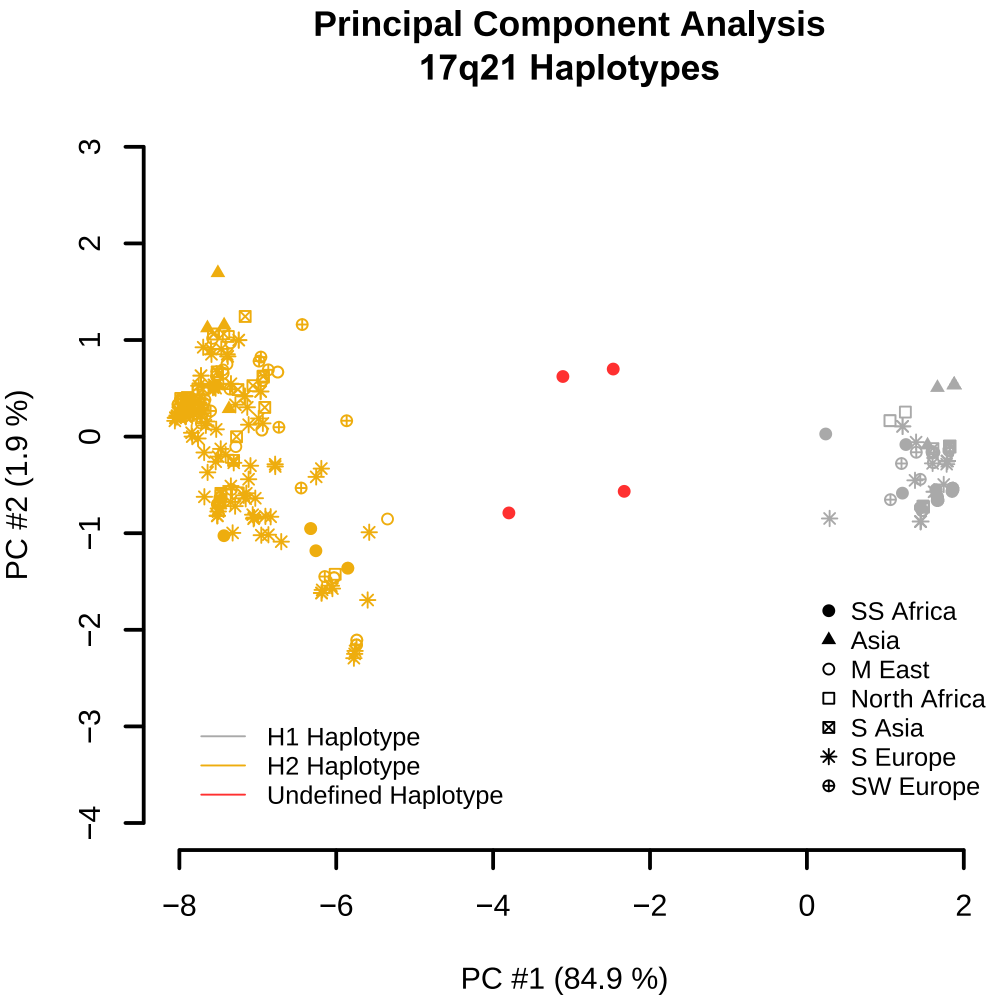


Figure 3: **Global genetic stratification at the 17q21 region (H2-Specific Panel)** - *Principal Component Analysis* (PCA) performed on 1600 statistically-derived chromosomes (800 Individual samples) from our final Dataset. A total of 80 H2-specific SNPs were used. Each dot corresponds to one chromosome, with distinct symbols representing geographical-specific groups. Distinct colors differentiate the two major haplotypes at 17q21 (*i.e.* H1 and H2). Red dots highlight individual chromosomes with undefined haplotype (all due to phasing errors). The first principal component (*i.e.* horizontal axis) illustrates the strong genetic differentiation between the two oppositely oriented haplotypes.

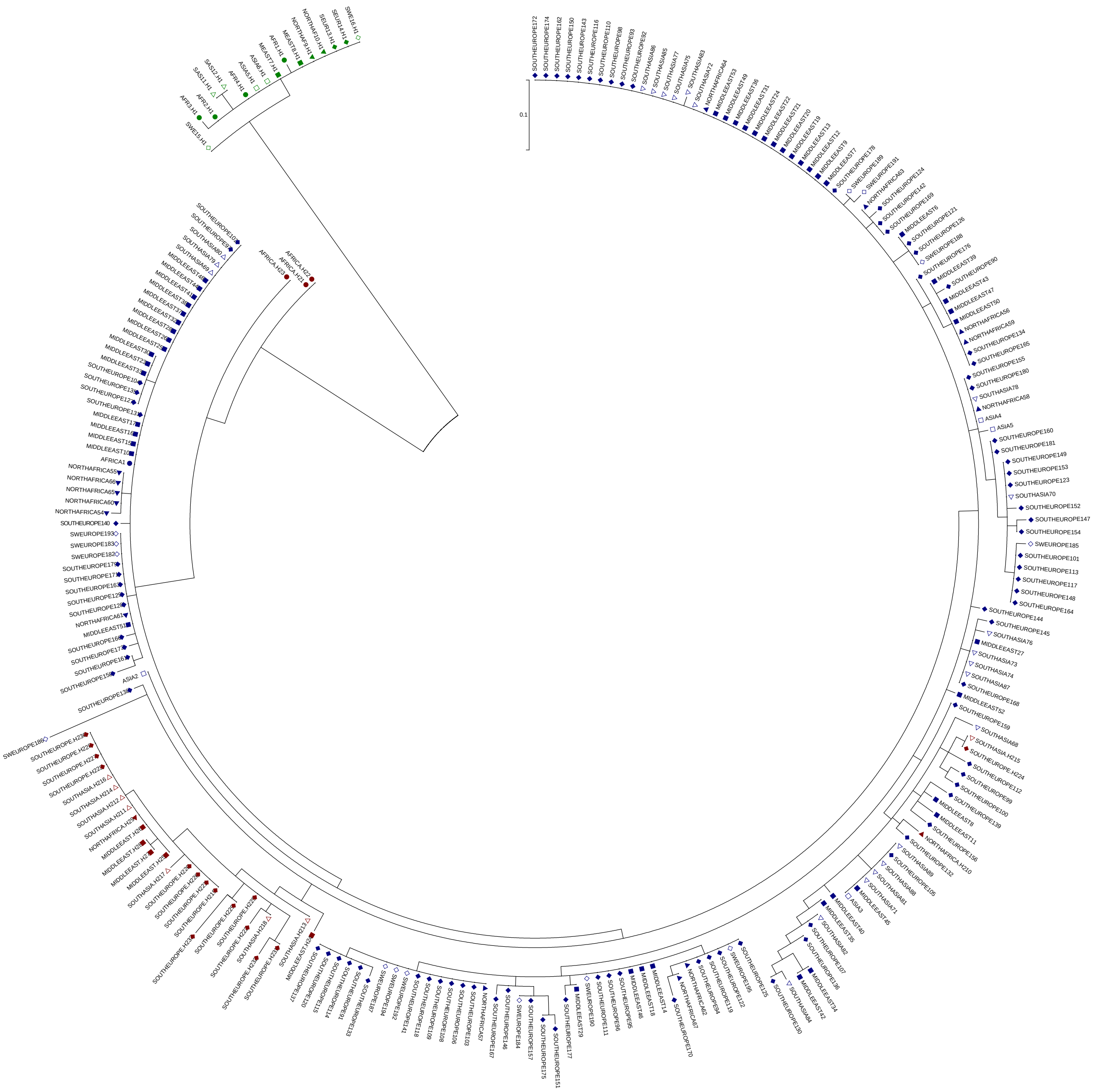
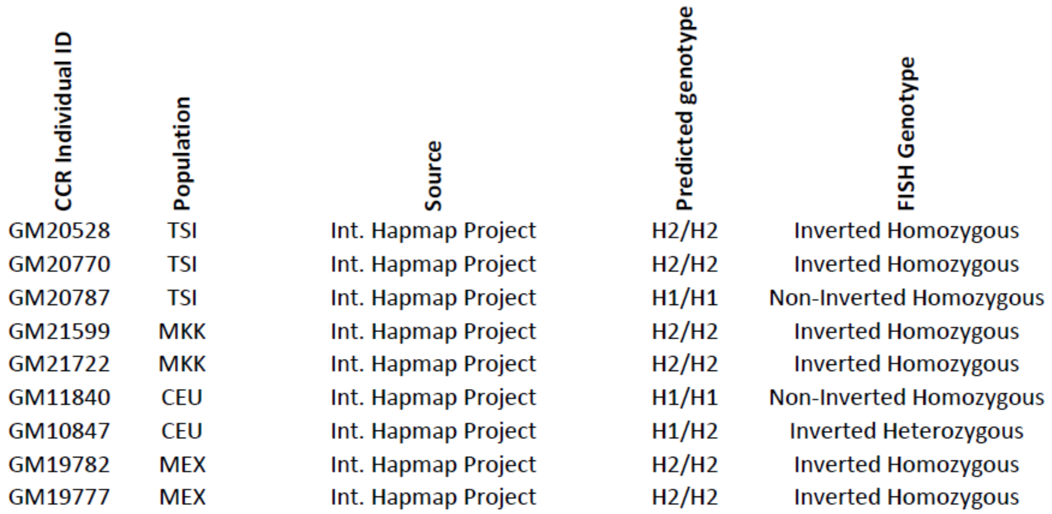


Figure 4: **Maximum-Likelihood phylogenetic reconstruction of 17q21 haplotypes**

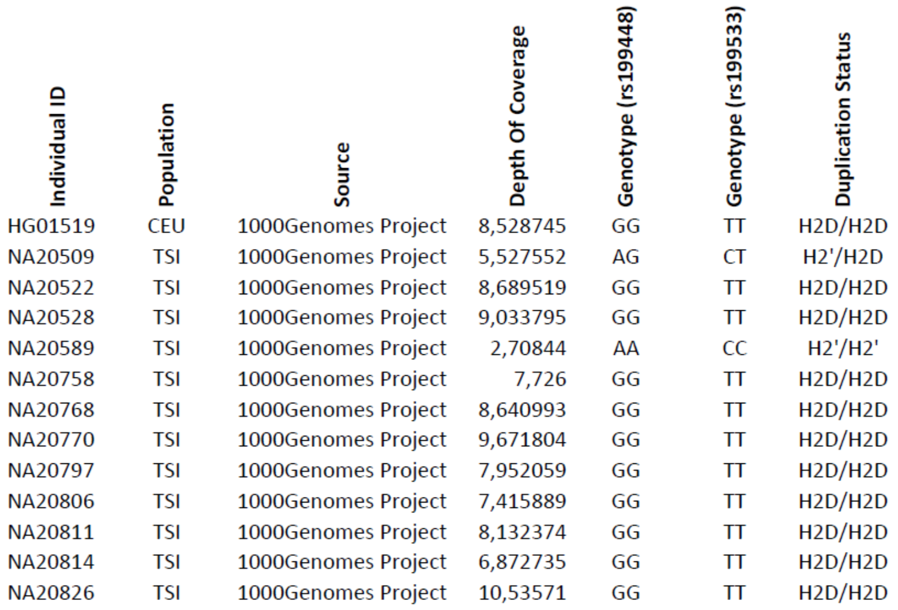
INDIVIDUAL	POPULATION	SOURCE	H2.RICH SNP SET	GENOTYPE DATA (Affy arrays)	DUPLICATION INFO	GENOTYPE	H2D (?)
HGDP01406	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01411	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01412	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01414	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01415	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01417	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01418	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00993	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00994	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01030	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01034	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01033	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01028	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01035	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01031	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00454	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00455	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00457	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00458	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00459	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00460	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00461	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00464	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00465	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00466	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00469	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00470	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00472	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00473	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00475	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00479	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00986	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01086	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01090	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP01094	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00904	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00905	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00906	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00907	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00908	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00910	AFRICA	HGDP PANEL		X	X	H1.H1	..
HGDP00912	AFRICA	HGDP PANEL		X	X	H1.H1	..

Table I: **List of Individuals, and corresponding geographical group, used in the current study.** Due to the large number of individuals in the table, the current list represents only a small subset extracted from the original table. The full table is accessible online and can be downloaded here: https://docs.google.com/spreadsheets/d/14xdfSXVnUGh86Oq3PVWge9C3TkSHgnEY-twcksrUOzI/edit?usp=sharing



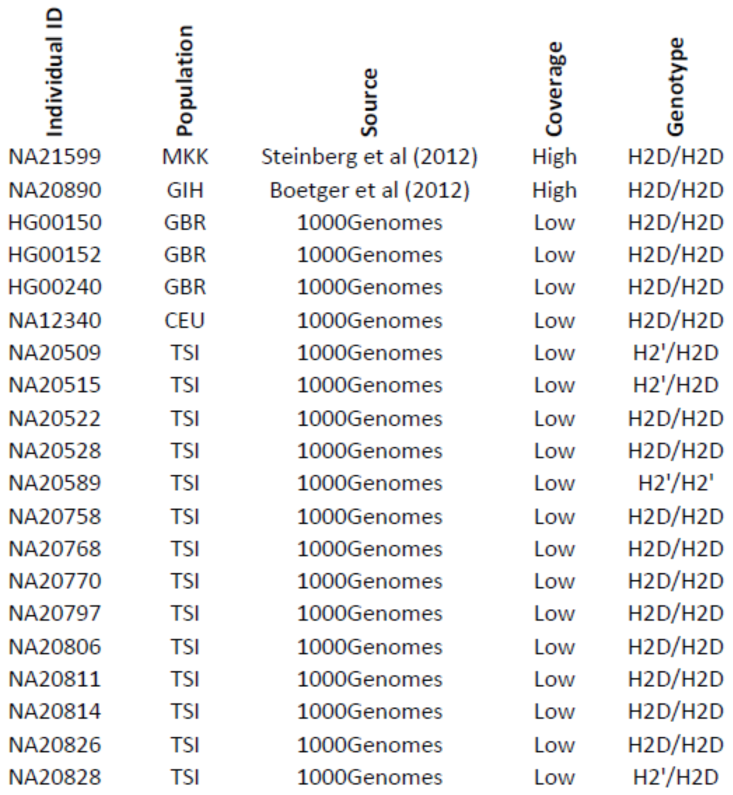
CCR Individual ID	Population	Source	Predicted genotype	FISH Genotype
GM20528	TSI	Int. Hapmap Project	H2/H2	Inverted Homozygous
GM20770	TSI	Int. Hapmap Project	H2/H2	Inverted Homozygous
GM20787	TSI	Int. Hapmap Project	H1/H1	Non-Inverted Homozygous
GM21599	MKK	Int. Hapmap Project	H2/H2	Inverted Homozygous
GM21722	MKK	Int. Hapmap Project	H2/H2	Inverted Homozygous
GM11840	CEU	Int. Hapmap Project	H1/H1	Non-Inverted Homozygous
GM10847	CEU	Int. Hapmap Project	H1/H2	Inverted Heterozygous
GM19782	MEX	Int. Hapmap Project	H2/H2	Inverted Homozygous
GM19777	MEX	Int. Hapmap Project	H2/H2	Inverted Homozygous

Table II: **List of Hapmap individuals used in the Fluorescent *in situ* hybridization analysis**



Individual ID	Population	Source	Depth Of Coverage	Genotype (rs199448)	Genotype (rs199533)	Duplication Status
HG01519	CEU	1000Genomes Project	8,528745	GG	TT	H2D/H2D
NA20509	TSI	1000Genomes Project	5,527552	AG	CT	H2'/H2D
NA20522	TSI	1000Genomes Project	8,689519	GG	TT	H2D/H2D
NA20528	TSI	1000Genomes Project	9,033795	GG	TT	H2D/H2D
NA20589	TSI	1000Genomes Project	2,70844	AA	CC	H2'/H2'
NA20758	TSI	1000Genomes Project	7,726	GG	TT	H2D/H2D
NA20768	TSI	1000Genomes Project	8,640993	GG	TT	H2D/H2D
NA20770	TSI	1000Genomes Project	9,671804	GG	TT	H2D/H2D
NA20797	TSI	1000Genomes Project	7,952059	GG	TT	H2D/H2D
NA20806	TSI	1000Genomes Project	7,415889	GG	TT	H2D/H2D
NA20811	TSI	1000Genomes Project	8,132374	GG	TT	H2D/H2D
NA20814	TSI	1000Genomes Project	6,872735	GG	TT	H2D/H2D
NA20826	TSI	1000Genomes Project	10,53571	GG	TT	H2D/H2D

Table III: **List of 1000 Genomes Project samples with read-depth coverage information** Average Depth of Coverage for each individual shown, as well as the Genotype of the *Duplication tagging* SNPs.



Individual ID	Population	Source	Coverage	Genotype
NA21599	MKK	Steinberg et al (2012)	High	H2D/H2D
NA20890	GIH	Boetger et al (2012)	High	H2D/H2D
HG00150	GBR	1000Genomes	Low	H2D/H2D
HG00152	GBR	1000Genomes	Low	H2D/H2D
HG00240	GBR	1000Genomes	Low	H2D/H2D
NA12340	CEU	1000Genomes	Low	H2D/H2D
NA20509	TSI	1000Genomes	Low	H2'/H2D
NA20515	TSI	1000Genomes	Low	H2'/H2D
NA20522	TSI	1000Genomes	Low	H2D/H2D
NA20528	TSI	1000Genomes	Low	H2D/H2D
NA20589	TSI	1000Genomes	Low	H2'/H2'
NA20758	TSI	1000Genomes	Low	H2D/H2D
NA20768	TSI	1000Genomes	Low	H2D/H2D
NA20770	TSI	1000Genomes	Low	H2D/H2D
NA20797	TSI	1000Genomes	Low	H2D/H2D
NA20806	TSI	1000Genomes	Low	H2D/H2D
NA20811	TSI	1000Genomes	Low	H2D/H2D
NA20814	TSI	1000Genomes	Low	H2D/H2D
NA20826	TSI	1000Genomes	Low	H2D/H2D
NA20828	TSI	1000Genomes	Low	H2'/H2D

Table IV: **List of individuals used to search for H2-specific polymorphic variants**

Chromosome	Genomic Position (GRCh 37)	SNP rsID	Reference allele	Alternative Allele
17	43653129	rs141461397	A	G
17	43654468	rs149133346	C	T
17	43663780	rs183968659	C	G
17	43667792	rs113083040	C	T
17	43667836	rs112193661	C	A
17	43674130	rs141928757	A	G
17	43688387	rs148732400	A	G
17	43692338	rs146145951	T	C
17	43701023	rs149553060	T	C
17	43707619	rs116916717	C	A
17	43711678	rs143533902	C	T
17	43718027	rs117618829	C	T
17	43722604	rs118084908	A	G
17	43723929	rs188115863	G	A
17	43735478	rs116961033	A	G
17	43742298	rs117003064	T	A
17	43757450	rs117795902	C	A
17	43757777	rs118171450	C	T
17	43761856	rs116870912	C	T
17	43773124	rs117403175	A	C
17	43786614	rs118076411	T	C
17	43789710	rs117556827	C	G
17	43798308	rs117615688	G	A
17	43806925	rs114967794	G	A
17	43812737	rs185046589	T	G
17	43825247	rs140493983	G	T
17	43828764	rs115283395	G	A
17	43836953	rs117521065	A	C
17	43839951	rs117450815	A	T
17	43848638	rs117024240	G	A
17	43862593	rs144817232	G	A
17	43897855	rs151064014	G	T
17	43902541	rs4327090	A	G
17	43923410	rs150431364	C	T
17	43926149	rs112907840	A	G
17	43927290	rs117194038	G	A
17	43932797	rs117051720	G	C
17	43939214	rs140843301	T	C
17	43946112	rs118050707	T	C
17	43946291	rs117908675	A	G
17	43949342	rs117984054	G	A
17	43953475	rs183583287	C	T
17	43954686	rs117515986	G	A
17	43968246	rs117293754	T	C
17	43977730	rs185273085	G	A
17	43994870	rs183374586	T	C
17	43998278	rs117852805	C	T
17	44001017	rs188817542	G	A
17	44019199	rs191419241	A	T
17	44039691	rs117155798	A	G
17	44041562	rs117932281	G	T
17	44045585	rs191929402	C	A
17	44051846	rs117965319	A	G
17	44052552	rs140753174	C	T
17	44073889	rs114553892	A	G
17	44078618	rs117499775	T	C
17	44101563	rs118104841	T	C
17	44102604	rs113373871	T	C
17	44105727	rs117379709	G	A
17	44117960	rs145209030	G	A
17	44141347	rs113010151	A	G
17	44142024	rs143590710	C	G
17	44156181	rs117540507	T	A
17	44175240	rs190618874	T	C
17	44196153	rs116875990	A	G
17	44201791	rs112073200	G	C
17	44217226	rs117980624	A	G
17	44220454	rs117122375	T	G
17	44226980	rs150079804	A	G
17	44236321	rs117991098	C	G
17	44256762	rs145723347	T	C
17	44292139	rs140034546	A	G
17	44296251	rs112830278	T	C
17	44314342	rs111265932	T	C
17	44314522	rs141211409	A	G
17	44335166	rs113630199	C	T
17	44357255	rs140176214	T	C
17	44357351	rs113377745	T	C
17	44809001	rs199448	A	G

Table V: **List of SNPs and genomic position of H2-polymorphic panel**

Appendix B: invertFREGENE Simulations

Supplementary Information

Using a modified version of the invertFREGENE software (O'Reilly *et al.* (2010)), we carried out a set of distinct simulations under a neutral model of evolution. We began by simulating a population of chromosomes of size 2 Mb in length to equilibrium without an inversion (10N generations). Afterwards, a 500 Kb inversion was introduced in one of the chromosomes (physical location: 750 - 1,250 Kb). The inversion would then be transmitted through consecutive generations until the target frequency was reached. Here, we set the target frequency of the inversion to 0.30 - frequency of the 17q21 in the Southern European populations.

We tested a total of 64 distinct scenarios, using 25 independent replicates for each parameter combination (*i.e.* using a different seed value for each replicate).

In all tested models, three important parameters were allowed to vary, namely Effective population size (Ne), mutation rate, and recombination rate. As suggested in O'Reilly *et al.* (2010), effective population size was decreased by a factor of ten to save computational time. Our Ne parameter was, thus, initially set to 1000. In contrast, the rates used for mutation and recombination were initially set to $2.3\text{x}10^{-07}$ and $1.25\text{x}10^{-07}$, respectively (*i.e.* values 10 times higher than the predicted per base mutation and recombination rate in humans).

These parameters were afterwards set to different values and our simulation framework included models where the mutation and recombination rate were set to 5 times higher (and/or lower) than average and the Ne parameter varied from 1000 to 2500.

The idea of using such unrealistic parameters may seem strange, but can often lead to important conclusions regarding the quality of the inferential method. For instance, if the genetic data generated from an unrealistic scenario is the one that best explains the data observed in natural populations, one will likely conclude that the chosen model is not the most appropriate.

All data (real and simulated) were then divided in genomic windows of 100 Kb in size. This approach allowed us to explore whether different regions within the inversion showed distinct patterns of genetic variation. Finally, the genetic data was summarized by a set of summary statistics, namely: (1) Number of Segregating Sites (S); (2) Nucleotide Diversity (Π); (3) Mean Heterozygostiy; (4) Tajima's D; (5) Haplotype Diversity.

A Principal Component Analysis (PCA) was subsequently performed using the obtained estimates to identify the relationship between the simulated data and the real data. The results of this analysis are illustrated in **Figure 1**. Interestingly, we have found that the simulated genetic data that produced genetic patterns that are more similar to the ones found in ”real” present-day populations, were generated using the standard values estimated from human data (see General Discussion).

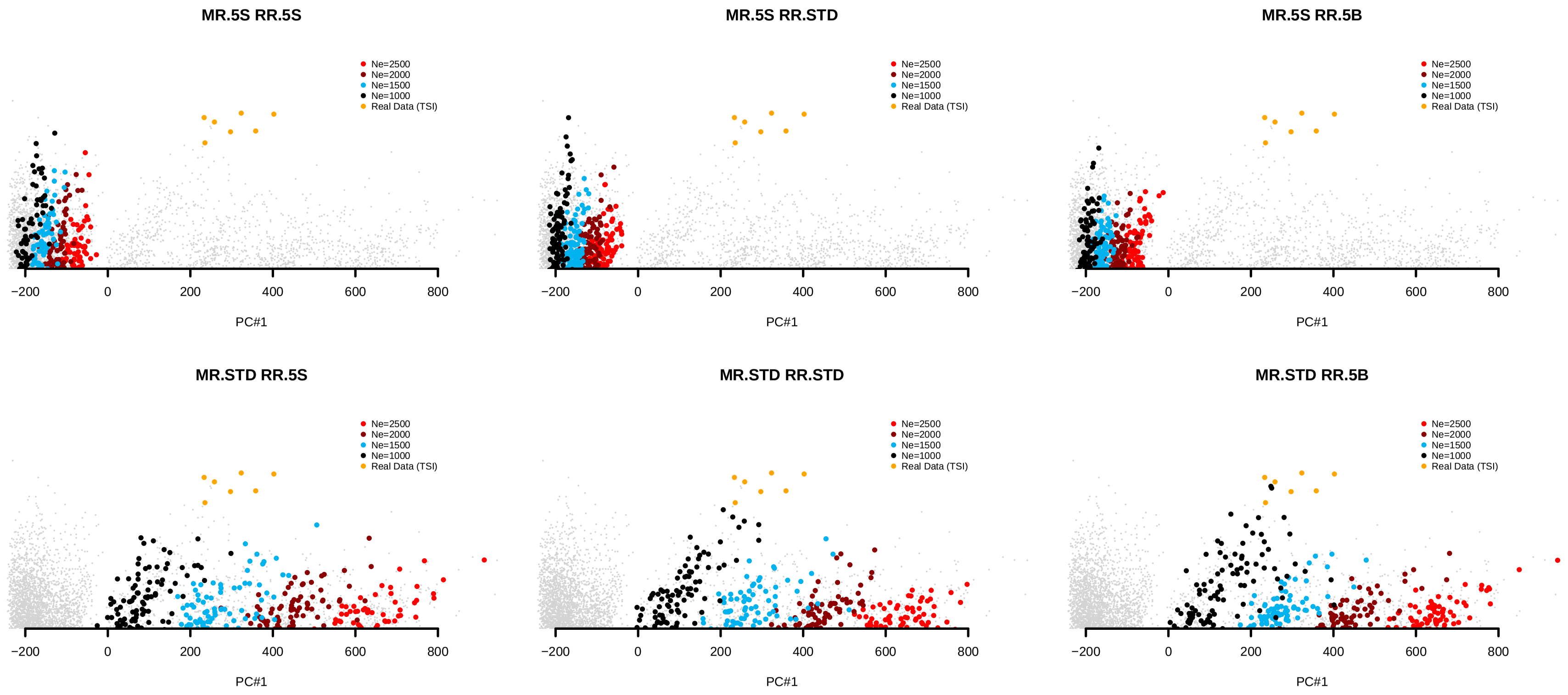


Figure 1: **Genetic patterns of Simulated data and Real data** - Results from a PCA using summary statistics of simulation generated data and real sequenced-derived data. Each plot represents a distinct model with variable mutation and recombination rates. Within each plot, the highlighted dots represent the obtained patterns for the different effective population sizes (Ne). Grey dots in the background correspond to the values obtained for the whole simulated set. Orange dots represent the values obtained for the TSI population.

Abbreviations: MR - Mutation Rate; RR - Recombination Rate; 5S - the used paramenter value was five times smaller than standard estimates; 5B - the used paramenter value was five times bigger than standard estimates; STD - the used paramenter value was equal to the standard estimates.

Appendix C:
Publications in Peer-Reviewed journals