Bioinformatic characterization and analysis of polymorphic inversions in the human genome

Author: Alexander Martínez Fundichely

DOCTORAL THESIS UPF / YEAR 2013

Supervisor: Dr. Mario Cáceres Aguilar

Tutor: Dr. Xavier Estivill Pallejà

Department of Experimental and Health Sciences

upf. Universitat Pompeu Fabra Barcelona

A thesis submitted for the degree of PhilosophiæDoctor (PhD) in Biomedicine. In the department of Experimental and Health Sciences of the Universitat Pompeu Fabra (UPF). By the doctoral candidate

Msc. Alexander Martínez Fundichely

Supervisor: Principal investigator of the Institut de Biotecnologia i de Biomedicina (IBB), Universitat Autònoma de Barcelona (UAB), Bellaterra, Barcelona, Spain. And member of the Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

Dr. Mario Cáceres Aguilar

Tutor: Senior investigator at the Center for Genomic Regulation (CRG), Barcelona, Catalonia in Spain. And associate professor of Pompeu Fabra University (UPF)

Dr. Xavier Estivill Pallejà

Day of the defense: 12/12/2013

To my Mom
To my maternal grandma
Especially to my missing grandparents
and
From the bottom of my heart, I wish share this achievement with my Dad (without you physically present, everything is harder than that we talked before, but step by step I'm keeping on the play).

ACKNOWLEDGEMENTS

First and foremost I offer my sincerest and utmost gratitude to my senior supervisor, Dr. Mario Cáceres, who has supported me throughout my graduate studies with his encouragement, guidance and resourcefulness. In addition to his experience in genomics that was a crucial factor in the success of this thesis, I have learned the requirements of being a good researcher from him. I thank him one more time for all his help.

I would like to give my regards and appreciations to Dr. Xavier Estivill, this whole story began with his acceptance to my visit to his lab. Starting from the first beginning, I would also like to thank my other collaborators in the CRG but especially thanks to the people in the lab of IBB, a real nice people who offered me friendship and collaborations and all of them without exception have been important in the development of this thesis. Their motivation was a huge driving force for me.

Especially thanks to my new family for the love, affection and support they gave me in all these years, that helps me to overcome the sadness of being far away from my loved ones and feel like one more catalan.

I would also like to thank all my friends every one are mentioned here although not seen the name list,(Cuban, Catalan, Spanish...) their help has been very important in this achievement.

Last, but not least, I thank to my mom and grandma, their support even at the distance has been essential to reach this goal, and thank to all the other Cuban family who are interested in me, and always helps my mother when she needs.

Alexander Martinez Fundichely

ABSTRACT

Within the great interest in the characterization of genomic structural variants (SVs) in the human genome, inversions present unique challenges and have been little studied. This thesis has developed GRIAL, a new algorithm focused specifically in detect and map accurately inversions from paired-end mapping (PEM) data, which is the most widely used method to detect SVs. GRIAL is based on geometrical rules to cluster, merge and refine both breakpoints of putative inversions. That way, we have been able to predict hundreds of inversions in the human genome. In addition, thanks to the different GRIAL quality scores, we have been able to identify spurious PEM-patterns and their causes, and discard a big fraction of the predicted inversions as false positives. Furthermore, we have created InvFEST, the first database of human polymorphic inversions, which represents the most reliable catalogue of inversions and integrates all the associated information from multiple sources. Currently, InvFEST combines information from 34 different studies and contains 1092 candidate inversions, which are categorized based on internal scores and manual curation. Finally, the analysis of all the data generated has provided information on the genomic patterns of inversions, contributing decisively to the understanding of the map of human polymorphic inversions.

RESUMEN

Dentro del estudio de las variantes estructurales en el genoma humano, las inversiones presentan retos específicos y han sido poco caracterizadas. Esta tesis aborda este problema a través de la implementación de GRIAL un nuevo algoritmo específicamente diseñado para detectar y localizar de forma precisa las inversiones a partir de datos de mapeo de secuencias apareadas PEM[1], que es el método más utilizado para estudiar la variación estructural. GRIAL se basa en reglas geométricas para agrupar los patrones de PEM correspondientes a los posibles puntos de rotura y refinar su localización para cada inversión. Los resultados de GRIAL nos permitieron predecir cientos de inversiones en el genoma humano. Además, gracias a la creación de índices de fiabilidad para las predicciones, se ha podido identificar patrones de inversión incorrectos y sus causas, descartando un gran número de predicciones posiblemente falsas. Por otra parte, se ha creado InvFEST, la primera base de datos dedicada a inversiones polimórficas en el genoma humano, la cual representa el catálogo más fiable de inversiones e integra toda la información asociada disponible de múltiples fuentes. Actualmente, InvFEST combina información de 34 estudios diferentes e incluye 1092 inversiones clasificadas según criterios internos y anotación manual. Por último el análisis de toda la información generada, nos ha permitido describir los patrones genómicos de las inversiones contribuyendo decisivamente a descifrar el mapa de las inversiones polimórficas humanas.

[1] del inglés paired-end mapping (PEM)

PREFACE

A little more than a decade pass from the completion of the Human Genome Project, the efficiency of DNA sequencing has drastically improved. The high throughput next generation sequencing (NGS) technologies have been reducing the cost and are increasing the capacity for sequence production of thousand of individuals at an unprecedented rate on within different international project [1000 Genomes Project *et al.*, 2012; International Cancer Genome *et al.*, 2010]. These newest sequencing technologies adding to traditional Sanger sequencing have significantly changed how genomic research is conducted, and have provided new insights on the genetic basis of phenotypic and disease-susceptibility differences between individuals through the uncovered an unprecedented degree of structural variation in the human genome.

Although the Database of Genomic Variants (DGV) is showing the great advances on identifying of several types of variations among individual genomes. The genomic inversions have been relatively disregarded compared to copy number variations (CNVs) due to their difficulty of study that makes the analysis of these variants on the sequenced genomes very challenging mainly due to the complex, and repetitive nature of human genomes that in particular become much harder when dealing with balanced rearrangements.

This thesis starts in chapter 1 with a general introduction on the study of structural variation with particular emphasizing in the inversions. This chapter exposed the statement of the scientific problem addressed and the main objective of the thesis project. The results have three integral parts that have a strong connection to each other from a methodological point of view, each part is addressing different levels of study of the inversions and are represented in the three manuscript of papers in chapters of results (2, 3, 4).

The first result in chapters 2 is focused on genome sequence analysis, in particular development of computational methods for structural variation discovery in sequenced genomes. We present our effort in developing novel paired-end mapping algorithm for identifying specifically inversions, adding also two scores to assess the reliability of the predictions, as well as a new dataset of inversion predictions discovered in multiple sample genomes. In this part, we also address some problems of paired-end sequencing experiment, based on fosmid clones. The second result of this thesis in chapters 3, is focused on develop a data base for a comprehensive integration of all information about the inversions in the human genome. In this case we create an alternative data source centered on human inversions studies, to store the predicted inversions and their accurate break points, validation status or population distribution among other data. The last third result of this thesis in the chapters 4 is, however, focused on the descriptive analysis of genomic patterns of the current information about human inversion polymorphisms. We discuss several features of the most reliable catalogue of inversion that represent a preliminary characterization of this variant in the human genome that is paramount to better understanding of the possible biases and trends in the detection of inversions in human genome.

The chapter 5 is a general discussion that presents in a integrated global result the three topic developed in the thesis project and their contributions. The thesis concludes with the chapter 6 that are a general conclusion of the thesis. The different results of this work have been presented at the RECOMB/ISCB 2012 and ISMB/ECCB 2012 and 2013 conferences. The chapter 3 is already published in Nucleic Acids Research (NAR) 2013. Finally in the appendix part are also presented two scientific articles in which the developing of this thesis has collaborated.

CONTENTS

Appendixes

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

aCGH	Microarray comparative genomic hybridization
alt-EJ	Alternative end-joining
BP	Breakpoint
CNP	Copy number polymorphism
CNV	Copy number variant
DECIPHER	Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources
DGV	Database of Genomic Variants
DGVa	Database of Genomic Variants archive
DMSO	Dimethyl sulfoxide; organic solvent, readily passes through skin, cryoprotectant in cell culture
DSB	Double-strand break
DSBR	Double strand break repair
ENCODE	The Encyclopedia of DNA Elements Project
eQTL	Expression quantitative trait loci
FISH	Fluorescence in situ hybridization
FoSTeS	Fork stalling template switching and microhomology mediate break induced replication
GRC	Genome Reference Consortium
GWAS	Genome-wide association studies
HGMD®	Human Gene Mutation Database
HGP	The Human Genome Project
HTS	High-throughput sequencing
indel	Short insertions and deletion
ISV	Intermediate-sized structural variation
LCR	Low copy repeat
LD	Linkage disequilibrium
LSV	Large-scale structural variation
MEPS	Minimal efficient processing segments
MMBIR	Microhomology-mediated break-induced replication
MMEJ	Microhomology end-joining and alternative end-joining
MMEJ	Microhomology-mediated end-joining
NAHR	Non-allelic allelic homologous recombination
NGS	Next generation sequencing
NH	Non-homology
NHEJ	Non-homologous end-joining
PEM	Paired-end mapping
PR	Paired Read
RD	Read depth

SD	Segmental duplication
SDSA	Synthesis dependent strand annealing
SNP	Single nucleotide polymorphism
SP	Split Read
STR	Short tandem repeat
SV	Structural variant

CHAPTER 1

GENERAL INTRODUCTION

1.1 The study of the human genome

One decade since the completion of "The Human Genome Project" (HGP) [Venter *et al.*, 2001; Lander *et al.*, 2001], biological sciences face at present with great impetus their main challenges, the deciphering of genome functions and understanding the complex way in which the genome sequences are translated into a big variety of phenotypic characteristics of individuals. Furthermore, the biomedical interest has intensified the thorough investigations of individual genome variation.

A wide variety of large-scale projects have been already launched to investigate the human genome from diverse perspectives and are focused on different aspects. One of the main targets was to find and annotate all functional elements in the human genome. With this goal, a public research consortium was created which launched "The Encyclopedia of DNA Elements Project" (ENCODE). The release of the initial results of this project has provided a complete map of the identification and detailed annotation of a wide variety of functional elements in the human genome. The analysis began from a little percentage (1%) of the human genome sequence [ENCODE Project *et al.*, 2007], but it has scaled up to the study of the entire genome [ENCODE Project *et al.*, 2012].

This knowledge is very important for the study of gene functionality, the complexity involved in the regulation of gene expression levels [Myers *et al.*, 2007] as well as in the disease association studies [Estivill and Armengol, 2007] that will certainly enable us to discover potential drug targets and to develop personalized medicine in the future.

Another important scientific aim after the completion of the human genome is the understanding of the nature and patterns of variation within the human species, including both common and rare variants, and its use as markers in linkage and association analysis. Initially, the focus of variation discovery was targeted on single nucleotide polymorphisms (SNPs), which are changes in one base between sequences. "The HapMap Project" [International HapMap Project, 2003] was launched with the goal of developing high-density SNP genotyping technology to provide the scientific and medical community ample information about common SNPs and identify haplotype blocks for the analysis of human variation and their potential associations with human complex traits and diseases [International HapMap I Project, 2005; Hinds *et al.*, 2005]. From the estimated 15 million places along our genomes where one base can differ from one person or population to another, around three million (3.1×10^6) such locations have already been validated and characterized as SNPs in the second phase of the project [International HapMap II Project *et al.*, 2007]. In addition, they have been charted using genotyping assays in 270 individuals from 4 geographically diverse human populations [International HapMap II Project *et al.*, 2007]. The project has continued evolving by extending the reference panel on 7 additional populations, and in the third phase 1.6 million common SNPs were genotyped in 1184 reference individuals from a total of 11 global populations, and ten regions of 100 kb were sequenced in 692 of these individuals [International HapMap III Project *et al.*, 2010]. This resulted in the characterization of population-specific differences among low-frequency variants, and the improvement of imputation accuracy, especially for variants with a minor allele frequency ($\leq 5\%$).

The convergence of new technologies that can genotype hundreds of thousands of SNPs markers, together with comprehensive annotation of genetic variation and functional elements, has contributed to "the genome-wide association studies" (GWAS).

This has generated several very active lines of research, such as the "expression quantitative trait loci" (eQTL) mapping studies, that have become a widely used tool for identifying DNA sequence variations that cause changes in regulation of gene expression, which in turn could have profound effects on cellular states [Nica and Dermitzakis, 2013; Ackermann *et al.*, 2013]. In these studies, expression levels are viewed as quantitative traits, and gene expression phenotypes are mapped to particular genomic loci by combining data of gene-expression variation patterns with genome-wide genotyping [Gilad *et al.*, 2008]. Results from recent eQTL mapping studies have revealed substantial heritable variation in gene expression within and between populations. These variations could affect tissue development and may ultimately lead to pathological phenotypes [Zhong *et al.*, 2010; Bossé, 2013].

The large efforts on the application of GWAS for the analysis of genome function, especially in the context of studies of genome variation has also allowed the discovery of regions of the genome that harbor genetic variants that confer risk to different types of complex diseases [Kingsmore *et al.*, 2008]. The GWAS provide encouraging successes in research on several types of cancer disease [Chen *et al.*, 2013; Chung *et al.*, 2010] as well as coronary heart disease and diabetes disease [Qi *et al.*, 2013] and also neurodegenerative disorders such as alzheimer disease and parkinson disease [Chung *et al.*, 2013].

In more recent years, the improvement of molecular analysis techniques have raised to a new level comparative genomic assays. In particular, the great advances in sequencing technology, referred to as next-generation high-throughput sequencing has spurred the race to sequence genomes for individuals and tissues as well. In the last years, many genome sequences of new individuals have been published [Levy *et al.*, 2007; Bentley *et al.*, 2008; Wang *et al.*, 2008; Wheeler *et al.*, 2008; Fujimoto *et al.*, 2010; Lilleoja *et al.*, 2012; Gupta *et al.*, 2012; Shen *et al.*, 2013; Azim *et al.*, 2013]. Additionally, there are active projects such as "The 1000 Genomes Project" for sequencing the genome of many more individuals, which recently made the announcement of the official release of the phase3 [1000 Genomes Project *et al.*, 2010]. The current outcomes of this project, it has been described the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. This resource captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, providing a validated haplotype map of 38 million SNPs, but also 1.4 million short insertions and deletions (indels) or copy number polymorphism (CNP), and more than 14,000 structural variants. This data enables analysis of common and low-frequency variants in individuals from diverse populations, showing, by characterizing the geographic and functional spectrum of human genetic variation, that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which may be further increased by the action of purifying selection. [1000 Genomes Project *et al.*, 2012]. Moreover, at present there are many more available data from exome sequencing, and also several other projects focused on sequencing the genomes of different types of cancer tissues and complex diseases[International Cancer Genome *et al.*, 2010].

Although the International Human Genome Reference Consortium provides us a powerful tool with the human reference sequence. Still around ~240 Mb (8%) of sequence is missing in the reference assembly (GRCh37 stats) [International Human Genome Sequencing, 2004]. The gaps in sequences can be hiding the detection of new structural variations. In addition, around ~5% of human genome is covered by segmental duplications (SDs), which are regions of interest for the study of SV but are hard to analyze [Bailey *et al.*, 2001; Marques-Bonet *et al.*, 2009; Uddin *et al.*, 2011].

This summary illustrates the extraordinary amount of information that is being accumulated, and the use of this information is one of the present main challenges of scientific community to address interesting questions related to human evolution, variation and disease.

1.2 Genomic structural variation

During the last years only, on the trail of two groundbreaking studies [Sebat *et al.*, 2004; Iafrate *et al.*, 2004], the discovery of an unexpected abundance of submicroscopic structural changes has expanded the paradigm of the variation in the human genome (see Table 1.1). These findings lead researchers to predict that genomic "structural variants" (SVs) are as important as SNPs, short tandem repeats (STRs) and other small changes in their contribution to genome variation. Furthermore, SVs may affect a wide range of the genome, containing even entire genes and their regulatory regions. As with SNPs, probably most of these types of variant are neutral and in many genomic regions have no obvious phenotypic consequence on the individuals that carry them. However, this other level of variation has resulted in a great interest in the study of structural variation.

Table 1.1: Genetic variation in the human genome.

Genomic variation	Description	Size rage[1]
Changes in a single base-pair	Single nucleotide polymorphisms, point mutations	1 bp
Small InDel	Insertion/Deletion events of short sequences usually <10 bp in size	$1-50$ bp
Short tandem repeats	Microsatellites, Microsatellite and other simple repeats	$\sim$$10-500$ bp
Fine-scale variation	Deletions, duplications, tandem repeats, inversions	50 bp to 1 kb
Structural variation	CNV, inversions, translocations	$\sim$1 kb to several Mb
Chromosomal variation	Euchromatic variants, huge[2] deletions, duplications, translocations, inversions, and aneuploidy	>5 Mb to entire chromosome

[1] These size ranges quoted are indicative only of the scale, not an strictly definition.

[2] Cytogenetically visible.

The general definition of SVs is a change in the DNA sequence of a region of genome ranging in size approximately between 500 bp and 5 Mb in size, usually greater than 1 kb for operative purposes [Sharp *et al.*, 2006]. These changes could be balanced or unbalanced genomic rearrangements, depending if there is gain or loss of DNA, such as in deletions, duplications, and insertions, three types which are usually referred to as copy number variants (CNVs), or not, such as those that involve a change in orientation, referred to as genomic inversions, and translocations, which represent the transference of DNA sequence to a nonhomologous region of another chromosome (reciprocal or non-reciprocal) (see Table 1.2).

Table 1.2: Description of genomic structural variation.

Structural Variation type	**Definition**
Copy-number variant (CNV) or Copy-number polymorphism (CNP)	A change that involves a segment of DNA ≥ 1 kb which is present at a variable copy number in comparison with a reference genome. If a CNV occurs in more than 1% of the population is referred as CNP. Classes of CNVs include insertions, deletions and duplications. Segmental duplications or low-copy repeats can therefore also be CNVs that occurs in two or more copies per haploid genome, with the different copies sharing $>90\%$ sequence identity
Translocation	A change in position of a chromosomal segment within a genome that involves no change to the total DNA content. Translocations can be intra- or inter- chromosomal
Inversion	A segment of DNA that is reversed in orientation with respect to the rest of the chromosome. Pericentric inversions include the centromere, whereas paracentric inversions do not.

The early knowledge about the chromosome structural organization was mainly based on observation at the microscopy of rare changes in the quantity and structure of chromosomes. These included aneuploidies [Edwards *et al.*, 1960; Smith *et al.*, 1961], chromosome aberrations [Bobrow *et al.*, 1971; Jacobs *et al.*, 1978], other heteromorphisms [Maegenis *et al.*, 1978; Verma *et al.*, 1978] and chromosome fragile sites [Lubs, 1969]. However, most of these changes are often associated with syndromes. Since them, the great advances in genomic approaches and DNA sequencing techniques have allowed the discovery of an increasing number of submicroscopic changes in the DNA [Feuk *et al.*, 2006] affecting between few base pairs, including variable numbers of short repetitive elements such as microsatellites and minisatellites and small indels (insertion-deletion) polymorphisms [Mills *et al.*, 2006], to several kilobases or megabases, including hundreds of large-scale copy number variations (CNVs) [Alkan *et al.*, 2011], inversions, translocations [Tuzun *et al.*, 2005; Feuk *et al.*, 2005; Kidd *et al.*, 2008; Korbel *et al.*, 2007; Levy *et al.*, 2007], and more complex rearrangements like those generated during chromotripsis [Liu *et al.*, 2012]. Most of them involve segments that are smaller than those recognized microscopically [Iafrate *et al.*, 2004; Sebat *et al.*, 2004], allowing the analysis of many SVs not studied before [Haraksingh and Snyder, 2013].

The characterization of genomic structural variation has highlighted the complexity of human genetic variations [Pennisi, 2007b], and has provided significant insight into the evolution of genomes and their dynamic and flexible nature. Therefore, the study of SV has opened a very productive line of research in the last years.

1.3 Copy number variation (CNV)

The most widespread type of structural variation detected are copy number variants (CNVs), also referred to as copy number polymorphisms (CNP) [Sharp *et al.*, 2005; Jakobsson *et al.*, 2008], which are changes that result in losses or gains of DNA segments. This type of SV has gathered most of the scientific interest [Redon *et al.*, 2006; Sebat *et al.*, 2004] and the results have put CNVs as the most frequent type of structural variant in the human genome. It is believed that somewhere around $5-25\%$of the human genome is copy number variable between individuals and CNVs represent 99% of all structural variation reported from 55 studies in the "Database of Genomic Variants" (DGV), at Aug 2013.

This was to a great extent possible thanks to the development of a new method based on array strategies for comparative genomic studies, the microarray comparative genomic hybridization (aCGH) technique [de Ravel *et al.*, 2007; Theisen, 2008]. This technique was the main approach for identifying unbalanced changes [1] whose application made possible to look for variation at the genome in a higher scale and with a resolution not seen before, even at the submicroscopic level. Therefore, the majority of the initial CNV studies relied on using (aCGH) [Dhami *et al.*, 2005]. Later, other techniques based on next-generation sequencing have enriched the discovery of the abundance of CNVs currently known [Alkan *et al.*, 2011].

[1] Net gain or loss of large segments of DNA

The unexpected abundance of CNVs has resulted in a great interest in their study to identify which of these variants have functional or evolutionary effects in the human genome. CNVs may be potentially related with changes in gene dosage, which might cause genetic disease, either alone or in combination with other genetic or environmental factors [McCarroll and Altshuler, 2007]. One of the simplest models for the functional impact of CNVs is the change in the levels of expression of genes within or surrounding the affected genomic region. An intuitive model suggests that an increase in the copy number of a specific gene will, on average, lead to a corresponding increase in the expression level of that gene, and vice versa. Moreover, it is likely that deletions or insertions might lead to a variety of effects that are more complex than the gene dosage level of expression expectation. Some studies have identified examples of CNVs that had a significant impact on the gene expression variation [Haraksingh and Snyder, 2013] either from a population perspective [Stranger *et al.*, 2007; Conrad *et al.*, 2010] or in a disease context [Aitman *et al.*, 2006; Li *et al.*, 2012].

According to the "Human Gene Mutation Database" (HGMD®), around 7% of all mutations associated with gene lesions responsible for human inherited disease are currently attributed to insertions or deletions. Moreover, several studies have found an association of CNVs with susceptibility to several human diseases, such as HIV-1/AIDS [Gonzalez *et al.*, 2005b], psoriasis [Eva *et al.*, 2011], systemic autoimmune diseases [Aitman *et al.*, 2006] and other complex diseases like mental retardation, Parkinsonism, Alzheimer or Schizophrenia [Estivill and Armengol, 2007]. There is also a CNV associated to functional differences on the amylase gene, with a big potential evolutionary impact [Perry *et al.*, 2007]. Otherwise a polymorphic CNV of the *CYP2D6* gene has been found associated with metabolic alteration in the activity of the cytochrome *P450 CYP2D6* drug-metabolizing enzyme, which is associated too with increased risk factors for laryngyal and lung cancers [Agùndez *et al.*, 2001]. Therefore, identify which of these structural variants do have functional consequences and which is their role in human evolution, diseases and phenotypic variation is a very important challenge at present.

1.4 Genomic inversions

However, despite of the great success in developing human genome maps of SVs, and the significant advances identifying which of these variants do have functional consequences, in the case of inversions similar detailed analysis are very limited and little has been revealed about the functional and evolutionary impact of inversions in the human genome.

Interestingly, genomic inversions were the first type of structural variant studied. They were discovered eight decades ago in heterozygous polytene chromosomes (the oversized huge chromosomes which are commonly formed in the salivary glands of the larval state of *Drosophila* flies). Since their discovery, the *Diptera* insect order remains the group in which large inversions can be most easily detected, thanks to these special chromosomes. This led to the discovery of an extraordinary rich inversion polymorphism in *Drosophila* species and opened a productive and interesting area of research on different experimental and theoretical aspects of inversion biology [Krimbas and Powell, 1992].

Unlike other types of structural variation, an inversion is theoretically presumed to be a balanced rearrangement that just change the orientation of a DNA segment and is not associated with either the gain or loss of genetic information. They tend to occur as a result of two simultaneous chromosome double strand breaks and the subsequent reorientation of the central fragment before the repair (rejoining) of the broken ends (see Figure 1.1), or an abnormal recombination process in which a segment of a chromosome is reversed end to end. However, in some cases during the rearrangement, the inversion also involves gain or loss of DNA material either at, or close to, the breakpoints, indicating that inversions are not always balanced events [Sharp *et al.*, 2006]. This lack or not of changes in the DNA content has important consequences in the methods used to detect inversions, and makes inversions much more difficult to study than the clear imbalanced changes like CNVs.

Inversions fall into two main different types: "**Pericentric inversions**" include the centromere and there is a breakpoint located in each separated chromosome arm (Figure 1.1 Left). "**Paracentric inversions**" do not include the centromere and both

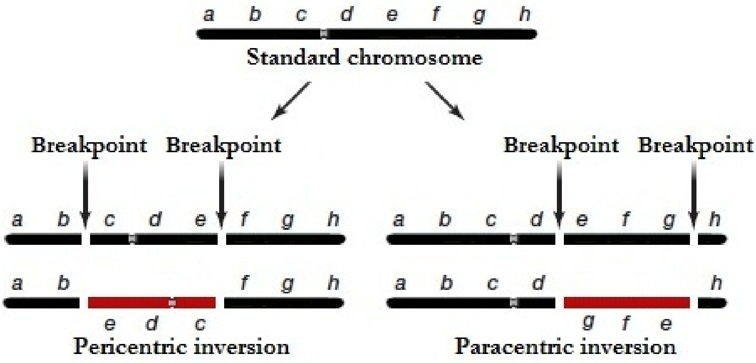


Figure 1.1: Types of genomic inversions - Schematic representation of the two different types of chromosomal inversions. Pericentric inversions (Left), if the centromere is located in between the two breakpoints. Paracentric inversions (Right), if the centromere is located outside of both breakpoints. The inverted region is highlighted in red

breakpoints are located in the same arm of the chromosome (Figure 1.1 Right). This classification has important consequences in the effect of recombination events inside the inverted region in heterozygotes. When synapsis occurs in inversion heterozygotes or heterokaryotypes (individuals with an inverted chromosome and a wild-type homolog), an inversion loop often forms to accommodate the point-for-point pairing along the chromosomes during meiosis (see section 1.4.2.4, inversion effects). A pericentric inversion will have the centromere located within the inversion loop, thus as a consequence of a crossover in the loop region, the recombination event yields two unbalanced recombinant chromatids, one with a duplication and other with a deletion. Conversely, a paracentric inversion will have the centromere located outside of the inversion loop. Thus as a consequence of a crossover in the loop region and the recombination event, two recombinant chromatids are also produced, but in this case one is dicentric and the other acentric.

There are another two major categories that inversions can be classified on the basis of the way that the inversion has been generated, and the event's evolutionary history (see Figure 1.2). If an inversion arises due to an stochastic process and this unique event continue segregating on the population, it is classified as a **monophyletic inversion.** On the detection of this type of inversions, the breakpoints sequences are almost

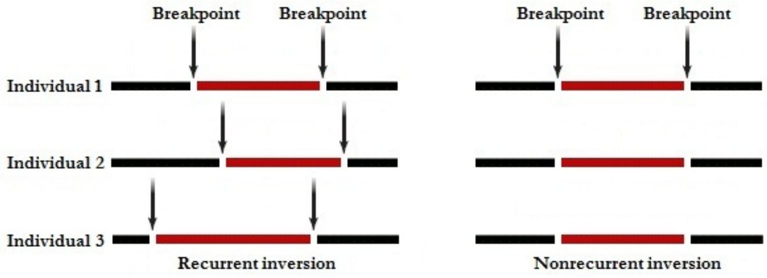


Figure 1.2: Schematic representation of recurrent and nonrecurrent genomic inversions - (Left) Polyphyletic inversion or recurrent inversion, are shown in different individuals having breakpoints relatively scattered around a hotspot, indicating that the inversion occurs in multiple unrelated individuals with same or slightly different breakpoints, and a common rearrangement interval size. (Right) The breakpoint of the nonrecurrent inversion are found in the same place and same sequence characteristics for all individuals, showing that inversion has occurred only one time on the evolutionary history, which means a monophyletic inversion.

identical for all individuals carrying the inverted allele (Figure 1.2 Right). If the inversion originates as a result of recurrent biological processes like recombination, and several occurrences of the inversion events are segregating together in the population, it is classified as a **poliphyletic inversion**. In this type of inversions it is possible to find differences between the respective breakpoints (Figure 1.2 Left) when several individuals carrying the inversion are analyzed. Traditionally, it has been considered that inversions found in natural population are monophyletic. Therefore, inversions have been used extensively for building phylogenies.

1.4.1 Origin of genomic inversions

Inversions can be generated by several molecular mechanisms shared with other SVs[1](see Table 1.3). Those mechanisms can be categorized into two main groups. First, there are those involving extensive stretches of high identity homologous sequence at the breakpoint junctions. Second, there are those occurring in absence of homology at the breakpoint junctions. Such mechanisms of formation may be mediated through DNA repair processes, replication, or recombination. In addition, inversion formation can

[1] For all abbreviations see the glossary on page xxii.

also be based on the dynamic process of transposable element movilization [Onishi-Seebacher and Korbel, 2011; Gu *et al.*, 2008].

Table 1.3: Different mechanisms involved in inversion formation.

Mechanism	**Description**	**Features**
NAHR[1]	Homologous recombination between non-allelic positions	Involves extensive DNA sequence homology
NHEJ[2]	Rejoins of double strand breaks without homology or with microhomology	No homology or microhomology, small sequence insertion possible
MMEJ/alt-EJ[3]	End-joining process that occurs with microhomoloy	Microhomology, small sequence insertion possible
FoSTeS/MMBIR[4]	Reestablishment of replication at collapsed or stalled replication forks using microhomology	Sequence insertion at junctions, generation of complex structural variants

[1] Non-allelic allelic homologous recombination.

[2] Non-homologous end-joining

[3] Microhomology end-joining and alternative end-joining

[4] Fork stalling template switching and microhomology mediate break induced replication

1.4.1.1 Homologous recombination mechanisms

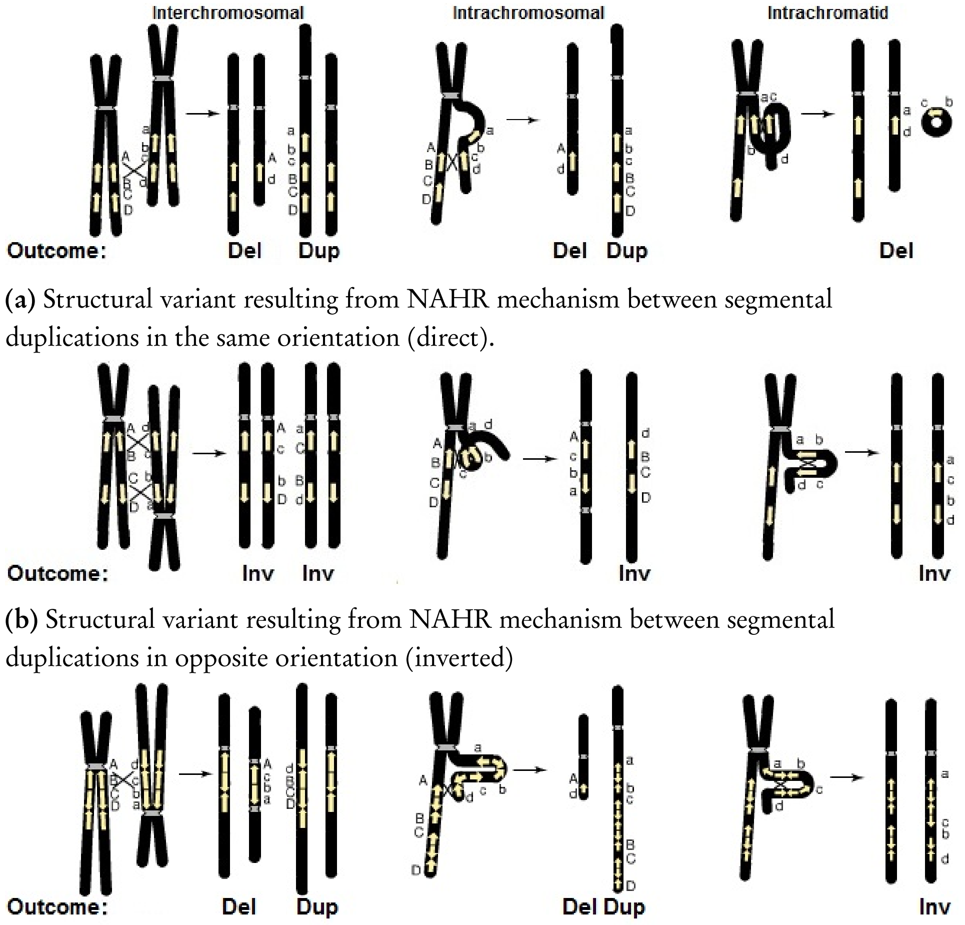
It is estimated that as much as $5-10\%$ of the human genome might be duplicated [Emanuel and Shaikh, 2001; Uddin *et al.*, 2011; International Human Genome Sequencing, 2004], including segmental duplications (SDs), also called low copy repeats (LCRs), and other types of repetitive sequences with a high sequence identity These sequences can provide significant lengths of sequence similarity as substrates for misalignment between alleles and mediate crossing-over between mismatched homologous regions, inter or intra chromosomally, also called as non-allelic homologous recombination (NAHR) [Gu *et al.*, 2008; Stankiewicz and Lupski, 2002; Lupski and Stankiewicz, 2005; Liu *et al.*, 2012] (see Figure 1.3). Substrates for homologous recombination appear to depend on genome architecture features. These features include repeat size, degree of homology (usually greater than $\sim 95\%$), distance between the sequences, and orientation with respect to each other [Stankiewicz and Lupski, 2002; Liu *et al.*, 2012].

If the NAHR process is between same orientation (direct) SDs, it provokes reciprocal deletions plus duplications of genomic segments (Figure 1.3a), whereas NAHR between oppositely oriented or inverted SDs causes an inversion of the genomic loci in between (Figure 1.3b). When genomic architecture has a complex SD structure, consisting of both direct and inverted subunits, they can serve as NAHR substrates leading to complex, genomic deletions/duplications and inversions (Figure 1.3c).

Some genome sequence elements that have been associated with double-strand breaks (DSB), such as minisatellites, transposons or palindromic segments, have often been found near regions of evidence of NAHR. This suggests a potential link between NAHR and the double strand breaks repair (DSBR) and synthesis dependent strand annealing (SDSA) pathways to repair double-strand breaks in DNA using recombination based methods [Gu *et al.*, 2008].

The crossing-over between strands during NAHR is located into a restricted group of narrow hotspots within the SDs and it is typically never evenly distributed along the SDs [López-Correa *et al.*, 2001; Bi *et al.*, 2003]. Another important feature of the NAHR process is that the hotspots must be minimal efficient processing segments (MEPS). This means that there must be segments of specific minimal length sharing extremely high similarity or identity between the SDs for NAHR to occur. The common limit of the MEPS length is over 100 bp, but there are known NAHR events mediated by matching fragments smaller than 50 bp [Gu *et al.*, 2008]. MEPS differences become important in case of interchromosomal or intrachromosomal rearrangements [Gu *et al.*, 2008]. The proximity between two SDs is one of the genomic architecture features with most influence in the MEPS and the efficiency of NAHR. Thus, bigger sized genomic rearrangements generated by SDs located further apart, often correlate with larger SDs [Lupski, 1998; Stankiewicz and Lupski, 2002].

On the other hand, NAHR occurs both during meiosis and mitosis. On meiosis, the NAHR process leads to constitutional genomic rearrangements in germ line cells, which can be either inherited if they continue to segregate across generations or sporadic if they always occur de novo. In humans, the demand of MEPS on meiosis appears



(a) Structural variant resulting from NAHR mechanism between segmental duplications in the same orientation (direct).

(b) Structural variant resulting from NAHR mechanism between segmental duplications in opposite orientation (inverted)

(c) Structural variant resulting from NAHR mechanism between complex distribution of segmental duplications

Figure 1.3: SVs mechanisms mediated by homology recombination - Schematic representation of genomic SV formation based on non-allelic homologous recombination (NAHR) mechanisms between segmental duplications. The yellow arrows depict segmental duplication within black chromosomes. The SVs products of recombination are shown according to orientation and structure of the SDs, and the figures depicts rearrangement separated by types of recombination (interchromosomal, intrachromosomal, and intrachromatid):(a) deletions and duplications resulting from NAHR mechanism between segmental duplications in the same orientation (direct), in the case of intrachromatid recombination can result in deletion and an acentric fragment.(b) Inversions resulting from NAHR mechanism between segmental duplications in opposite orientation (inverted). (c) Examples of deletions, duplications and inversions resulting from NAHR mechanism between complex distribution of segmental duplications. (Adapted from [Stankiewicz and Lupski, 2002]).

to require a minimum range of $\sim 300-500$ bp in length of uninterrupted homology [Stankiewicz and Lupski, 2002]. On mitosis, the NAHR process leads to mosaic populations of somatic cells carrying abnormal genomic rearrangements. MEPS requirements on mitotic process may be slightly lower $\sim 200-300$ bp than meiotic NAHR [Stankiewicz and Lupski, 2002].

Finally, mechanisms mediated by sequence homology are frequently associated with recurrent rearrangements in human diseases. Specifically, for inversions NAHR is the molecular mechanism that has been shown to be responsible for the vast majority of the recurrent rearrangements[Shaw and Lupski, 2004].

1.4.1.2 Non-homologous and microhomology mechanisms

By contrast, non-homology (NH) based mechanism are thought to use either non-homologous DNA sequences or very short homologous sequences (less than ~ 10 bp) also known as microhomology. This category includes non-homologous end-joining (NHEJ), as well as complementary pathways, such as alternative end-joining (alt-EJ) and microhomology-mediated end-joining (MMEJ) [Onishi-Seebacher and Korbel, 2011]. These mechanisms have been reported as another molecular mechanism involved on repair of DNA double strand breaks (DSB). Both, natural DSB, such as in somatic recombination, and accidental DSB, such as those caused by ionizing radiation or by free radicals, could be responsible of nonrecurrent genomic rearrangements [Lieber *et al.*, 2003; Gu *et al.*, 2008].

The process of NHEJ occurs in four steps (see Figure 1.4). Once the DSB is detected, both broken DNA ends are bridged together, followed by the modification of the ends to make them compatible for the final ligation step [Weterings and van Gent, 2004]. For the process it has been described a 'canonical' pathway, also referred as classical pathway, that utilizes the DNA-PK (which includes the *Ku70/Ku80* heterodimer and DNA-PKcs) and DNA-ligase *IV/XRCC4/XLF* complexes and Artemis. However, it is recognized that the end-joining can occur in the absence of canonical pathway repair factors, such as DNA-ligase IV and *Ku70/Ku80*, and then the mechanism is referred as Alt-NHEJ [Sankaranarayanan *et al.*, 2013; Bennardo *et al.*, 2008]. Another mechanism for DSB repair is the microhomology-mediated end-joining (MMEJ) (see

Figure 1.4). The most important and distinguishing property of MMEJ is the requirement and use of $5-25$ bp microhomology during the alignment of broken ends before joining, resulting in deletions flanking the original break [McVey and Lee, 2008; Sankaranarayanan *et al.*, 2013].

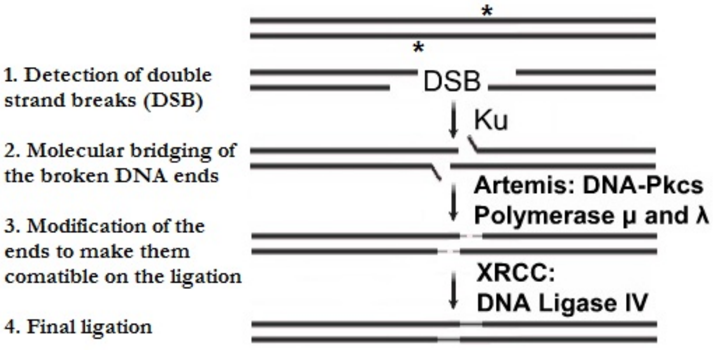

1. Detection of double strand breaks (DSB)
2. Molecular bridging of the broken DNA ends
3. Modification of the ends to make them comatible on the ligation
4. Final ligation
*
*
DSB
Ku
Artemis: DNA-Pkcs
Polymerase μ and λ
XRCC:
DNA Ligase IV

Figure 1.4: NHEJ mechanisms - Schematic representation of different steps of NHEJ-based mechanisms for generation of genomic rearrangements. (Adapted from [Gu *et al.*, 2008]).

NHEJ may also be stimulated by genome architecture, such as the presence of LINE, Alu and MIR elements among others, but does not require obligatorily SDs neither minimal efficient processing segments (MEPS) to mediate the recombination. In some cases, NHEJ leaves an 'information signature or scar', consisting in that the rejoining site often contains within the product of the repair a microdeletion and nucleotides addition as molecular footprint of the DNA end junction [Lieber, 2008; Gu *et al.*, 2008].

Other replication-based mechanisms for DNA repair is Fork Stalling and Template Switching (FoSTeS) [Lee *et al.*, 2007] which is associated with non-recurrent genomic rearrangements by error like one-ended DSB resulting from a collapsed DNA replication fork during DNA synthesis. The FoSTeS model has been further generalized with more molecular mechanistic details in the microhomology-mediated break-induced replication (MMBIR) model [Hastings *et al.*, 2009], that is used to repair the damage in one single double strand end, when stretches of single-stranded DNA are available and share microhomology with the 3' single-strand end from the collapsed fork [Hastings *et al.*, 2009]. These mechanisms act under circumstances when NHEJ is not an option, because after replication fork breakage, there is only a single end with no second end to which the one end can be annealed or ligated. Thus these mechanisms that repair single DNA ends are more appropriately invoked for spontaneous damage during replication than the mechanisms that act on two-ended DSBs [Sankaranarayanan *et al.*, 2013].

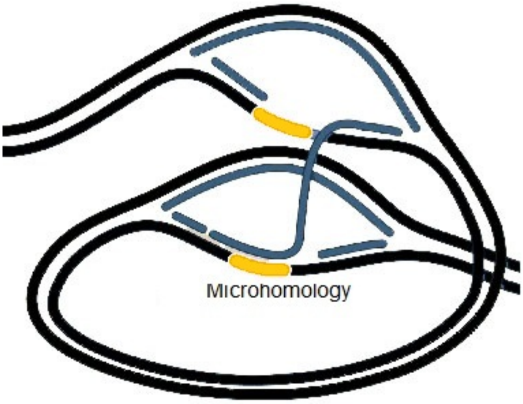


Figure 1.5: FoSTeS mechanisms - Schematic representation of the FoSTeS-based mechanism for genomic rearrangements.

FoSTeS occurs (see Figure 1.5), when the active replication fork stalls and switches templates using complementary template microhomology to anneal and prime DNA replication. As result, there are interrupted duplications in which stretches of DNA of normal copy number were punctuated by stretches of DNA that were amplified two or three times [Lee *et al.*, 2007]. The FoSTeS events occur preferentially in regions of complex genomic architecture that contain abundant low-copy repeats with high sequence identity and in various orientations that might bring into proximity highly similar DNA segments or repetitive sequences that normally lie far apart [Branzei and Foiani, 2007]. This could favor replication long-distance template-switching models between different replication forks stalling and slippage and, consequently, enables the joining or template-driven juxtaposition of different sequences from discrete genomic positions, generating complex genome structural rearrangements, including inversions [Lee *et al.*, 2007]. Unlike the DNA double-strand break-induced genome rearrangement model involving NAHR or simple NHEJ, the long-distance template-switch model for genome amplifications suggests a single-strand DNA lesion as the initiating trigger [Slack *et al.*, 2006; Lee *et al.*, 2007].

1.4.2 Inversion effects

As other structural variants, inversions could have important consequences on the genome. Theoretically the inversions, as balanced rearrangement, do not involve the quantitative alteration in the content of cellular DNA, but in particular, the reorganization of a genomic segment is characterized by having three major genetic effects which may have several repercussions [Alves *et al.*, 2012], including direct or indirect mutations, affect the positional distribution of genes, and exert influence in the recombination process [Feuk, 2010]. These effects are relatively different from those of other structural variations, like copy number variations, which are mainly related with changes in gene dosage [Stankiewicz and Lupski, 2010].

1.4.2.1 Mutational effect

One of the possible obvious consequences of the inversions is the mutational effect at the breakpoints. This is a direct effect of breaking the DNA molecule and the consequences of this break will depend on its location with respect to functional sequences, as for example in a break within a coding sequence. It is evident that the rupture of an exon by the breakpoint leads into the disruption of the functionality of the gene and tends to be a deleterious change. Furthermore, the mutational changes that alter the coding structures, not by directly breaking an exon, but by breaking within the introns and subsequent reordering of the distribution of exons within the gene, is highly likely that might lead into genomic disorder as well. A good example is the study that proved the loss of expression of *Hoxd* genes during limb development and phenotypic alterations in mice by inducing an inversion that split the mammalian *Hoxd* gene cluster into two independent pieces [Spitz *et al.*, 2005]. In humans, there are various evidences of inversions that involve genes and are related to diseases, such as the X-linked disorder caused by an inversion that breaks the factor *VIII* gene, which gives rise to hemophilia A [Lakich *et al.*, 1993; Antonarakis *et al.*, 1995], or the inversion that breaks the *IDS* iduronate 2-sulfatase gene that causes Hunter syndrome [Bondeson *et al.*, 1995].

1.4.2.2 Positional effect

Even when the location of the inversion within the genome does not break a functional element, it is important to appreciate that they can have a significant effect at a distance [Sharp *et al.*, 2006]. Positional effect is a direct consequence of the inversion due to the movement of genomic segments from one region to another. Position effects can be caused by translocation of a gene into a heterochromatic region, resulting in the methylation of promoter regions and consequent down-regulation of expression [Kleinjan and van Heyningen, 1998], or by intergenic genomic rearrangements that detach a gene from its transcriptional regulatory elements or that bring a gene into close proximity to another regulatory element, altering gene expression [Spitz *et al.*, 2005].

Although the current estimated fraction of the genome that is evolutionary conserved through purifying selection represents an small portion, around $\sim 10\%$, the recent polemic results of the ENCODE project suggest that there is now substantial evidence that many other hidden elements are potentially functional, most of which with a role in gene regulation[ENCODE Project *et al.*, 2007; Graur *et al.*, 2013]. Thus, inversions cannot be presumed to be functionally harmless or neutral because they encompass only non-coding segments, but instead a careful assessment of nearby genes that may be affected via a positional effect mechanism also needs to be considered.

1.4.2.3 Predisposition to further rearrangements

The potential effect of inversions might not be directly associated to the alteration of gene expression, either by disrupting coding regions that span the breakpoints or by position effects acting on genes adjacent to the breakpoints. Instead, the real effect of an inversion could be that it can act as a risk factor for other genomic changes [Sharp *et al.*, 2006]. That is the case of several polymorphic inversions generated between flanking duplications which not have any direct consequence but it is thought that they result in abnormal meiotic pairing, leading to an increased susceptibility to unequal NAHR. In this situation, the inversions presumably predispose to secondary rearrangement by switching the orientation of large, highly identical stretches of sequence on homologous chromosomes, thus allowing their subsequent misalignment during synapsis and hence facilitating illegitimate recombination [Sharp *et al.*, 2006].

Therefore, these inversions have been associated with an increased susceptibility to rearrangements at these loci [Giglio *et al.*, 2002; Sharp *et al.*, 2006; Giglio *et al.*, 2001]. So far, there is growing evidence for several polymorphic inversions flanked by highly homologous segmental duplications, for which, parents that carry the inversions in heterozygosis confer a predisposition to further deletion of the inverted segment in subsequent generations. Most of these cases have been described as microdeletion syndromes in the offspring of inversion heterozygotes, such as Sotos syndrome [Visser *et al.*, 2005], Angelman syndrome [Gimelli *et al.*, 2003], Williams-Beuren syndrome [Osborne *et al.*, 2001].

1.4.2.4 Effect on recombination

Alleles carrying inversions usually do not cause any abnormalities as long as the rearrangement is balanced, without missing or gaining DNA material, except for the cases described above. However, one of the main effects of inversions is their influence as a suppressor of recombination in inversion heterozygotes [Kirkpatrick, 2010]. This effect can be by two different mechanisms. The first is a real suppression of recombination due to the difficulty of complete synapsis between the two homologous in the regions at the ends of the inversion loop during meiosis (see Figure 1.6). This means that inversions hinder the recombination at the inversion boundaries, and the closer to the breakpoints, the more reduction in the crossover frequency. The second mechanism is an apparent suppression of recombination within the inverted regions (see Figure 1.6). In reality, recombination could occur at a fairly normal frequency within the inversion region relative to the same region in a homozygous individual or other not inverted region. However, the gametes produced from recombination within the inverted region in heterokaryotypes are usually unable to produce viable offspring.[Stevison *et al.*, 2011; Adi *et al.*, 2011; Coyne *et al.*, 1991; Navarro and Ruiz, 1997].

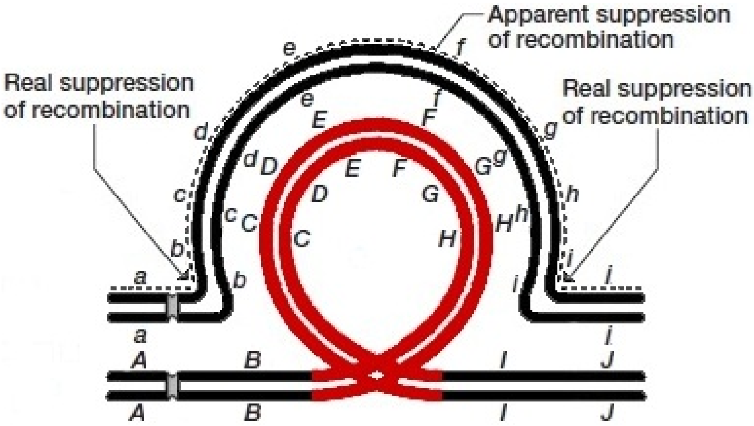
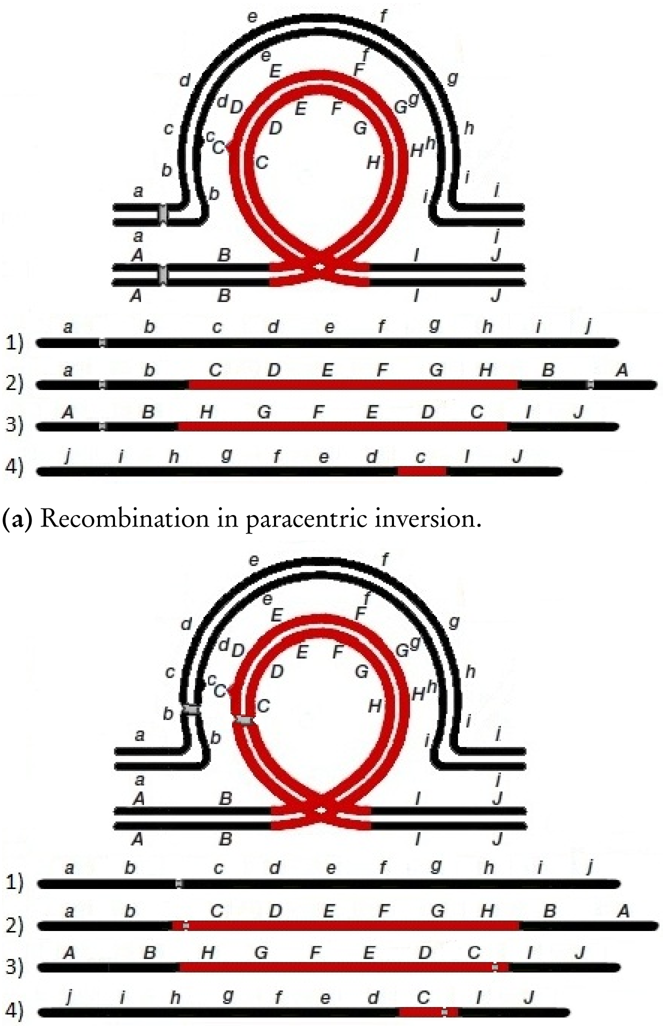

Apparent suppression of recombination
Real suppression of recombination
Real suppression of recombination
a b c d e f g h i j
a b c d e f g h i j
A B C D E F G H I J
A B C D E F G H I J

Figure 1.6: Suppression of recombination - Schematic representation of the synapsis of a paracentric inversion heterozygote. During the process, the inverted region in red color, is incorporated into a loop to maximize synapsis along the length of both chromosomes. A real suppression of recombination occurs near the breakpoints (at the base of the loop) due to the difficulty of synapsing in this region. An apparent suppression of recombination occurs in this region due to the formation of inviable recombinant chromatids.

The different types of inversions have an effect over recombination but the specific mechanism results in a slightly different outcome. In the case of paracentric inversions, the crossover within the inversion loop in a heterozygote (see Figure 1.7a) results in that the two nonsister chromatids that are not involved in the crossover will end up in normal gametes (carrying either the standard or inverted allele). However, the products of the crossover, rather than being a simple recombination of alleles, are a dicentric and an acentric chromatid. The acentric chromatid is not incorporated into a gamete nucleus and this recombinant chromatid will be lost. The dicentric chromatid begins a breakage-fusion-bridge cycle, as the two centromeres are pulled to opposite centrosomes during meiosis I. Ultimately, the dicentric chromosome randomly breaks between the two centromeres and each chromatid, containing deletions, produces a genetically unbalanced gamete. Thus, the gametes derived from the recombinant chromatids are unable to produce viable offspring. In case of pericentric inversions, the crossover within the inversion loop in a heterozygote (see Figure 1.7b) results in that all four chromatid products from a single crossover within the loop will have centromeres and



(a) Recombination in paracentric inversion.

(b) Recombination in pericentric inversion.

Figure 1.7: The schematic representation of the consequence of a crossover in a inversion heterozygote shows that both types of inversions produce recombinant chromatids which result in gametes that are genetically unbalanced and unable to yield viable progeny, and thus recombination appears to be blocked. (a) Recombination in paracentric inversion yields two non-recombinat parental chromosomes that contain either the standard (1) or inverted allele (3) and also produce two recombinan chromatids, one unbalanced dicentric chromatid (2) and another unbalanced acentric chromatid (4). (b) Recombination in pericentric inversion yields two non-recombinant parental chromosomes that contain either the standard (1) or inverted allele (2), and also produce two recombinant chromatids (2, 4) both of which are not balanced containing a reciprocal duplication and deletion.

are therefore incorporated into the nuclei of gametes. However, the two recombinant chromatids are not balanced, and both have duplications and deletions. According to the magnitude of the duplications and deletions, these gametes tend to form non-viable zygotes. [Stevison *et al.*, 2011; Adi *et al.*, 2011].

Therefore, both paracentric and pericentric inversion heterozygotes will not show recombinant gametes if a crossover occurs within the region of inversion. This is the cause of the apparent suppression of recombination. The inversion realy hinder recombination around the breakpoints at the base of the inversion loop but actually the recombination can occurs at normal frequencies outside these breakpoint regions. For this reason, recombination is only suppressed very near to or within the inversion loop in an inversion heterozygote.

This effect leads to interesting consequences. One is the reduction in fertility due to some gametes forming non-viable zygotes in the progeny. Therefore, inversions will have a lower fitness in heterozygosis and this will make the inversions under-dominant. In the case of *Drosophila* species, male individuals do not recombine and the exclusion of recombinant offspring from the gametes in heterozygote female possibilitates the great number of inversions found in these species [Andolfatto *et al.*, 2001; Stevison *et al.*, 2011].

Another consequence of suppression of recombinant offspring within the inversion region of inversion heterozygotes is the putative role of inversions in population divergence, reproductive isolation and speciation phenomena. Based on these inversion potential effects, while in the classic models, inversion might be a mechanisms by which dysfunction in hybrid fertility is generated, other theoretical hypothesis have been proposed based on the gene flow interruption and accumulation of differences between the two chromosomal configurations by the suppression of recombination caused by an inversion [Hoffmann and Rieseberg, 2008]. This hypothesis has been supported by varied observations, in a wide range of species that include from fungi and plants to animals (insects, birds and mammals). Observations in primates and human are still under investigation [Zhang *et al.*, 2004; Adi *et al.*, 2011].

1.4.3 Adaptive value of inversions

Since early studies in *Drosophila* it was shown that inversions were adaptive. However, it was not clear by which mechanisms inversions have this adaptive value. The potential genomic effect of inversions over populations could lead to positive consequences that have generated two main different hypotheses to explain the adaptive value of inversions. [Kirkpatrick, 2010].

- The value of positive mutation/position effect. This hypothesis postulates the potential advantage of inversions whose breakpoints cause a mutation or position effect with beneficial consequences.
- The coadaptation hypothesis. This hypothesis is focused on the potential suppression of recombination and suggests that inversions could protect coadapted combinations of alleles that have functional advantages if they remain working together.

However, the two hypotheses are not mutually exclusive and it is also possible that both, mutational and coadapted effects, contribute at the same time to the increase of frequency of the inversion in the population. In particular, inversions have been proposed to be involved in the rapid adaptation of populations to local environmental conditions [Krimbas and Powell, 1992]. In addition, reduced recombination between alternative arrangements in heterozygotes may protect sets of locally adapted genes, promoting ecological divergence and potentially leading to reproductive isolation and speciation [Kirkpatrick and Barton, 2006].

Local adaptation is an excellent case of the adaptive value of inversions. This is the phenomenon in which some combinations of genes are favored in different environments. If an inversion, through the recombination suppression effect, captures and holds together a group of alleles that are better adapted to the local environmental conditions than other ancestral alleles, then it has a selective advantage that can cause its spread in the population [Kirkpatrick and Barton, 2006].

A good example of the contribution of the inversions to the local adaptation is the inversion "In(3R)Payne" in *Drosophila melanogaster*, which shows parallel latitudinal clines on three continents [Hoffmann *et al.*, 2004; Anderson *et al.*, 2005]. Another study of an inversion in the mosquito *Anopheles funestus*, which is one of the most important and widespread malaria vectors in Africa, also shows that the inversion has an important role for environmental selection in shaping their population distribution[Ayala *et al.*, 2011].

Finally, human inversions could also have important evolutionary consequences, as in the case of the 17q21.31 inversion that has been related with increased female fertility and positive selection [Stefansson *et al.*, 2005].

1.5 Methods of detection

The methods that can be used for structural variation detection depend upon the length of the variants (see Figure 1.8). Besides the resolution limitation, the method depends on the type of the genomic abnormality studied, and the methods specificity over the particular type of structural variant that is studied. In addition, the method depends also of the type of detection, that is, if it is a target-manner detection in which the location of the rearrangement is known and the assay is just to confirm the presence or not of the variant, or if it is a prediction of new rearrangements based on the comparison to a reference.

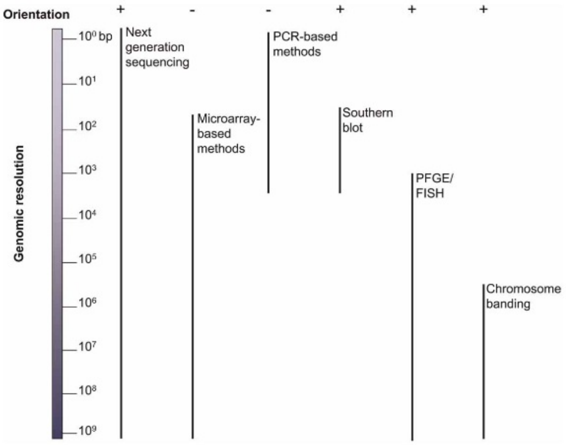


Figure 1.8: Methods for SV detection - Methods for SV detection. (Adapted from [Carvalho *et al.*, 2011]).

1.5.1 Cytogenetic methods

At microscopic resolution level, structural variations of large chromosomal segments have long been possible to be detected cytogenetically. G-banded karyotyping is the standard method for detecting rearrangements of large chromosomal regions of around 5–10 Mb. Inversions are the most common cytogenetically detectable rearrangements, particularly the pericentric ones [Feuk *et al.*, 2006; Feuk, 2010]. Although this method is geared to identification of big variants, some significantly large inversions still could remain undetectable if the inverted segment leads to little difference in the banding pattern. With the advent of additional chromosome-banding techniques and the ability to work with elongated prometaphase chromosomes, more discrete structural abnormalities became apparent. Furthermore, the advances in fluorescence in situ hybridization (FISH) analysis allowed a more refined characterization of the extent of these variants [Sharp *et al.*, 2006; Feuk, 2010].

1.5.2 Microarray-based methods

At the submicroscopic resolution level the methods for SVs detection are much more recent. The hybridization based methods such as SNP microarrays and aCGH techniques, which is similar to SNP arrays but it is a more appropriate method for analysing unbalanced sequences, have been deeply used for detecting CNVs. The main objective in using aCGH is to determine the ratio between the amount of DNA of a specific region of two different samples (for example affected vs. control). The oligonucleotides or BACs of the genomic region of interest are immobilized in a microarray, and the DNA samples from the two individuals are marked with two different fluorescent colors and hybridized into the array. Finally, a special scanner compares the difference between the signal intensity of the two colors to measure the ratio of copy number differences between samples. However, other forms of variation without any DNA gain or loss cannot readily be detected with microarrays. Thus, it does not apply to inversion detection.

1.5.3 Sequencing-based methods

Until relative recently, variation discovery by sequence analysis was done using low-coverage Sanger-based sequencing. However, in the recent years, the emergence of several high-throughput sequencing (HTS) technologies, also commonly known as next generation sequencing (NGS), are revolutionizing the field of genomics by making it possible to generate billions of short $\sim 35 - 250$ bp sequence reads [Mardis, 2006] using several different technologies such as Ilumina, Roche/454, or SOLiD. Therefore, the ability to sequence genomes with high coverage and low cost has made feasible to perform comprehensive and detailed studies of rearrangement detection and analysis of genomic variants of different individuals. This new technology has significantly changed how genomics research is conducted and has increased the demand for computational tools to optimize the utilization of the data generated by the sequencing platforms [Mardis, 2006].

Now next generation sequencing makes possible to overcome the major limitations in the characterization of global genome variation. It give us the opportunity to study a high number of individuals across many human populations, allowing us to complement the human reference genome. Evidence of this is the publication of the complete genome sequences of an increasing number of individuals [Levy *et al.*, 2007; Bentley *et al.*, 2008; Wang *et al.*, 2008; Wheeler *et al.*, 2008; Ahn *et al.*, 2009; Azim *et al.*, 2013; Lilleoja *et al.*, 2012; Shen *et al.*, 2013; Gupta *et al.*, 2012; Fujimoto *et al.*, 2010].

The strategy of *De novo* assembly and direct comparison between different genomes (see Figure 1.9) is theoretically the most complete method for the study of genomic structural variation [Li *et al.*, 2011]. This approach gives the most valuable result because it overcomes the possible limitations of the genome reference, allowing to detect all types of variation (SVs, small variants and SNPs), as well as to discover novel sequences. In addition, it gives the exact location of the variants, allowing to resolve the breakpoints to the nucleotide resolution.

The strategy has two phases

1. *De novo* assembly: Using the power of assembler, the reads from the sample genome are concatenated to obtain large contigs/scaffolds.

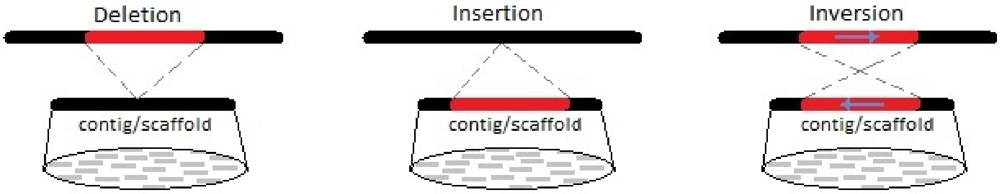


Figure 1.9: ***De novo* assembly and direct comparison strategy** - Signatures for different types of structural variation events such as deletion, insertion, inversion.

2. Alignment: The different structural variants can be detected through the alignment of the contigs produced in the first step against the assembled sequences of the same genome region in other individual or the reference genome.

The much deeper coverage of short-read sequencing projects does not entirely compensate for the shorter read length because the assemblies with longer reads have still far better contiguity than the NGS short-read assemblies [Gnerre *et al.*, 2011]. Therefore, the current most useful method of this strategy has been sequencing the entire fosmid or BAC clones of the region of interest with the traditional Sanger-based capillary sequencing. However, the high cost of capillary sequencing is unassumable to study large number of individuals. This highlights that assembling large mammalian genomes from short reads remains an extremely challenging problem, albeit there has been considerable progress represented in the wide variety of *De novo* assembly algorithms [Li *et al.*, 2010a,b; Reinhardt *et al.*, 2009]. Another obvious handicap of this strategy for detecting variation is that it requires further processing before comparing the sample sequences and the assembly softwares are time consuming and require high power computational resources.

Due to these limitations, currently the most used methods to predict all kinds of structural variants consist of an intermediate strategy. These methods are based on mapping reads taking the reference sequence as an intermediate guide and have been considerably developed in the last years [Xi *et al.*, 2010; Medvedev *et al.*, 2009].

This approach also has two main steps.

1. Mapping: The initial mapping of the library of reads from the sample genome

against the reference genome sequence. The identity of the alignment defines the threshold to consider as putative regions where reads were sequenced from.

2. Prediction: Through the analysis of the mapping pattern, the next step is to locate regions of discordance or abnormalities in the aligned signature, and then predict the SVs in the sample genome that could be the cause of discordance.

These methods are known as mapping-based strategies and several computational tools have been developed for characterizing structural variation among different individuals, using next generation sequencing platforms. The algorithms can be classified into three major strategies, attending to the mapping signature used for the analysis.

Read depth (RD):

This approach uses the profile of depth of coverages per region, defined as the average number of reads which map in that region on the reference genome. The strategy has the implicit assumption that the probability of mapping reads per region obeys a Poisson distribution. Then, the mean value of the distribution is the expected depth of read coverage in the region [Alkan *et al.*, 2009].

The depth of coverage of a region is proportional to the number of times that this region appears in the sample genome. The identification of a significant divergence (undercount or overcount) regarding the expected depth of coverage could be associated with CNVs in that region on the sample genome (see Figure 1.10). An increase in the read depth of a region indicates a greater number of locus copies of the sequence (insertion/duplication) in the sample genome in comparison to the reference genome; while a reduction of the read depth indicates a smaller number of locus copies of the sequence (deletion) in the sample genome. Moreover, in the case of inversion breakpoints, that region also could be associated with a reduction in read depth due to the problem of mapping the reads spanning the breakpoints, although this signal tends to be very low and it is useful only to corroborate breakpoints but not for prediction.

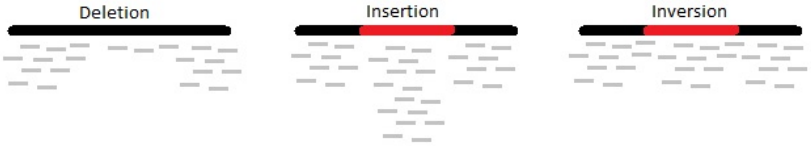


Figure 1.10: Read depth (RD) strategy - The read depth method allow us to detect insertion and deletion and helps to pinpoint the breakpoints of the inversion.

Split Read (SP):

This approach uses the profile of incomplete aligned reads to pinpoint the exact breakpoints of structural variant events [Ye *et al.*, 2009]. It is based on the pattern of mapping of reads from a sample genome which span breakpoints that will be mapped partially between both sides of the breakpoint in the reference genome. That means the read will be broken into two segments (see Figure 1.11).

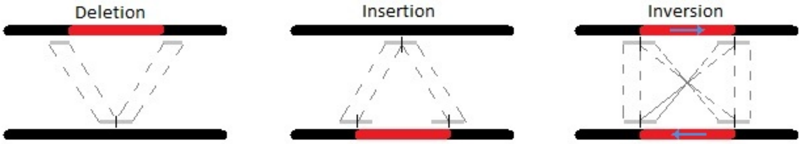


Figure 1.11: Split reads (SR) strategy - The split read method helps to pinpoint the breakpoints more precisely for different types of structural variation events, such as deletion, insertion, or inversion.

Since a split read signature indicates a breakpoint, a deletion in the sample genome will be associated to split reads mapping with a inner gap that represents the extra sequence in the reference genome. Insertions in sample genome will be associated with a set of reads that map partially to the reference genome, with just the left or right extreme aligned, depending on which breakpoint (left or right) in the sample genome is bridged by the split reads. In the case of inversions, the reads spanning the breakpoints in the sample genome will be associated with read mapping divided into two fragments in relative inverted orientation one to the other and spanning an inner gap, which in this case represents the inverted sequence in the sample genome. Split read strategy is able to detect breakpoints of SV with very high resolution, especially in unique regions.

Theoretically, the split read approach should be able to detect the exact breakpoint at nucleotide resolution. This approach is more useful the longer the reads sequenced reads, and the development of NGS technologies is continually improving the lengths of the reads obtained. Thus, this increases the possibility to predict structural variants and pinpoint their exact breakpoints by means of the split read strategy. However, it has the limitation of the presence of inverted repeats in the breakpoints of the structural variants generated by NAHR.

One of the first algorithms using split reads approaches to identify structural variants is "Pindel" [Ye *et al.*, 2009]. This tool only allows unique mappings, and uses a pattern growth approach to search for unique substrings of unmapped reads in the genome. The algorithm then checks whether a complete unmapped read can be reconstructed combining the unique substrings found in the previous step. Another recent algorithm that uses the split read approach is "Splitread" [Karakoc *et al.*, 2012]. In this case multiple mappings are clustered based on the maximum parsimony method. Finally, there are several other algorithms which use the split read approach applied to specific features, such as "TopHat" [Trapnell *et al.*, 2009] and "Dissect" [Yorukoglu *et al.*, 2012], which are specialized in detection of transcriptome structure analysis using RNA-Seq.

Paired Read (PR):

One of the methods most commonly used for the detection of structural variants is the analysis of the mapping of paired reads. This approach takes advantage of the technologies that produce paired reads by the sequencing of both extremes of the sample fragments. The method is based on aligning the paired-end reads to the reference genome and then uses the paired-end mapping profile of the fragment library of the sample genome to study the discordant mapping pattern [Tuzun *et al.*, 2005; Volik *et al.*, 2003]. After the alignment phase, the insert size fragment distribution stats (minimum, maximum, mean, and standard deviation of length) are computed. This step sets the expected range of insert size between the paired ends. Next, the structural variant prediction is based on the localization of regions with significant difference (also called discordance) regarding the expected mapping pattern profiles, either in the distance between reads or in orientation.

Given a data set of paired-end reads from a specific region of the sample genome, the expected pattern of mapping (also referred as concordant signal) is that which fulfills the threshold of the insert size fragment distribution. Moreover, both paired-end reads must align at the same chromosome and they should map in the expected relative orientation, which depending on the sequencing methods, it could be one read in the forward and the other in the reverse strand (also known as +/–) (see Figure 1.12) or both ends in the forward or reverse strand also known as +/+ and –/–.

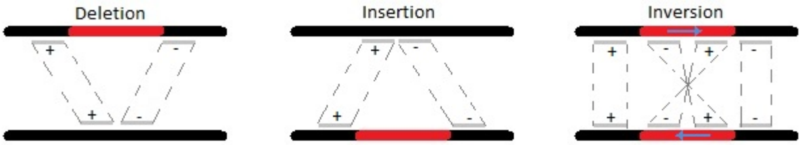


Figure 1.12: Paired read (PR) strategy - Signatures for different types of structural variation event, such as deletion, insertion, and inversion.

The different discordant patterns indicate different structural variants. A deletion in the sample genome will be associated to a discordant mapping pattern in which the orientation between paired ends is correct, but it represents an insert size greater than the threshold expected. An insertion will be associated to a discordant mapping pattern in which the orientation between paired ends is correct, but represents an insert size lower than the threshold expected. In both cases the alignment is at the same chromosome and the orientation of the two reads is concordant according to the sequencing technology. A translocation will be associated to a discordant mapping in location, where one read maps in another chromosome. Finally inversions will be associated to a discordant mapping pattern that does not fulfill the expected orientation, with the paired end reads aligning in the opposite orientation, and also the distance between paired-end reads does not have to be necessarily within the expected range.

> The paired-end mapping (PEM) methods perform well predicting a wide variety of SVs. With the advent of next-generation sequencing technologies, many groups have identified structural variants using high throughput sequencing that implemented this strategy from different points of view. This has resulted in the development of diverse algorithms based in different strategies for the prediction of structural variants from the discordant patterns
>
> For example, there are some algorithms that employ a "hard clustering" approach using only the best location for each mapped paired end read for finding structural variants, such as PEMer [Korbel *et al.*, 2009], GASV [Sindi *et al.*, 2009] and BreakDancer [Chen *et al.*, 2009]. Alternatively, other structural variant detection algorithms use multiple mappings for each paired end read and employ a soft clustering method through a combinatorial optimization framework [Lee *et al.*, 2008; Hormozdiari *et al.*, 2009] and maximum parsimony or a heuristic approach. Examples of this type of algorithms are VariationHunter [Hormozdiari *et al.*, 2010] and Hydra [Quinlan *et al.*, 2010], amount several others.

The analysis approaches are significantly different for each mapping strategy of detection of structural variants. Nonetheless, recently some integrative methods have been developed in which multiple signals are used in order to achieve improvements on the structural variants discovery. The implementation of multi-approach algorithms that integrate the analysis of varied patterns of read mapping stems from the need to improve the accuracy of structural variants discovery methods, because none of the single approaches perform in a comprehensive way. Most of these integrative algorithms combine paired end reads and read depth patterns. One of them is "GASVPro" [Sindi *et al.*, 2012], which uses paired end read patterns to find candidate structural variants and then uses the read depth pattern as a posterior filtering. "Novelseq" [Hajirasouliha *et al.*, 2010] utilizes *De novo* assembly together with paired end mapping pattern to find structural variants and insertions of novel sequence (sequence in sample but missing in the reference). "Delly" [Rausch *et al.*, 2012] combines paired ends read and split read approaches, using read pair signatures to detect candidate structural variants and then refine as much as possible the breakpoints using split read information.

1.5.4 Challenges of inversion detection

In particular, the detection of inversions is especially problematic due to that inversions are frequently located between inverted repeats [Feuk *et al.*, 2005]. Non-allelic homologous recombination between highly identical (and presumably inverted) segmental duplications is considered the primary mechanism by which most of the largest inversions are formed [Feuk, 2010; Kidd *et al.*, 2010]. Therefore, the breakpoints of the polymorphic inversions are more likely to occur where they are less likely to be detected, namely in repetitive sequences.

In the regions of segmental duplications is more difficult to map reads uniquely. Thus, many reads sequenced across inversion breakpoints are mismapped concordantly, and the power of PEM methods to detect them is significantly reduced. In addition, many reads sequenced from regions without an inversion could be mismapped discordantly, and the number of false positives detected is significantly increased. In the case of the paired end mapping method, for this reason, it has a better performance if the insert size used encompass completely the segmental duplication with the potential breakpoint. This means, that when detecting inversions, longer template insert sizes always improve sensitivity and specificity [Lucas Lledó and Cáceres, 2013]. But the insert size of most used next generation sequencing technologies is yet under 3 kb [Mardis, 2006]. Thus, a big fraction of the polymorphic inversions in the human genome remain difficult to discover by paired end mapping using this data. Other feature that makes inversion detection more difficult is that a good definition of an inversion must detect two paired breakpoints and locate the inverted region in between. Thus the detection has to find two sets of discordant mappings, as well as refine both breakpoints loci.

1.6 Inversions in the human genome

Historically, inversions in humans have remained relatively poorly studied, with regard to copy number variants. This is mainly due to the technical difficulty of inversion detection. For example, the first widely used array-based technology for the study of structural variation is not suitable for balanced rearrangements. This feature of inversions also downplayed its initial clinical interest, due to that the study of regions that change the amount of genetic material promised more productive outcomes attending to their functional effect. Despite this, at present several polymorphic inversions have been identified in humans, which highlights the importance of this type of structural variation, and the interest in them is growing once many more inversions have been predicted (see Table 1.4) mostly based on next generation sequencing technology.

Table 1.4: Studies that predict inversions in humans.

Study	**Predicted inversions**	**Method of detection**	**Sequencing method (insert size)**
Tuzun *et al.* [2005]	56	PEM	Sanger, Fosmid 40 kb
Korbel *et al.* [2007]	122	PEM	NGS, 3 kb
Levy *et al.* [2007]	90	full genome sequencing comparison	Sanger, HuRef assembly
Wang *et al.* [2008]	17	PEM	NGS, (135-440) bp
Kidd *et al.* [2008]	224	PEM	Sanger, Fosmid 40 kb
Ahn *et al.* [2009]	415	PEM	NGS, (100-300) bp
McKernan *et al.* [2009]	91	PEM	NGS, 3.5 kb
Pang *et al.* [2010]	105	PEM	Sanger, (2-37) kb

Another important study was the one that performed a cross-species comparison between the human and chimpanzee genomes assemblies [Feuk *et al.*, 2005]. This study identified ~1500 putative inversion regions, covering more than 154 Mb of DNA. From those, it was experimentally validated 23 of the 27 semi-randomly chosen regions, and 13% (3/23) of the chosen inversion were polymorphic in a panel of human samples. This three polymorphic inversions include fragments of 730 kb (at 7p22), 13 kb (at 7q11), and 1 kb (at 16q24) and their minor allele frequencies are 5%, 30%, and 48%, respectively. These results suggest that inversions may be a more common feature of the human genome than it was thought and an important source of variation in primate genome evolution.

1.6.1 Inversion polymorphism in the human genome

Despite at present several hundred inversions have been reported in the human genome [Feuk, 2010], the real knowledge about human inversions has lagged behind and just a small number of inversions (around ∼15) have been characterized in greater detail [Antonacci *et al.*, 2009; Bosch *et al.*, 2009; Deng *et al.*, 2008; Entesarian *et al.*, 2009; Feuk, 2010; Giglio *et al.*, 2002; Gilling *et al.*, 2006; Gimelli *et al.*, 2003; Martin *et al.*, 2004; Osborne *et al.*, 2001; Pang *et al.*, 2013; Salm *et al.*, 2012; Starke *et al.*, 2002; Stefansson *et al.*, 2005]. The result of this targeted inversion studies, together with other studies that carry out a more general characterization of inversions [Kidd *et al.*, 2008, 2010; Korbel *et al.*, 2007; Feuk, 2010] have shown that size distribution of the current map of inversions in the human genome is slightly different compared to the size distribution of the copy number variation. Most of the inversions discovered to date are in the $\sim 10-100$ kb interval [Feuk, 2010], and this average is greater than the average of copy number variants, in around $\sim 1-10$ kb [Feuk, 2010]. Moreover, two main processes are considered the primary mechanism by which inversions are generated: breaks in relatively simple regions that are joined in opposite orientation by non-homologous mechanisms [Onishi-Seebacher and Korbel, 2011] and non-allelic homologous recombination between inverted repeats or segmental duplications [Feuk, 2010; Kidd *et al.*, 2010; Feuk *et al.*, 2005].

However, important features of inversions remain unknown, such as their frequency and population distribution, because most studies were limited to a handful of individuals. So far, there are six large inversions studied by FISH in 27 individuals of three populations [Antonacci *et al.*, 2009], the worldwide genotyping of the 8p23 inversion distribution based on SNP data and genetic substructure [Salm *et al.*, 2012], and the recent analysis of eight simple inversions in 42 human samples of diverse origins, including one inversion genotyped in 57 populations [Pang *et al.*, 2013]. Finally, the most intensely studied human inversion polymorphism is a fragment of ∼900 kb at 17q21.31 [Stefansson *et al.*, 2005; Zody *et al.*, 2008]. In-depth analysis of the refined physical map of this chromosome showed that the alternative orientations of this inversion correlate perfectly with two highly divergent (since ~ 3 mya) haplotype lineages in European population (H1 and H2), which have strong linkage disequilibrium (LD). The local recombination suppression by the inversion explains the divergent haplotype structure at this locus. Other interesting finding was that the H2 lineage is rare in Africans and almost absent in East Asians, but it is found at a frequency of 20% in Europeans, in whom the haplotype structure suggests it is undergoing positive selection. Detailed analysis of 29,137 individual genotypes, (16,959 women and 12,178 men) from Iceland showed a significant increase in fertility in female carriers of either one or two copies of the inversion, explaining likely its increase in frequency in European [Donnelly *et al.*, 2010; Steinberg *et al.*, 2012; Boettger *et al.*, 2012]. The exact mechanism by which this inversion causes the elevation of fertility is still not totally clear, but a plausible explanation may be that the significantly higher recombination rate along the chromosome, observed in 23066 studied individuals [Kong *et al.*, 2004], might lead into a reduction in the rates of maternal non-disjunction, the leading cause of pregnancy loss due to aneuploidy in the fetus [Kong *et al.*, 2004]. However, it has also been found that there are copy number variations in genes associated to the inversion and gene expression change [de Jong *et al.*, 2012].

1.6.2 Inversions in human disorders

Theoretically most inversions are not associated with alterations in the amount of DNA material, and thus they are more likely to be apparently neutral and may not cause an obvious phenotypic consequence. Moreover, since very little was known about inversions in humans until relatively recent, it is often problematic to assess whether the inversion present in a patient is actually associated with the disease or just a polymorphism. It has been reported that *CRHR1* gene variants within a 900 kb inversion are associated with inhaled corticosteroid response in asthma complex disorders [Tantisira *et al.*, 2008].

The mutational effects that change the gene coding structures, such as the break within an intron and the reordering of the distribution of exons within a gene, might lead into a genomic disorder. An example is the case of the inversion associated to an X-linked disorder caused by the disruption in the factor *VIII* gene, which gives rise to hemophilia A. This is a recurrent inversion that spans a fragment of approximately 400 kb and has been found present in ~20 −45% of patients on families with severe disease [Lakich *et al.*, 1993; Antonarakis *et al.*, 1995]. As is usual in recurrent inversions, it is mediated by two inverted segmental duplications, one of which is located in intron 22 of the factor *VIII* gene, with two other copies being located ~400 kb distal to the gene. Other recurrent inversion generated from recombination events between *IDS gene* and a second *IDS* locus (*IDS-2*) located within 90 kb, has been shown to lead into a disease phenotype on 13% of patients with the Hunter syndrome. In this case, the effect of the inversion results in a disruption in the intron 7 of the *IDS* gene [Bondeson *et al.*, 1995].

A specific category of inversions associated with genetic disorders are those that are not directly causative, but rather increase the risk of further rearrangements that cause disease. Such is the case of the inversion of ~3.5 Mb at chromosome 4 and the inversion of ~6 Mb at chromosome 8. Both of these inversions have breakpoints that fall in clusters of olfactory receptor *OR* genes of high identity on both 4p16 and 8p23 and might be involved in the origin of the t(4;8)(p16;p23) translocation, possibly related with Wolf-Hirschhorn syndrome and dysmorphic/mental retardation syndrome [Giglio *et al.*, 2002].

An inversion in the 48 kb region of the filamin (*FLN1*) and emerin genes (*EMD*), that was found in heterozygosis in the 33% of females studied, helps to explain some cases of the X-linked disorder that leads into Emery-Dreifuss muscular dystrophy (*EMD*) [Raffaele Di Barletta *et al.*, 2000], by inducing the deletion of the *EMD* gene and also a partial duplication of the nearby *FLN1* gene. The inversion is generated by non-allelic homologous recombination among two large inverted segmental duplications (11.3 kb with $>$ 99% sequence identity), which flank this region [Small *et al.*, 1997; Small and Warren, 1998].

Several polymorphic inversions confer a predisposition to further chromosomal microdeletion in subsequent syndrome-affected generations (see Table 1.5). The inversion at the 7q11.23 region was found in parents of 33% of the patients of Williams-Beuren syndrome which carried the 1.5 Mb hemizygous microdeletion putatively causing the disease [Osborne *et al.*, 2001]. The inversion seems to be related to the disease, but not directly associated with the abnormal phenotype in itself, since the frequency in the general population is approximately ~5% [Tam *et al.*, 2008] and the carrier parents

Table 1.5: Polymorphic inversions that predispose to microdeletion in offspring affected by genomic syndromes.

Cytogenetic locus	Frequency[a]	Inversion Length	related syndrome
5q35	Unknown	∼1.3 Mb	Sotos
15q11-q13	9%	∼4.0 Mb	Angelman
7q11.23	5%	∼1.5 Mb	William-Beuren

[a] Frequency in population. Higher values has been found in parents of patients with microdeletion.

are normal. Similarly, an inversion located at 15q11-q13 was found on heterozygosis in 67% of mothers of the Angelman syndrome patients carrying microdeletion in this region [Gimelli *et al.*, 2003]. In the normal population, the incidence of the inversion is ∼9%. This difference in the frequency of the inversion suggests that the inversion could be an intermediate state that facilitates the occurrence of 15q11-q13 deletions in the offspring [Gimelli *et al.*, 2003]. In the case of Sotos syndrome, which is also often caused by microdeletion in the two patients with a deletion in the maternally derived chromosome, all four parents were heterozygous for the inversion of segment ∼1.3 Mb at 5q35 region [Visser *et al.*, 2005].

In these examples, inversion of the region between the flanking duplications is thought to result in abnormal meiotic pairing, leading to an increased susceptibility to NAHR. Thus, these inversions have so far only been associated with an increased susceptibility to deletions at these loci. The syndrome studies therefore highlight the inversion as a risk factor, since these events apparently occur at increased frequencies when the transmitting parent carries an inversion of the segment that is deleted in the affected offspring.

1.7 Storage projects and databases of structural variants

The increased number of inversions that are being predicted are currently stored together with the other structural variants in different databases that provide stable and traceable identifiers for their analysis (see some examples in Table 1.6). Most of these projects have been developed to store structural variants that are linked to different phenotypes and related with diseases, and because of this inversions are little represented. However, the ubiquity of structural variants on human genomes has impulsed some other projects that support public access to a much broader information on structural variants that generally are not known to cause diseases but are very useful for biomedical studies.

The Human Gene Mutation Database (HGMD®) represents an attempt to collate known (published) gene lesions responsible for human inherited disease [Stenson *et al.*, 2009]. This database has now acquired a broad utility in that it embodies an up-to-date and comprehensive reference source to the spectrum of inherited human gene lesions, and stores valuable data mainly of copy number variants. Thus, HGMD provides information of practical diagnostic importance to: (i) researchers in human molecular genetics, (ii) physicians interested in a particular inherited condition in a given patient or family, and (iii) genetic counsellors.

DECIPHER is a Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources [Swaminathan *et al.*, 2012]. The primary purposes of the DECIPHER project are to: (i) Increase medical and scientific knowledge about chromosomal microdeletions/duplications; (ii) Improve medical care and genetic advice for individuals/families with submicroscopic chromosomal imbalance; (iii) Facilitate research into the study of genes which affect human development and health. Known and predicted genes within an aberration are listed in the DECIPHER patient report, common copy-number changes in healthy populations are displayed, and genes of recognized clinical importance are highlighted. It is expected that the data generated from the project will be used by others, such as researchers interested in developing new analytical methods, in understanding patterns of polymorphism, and in refining critical intervals to map genes involved in specific phenotypes and diseases.

The Human Genome Structural Variation Project [Eichler *et al.*, 2006; Human Genome Structural Variation Working Group *et al.*, 2007] includes the discovery of variants through development of clone resources, sequence resolution of variants, and accurate typing of variants in individuals of African, European or Asian ancestry. This project has employed a clone-based method to systematically identify and sequence structural variants genome wide. It is resulting in an integrated database of

Table 1.6: Databases and resources for structural variation studies.

Name	Description
Database of Genomic Variants (DGV)	A curated catalogue of human genomic structural variation. The content represents structural variation (larger than 50bp) identified in healthy control samples.
Human Genome Structural Variation Project	A catalogue of human genomic polymorphisms ascertained by experimental and computational analyses. The data are mapped against the UCSC Human Genome Browser.
Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER)	It is an interactive web-based database which incorporates a suite of tools designed to aid in the interpretation of submicroscopic chromosomal imbalances. The project enhances clinical diagnosis by retrieving information from a variety of bioinformatics resources relevant to the imbalance found in the patient.
The Human Gene Mutation Database (HGMD®)	It represents an attempt to collate published gene lesions responsible for human inherited disease. The project seeks to include DNA sequence variants that are either (i) disease-associated and of likely functional significance, or (ii) of clear functional significance even though not associated clinical phenotype may have been identified to date.
Database of Genomic Variants archive (DGVa)	This project gives support the DGV project and also is a repository that provides archiving, accessioning and distribution of publicly available genomic structural variants, in all species. It is integrated with EMBL-EBI resources and the Ensembl genome browser
NCBI database of genomic structural variation (dbVar)	dbVar stores all types of structural variants and accepts data from all species. It is integrated with Entrez and other NCBI resources

structural variation polymorphisms ascertained by experimental and computational analyses. This database includes large-scale structural variation (LSV), copy number polymorphisms (CNPs) and intermediate-sized structural variation (ISV), mostly determined by fosmid paired-end sequence analysis [Tuzun *et al.*, 2005; Kidd *et al.*, 2008]. The data are represented against the UCSC Human Genome Browser and related with SNPs using the same DNA samples used by the HapMap Project.

The necessity to integrate and archive the explosion of public structural variation data from several large-scale projects such as the Human Genome Structural Variation or the 1000 Genomes Project, and include other studies that report SVs in a separate sample genome has been satisfied through the development of two main official projects that are serving this role to the scientific community: the NCBI database of genomic structural variation (dbVar) and the Database of Genomic Variants archive (DGVa) [Lappalainen *et al.*, 2013; Church *et al.*, 2010]. Although dbVar and DGVa, both are managing more or less the same data source, they are providing complementary value-added tools and data access. The dbVar stores all types of structural variants and accepts data from all species, including clinical data of human samples of healthy controls and diseased patients. It also identifies variant prediction artifacts and provides some curation through cross referencing of its data with information from the Genome Reference Consortium (GRC). dbVar is integrated with Entrez and other NCBI resources. Meanwhile, the DGVa catalogue, stores and freely disseminates this important class of variation also for all species, and it is integrated with EMBL-EBI resources and the Ensembl genome browser. Both projects are providing a valuable resource to a large community of researchers.

Moreover, DGVa has been designed to facilitate the curatorial work of the major database project focused on human structural variation, the Database of Genomic Variants(DGV) [Iafrate *et al.*, 2004]. The main goal of this database project is to provide a useful catalogue of curated data, and to facilitate the interpretation of structural variants within the studies aiming to correlate genomic variation with phenotypic data. DGV has served a very important role collecting and analyzing structural variation data. Unlike the previous two databases, DGV is not designed for a prompt updating of the newer data. It has by contrast a discontinued submission of selected and preprocessed studies. Currently the database stores 55 studies that predict all the variety of structural variants, in which there are 2304349 of predicted events of CNV and 3380 of inversion events. The predictions with similar boundaries across the sample set are merged to form a representative variant that highlights the common variant found in the study. At this merge level, the number of CNVs gets down to 109863 and for inversions to 238 events.

1.8 Statement of the scientific problem

Over the last years there has been a major drive in genomic research to address one of the main scientific breakthroughs [Pennisi, 2007a], the comprehensive identification of structural variation in the human genome and their role in phenotypic variation. As a result of this, so far there is already a great accumulation of information over the structural variation and their potential effects. Despite that the early findings were a great success in developing human genome maps of CNVs, the mapping of inversions has been lagging behind. However, inversions have been increasingly recognized as a relatively common source of variation and an important genomic force in the human genome, contrarily to early predictions from classical cytogenetics.

The current dilemma in studying human inversions is the high level of false positive predictions, and the absence of a consolidated catalogue of the most reliable inversions and their associated information. To face these challenge will undoubtedly help the experimental validation and characterization of inversions and will pave the way for the main goal, to enhance our understanding of the functional and evolutionary consequences of the inversions, their real impact in the human genome, and their role in the phenotypic differences between individuals. In addition, this will open the gate to further biomedical lines of research.

1.9 Objectives

The main goal of the thesis is to improve the performance of PEM methods for inversion prediction. Furthermore, the thesis aims to obtain a better insight into the global knowledge of human inversions by the full integration of the information currently available. The study is divided into four objectives complementary to each other, that are focused on different aspects of the study of inversions. The specific objectives are:

1. Study the theoretical framework of the paired-end mapping (PEM) for inversions and develop a new algorithm focused on improving the reliability and the accuracy of the predictions.

2. Generate reliable polymorphic inversion predictions from available PEM data in the human genome.

3. Design the first database of polymorphic inversions in the human genome and generate the best non-redundant catalogue of independent inversions linked to all available related data.

4. Describe the global pattern of human inversion polymorphism based on the most reliable information stored in our database.

Manuscript in preparation
Alexander Martínez-Fundichely[1], *Meritxell Oliva*[1], *David Vicente-Salvador*[1], *Cristina Aguado*[1], *David Izquierdo*[1], *Sergi Villatoro*[1], *Angel Novoa*[1], *Xavier Estivill*[2], *José Ignacio Lucas-Lledó*[1], *Marta Puig*[1], *Juan R. González*[3], *Evan E. Eichler*[4], *Sònia Casillas*[1], *and Mario Cáceres*[1,5,*].

[1] *Institut de Biotecnologia i de Biomedicina,*
Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain.
[2] *Center for Genomic Regulation (CRG-UPF),*
Barcelona, Spain; 3 Center for Research in Environmental Epidemiology (CREAL)
[3] *Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain*
[4] *Department of Genome Sciences, University of Washington, USA*
[5] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.*

* *To whom correspondence should be addressed.*

Tel: +34 935868726; Fax: +34 935812011; Email: mcaceres@icrea.cat

CHAPTER 2

ACCURATE CHARACTERIZATION OF INVERSIONS IN THE HUMAN GENOME FROM PAIRED-END MAPPING DATA WITH THE GRIAL ALGORITHM

2.1 Summary

During the last years there has been a great interest in the characterization of genomic structural variation, and paired-end mapping (PEM) is the most widely used method to detect different types of these variants. However, compared to insertions and deletions, inversion prediction presents unique challenges. GRIAL is a new algorithm developed specifically to detect and map accurately inversions. It is based on geometrical rules

derived from expected inversion PEM patterns to cluster individual mappings belonging to each breakpoint, merge clusters into inversions, and refine breakpoint location. Using available fosmid PEM data from 9 different individuals, we have been able to predict several hundred inversions in the human genome, including many highly overlapping predictions that are consistent with complex events. In addition, by combining published data and experimental validation in the same individuals, we have created a gold standard of 56 polymorphic inversions in humans. Thanks to different quality scores to assess the reliability of the predictions according to the expected breakpoint support included in GRIAL, we have been able to identify misleading PEM patterns and their causes, and discard a big fraction of the predicted inversions as false positives. Among the main causes of inversion prediction errors are sequencing artifacts and individual sequence differences between repeats due to gene conversion. The comparison of GRIAL predicted inversions with other programs shows that GRIAL has higher sensitivity and precision in breakpoint location. Therefore, this analysis has allowed us to identify around 100 reliable human polymorphic inversions, which is the first step to determine their main characteristics and their functional impact.

2.2 Introduction

Genomic studies have discovered a large amount of structural variation (SV) in humans and other organisms [Feuk *et al.*, 2006; Weischenfeldt *et al.*, 2013; Yalcin *et al.*, 2012; Sudmant *et al.*, 2013]. However, the level of characterization of the different types of SVs is still quite heterogeneous. Most studies have focused on unbalanced SVs, including insertion-deletion (indels) or copy number variants (CNVs) [1000 Genomes Project *et al.*, 2012; Conrad *et al.*, 2010; Redon *et al.*, 2006], which can be easily validated and genotyped in multiple individuals using arrays. Due to the difficulty of its detection both at the bioinformatic and experimental level, other SVs like inversions have been poorly characterized. Inversions per se are typically balanced SVs and just change the orientation of a part of the genome, without resulting in alterations of the amount of DNA. Therefore, despite the early interest in inversions from studies in Drosophila, which discovered thousands of inversions and showed that they could have important effects in genetic recombination and adaptation [Hoffmann and Rieseberg, 2008; Kirkpatrick, 2010; Krimbas and Powell, 1992], until recently just a handful of polymorphic inversions had been characterized in some detail in humans [Feuk, 2010]. Nevertheless, many of them have shown to have functional consequences, either by the increased susceptibility to other rearrangements in parents of individuals with genomic syndromes [Antonacci *et al.*, 2009; Osborne *et al.*, 2001; Small and Warren, 1998], association to certain human diseases [Lakich *et al.*, 1993; Salm *et al.*, 2012], or increased female fertility [Stefansson *et al.*, 2005].

Thanks to the advances in DNA sequencing, the most used method for the genome-wide prediction of all kind of SVs is that of end-sequence profiling (ESP) or paired-end mapping (PEM) [Alkan *et al.*, 2011], which was first introduced to identify polymorphic SVs in the human genome [Tuzun *et al.*, 2005; Volik *et al.*, 2003]. This technique is based on the analysis of the mapping profile to a reference genome of the two end sequences (also known as paired-ends or mate-pairs) of a large number of fragments of known length. Abnormal or discordant PEM patterns could be evidence of putative SV breakpoints, and PEM has been used extensively with libraries of different fragment sizes [Ahn *et al.*, 2009; Kidd *et al.*, 2008; Korbel *et al.*, 2007; McKernan *et al.*, 2009; Pang *et al.*, 2013; Wang *et al.*, 2008]. The main advantage of this technique is that it has higher sensitivity to detect balanced SVs over array-based methods. In addition, it locates breakpoints with relatively good resolution, which depends on the size of the fragments used, giving us the opportunity to gain information on the effects and mechanism of generation of these changes. PEM studies have therefore resulted in the prediction of several hundred inversions in the different individuals analyzed [Ahn *et al.*, 2009; Kidd *et al.*, 2008; Korbel *et al.*, 2007; McKernan *et al.*, 2009; Pang *et al.*, 2013; Wang *et al.*, 2008]. However, we do not yet understand well the source and amount of false positives and false negatives generated by these analyses [Lucas Lledó and Cáceres, 2013; Onishi-Seebacher and Korbel, 2011].

One of the critical steps in SV prediction by PEM is the bioinformatic analysis of the data, and the algorithms to predict SVs are constantly growing both in number and sophistication. Briefly, three main strategies have been used to identify SVs both alone or in combination [Medvedev *et al.*, 2009]: the clustering of consistent discordant PEM evidences that support the same SV (*e.g.* PEMer [Korbel *et al.*, 2009], BreakDancer [Chen *et al.*, 2009], SVDetect [Zeitouni *et al.*, 2010] or GASV [Sindi *et al.*, 2009]), the analysis of the read-depth of concordant and discordant mappings per region (*e.g.* GASVpro [Sindi *et al.*, 2012], PESV-Fisher [Escaramís *et al.*, 2013], or HYDRA [Quinlan *et al.*, 2010], with local de novo assembly), and the identification of breakpoints using split-reads (*e.g.* DELLY [Rausch *et al.*, 2012]). Other algorithms are capable of managing multiple possible read locations and reduce the mapping mistakes by using probabilistic methods to define the correct location of each paired-end (*e.g.* Variation-Hunter (VH) [Hormozdiari *et al.*, 2009, 2010] or GASVpro [Sindi *et al.*, 2012]. However, most of these methods have been designed with several types of SVs in mind and have not devoted special attention to inversion characteristics in order to improve their prediction. In addition, it is suspected that the rate of false positives in most programs could be quite high [Handsaker *et al.*, 2011]. Thus, although in theory PEM should be an excellent technique for SV detection, the low reliability of the predictions reduces considerably the usability of the data.

In particular, inversions could be more difficult to predict than other types of SVs, like indels or CNVs [Lucas Lledó and Cáceres, 2013; Onishi-Seebacher and Korbel, 2011]. First, inversions are identified so far mainly just by a specific signature of one paired end being mapped in the unexpected orientation, which could be also due to other causes. Second, inversions should be defined by a combination of different orientation-discordant PEMs that correspond to the two breakpoints. Finally, inversions usually occur between highly-identical inverted repeats (either segmental duplications (SDs) or repetitive elements) that are especially challenging for PEM methods [Lucas Lledó and Cáceres, 2013; Onishi-Seebacher and Korbel, 2011]. The difficulty of inversion detection is exemplified by the recent analysis of the 1000 Genomes data, which described insertions and deletions but did not attempt to predict inversions [1000 Genomes Project *et al.*, 2012]. Therefore, there is a need for improved detection methods that provides a more accurate and reliable picture of the inversion polymorphism in the human genome. It is also important to have a good validated inversion set to benchmark the different SV prediction algorithms.

In this study we have developed a new specialized algorithm, GRIAL, based on inversion-specific characteristics to define inversions accurately from PEM data, eliminate most false positives through a prediction scoring system, and refine the breakpoints to the minimum interval. For that, we have used the most complete PEM dataset so far, including fosmids of 9 individuals [Kidd *et al.*, 2008]. We have identified a few hundred reliable inversions and uncovered some of the most common sources of error in PEM predictions. We have also benchmarked our algorithm against a gold standard of experimentally validated inversions, including many validated in this work, and found that it performs considerably better than other commonly used programs.

2.3 Materials and methods

2.3.1 Theoretical framework for inversion PEM

The inversion process generates a particular PEM pattern characterized by discordant orientation of the two mapped ends, and the exact orientation observed will be different depending on the breakpoint spanned (either BP1 or BP2) (Figure 2.1). Inversions specific characteristics allow us to define a set of mathematical rules that all the PEMs supporting an inversion with the same breakpoints should fulfill. These rules are mainly based on the geometrical relationship between the pairs of mapped ends derived from the change of orientation of the segment between the breakpoints. Contrarily to other SVs, each inversion has two breakpoints that are identified by different sets of paired-end reads discordant in orientation. Therefore, the rules for inversion PEM are divided in: (1) identification of a discordant region that is evidence of a single inversion breakpoint; (2) merging of two breakpoints of the same inversion to avoid redundant predictions, and (3) refinement of the intervals where the breakpoints are most likely located. The formalization of the expected inversion patterns is very important to determine more accurately which PEMs are really indicating the presence of a valid inversion and separate them from discordant mappings caused by other changes or complex regions with multiple SVs.

Different sequencing strategies have been used to obtain the pairs of end sequences, which sometimes have been distinguished (a bit confusingly) as paired-ends or mate-

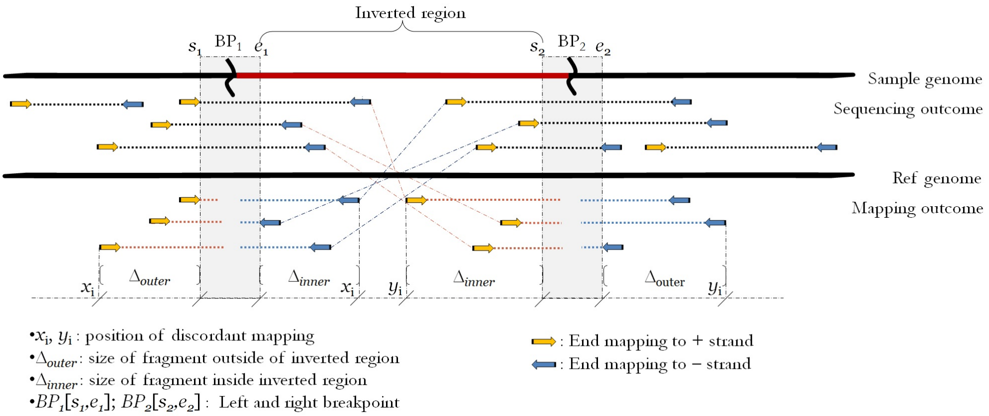

Inverted region
s_1 BP_1 e_1
s_2 BP_2 e_2
Sample genome
Sequencing outcome
Ref genome
Mapping outcome
x_i Δ_{outer}
Δ_{inner} x_i
y_i Δ_{inner}
Δ_{outer} y_i
•x_i, y_i : position of discordant mapping
•Δ_{outer}: size of fragment outside of inverted region
•Δ_{inner}: size of fragment inside inverted region
•$BP_1[s_1,e_1]$; $BP_2[s_2,e_2]$: Left and right breakpoint
: End mapping to + strand
: End mapping to − strand

Figure 2.1: Schematic diagram of an inversion and the resulting paired-end mappings. - The diagram shows the pattern of paired-end mapping that results from the inversion (red fragment). The yellow arrows represents reads that maps in positive strand, and blue arrow represent reads that maps in negative strand.

pairs [Alkan *et al.*, 2011; Medvedev *et al.*, 2009]. These sequencing strategies lead to different relative orientation of the two end sequences spanning each breakpoint. For simplicity, for the rest of the paper we describe the results from a typical Sanger [Tuzun *et al.*, 2005] or Illumina paired-end sequencing experiment, which sequences each end of a fragment from a different DNA strand (the methods described apply to the two sequencing strategies, just changing the expected mapping orientation). According to this, both end reads of discordant pairs will map in forward orientation (+/+) if they correspond to BP1, or in reverse orientation (−/−) if they correspond to BP2 (see Figure 2.1). A brief description of the geometrical rules for inversion definition is listed below according to the terms described in the Figure 2.1.

(1) Identification and clustering of discordant PEMs that are consistent with the same inversion breakpoint:

Rule 1.1: For 2 PEM to belong to the same inversion BP, all should have the same orientation and the distance between the two ends have to be within the variation of the library size.

$$\Delta_{inner} \text{ and } \Delta_{outer} = \begin{cases} |x_i - x_j| \\ |y_i - y_j| \end{cases} \leq \max L_f$$

Rule 1.2: Sum rule: The sum of the position of the two ends must range around the variation of the fragment template length therefore, it can not vary more than the maximum variation of the insert size in the library.

$$|\delta(x_i + y_i)| = |\delta L_f| \leq |\max \Delta L_f|$$

(2) Merging the clusters of the two breakpoints of the same inversion:

Rule 2.1: Maximum distance between the limits of the clusters must not exceed twice the library length plus the expected variation ($\max L_f$), and the beginning of the positive cluster Cls^{++} must not exceed the beginning of the negative cluster Cls^{--}).

$$|\Delta Cls| = \begin{cases} BP1 = |Cls^{++}_{\min x} - Cls^{--}_{\max x}| \\ BP2 = |Cls^{++}_{\min y} - Cls^{--}_{\max y}| \end{cases} \leq 2 \max L_f$$

Rule 2.2: The difference between the sum of the positions of the two ends of PEM+ and PEM− clusters should be within the range of the double of minimum and the double of maximum of insert size in the library.

$$2 \min L_f \leq |(x_i + y_i)_{++} - (x_j + y_j)_{--}| \leq 2 \max L_f$$

(3) Refinement of the intervals where the breakpoints are most likely located:

Rule 3.1: Breakpoints are defined by the position of the most internal mapping outside the inversion and the closest within (either + or −).

$$BP_1 \begin{cases} s = & Cls^{++}_{\max x} \\ e = & \min(Cls^{++}_{\min y}, Cls^{--}_{\min x}) \end{cases}$$

$$BP_2 \begin{cases} s = & \max(Cls^{++}_{\max y}, Cls^{--}_{\max x}) \\ e = & Cls^{--}_{\min y} \end{cases}$$

Rule 3.2: Alternatively, the limit can be defined by adding to the closest PEM, the difference between the maximum library length minus the distance between the closest and the furthest PEM.

$$BP_1 \begin{cases} s = \max(Cls^{++}_{\max x}, Cls^{--}_{\max x} - \max L_f) \\ e = \min(Cls^{++}_{\min x} + \max L_f, Cls^{--}_{\min x}) \end{cases}$$

$$BP_2 \begin{cases} s = \max(Cls^{++}_{\max y}, Cls^{--}_{\max y} - \max L_f) \\ e = \min(Cls^{--}_{\min y}, Cls^{++}_{\min y} + \max L_f) \end{cases}$$

Several of these rules are applicable to other types of SVs and have been previously described (see for example, [Hormozdiari *et al.*, 2009]). However, the newly derived paired-end Sum rule to define between which PEMs belong to the same breakpoints (1.2) and to merge the clusters of the two inversion breakpoints (2.2) are specific of this work. In addition, we have also defined new rules to define breakpoints more accurately. It is important to note that these rules just depend on the expected size of the library fragments generated for the PEM and should be very robust. However, one problem is that large indels within the discordant fragments above the expected error of the library size distribution would appear as inversions with slightly different breakpoints.

2.3.2 Inversion prediction scores

The amount of discordant support and the read-depth profile in a particular region have been used previously to assess the reliability of SV detection [Escaramís *et al.*, 2013; Sindi *et al.*, 2012]. The accurate breakpoint definition provided by the above geometrical rules allowed us to develop two new scores to measure the quality of inversion predictions and eliminate additional false positives.

1. **Discordant/Concordant ratio (*D/C-score*).** Assuming a diploid genome, the ratio of the number of discordant fragments supporting a predicted breakpoint interval with respect to the total number of mappings (both concordant and discordant) across the same interval (Equation 2.1) is expected to be 1 for homozygote inversions and 0.5 for heterozygote inversions (half the reads mapped at each side of the breakpoint interval should be concordant and the other half should be discordant). Deviations from these expected values could be considered signal of erroneous inversion predictions.

$$\mathit{DC\text{-}ratio} = \frac{\sum_{i=1}^{N} Disc_i}{\sum_{i=1}^{N} Disc_i + Conc_i} \tag{2.1}$$

   Where, *Disc* is the total number of Discordant PEMs supporting the inversion at each side of the BP1 and BP2 intervals, and *Conc* is the number of PEMs not supporting the inversion at each side of the the BP1 and BP2 intervals.

   To calculate this score, the observed *DC-ratio* was calculated for all (N) individuals together according to Equation 2.1. The expected *DC-ratio* was calculated assuming that individuals with only discordant PEM are homozygous for the inversion and those with at least 1 discordant PEM are heterozygous for the inversion, which is a conservative estimate. In addition, to avoid overestimating the number of concordant PEMs in breakpoint regions located within highly identical inverted SDs, we removed all concordant PEMs in which one end is completely located in a region of 100% identity between the SDs according to the alignments available in UCSC (non-informative PEM). Then, inversion predictions for which the expected *DC-ratio* is more than double the observed one and the expected vs observed ratio test *P*-value was less than 0.05 were filtered.

This score has the advantage that uses the concordant mapping in the same region to estimate the expected number of discordant PEM depending on the predicted inversion genotype, and therefore it is not sensitive to specific characteristics of the region that affect the number of mapped reads. In addition, for valid inversions, the *DC-ratio* for each individual can be also used to predict the inversion genotype.

2. **Discordant support (DS) score**. The second score calculates the expected support for the inversion considering a homogeneous read-depth for each chromosome taking into account the predicted inversion size, the size of both breakpoint intervals, the length of the library, and the mappability in the region according to the presence of SDs and repetitive elements. In this case, we filtered all predictions for which the Poisson distribution *P*-value of the observed discordant support was less than 0.001.

2.3.3 Implementation of the inversion prediction algorithm

In order to predict inversions accurately from PEM data, the above rules and scores have been implemented into a Perl package named GRIAL (Geometric Rule Inversion Algorithm; Spanish for Grail). The GRIAL package is distributed under the GNU General Public License (GPL) and can be downloaded and run locally without limitations from http://grupsderecerca.uab.cat/cacereslab/grial. It has been tested under different Linux distributions, Microsoft Windows and Mac OS . The different steps of the algorithm are represented in Figure 2.2 and a complete description of the program can be found in the GRIAL user's manual. The work-flow of the program has been divided in 5 parts that include: (1) the read of input data and the configuration file, where the user sets the parameters that GRIAL will use throughout the process; (2) the clustering of compatible discordant PEMs, in which all possible candidate clusters are considered and the best candidate clusters (*i.e.* those with the highest support, or the lowest variance of the sum coefficient) are selected to create an unbiased non-redundant set; (3) the prediction of inversions from matching $+/+$ and $-/-$ PEM clusters of the two breakpoints or single breakpoint clusters ($+/+$ or $-/-$) and the refinement of the location of breakpoints; (4) the calculation of the scores to assess the reliability of the inversion predictions; and (5) the writing of the output data and the generation of the final report of predicted inversions. Due to the parallelization of the initial process by chromosome and DNA strand, the program runs relatively fast.

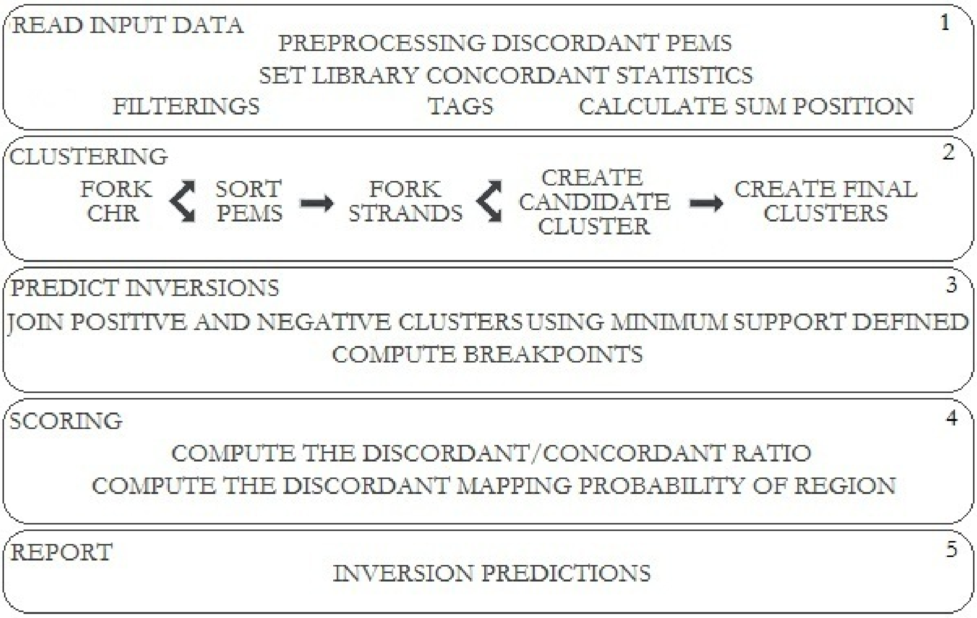

READ INPUT DATA
1
PREPROCESSING DISCORDANT PEMS
SET LIBRARY CONCORDANT STATISTICS
FILTERINGS
TAGS
CALCULATE SUM POSITION
CLUSTERING
2
FORK CHR
SORT PEMS
FORK STRANDS
CREATE CANDIDATE CLUSTER
CREATE FINAL CLUSTERS
PREDICT INVERSIONS
3
JOIN POSITIVE AND NEGATIVE CLUSTERS USING MINIMUM SUPPORT DEFINED
COMPUTE BREAKPOINTS
SCORING
4
COMPUTE THE DISCORDANT/CONCORDANT RATIO
COMPUTE THE DISCORDANT MAPPING PROBABILITY OF REGION
REPORT
5
INVERSION PREDICTIONS

Figure 2.2: Work-flow of the different steps and processes in the GRIAL algorithm. - Description of the main tasks to predict inversions from paired-end mapping

2.3.4 Calculation of inversion prediction complexity

Determining the level of complexity of the predictions is also very important for the interpretation and experimental validation of inversions. In some regions, inversion discordant PEMs are not actually compatible with a single inversion and give rise to multiple predictions with different degree of overlap. To establish the complexity we have defined the equivalent prediction $pred^*$ (Equation (2.2)), that gives an idea of really how many different inversions could be present in a region. It is calculated by computing the shared fraction between the overlapped predictions with respect to the potential total as:

$$pred^* = \frac{pred \times L}{\sum\limits_{i=1}^{pred} l_i} \tag{2.2}$$

Where "$pred$" is the total number of predictions overlapping in the region, "L" is the total length of the region, "l_i" is the length of the different predictions. This value is 1 if all predictions overlap completely and 1.5 if two predictions overlap reciprocally by 50%.

Locations with more than 4 predictions in which $pred^* > 2$ were defined as high complex regions. Locations in which $pred^* < 1.5$ and there are 2 or 3 predictions with an overlap between them higher than 70% were defined as simple regions, suggesting that there is likely only one inversion. The predictions in between are classified as low complex regions, indicating that there are at most two different inversions.

2.3.5 Data set used for inversion definition.

To test and optimize the algorithm we used the extensive paired-mapping resource from fosmids of 9 individuals of diverse origin [Kidd *et al.*, 2008; Tuzun *et al.*, 2005] belonging to the Human Genome Structural Variation Project (http://hgsv.washington.edu/). This corresponds to Sanger sequences ($\sim 300 - 800$ bp) of the two ends of $\sim 35 - 40$ kb fragments. Previous mapping data from the concordant and discordant fosmids of the 9 individuals [Kidd *et al.*, 2008] to the HG18 human genome assembly were downloaded from http://mrhgsv.gs.washington.edu/cgi-bin/hgTables. In order to increase the power of inversion detection with GRIAL, all libraries were merged in a unique big dataset, since no significant differences in the size of the fosmids were found. For all the libraries combined, the average fosmid length is 39222 bp, the standard deviation is 2691 bp, the minimum and maximum lengths are respectively 25163 bp and 49224 bp, and there are a total of 12162 inversion discordant fosmids.

Before applying the GRIAL algorithm, several filtering steps were applied to the data to eliminate false positives. First, we identified discordant fosmids that could have been artifactually duplicated during the construction of the library and have virtually the same location [Tuzun *et al.*, 2005]. According to the different sequencing strategy used and the analysis of the distance between the fosmid mapped ends in each library, the criteria to identify duplicated fosmids was a distance between the two mapped ends of fosmids of the same individual and same orientation of ≤ 17 bp for G248 and ≤ 50 bp for ABC7-14 libraries. This identified 1624 fosmids as potentially duplicated, and several of them have been shown to be identical by whole fosmid sequencing. Fosmids with this error have been weighted down proportionally to the number of fosmids that could have been duplicated since they provide redundant information and overestimate the discordant signal in the region. In addition, we found 3366 discordant fosmids in which the best mapping of the two ends overlapped by $> 50\%$. Many of those were due to incorrect or partial mappings and were excluded from the input dataset. Finally, all discordant fosmids mapping at random chromosomes were also eliminated.

2.3.6 Benchmarking of GRIAL against other methods and real inversion data

To check the performance of GRIAL, we compared the inversions predicted by Kidd *et al.* [2008] and those obtained by other available methods. These predictions were also compared to the gold-standard of previously validated inversions and our own set of validated inversions using bioinformatic and experimental methods (see below). In order to compare the 251 inversions predicted by Kidd *et al.* [2008] using as reference the HG17 human genome assembly, the HG18 positions of the fosmids included in Kidd *et al.* [2008] predictions were used to determine the inversion limits. The other algorithms were run using default parameters and the same original data set as GRIAL (before the filtering process). In all cases the minimum discordant support for inversion prediction was 2. In addition, since each method predicts breakpoint coordinates differently, they were modified in order to make predictions comparable by adding the maximum fosmid length to create an interval where the breakpoint should be located.

2.3.7 Validation of predicted inversions and building of inversion gold-standard data set

To obtain more information on the methods performance and to generate a catalogue of human polymorphic inversions, we validated inversion predictions in two different ways. First, to discard possible false positives, PEM support of some predictions was manually re-analyzed using all available data, including comparison with human assemblies HG19 (plus additional patches) [Church *et al.*, 2011], HuRef [Levy *et al.*, 2007] and additional BAC and fosmid sequences of the same region. This was done by re-mapping of the fosmid end sequences and by local alignment of sequences with Blat [Camacho *et al.*, 2009] and Megablast [NCBI Resource, 2013]. Second, experimental validation of some of the predicted inversions was carried out by PCR, especially those differentially predicted by GRIAL. Two pairs of primers were designed at each side of the two predicted breakpoints using Primer3 [Rozen and Skaletsky, 2000] and the two orientations, HG18 reference (standard or *Std*) and inverted (*Inv*), were tested by PCR amplification of DNA from the same 9 individuals from which the fosmids were derived and HuRef (J. Craig Venter) DNA. PCR was performed in 25 μl reactions with $50-100$ ng of DNA, 1.5 U of Taq DNA polymerase (Biotherm or Roche), 0.4–0.8 μM of each primer, 0.8 mM dNTPs, 1.5 mM $MgCl_2$, and 1× Taq DNA polymerase buffer, by an initial denaturation of 5 min at 95°C, followed by 35 cycles at 95°C for 30 s, $59-62$°C for 30 s, and 72°C for $30-120$ s depending on the template size, and a final extension at 72°C for 7 min. PCR products were analyzed by gel electrophoresis on $1.5-2\%$ agarose gels stained with ethidium bromide and for those inversions in which the *Inv* orientation sequence was not available, the amplification products were purified and sequenced to determine the exact location of the breakpoints. DNA was extracted from Epstein-Barr virus-transformed B-lymphoblastoid cell lines of each individual as previously described [Aguado *et al.* 2013 submitted (Appendix C)] or obtained directly from Coriell Cell Repositories (Camden, New Jersey, USA) (HuRef).

The gold-standard inversion data set was generated from all the validated inversions found in the literature and those validated in this work. Due to possible assembly errors or chimeric fosmids, no inversions were considered validated based only in sequence information, such as that of the HuRef genome [Levy *et al.*, 2007] or whole-sequenced fosmids [Kidd *et al.*, 2010], without additional independent experimental validation. In addition, only previously PCR validated inversions with sequence support for the breakpoints were considered [Korbel *et al.*, 2007; Lam *et al.*, 2010]. Breakpoint positions were refined by the comparison of the *Std* and *Inv* sequences. For inversions with inverted repeats (IRs) at the breakpoints, the IR sequences in *Std* and *Inv* orientation were aligned using Muscle [Edgar, 2004] to identify sequence changes between the paralogous copies of the repeats and the point where these variants get exchanged due to the inversion. In these cases, the breakpoint intervals were defined by three or more consecutive paralogous sequence variants (PSVs) indicating a recombination between the IRs in the inverted sequences. Finally, four identified assembly errors in HG18 were also included in the comparison to increase the sample size.

2.4 Results

2.4.1 Prediction of human polymorphic inversions with GRIAL

In order to get a better idea of the real inversion set in the human genome and predict accurately their breakpoints, we tested our newly developed inversion prediction algorithm GRIAL with the PEM data of fosmid libraries of 9 different individuals already mapped to HG18 [Kidd *et al.*, 2008]. These libraries were previously used to predict a total of 251 inversions [Kidd *et al.*, 2008] and are one of the most complete data sets for SV detection in humans. After doing an ANOVA with the previously identified concordant set to ensure that the fosmid libraries do not show significant differences in template length, we made a unique pool to increase the coverage of the human genome and the power to detect inversions, especially for small ones [Lucas Lledó and Cáceres, 2013]. Then, we selected the previously identified 12162 inversion discordant fosmids, and 3366 discordant PEM were discarded because of high overlap between the mappings of the two ends and 1624 were marked as putative duplicated copies during library construction in the GRIAL preprocessing step (see Materials and Methods).

Considering a minimum support of two discordant fosmids, GRIAL predicted 636 inversions located at 306 regions. Of those, 220 (34.6%) inversion predictions have PEM support of both breakpoints and in 416 (65.4%) only one breakpoint has been detected. In addition, 201 inversions have only support of a single PEM in each breakpoint or a single PEM in two individuals, and are identified just by the merging of all the information. To check the reliability of the predictions we used two filtering scores based in the ratio of the number of discordant and concordant PEM that support one or the other orientation (*DC-score*) and the expected discordant support (*DS-score*) (see Materials and Methods.) After applying GRIAL with the DC and DS scores, also named as GRIAL+, 414 (65.1%) predictions were filtered and the final high quality set consists of 222 (34.9%) reliable predictions located in 187 regions. This suggests that the number of false positive inversion predictions from PEM data is extremely high.

One interesting observation from GRIAL results is that there are many regions with several overlapping inversion predictions supported by PEM that are not consistent with the same inversion breakpoints. These complex predictions have been classified as unique and simple regions with only one putative inversion (47.5%), low-complex regions with two putative inversions (13.2%), and high-complex regions if there appear to be above three putative inversions (39.1%) (Table 2.1). On the GRIAL+ results this degree of complexity is considerably reduced, with clearly most predictions (70.3%) at simple regions, and only 18.0% at high complex regions (Table 2.1). In addition, when the association of these regions with gaps in the genome reference assembly was examined, we observed a clear enrichment of assembly gaps within or at a distance smaller than the template length of high complex regions (Table 2.1). This suggests that many of these predictions could be due to wrong PEM signals, caused by problems in the genome assembly, and gives support to GRIAL+ score results.

Table 2.1: Summary characteristics of inversions predicted by GRIAL and GRIAL+.

Region type	GRIAL predictions (GRIAL+)	Inversion regions (GRIAL+)	Predictions in gaps (GRIAL+)
Unique and simple	302 (156)	253 (151)	14 (7)
Low complex	84 (26)	25 (17)	16 (7)
High complex	250 (40)	28 (19)	128 (11)
Total	636 (222)	306 (187)	158 (25)

To assess GRIAL performance, we compared the results to the 251 inversion predictions of Kidd *et al.* [2008] after translating the coordinates to HG18. In general, overlap of Kidd *et al.* [2008] predictions with those of GRIAL was good, but GRIAL generated many more predictions. This is expected by the use of geometrical rules that refine the inversion predictions in complex regions and the merging of the different individual libraries in a single dataset that improves the capacity of detection of small inversions with low support in each separate library. Figure 2.3 illustrates the difference between the prediction strategies. Of Kidd *et al.* [2008] predictions, 72.5% (182) were detected by 233 GRIAL predictions. The remaining 27.5% (69) were not detected by GRIAL due to filtered PEMs in the GRIAL's preprocessing step (53) (mainly predictions supported only by possibly duplicated fosmids) or predictions that do not match with GRIAL defined inversions due to differences in the clustering methods (16), mainly in complex regions. On the other hand, there were 403 predictions found only in GRIAL. As previously mentioned, the majority of them have a low fosmid support from different individuals or different breakpoints that were not considered by the conservative threshold of Kidd *et al.* [2008]. The rest (158) are mostly predictions in complex regions, in which GRIAL has generated additional overlapping predictions due to the inconsistency between supporting PEM, whereas they were originally merged in one prediction [Kidd *et al.*, 2008]. When only the most reliable predictions after the scoring process are considered, the number of GRIAL+ unique predictions is reduced by 63.4% and only 40.2% of Kidd *et al.* [2008] predictions correspond to those of GRIAL+, with the rest being filtered out. Therefore, although the final number of inversion predictions from GRIAL+ and Kidd *et al.* [2008] are similar, the actual subsets of inversions are quite different.

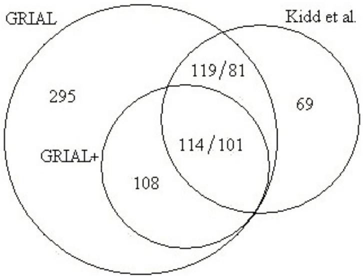


Figure 2.3: Venn diagram of the comparison of the inversion predictions of Kidd *et al.* [2008] and GRIAL. - Numbers of common and specific predictions by each method are represented inside the circles, with the numbers corresponding to the GRIAL and Kidd *et al.* [2008] subsets separated by "/" . The GRIAL+ results are represented by a smaller circle within that of GRIAL.

2.4.2 Validation of predicted inversions and main types of errors in inversion prediction

So far, limited information exists about validated polymorphic inversions in the human genome. This includes ~40 inversions detected with different techniques and different degree of precision in the breakpoint definition [Antonacci *et al.*, 2009; Feuk, 2010; Feuk *et al.*, 2005; Giglio *et al.*, 2002; Gilling *et al.*, 2006; Gimelli *et al.*, 2003; Korbel *et al.*, 2007; Martin *et al.*, 2004; Osborne *et al.*, 2001; Pang *et al.*, 2013; Stefansson *et al.*, 2005] and [Aguado *et al.* 2013 submitted (Appendix C)]. In addition, in this work we developed PCR assays to analyze the two BPs of 23 predicted inversions by GRIAL+, of which 10 are new and 13 had been validated to some degree previously. Taking advantage of these assays, the inversions were genotyped in the same individuals were the fosmids come from. The genotyping information was combined with BP sequences, either previously available or generated during this work, to locate accurately the inversion breakpoints. Of the analyzed inversions, 22 were validated and one turned out to be an inverted duplication of the *SLC2A14* gene. In particular, we tested 7 specific GRIAL predictions with minimum support, of which 4 were supported by one fosmid in each breakpoint and 3 by one fosmid of the same breakpoint in different individuals and all of them were validated. This shows that the combination of all the data, together with the use of strict rules and scores, can predict accurately additional inversions.

On the other hand, by bioinformatic analysis of all available sequences (including remapping of paired-end sequences and analysis of fully sequenced fosmids or other available human sequences) or by PCR amplification of predicted BPs we were able to identify a series of false discordant PEMs that resulted in 53 incorrect inversion regions. For the sequence analysis we tried to use as much of the data available as possible, although this set was biased to inversion predictions that were filtered by GRIAL or GRIAL+. In addition, we tested by PCR three inversion breakpoints validated by fully sequenced fosmids that were filtered by GRIAL because they were supported only by fosmids that seemed to have been duplicated. None of them corresponded to a real inversion, suggesting that they were caused by the formation of chimeric fosmids. Therefore, these results indicate that PEM analysis, particularly in the case of inversions, has a considerable amount of false positives due to different sources and it is important to apply appropriate filters to reduce it. Furthermore, thanks to the exhaustive validation process, we were able to determine the most common errors in inversion prediction, some of which are general of the PEM method and others are specific to the technique used.

First, there are several errors associated with incorrect mapping due to sequence differences between individuals or errors in the genome reference sequence (not counting regions assembled in the opposite orientation). For example, there is a GRIAL specific inversion prediction supported just by one PEM of each breakpoint caused by a polymorphic repetitive element insertion present in the analyzed individual and HuRef, but not in HG18. Similarly, there could be incorrect mappings due to missing sequences in the reference genome. This is the case of two inversion predictions associated with gaps in HG18 that are filled in HG19, generating a new copy of a duplicated sequence and resulting in a fully concordant PEM pattern. Also, there are four inversion predictions supported by discordant PEM patterns that disappear in a HG19 patch including 75 kb of extra sequence that extends the previously identified SDs [Aguado *et al.* 2013 submitted (Appendix C)]. In addition, the most common source of mapping errors is the sequence divergence between homologous and paralogous IRs. If gene conversion happened between two IRs in the target genome, the two inverted copies will be virtually identical, at least along the conversion tract, in contrast to the situation in the reference genome, where the two copies may be distinguishable. Then, sequenced reads originated from one copy from the target genome could be incorrectly mapped to the alternative copy in inverted orientation, which is a clear sign of an inversion breakpoint. Something similar could happen if one of the IR copies has a specific indel or increased divergence in the reference genome. Thus, for all predictions supported only by PEM with one end mapping completely within IRs, is difficult to know the correct origin of the end sequence and are low reliable. Another type of mapping error is due to the previously detected mixing of two haplotypes in the HG18 assembly [Antonacci *et al.*, 2010].

Second, in a few other cases we have been able to identify particular PEM patterns that look like inversions but they are not. The most typical one is the inverted duplication in tandem, that usually creates an inversion PEM pattern in just one breakpoint corresponding to the point of insertion, and depending on the size of the duplication could be accompanied by an insertion signal or not. Two examples are a ∼20 kb inverted duplication of the 5' end of the *SLC2A14* gene that was validated by PCR and analysis of fosmid sequences, and a 300 kb inverted duplication in Chr. 16 validated by optical mapping [Teague *et al.*, 2010], although there could be other small ones.

Finally, we have detected several inversion predictions with support of only one breakpoint from PEMs that all share the location of one end. By analysis of whole sequenced fosmids it was found that they mapped 100% concordantly in the human genome and that the end read was actually generated from an internal region of the fosmid, creating the fictitious discordant mapping. In fact, practically in all the cases the conflicting sequence was generated with the reverse primer and a hit for the primer was located close to the beginning of the sequence. Therefore, these false positive predictions were apparently caused by misspriming during sequencing and the generation of an internal sequence instead that from the fosmid end. In addition, this type of misspriming could be also responsible of some apparent deletion calls in the fosmid PEM data.

2.4.3 Benchmarking of GRIAL against other methods and real inversion data.

To check the performance of GRIAL we compared its results with those obtained in three other available algorithms: PEMer [Korbel *et al.*, 2009], Variation Hunter (VH) [Hormozdiari *et al.*, 2009] and GASV [Sindi *et al.*, 2009]. Since each program defines the breakpoints a bit differently, we modified the coordinates taking into account the maximum length of the fosmids to make all of them comparable. Also we considered that PEMer predictions and most of VH predictions just address one breakpoint, which means that these algorithms have in general two predictions for each inversion detected.

First, we compared directly the total number of predictions from each method, which is more or less similar between all of them (Table 2.2). Apart from GASV that shows a lower degree of overlap, more than 80% of GRIAL predictions are found with the other methods, with PEMer being the program that matches more closely

GRIAL predictions. However, for every program there is a fraction of predictions that is exclusive and different from that of GRIAL. The only exception is GASV-max, which includes almost all GRIAL predictions, at the cost of making many more predictions. The main differences between algorithms are related to the way the PEM clusters are done, which results in slightly different inversions predicted, and specific criteria for some programs, like the elimination of long inversions by VH [Hormozdiari *et al.*, 2009]. In addition, as already mentioned, GRIAL has the ability to predict low support inversions. In GASV and GASV-max there are considerably more specific predictions, which are probably related to the different strategy used and the higher constraint of GRIAL geometrical rules. In the case of GRIAL+, the different programs identify most of its predictions, with just a limited number of GRIAL+ specific inversions, but in all programs there is a high proportion of predictions (53−74%) that are filtered by GRIAL+.

Table 2.2: Comparison of inversions predicted by GRIAL and GRIAL+ with those of four other methods.

		GRIAL (636)		GRIAL+ (222)	
Method	Predictions	Common	Different	Common	Different
PEMer	720	659 (601)	61 (35)	197 (215)	523 (7)
VH	633	559 (521)	74 (115)	265 (207)	368 (15)
GASV	731	422 (444)	309 (192)	180 (179)	551 (43)
GASV-max	1398	811 (629)	587 (7)	289 (222)	1109(0)

For GRIAL and GRIAL+, the number of predictions within each category is shown in parenthesis.

Next, in order to check to what extent the predictions of the different methods are reliable, we have compared them with the above gold standard of 56 validated inversions as well as 53 regions with discordant PEM not associated with a real inversion (Table 2.3). To establish whether an inversion has been detected, for each method it was recorded if the predictions were located within the known region of the breakpoint for each inversion plus and minus the maximum fosmid length. The inversion is considered well detected if both predicted breakpoints are identified. If the prediction identifies one breakpoint but not the other, the detection is incomplete. Theoretically, each inversion should have only one prediction, and if there are more predictions they are counted as over detection and result in lower detection efficiency (ratio between the number of inversions detected and the number of predictions made). Of the 56 polymorphic inversions, 53 were detected at least in part by the two GASV configurations, and 3 could not be detected by any of the programs because they do not have a PEM signal in the libraries used. GRIAL was the second program that detected most inversions (51), but 9 of them were under the threshold of the GRIAL+ scores and were filtered. Nevertheless, regarding to the usefulness of the predictions is very important the detection efficiency, which is maximum for GRIAL+ (0.95), GASV (0.90), and GRIAL (0.78). On the opposite end, the lowest detection efficiency is that of GASV-max (0.52), in which the screening of all possible PEM patterns comes at a cost of generation too many predictions.

Another measure of the accuracy of the inversion predictions is the fraction of well-detected breakpoints and the distance between the breakpoints predicted and the real ones. The subset of 40 inversions that was detected by all algorithms allows us to compare the breakpoint precision of the methods. The median and the standard deviation of the breakpoint error distance shows that GRIAL is more accurate in breakpoint location than GASV-max and GASV. In all comparisons, PEMer and VH perform worse than the other algorithms, as could be expected because sometimes they only predict separately big intervals in which a breakpoint is located, instead of inversions with two defined breakpoint regions.

Table 2.3: **Detection performance of 56 validated inversions and 53 false inversion regions by different methods.**

	Validated inversions (56)						Inversion error regions (53)				
					BP error distance*			Incorrect disc. fosmids			
Method	Well det	Part det	Over det	Det eff	Median	SD	Pred	In	Out	Sensit	FPR
GRIAL	48	3	14	0.78	26268	11421	41	118	12	0.96	0.45
GRIAL+	39	3	2	0.95	26268	11421	21	40	90	0.79	0.33
GASV	47	6	6	0.90	27271	18122	47	130	0	1.00	0.47
GASV max	51	2	49	0.52	27000	11944	84	130	0	1.00	0.61
VH	29	12	28	0.59	30101	95642	48	124	6	0.77	0.54
PEMer	33	15	44	0.52	30101	140461	44	118	12	0.91	0.48

* The breakpoint error distance between the predicted and the real breakpoints was calculated using only the 40 inversions detected by all algorithms. Well det: Well detected; Part det: Partially detected; Over det: Over detection. Number of predictions that exceeds the number of inversions; Det eff: Detection efficiency; SD: Standard deviation; Pred: Predictions; In : Included fosmids; Out: Filtered fosmids; Sensit: Sensitivity; FPR: False positive rate

Finally, it is also very important to determine the false discovery rate of each method, and we have calculated the number of predictions including fosmids with erroneous discordant PEM that do not correspond to a real inversion. The reliability scoring process allows GRIAL+ to reduce drastically the number of false predictions, by more than half for most algorithms and four-times for GASV-max, and filters out 69% of the fosmids with incorrect inversion PEM patterns. Therefore, GRIAL+ provides the most reliable predictions, with a slightly lower sensitivity of 79%, and a false positive rate of 33% compared to the ~50% of the other methods.

2.4.4 Generation of final set of reliable inversions in the human genome

Despite the good performance of GRIAL+, there are still several false positives that could not be eliminated by the scoring process (21). This includes mainly misspriming errors that were identified by sequencing of whole fosmids (16), which are very difficult to detect statistically. To eliminate other possible inversion predictions with the same error, a subset of PEMs with the same orientation and one mapped end within a distance of < 400 bp was selected. For those PEMs, the blastn-short [Camacho *et al.*, 2009] program was used to identify hits to the primers used for the library sequencing (bitscore ≥ 15.7 and only one missmatch or gap in the 8 3' bases of the primer) within 100 bp of the beginning of the mapped sequence. A total of 93 inversion predictions, in which all the supporting PEMs (or all except one) had a hit internally within the fosmid which could generate an incorrect end sequence, were identified. These included the 28 incorrect predictions previously identified by fosmid sequence analysis, and all these predictions possibly affected from misspriming were eliminated.

The remaining predictions also include several other identified PEM errors (5) that have been previously mentioned, such as mapping errors caused by sequence differences between SDs, missing sequence in the assembly, or polymorphic indels. In addition, there are other four predictions that appear to correspond to inverted duplications and not real inversions.

Finally, a perfectly predicted inversion by PEM could be caused by the assembly of the region in the incorrect orientation in the human reference genome. In fact, the GRIAL+ list includes four assembly errors in HG18 that have been corrected in the

HG19 assembly or subsequent patches. To identify other potential assembly errors, we have carried identified inverted regions for which there is not PEM support for the *Std* orientation in any of the analyzed individuals. This analysis has identified 24 other problematic regions that should be experimentally analyzed using the DNA of the same BAC represented in the genome sequence to differentiate if they are low frequency inversions or assembly errors.

2.5 Discussion

Inversions are a type of structural variant that has been traditionally difficult to detect and validate. The PEM strategy offered an excellent tool for the genome-wide detection of inversions of an individual based on the mapping of the two ends in incorrect orientation [Alkan *et al.*, 2011]. However, despite the apparent simplicity of this signal, genomic data could be very noisy, which can lead to false inversion predictions or incorrect location of the real ones. To solve this problem we have developed a robust theoretical framework for the reliable detection of inversions by PEM and implemented it in a new algorithm specialized in inversion prediction. In addition, by extensive sequence analysis and experimental validation, we have built a gold standard of real polymorphic inversions in humans and regions associated with PEM errors to which the performance of different inversion prediction methods can be compared.

One of the main advantages of the GRIAL algorithm is that it uses a complete set of geometrical rules based on inversion characteristics and an exhaustive process of selection of the best clusters to predict inversions accurately. In particular, determining the right PEM clusters is crucial to be able to merge them and define the potential inversion breakpoints as precisely as possible. This accurate refinement of breakpoints has allowed us to develop two scores based in the expected support for a real inversion, which work well together for the elimination of many of the different types of false positives obtained from PEM predictions. We have seen that our method performs better than other available programs in several key aspects. First, the standard GRIAL algorithm predicts inversion as well or even better than the other available programs and shows always higher accuracy in breakpoint definition. Second, thanks to the scoring process, GRIAL+ has much lower false positive rates. Third, the use of strict rules and scores makes possible to reduce the minimum support needed to predict reliably an inversion from merged libraries of different individuals, avoiding an excess of wrong predictions and increasing the power to detect inversions at low coverage. Finally, it is also important to mention that GRIAL has shown good performance on inversion prediction from simulations of next-generation sequencing PEM data as well [Lucas Lledó and Cáceres, 2013].

The use of the geometric rules and scoring process to increase reliability comes also at a cost. In any prediction process is important to have a good compromise between sensitivity and specificity. In the case of GRIAL+, a few true inversions are filtered out based on the score results. However, most of them are mediated by highly identical SDs of > 100 kb and are very difficult to predict with 40 kb fosmids. In fact, these inversions are supported just by a few PEMs located within the SDs that generate multiple incompatible predictions (up to 7 for the chr.8p23 inversion). Therefore, the resulting PEM pattern does not really fit that of a valid inversion. On the other hand, the scores are not able to eliminate all the false positives either. In particular, one of the main limitations are small inversions with low support and breakpoint intervals close to the template size. In any case, the elimination of some real inversions is a reasonable price to pay to ensure the accurate prediction of the rest of inversions. In contraposition, other methods such as GASV-max detect all the inversions at the expense of generating a very high number of predictions and increase considerably the noise.

Another problem is that sometimes PEMs belonging to the same inversion are not merged correctly, generating several different predictions. For example, the presence of large indels close to the breakpoints could result in the separation of PEM from the same inversion in two predictions, as happened in HsInv1052 which has a 5 kb deletion close to BP1. Also, sometimes inaccurate mapping of some PEM due to divergence between SDs in different individuals can make that clusters are not compatible and affect the precision of the breakpoints and the scores. One case of that is HsInv1051, which is filtered in GRIAL+ because there are two predictions that include multiple PEMs of 6 individuals that do not have the inversion However, we have solved that by merging simple predictions with a high overlap that likely belong to one single inversion.

Other than that, many programs point to a complex landscape of human inversions predicted by PEM, with many PEMs being consistent with different inversion predictions in the same region. The cause of this complexity is not really known. Assuming that everything is correct, the immediate conclusion is that this complex regions point to different inversions with similar breakpoints. The phenomenon of breakpoint reuse and breakage hotspots and fragile sites is a common theme in genome evolution in mammals and *Drosophila* [González *et al.*, 2007; Pevzner and Tesler, 2003]. However, so far all the studied examples are related to complex SDs or problems of the region, such as gaps or missing sequence in the human genome [Aguado *et al.* 2013 submitted (Appendix C)]. Therefore, until further characterization can be carried out all regions with many overlapping predictions have to be considered suspicious.

The use of the PEM technique has extended exponentially for the detection of SVs in multiple normal and disease genomes. Our exhaustive analysis has allowed us to identify some of the common causes of false inversion PEM predictions. Apart from errors in the human assembly and a specific problem during sequencing, we have seen that one of the main limitations for PEM is the representativity of the human reference genome, and that in some regions the human genome does not represent the most common sequence in human populations. In particular, as we have mentioned, SDs could be the subject to rapid evolution among individuals by mechanisms such as gene conversion . Gene conversion is known to happen between IRs with appreciable frequency, at least on human chromosome Y [Rozen *et al.*, 2003] and could cause spurious PEM signals. In addition, this is especially problematic for next generation sequencing data generated from short reads and short templates, in which discordant mapping could be based in a few paralogous sequence variants between IRs. Therefore, it is important to find ways in the mapping and scoring step to minimize these possible errors and filter them out. One possibility would be to use not just one, but several well assembled genomes for PEM experiments. In the mean time, a mask of potentially problematic regions prone to erroneous predictions in PEM could be constructed.

The accurate prediction of SVs in general, and inversions in particular, is a key step for follow-up studies or otherwise the generated information is of little use. In this work we have shown that there is a potentially extremely high rate of false positives in inversion PEM predictions, which are higher than 50% in many cases. Our analysis indicates that the number of real polymorphic inversions in the human genome estimated from this data is probably around 100. This is a considerable lower estimate that the one initially predicted by other studies [Kidd *et al.*, 2008; Sindi *et al.*, 2009], but allows us to get a better picture of some of the main characteristics of inversions in the human genome and emphasizes the need of a careful analysis before making reliable conclusions of this kind of data. As an example, the number of inversions with breakpoint affecting the exonic sequence of genes has been drastically reduced in the curated data set. Another interesting question is to what extent the PEM predicted inversions are a good representation of human inversions. In this regard, only part of previously known inversions can be reliable detected due to technical limitations caused by the size of the IRs found at the breakpoints. In addition, inversions are more sensitive to variations in the physical coverage than indels and some of the smaller variants might be missed. Finally, many of the inversions described here are found just in one individual, which suggest that there could be many more inversions at low frequencies. Therefore, additional studies with more individual and a broader range of library sizes, from few kb to BACs might be necessary to capture the full set of polymorphic inversions in the human genome.

2.6 Acknowledgements

We thank Antonio Barbadilla for help with the mathematical and statistical aspects and Marta Morell and Raquel Rubio for the lymphoblastoid cell lines culture.

2.7 Funding

This work was supported by the European Research Council (ERC) Starting Grant 243212 (INVFEST) under the European Union Seventh Research Framework Programme (FP7) to M.C., a MAEC-AECI doctoral fellowship from the Ministerio de Asuntos Exteriores y Cooperación (Spain) to A.M.F., a FPI doctoral fellowship from the Ministerio de Educación y Ciencia (Spain) to M.O., and a Beatriu de Pinós Postdoctoral fellowship from the Generalitat de Catalunya (Spain) to J.I.L.LL.

Alexander Martínez-Fundichely[1,4], *Sònia Casillas*[1,2,4], *Raquel Egea*[1,2], *Miquel Ràmia*[1], *Antonio Barbadilla*[1,2], *Lorena Pantano*[1], *Marta Puig*[1,2], *and Mario Cáceres*[1,3,*]

Nucleic Acids Research 2013; doi: 10.1093/nar/gkt1122

[1] *Institut de Biotecnologia i de Biomedicina,*
Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain.
[2] *Departament de Genètica i de Microbiologia,*
Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain.
[3] *Institució Catalana de Recerca i Estudis Avançats (ICREA),*
Barcelona, Spain.
[4] *The authors wish it to be known that, in their opinion,*
the first two authors should be regarded as joint First Authors

[*] *To whom correspondence should be addressed.*

Tel: +34 935868726; Fax: +34 935812011; Email: mcaceres@icrea.cat

CHAPTER 3

INVFEST, A DATABASE INTEGRATING INFORMATION OF POLYMORPHIC INVERSIONS IN THE HUMAN GENOME

Martínez-Fundichely A, Casillas S, Egea R, Ràmia M, Barbadilla A, Pantano L, Puig M, Cáceres M. InvFEST, a database integrating information of polymorphic inversions in the human genome. Nucleic Acids Res. 2014 Jan;42(Database issue):D1027-32. doi: 10.1093/nar/gkt1122.

Manuscript in preparation
Alexander Martínez-Fundichely[1]*, José Ignacio Lucas-Lledó*[1] *and Mario Cáceres*[1,2,*]*.*

[1] *Institut de Biotecnologia i de Biomedicina,*
Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain.
[2] *Institució Catalana de Recerca i Estudis Avançats (ICREA),*
Barcelona, Spain.

[*] *To whom correspondence should be addressed.*

Tel: +34 935868726; Fax: +34 935812011; Email: mcaceres@icrea.cat

CHAPTER 4

DESCRIPTIVE ANALYSIS OF GENOMIC PATTERNS OF HUMAN POLYMORPHIC INVERSIONS

4.1 Summary

Inversions are a particular type of structural variation in the genome with increasing biomedical interest. So far, there are several reports of disease associations involving inversions, making inversions currently a "hot" topic. The fact that most inversions are copy-number balanced rearrangements and do not affect the genomic dosage, had resulted in a poor characterization of these variants because the earlier array-based studies were unable to detect them. Nevertheless, the boom of next-generation sequencing based studies have provided data and tools for the complete characterization of inversions in the human genome and for a better understanding of their structural and evolutionary role. This study makes a descriptive analysis of the genomic patterns of the

current information about human inversion polymorphisms. The results show that inversions are distributed relatively random among the chromosomes with a significant excess or defect of inversion in some of them. Most of the inversions in the human genome are small events that occur in locations where the genes are not affected. In addition, there is a trend of bigger inversions to have low minor allele frequency. However, the map of human inversions is still quite limited because most of the detected inversions remain without independent experimental validation and genotyping assay. It is also important to note that our understanding about the real number of inversions and their size distribution is probably biased by the genomic and large scale approaches used for the identification of these variants.

4.2 Introduction

The fast development of powerful molecular biology techniques, as the novel platforms of high-throughput sequencing, together with the prompt improvements in computational and statistical tools for the analysis of the massive results, have generated an increasing amount of high-quality whole-genome sequencing data from where it is possible to extract an unprecedented degree of information of structural variation (SV) in the human genome [Ahn *et al.*, 2009; Kidd *et al.*, 2008; Korbel *et al.*, 2007; McKernan *et al.*, 2009; Wang *et al.*, 2008; Redon *et al.*, 2006; Pinto *et al.*, 2011]. Thanks to this productive line of research, it has become obvious that genomic variation is far more complex than previously thought. Nowadays for our understanding of genome evolution and its functionality, it is very important the discovery, validation and characterization of the wide variety of rearrangements, ranging from unbalanced copy number variants (CNVs), to balanced inversions and translocations [Feuk *et al.*, 2006; Stankiewicz and Lupski, 2010; Sharp *et al.*, 2006].

The possible clinical functional impact of the SVs has been the main biomedical focus in some genomic studies [Hurles *et al.*, 2008], highlighting the role of various rearrangements in several human diseases. This includes rare disease such as autism [Marshall *et al.*, 2008], mendelian disease such as haemophilia A [Lakich *et al.*, 1993], or more common diseases such as psoriasis [Bassaganyas *et al.*, 2013], HIV susceptibility

[Gonzalez *et al.*, 2005a] and cancer [Bell *et al.*, 1993]. Despite the encouraging results that link SVs with genomic disorders, most of the existing SVs have been detected in healthy individuals [Iafrate *et al.*, 2004]. That means that the contribution of these polymorphisms to the functionality and phenotype differences between individual genomes remain relatively unknown yet.

Specifically, for genomic inversions, the information about their involvement in the functional or evolutionary process in humans, as well as their relation with diseases is often scarce. Nonetheless, the possible inversion effects could have considerable importance, due to that some inversions have been associated to potential diseases either as direct cause [Lakich *et al.*, 1993] or indirectly as risk enhancer of further rearrangements that cause disease [Giglio *et al.*, 2002; Bondeson *et al.*, 1995; Antonarakis *et al.*, 1995; Small *et al.*, 1997; Osborne *et al.*, 2001; Gimelli *et al.*, 2003; Visser *et al.*, 2005]. Moreover, interest in inversion goes beyond its implication in disease, because they could have a major role in important evolutionary processes such as fertility in humans [Zody *et al.*, 2008; Stefansson *et al.*, 2005], and in other species it has been shown their role on phenotypic variability [Joron *et al.*, 2011], adaptive divergence within species [Ayala *et al.*, 2011; Kirkpatrick and Barton, 2006; Navarro and Barton, 2003], reproductive isolation [Lowry and Willis, 2010] and sex chromosome evolution [Kirkpatrick, 2010].

Genomic inversions are increasingly recognized as a common source of variation in the human genome [Feuk, 2010; Alves *et al.*, 2012]. The accumulation of data already available are starting to require a global descriptive analysis with a general characterization of the inversion pattern to evaluate the current impact of inversion events in humans and assess the potential limitation or biases in the current information. This work aims to bring out the state of knowledge about human inversion polymorphisms at the present, taking advantage of the last compilation of the most accurate catalogue of human polymorphic inversions stored in the InvFEST database [Martinez-Fundichely *et al.* submitted]. The achievement of this objetive will indicate the strengths and weaknesses in the current data of polymorphic inversions and it will point out new interesting directions of research, or the necessity to fill gaps in the information of the human genome.

4.3 Materials and methods

4.3.1 Source data

The data source used in this work is available on the first release of the Human Polymorphic Inversion DataBase InvFEST[1] [Martinez-Fundichely *et al.* submitted]. It represents the widest catalogue of human polymorphic inversion, while simultaneously contains the most accurate information and reliable classification. From the aims of the InvFEST project, the total number of inversions known is constantly updating. This is done not only by the addition of new inversion predictions, but also by the curation and validation process that continuously updates the *Validated* or *False* prediction subset, as well as the fraction of *Unreliable predictions*.

Despite the data represents a non-redundant catalogue, there are some regions where there are located several inversions. This could represent several different isolated events, but also inaccurate predictions in regions still not well curated. Therefore, To estimate the total amount of independent inversion regions, the shared fraction between the overlapped inversions was computed (see equation (4.1)), which can be interpreted as the number of independent equivalent inversions (inv^*) per regions.

$$inv^* = \frac{inv \times L}{\sum\limits_{i=1}^{inv} l_i} \tag{4.1}$$

Where "inv" is the total number of inversions in the region, "L" is the maximum size of the overlapping region, "l_i" is the length of the different inversions.

Since there is not specific information of the mechanism of formation per each inversion, the dataset was classified using the "BreakSeq" pipeline [Lam *et al.*, 2010], in which the breakpoint sequences are characterized with respect to genomic landmarks, chromosomal location and physical properties.

[1] Website available at http://invfestdb.uab.cat

4.3.2 Genomic information

All data used to establish the relationship between the inversions and the genomic elements (genes and segmental duplication) was obtained from the UCSC Genome Browser database. The reference assembly used for all analysis was NCBI36/HG18, because the most reliable inversion information is on this human reference assembly [Martinez-Fundichely *et al.* submitted]. For the segmental duplications we used the subset that map in the same chromosome in inverted orientation with respect to each other, and the fraction of matching bases including indels exceeds 95%.

The relation with genes considers the relative position of the inversion breakpoints loci: (i) If both breakpoints are outside a gene, the inversion is classified as "Intergenic". (ii) If both breakpoints are inside a gene but within intron, the inversion is classified as "Intronic". (iii) Otherwise, the inversion is classified as "breaking genes" when its breakpoints breaks within an exon or reorders part of a gene.

4.4 Results and discussion

The description of the inversion patterns is always carried out from the InvFEST most reliable inversion class: (i) inversions already validated, and (ii) inversions predicted [Martinez-Fundichely *et al.* submitted]. In addition, the analysis tries to characterize the data set of the genomic distribution of inversions, pointing out possibles biases or gaps in the information.

4.4.1 The current landscape of inversion discovery

The rise of next-generation sequencing (NGS) technologies, together with the improvement of the computational tools for the analysis of the huge amounts of data obtained, have brought us from the few inversions early detected by target-focused strategies, to an increasing number of whole-genome sequencing studies that report a considerable amount of inversion predictions in the human genome [Feuk, 2010; Korbel *et al.*, 2007; Pang *et al.*, 2013; McKernan *et al.*, 2009; Ahn *et al.*, 2009; Kidd *et al.*, 2008; Wang *et al.*, 2008]. From the most comprehensive integration of inversion data in humans within the InvFEST database [Martinez-Fundichely *et al.* submitted], using the most reliable an updated data (without taking into account the subsets classified as *unreliable prediction* and *False*), at the present there are a total of 611 inversion events reported. However, there are still 326 inversions located in a total of 66 overlapping regions (see Table 4.1). This overlapping regions in some cases include two predictions pointing at nearly the same location, which actually could represent the same inversion locus. In other cases, there are predictions of big inverted regions which overlap with other relatively small inversions, and the maximum overlap per region was 31 inversions. To distinguish both cases, we calculated the number of equivalent inversions (Inv^{*}) per region (see materials and methods equation 4.1). Actually there are a total of 541 equivalent inversions, thus in the current catalogue of inversion there are likely a total of 541 non-redundant inversions.

This multiplicity of inversions per region could be due to real differences in specific breakpoints boundaries between individuals, to inaccurate breakpoint predictions from methods with low resolution, or to non-curated regions in which there still remain wrong predictions corresponding to false inversions. However, the bioinformatic identification of different overlapping predictions as independent inversions or as support of a common event is inaccurate because in most cases there is not independent data available to reanalyze the breakpoints and even the using of the same reference could lead to inconclusive situations.

Table 4.1: Summary of the inversion catalogue.

		Overlapped region		
	Single inversion	**Two inversion**	**More than two inversion**	**Total of inversions**
Number of inversion	258	86 (43)	240 (23)[a]	611

[a] The maximun of overlap was 31 inversions.

Furthermore, the important and laborious task of experimental validation of inversions still remains lagged behind. So far, there is a very small number of inversions, just 79, that have been validated. This represents that less than 15% of the predicted inversion regions are experimentally validated. This gap on the inversion information could be due to that the study of inversions has been mostly focused in cases related with disease susceptibility [Bondeson *et al.*, 1995; Antonarakis *et al.*, 1995; Small *et al.*, 1997; Giglio *et al.*, 2002; Osborne *et al.*, 2001; Gimelli *et al.*, 2003; Tam *et al.*, 2008], and thus so far they have not been deeply studied as a common source of human genetic variation, and few studies of validation of human inversion have been carried out [Pang *et al.*, 2013] and [Aguado *et al.* 2013 submitted (Appendix C)].

4.4.2 Chromosomal distribution of inversions

A previous study that discovered human inversion polymorphisms by comparative analysis between human and chimpanzee genome sequence assemblies [Feuk *et al.*, 2005] reported that chromosome distribution of inversions is related with the size of each chromosome. This result mildly corroborated the general idea that the random distribution of rearrangements tends to correlate to the size of the chromosome. Attending to the current distribution of the number of inversions per chromosome in the InvFEST database, there is a fairly strong positive relationship between the amount of inversions and the length of each chromosome (see Figure 4.2). The value of Pearson's correlation is ($r = 0.68,\ df = 22,\ p = 0.0003$, two-tailed test).

The current number of inversions per chromosome can not be explained only by the correlation to the length of chromosomes. There are seven cases (marked with an asterisk) in which it exists a significant deviation of the expected number of inversions corresponding to its chromosome size (see Figure 4.1). This could be due to the possibility that the chromosomal architecture, especially the SDs [Feuk *et al.*, 2005; Bailey *et al.*, 2001; Emanuel and Shaikh, 2001], instead of the size is leading the chromosomal distribution of inversions. This hypothesis is reinforced by the evidence of a significant correlation between the amount of inversions predicted and the number of inverted SDs in each chromosome (see Figure 4.3), the Pearson's correlation is ($r = 0.63, df = 22,\ p = 0.0010$, two-tailed test) though it is slightly shifted away from the linearity. This contrasts to the low correlation ($r = 0.29,\ df = 22\ p = 0.1682$, two-tailed test) between the chromosome size and the number of inverted SDs in each chromosome.

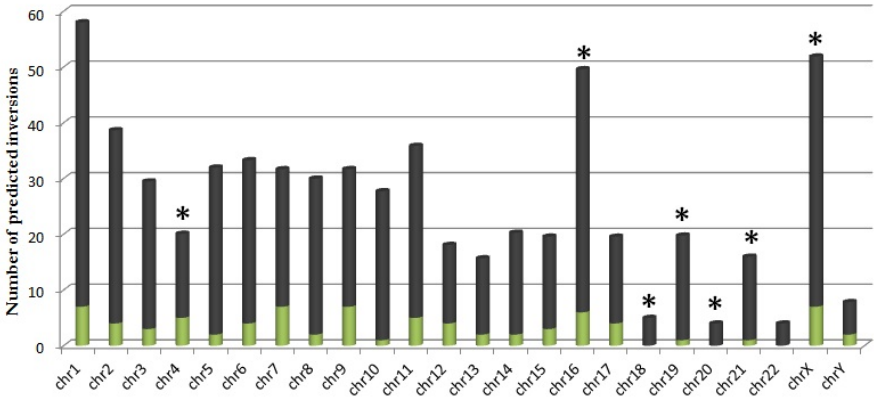


Figure 4.1: Chromosomal distribution of non-redundant inversions - The fraction in green is the number of inversions that are already validated. The fraction in gray represent the rest of inversions predicted per chromosome. Taking into account the expected number of inversions according to the chromosome size, the Z-test (Z score 1.96) shows for the cases of autosomal chromosomes 4, 18 and 20 (marked with an asterisk) the prediction proportions is less than expected. Chromosome 16, 19, 21 and X are also marked with an asterisk, because in this cases they have more inversions than the number expected for their corresponding size.

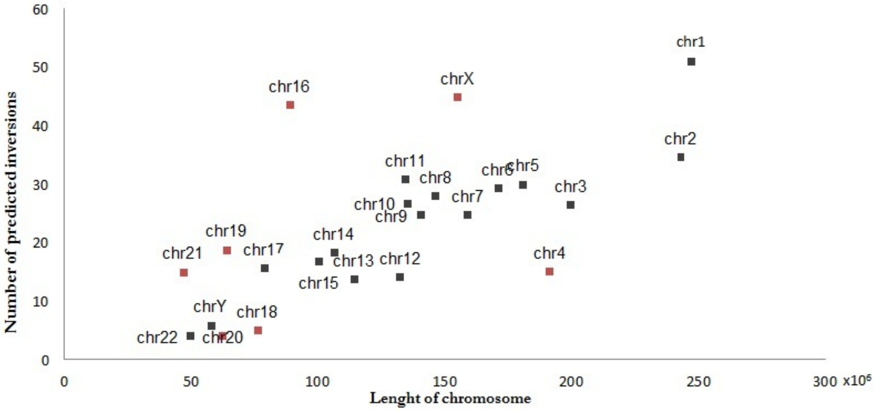


Figure 4.2: Correlation between the number of inversions and the length of the chromosome - The dispersion of the graphs show a clear positive relationship between the number of inversion and the size of the chromosome.

A possible explication is that the detection of inversion is biased to the region of SDs because the most common mechanism of inversion formation is non-allelic homologous recombination (NAHR) between highly similar inverted sequences [Stankiewicz and Lupski, 2010; Shaw and Lupski, 2004; Gu *et al.*, 2008; Lupski and Stankiewicz, 2005]. On the other hand, the presence of SDs could increase the number of wrong predictions because it is a known source of error of the paired-end mapping methods used for the prediction of most inversion [Onishi-Seebacher and Korbel, 2011] and [Martinez-Fundichely *et al.* chapter 2].

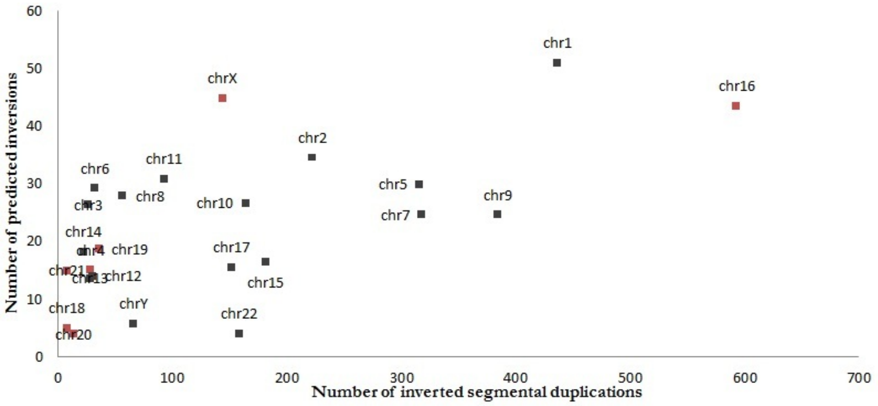


Figure 4.3: Correlation between the number of inversions and the number of segmental duplications per chromosome - The dispersion of the graphs show a clear positive relationship between the number of inversion and the number of SDs, in this case the relationship loses slightly the linearity.

Chromosome 16 has much more inversions than expected (Z score = 7.2) and it is remarkable that it is the most duplication-rich chromosome [Martin *et al.*, 2004]. Thus, this positive correlation might explain the observation of the unexpectedly higher number of inversions in this chromosome. In the case of the X chromosome, which also shows an excess of inversions (Z score = 3.5), the number of inversions is not explained by the effect of the correlation with the SDs. This pattern in which the X chromosome shows an increase of inversions compared to autosomes of corresponding size also was reported in the previous study of inversions between the human and the chimpanzee genomes [Feuk *et al.*, 2005]. Since males carry only one copy of the X chromosome, a plausible explanation might be that there is an increment of the intra-chromosomal recombinations within unpaired X chromosome (in males) that increases the probability to generate inversions by the NAHR mechanism. The chromosome 4, 18, and 20 have less inversions than expected and the chromosome 19 and 21 have more inversions than expected, but these deviations can not be explained by these simple correlation suggesting that the number of inversions per chromosome depends on a combination of factors.

4.4.3 Mechanism of origin

The inversion formation mechanisms, as for other structural variation, are categorized attending to the breakpoint features in those involving extensive stretches of sequence identity spanning the breakpoint junctions, called as sequence-homology mediated, and those occurring in the absence of homology [Onishi-Seebacher and Korbel, 2011; Gu *et al.*, 2008; Lupski and Stankiewicz, 2005; Shaw and Lupski, 2004]. Moreover, attending to the cellular event leading the process, the mechanisms of formation can be classified further. First, rearrangements can occur by recombination based mechanisms, of which non-allelic homologous recombination (NAHR) has been shown as the main mechanism of formation of polymorphic inversions [Pang *et al.*, 2013; Feuk, 2010]. The principal signal of this mechanism is the presence of highly identical and inverted sequences within the breakpoints. These are the locations in which the recombination event takes place and are commonly associated to segmental duplications or different types of repetitive elements. Since recombination could be a recurrent genomic process, this mechanism has been associated to some events of recurrent inversions [Cáceres *et al.*, 2007] and [Aguado *et al.* 2013 submitted (Appendix C)]. "Replication" based mechanisms involve DNA synthesis, such as in microhomology mediated break induced replication (MMBIR) [Hastings *et al.*, 2009]. Other suggested mechanism involved in creation of inversions during replication is fork stalling and template switching (FoSTeS) [Koumbaris *et al.*, 2011; Lee *et al.*, 2007]. Finally, there are mechanisms based on double strand repair pathways, which are nonhomologous end joining (NHEJ) [Davis and Chen, 2013] or microhomology-mediated end joining

(MMEJ) [Sankaranarayanan *et al.*, 2013]. However, in most cases it is difficult to determine the exact process that generate a rearrangement and they tend to be classified just in homologous and non-homologous.

Table 4.2: Mechanism of formation.

	NAHR	NH	Not determined
Predicted inversions	65	292	175
Validated inversions	35	41	3
Total of inversions	100	333	178

The precise mechanism of generation for most individual inversions is still unknown, because the majority of inversions are poorly studied beyond their location. Furthermore, the limited number of inversions with breakpoints located at nucleotide resolution or within a narrow interval available to date has prevented a thorough investigation of mechanisms and sequence motifs giving rise to inversions. Table 4.2 shows the current possible classification of the inversion dataset by the mechanisms of origin according to the BreakSeq pipeline [Lam *et al.*, 2010]. Based on the presence or absence of homologous sequence at the breakpoint region, the inversions are classified into NAHR or nonhomologous processes, in which all inversions that could be generated by the different replication or double strand break repair mechanisms are grouped.

In total, of the 611 inversions that could be classified, 54.5% presumably occur by non-homologous mechanisms and only 16.4% by NAHR. In the validated set this changes and there are almost half of the inversion generated by each mechanism. This could be due in part to that inversion mediate by large SDs often can not be predicted by PEM [Lucas Lledó and Cáceres, 2013], whereas those with simple breakz are easier to detect. Thus studied inversion could not be really a random sample.

4.4.4 Inversion length distribution

Next we examined the distribution of inversion lengths. The current length distribution of inversions (see Figure 4.4) is shifted towards smaller size variants. The majority of the inversions are in the interval of less than 5 kb, though some inversions extends to several megabases. This result changes the previous distribution of inversion lengths [Feuk, 2010] which reported that until that moment, the majority of the inversions were in the interval of 10 kb to 100 kb. This new inversion length distribution is concordat in form (decremental), with a theoretical length distribution of random inversions [Cáceres *et al.*, 1999] but the theoretical expected mean size is around ⅓ of chromosomes. Thus, it is evident that the size of the known inversions in human genome is much smaller than the expected, even considering as reference the smallest chromosome.

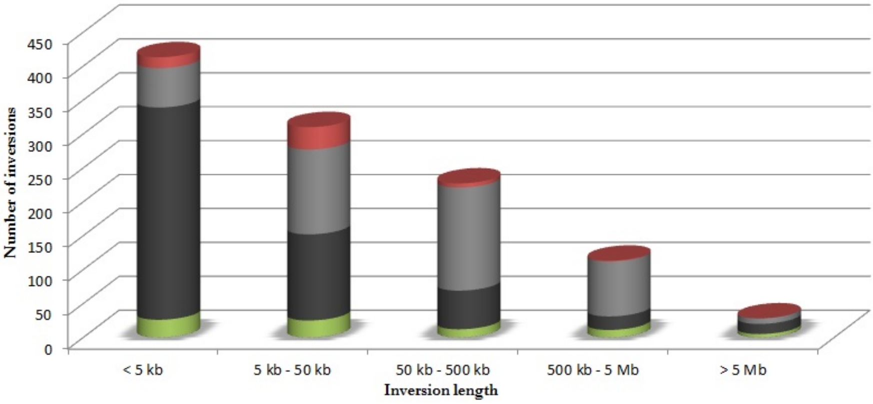


Figure 4.4: Length distribution of inversions - The length of the predicted inversions ranges from few inverted base pairs, to ~23 Mb or bigger for a validated inversion and other just predicted inversions. More than half of all inversions are less than 5 kb.

The shift towards a smaller size could be explained by the refining of the inversion region boundaries, and the curation process that filtered out all non-reliable predictions. The Chi-square test of association between the inversion length categories small (< 5 kb) and big (> 5 kb)inversion with the categories of reliable and unreliable inversions (true or predicted / false or filtered out) rejects the null hypothesis of independence ($\chi^2 = 183.44,\ df = 1,\ p = 0.0001$). This suggests that so far the filtering processes tends to eliminate the bigger inversions.

Moreover, this might reflect the bias on the resolution of the prediction method and sequence coverage. Since more studies have been searching structural variation using next generation sequencing and paired-end mapping methods, the relatively small insert size of the template fragments and the higher coverage tend to increase the power of detection of small inversions [Lucas Lledó and Cáceres, 2013]. On the other hand, bigger inversions tend to have larger inverted repeats at their breakpoint loci and might not be encompassed by the insert-size of the next generation sequencing PEM templates [Pang *et al.*, 2013]. Therefore, the new techniques might cover well small inversions with relatively simple breakpoints, while some of the bigger inversion with more complex breakpoints might still be missing.

Biologically, small inversions could be considered neutral, without obvious phenotypic consequences if the breakpoints do not interrupt the coding or regulatory sequence of the genes near the inversion. Then, in theory it is more important where the breakpoints are located, than the amount of chromosome spanned by the inversion. However, this size distribution shows actually that the big majority of inversions are small. This data supports the hypotheses that the size of inversions has a biological effect beyond the breakpoint location, but it is evident the need of more work in this area. There are studies about the influence of inversions in the overall recombination rate and the vital role of crossing over during meiosis for proper chromosome segregation [Adi *et al.*, 2011; Stevison *et al.*, 2011]. In particular, the large inversions could have a more negative effect on fitness due to formation of imbalanced gametes by crossing over in inversion heterokaryotypes during meiosis.

4.4.5 World-wide frequency distribution of inversions

The data of inversion frequency in the global human population (see Figure 4.5) suggests that even large inversions may by frequent in the human population without a strong negative effect on fitness, such as the inversions studied in [Antonacci *et al.*, 2009; Gilling *et al.*, 2006]. However, in general big inversions show a significantly lower minor allele frequency on human populations.

Within the 79 inversions that are experimentally validated, 45 cases are fully analyzed and include an estimate of the frequency in the population (see Figure 4.5). Due to the technical difficulties, the genotyping of polymorphic inversions in the human population is still little known and our understanding about the allele frequency distribution is probably biased to targets of biomedical interest for inversion validation.

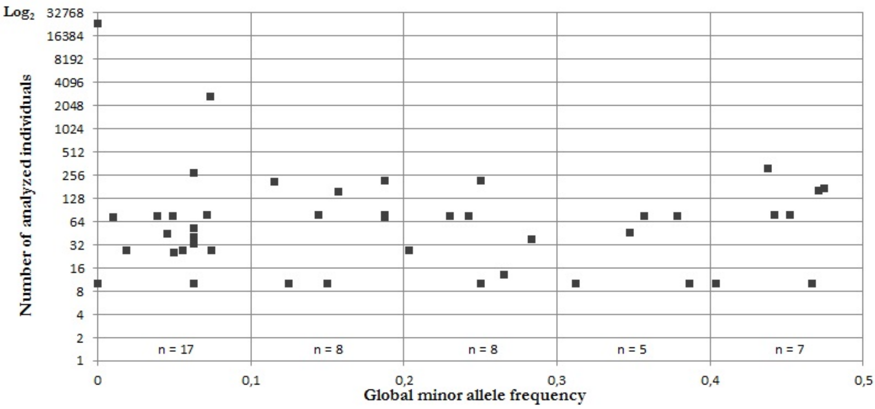


Figure 4.5: Number of individuals analyzed per range of frequency - Graph of the number of individuals analyzed per frequency in the population, which show more information for inversions with frequency less than 0.10, and the number of individuals analyzed ranges from 10 to 23500.

In most inversions, the complexity of breakpoint regions makes them usually hard to analyze and genotype in a large number of individuals. Just one inversion [Gilling *et al.*, 2006] has been genotyped in a big number of individual from several populations, although in this case it was a very large inversion routinely genotyped during karyotyping of samples by prenatal diagnostic labs. The association with a tag-SNPs has also allowed the prediction of the genotype of the chromosome 17 inversion in a large number of individuals (2700), with an emphasis on African populations [Steinberg *et al.*, 2012].

In the full data set of inversion frequencies, the minimum number of individuals analyzed for a big fraction of inversion is 10, but the real limitation to analyze the global frequency distribution is the reduced subset of inversion that have information. However, from this preliminary data it is interesting to note that the number of inversions which have less than 0.10 of minor allele frequency is nearly double than any other frequency ranges (see Figure 4.5). Another important trend that is shown from the current data is that the bigger the inversion size, the lower the frequency observed in the population is. According to the size of inversions, there are significant differences between the frequency observed in inversions with size less than 5 kb and the frequency

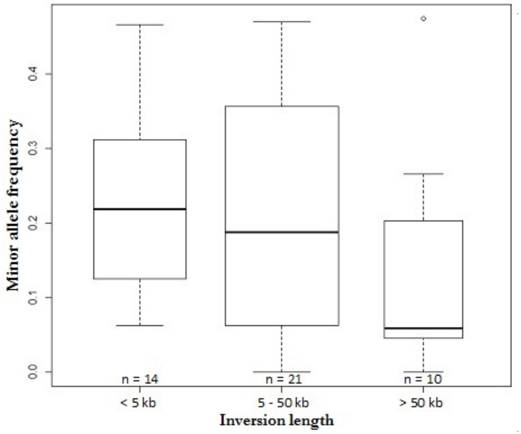


Figure 4.6: Boxplot of global frequency of inversions classified by inversion length - The boxplot shows the trend of bigger inversions to have low frequency.

observed in inversions with size greater than 50 kb. The Wilcoxon rank sum test with continuity correction which does not assume a normal distribution gave a p-value = 0.03484 (see Figure 4.6). This fits well the idea that bigger inversion could have more negative effects.

At present we can not extract other conclusions about global demography of inversions and evolutionary history in the human populations, but recent studies are starting to extensively characterize the inversions at the population level [Pang *et al.*, 2013] and [Aguado *et al.* 2013 submitted (Appendix C)] and it is expected that inversion frequency will be more informative in the near future, with hundreds of individuals of several populations analyzed [Villatoro and Cáceres, unpublished data].

4.4.6 Potential functional effects of inversions

The principal immediate possible genomic consequence of an inversion is the disruption of genes or their regulatory sequence, or even the change of the relative position of the elements nearby necessary for correct transcription. There are various evidences that such rearrangements could involve several genetic elements and could be associated to human disorders. One of the examples is the inversion associated to the disruption of the factor VIII gene, that is the cause of hemophilia A [Lakich *et al.*, 1993]. This is a recurrent inversion mediated by inverted segmental duplications located within an intron of the gene that has been found in approximately 43% of patients with this disease [Antonarakis *et al.*, 1995]. Another example is the case of the Hunter syndrome in which there is the disruption of the IDS iduronate 2-sulfatase gene [Bondeson *et al.*, 1995].

Therefore, the potential relation of the inversions with genes is an important information to analyze to determine their possible functional effects. The distribution of the positional relationship between the inversions and the UCSC genes data (see Figure 4.7). To asses the significance of the relation, we test the association between the categories of inversions affecting / not affecting genes and the reliable / unreliable inversions categories (true or predicted / false or filtered out). The Chi-square test ($\chi^2 = 10.49,\ df = 1,\ p = 0.001$) rejects the null hypothesis of independence between the inversion categorized according to gene effects and the classification of reliability. Therefore, the information of inversions that breaks genes are significantly enriched in false or unreliable predictions. This suggests that the expected number of real inversions that disrupts gene coding sequences should be low, which is a reasonable result since the inversion dataset comes from predictions on healthy individuals.

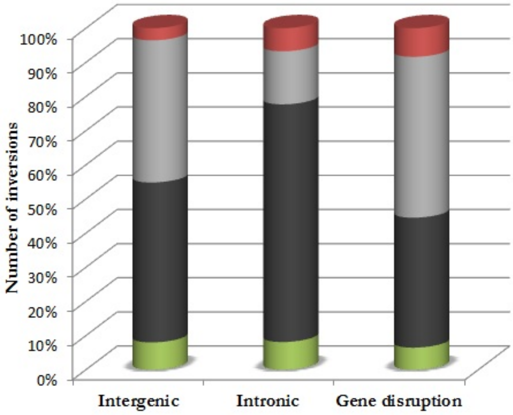


Figure 4.7: Distribution of the association between inversions and genes - The graph shows that more than half of the initial inversion predictions that break genes was filtered out in the false (red) or unreliable (gray) subset of inversions.

Nevertheless, although we have identified some interesting cases of inversions disrupting genes that are currently under investigation, most predictions that potentially are affecting genes remain without validation assay. This indicates that despite the evident potential clinical interest, those inversions are very difficult to analyze (see Table 4.3) and (Appendix A). Therefore, in future studies some of those inversions could be identified as false positive predictions. Moreover, the inversions with breakpoints falling within intronic regions should be analyzed carefully and this information taken into account on the discovery of thus far hidden functional regions.

Table 4.3: Summary of the effect of inversions on genes.

Inversion effect	Affected genes	Validated inversions
Breaks one gene	28	5
Breaks two genes	29	4
Rearranges part of a gene	6	0
Breaks a gene's exon	12	1

As mentioned before, the inversions also may be indirectly associated with genomic disorders [Antonacci *et al.*, 2009]. In this case, the presence of an inversion could lead to the generation of other types of rearrangements (commonly microdeletion) in the offspring, which could be the direct cause of the disease. For example, the Williams-Beuren syndrome, is reported as most commonly being the consequence of a microdeletion and it has been shown that in around 30% of the cases the parent carried an inversion. This indicates that the inversion may be increasing the risk of further rearrangement, but the global frequency of the inversion does not indicate that it may be associated with the syndrome in itself [Osborne *et al.*, 2001; Tam *et al.*, 2008]. Other examplee where inversions have been associated to a disease phenotype is the deletion of the emerin gene in Emery-Dreifuss muscular dystrophy associated to an inversion including the *FLNA-EMD* gene [Small *et al.*, 1997].

4.5 Conclusions

So far, inversion studies have been focused on their potential effect, but the current information about human inversion polymorphisms shows that most of the inversions in the human genome are relatively small events that occur in locations where the genes are not affected. The inversions tend to correlate with inverted segmental duplications. That means that the inversions could represent a recurrent change. However, non-homologous mechanisms are also frequently involved in the generation of inversions. The map of human inversions is still quite limited because most of the inversions already detected remain without independent experimental validation and genotyping assay. Thus, little is also known about their frequency in population, but the current distribution shows that the bigger inversions have low minor allele frequency. However, it is important to note that our understanding about the real number of inversions and their size distribution is probably biased by the genomic and large scale approaches used for the identification of these variants.

4.6 Acknowledgements

We thank Antonio Barbadilla for help with the mathematical and statistical aspects. This work was supported by the European Research Council (ERC) Starting Grant 243212 (INVFEST) under the European Union Seventh Research Framework Programme (FP7) to M.C., a MAEC-AECI doctoral fellowship from the Ministerio de Asuntos Exteriores y Cooperación (Spain) to A.M.F., and a Beatriu de Pinós Postdoctoral fellowship from the Generalitat de Catalunya (Spain) to J.I.L.LL.

CHAPTER 5

GENERAL DISCUSSION

In this dissertation we address the problematic question of the reliability of the prediction of structural variants in the human genome, mainly focused on inversion discovery using mapping-based strategies. As mentioned in the introductory chapter of this dissertation, the advent of high-throughput sequencing has completely changed the landscape of human genomic structural variation, giving the opportunity to thoroughly study chromosomal inversions. Thus, as noted in the scientific problem identified in this research, the development of efficient algorithms focused on the improvement of the accuracy of predictions, and bioinformatic systems for integration and analysis of inversion data are crucial for further studies.

5.1 Improving the accuracy of inversion predictions

In the first objective of this thesis, we considered the most appropriate approach for large-scale prediction of inversions. We selected the paired-end mapping method (PEM) [Tuzun *et al.*, 2005; Korbel *et al.*, 2007], because paired-end mapping has the advantage, over other SV-detecting approaches, of providing the best signal to discover balanced rearrangements, such as is the case of genomic inversions. In addition, the amount of studies predicting different types of structural variants, including inversions, using PEM combined with next generation sequencing, is expected to increase exponentially in the next years.

The thesis faces the challenge of dealing with the high proportion of false positive predictions, but also false negatives, produced by the PEM strategy [Chen *et al.*, 2009; Korbel *et al.*, 2007; Onishi-Seebacher and Korbel, 2011; Lucas Lledó and Cáceres, 2013] focusing exclusively on inversions. The research on the accuracy of inversion prediction led us to reinterpret the geometrical rules of the paired-end mapping pattern generated by an inversion and the related spurious signal that should be eliminated. Hence, a contribution of this work is the bolstering of the mathematical theoretical framework of inversion prediction [Lee *et al.*, 2008] that has allowed us to increase the constrictions of clustering algorithms behind the PEM prediction methods. In particular, one of the original ideas is the deduction of the *sum rule* [Rule 1.2 in chapter 2]. This rule constricts the differences of sum-coordinate among all pem_i (pem_{i++} or pem_{i--} on each case) spanning a same inversion breakpoint within the variation range of the length fragment distribution ($|\Delta L_{fi}|$). It represents an important extra criteria for the clustering constrictions that increases the sensitivity for the detection of complex patterns in region where the signal could come from real inversion, but also from spurious signals frequently associated to the forming mechanism [Onishi-Seebacher and Korbel, 2011; Lucas Lledó and Cáceres, 2013].

The improvement of the PEM rules has already been used on the collaboration to develop another pipeline of algorithms called PeSV-Fischer that uses a combination of patterns, based on paired-reads and read-depth strategies, for the detection of CNVs, translocations (intra and inter chromosomal), and inversions [Escaramís *et al.*, 2013]. PeSV-Fisher has been designed with the aim to facilitate the identification of somatic variation, and, as such, it is capable of analyzing two or more samples simultaneously, producing a list of non-shared variants between samples (see Appendix B).

Apart from this, the first remarkable result of the thesis is the development of GRIAL, a new tool based on a PEM hard-clustering algorithm to predict and refine the breakpoints of inversions as accurately as possible. The improvement in the "sensitivity" (proportion of true breakpoints correctly predicted), the "specificity" (proportion of false rearrangements which are correctly identified as such), and the average "precision" attained in the breakpoint refining has been emphasized in a subsequent study [Lucas Lledó and Cáceres, 2013]. In this work, GRIAL showed the best performance in inversion prediction compared with other SV-detecting algorithms, using a series of simulation data to look for the optimal sequencing strategy to detect inversion by the paired-end mapping method. In addition, with real fosmid PEM sequence data, we have also reported high accuracy, specificity and sensitivity in the inversion prediction compared to other algorithms. The result of the pairwise comparison showed that the power of detection between the different programs is similar, and most of the predictions not matching those of GRIAL were identified as a previous filtered region attending to our PEM reliability criteria and many have been shown to be false. Also as an example of the contribution of GRIAL to a higher specificity thank to its more constrict geometrical rules, there are many regions where the PEM pattern does not point to an specific inversion, and thus GRIAL does not make any prediction, although there are PEM signals. On the contrary, the unique predictions in GRIAL confirm the contribution of the GRIAL strategy to increase the power of low support prediction, since GRIAL, unlike other algorithms, can compute the lowest PEM supported pattern considering the PEM signal of the two breakpoints together, detecting more inversions in regions with low coverage.

The capacity of GRIAL's higher sensitivity to improve the precision of the inversion prediction by PEM methods is an important contribution. However, one of the main problems of inversion detection is the high number of false positive predictions. This is caused by genomic locations that can generate discordant PEM patterns similar to the inversion signal, including regions rich on segmental duplications in different orientations, or sequence differences between inverted repeats among individuals, due for example to gene conversion processes.

The advantage of GRIAL is that thanks to the refinement of breakpoints to a small interval, a good estimate of the expected support for a real inversion can be calculated and those prediction not fulfilling the expected threshold can be discarded. Therefore, one of the most relevant characteristic of GRIAL is that it not only improves the rules for predict and refine the breakpoint accurately, but it also provides two probability scores for assigning reliability to each candidate prediction: (i) the *DS-score*, the probability of the observed PEM "Discordant Support" attending to the size of the inversion and the characterestics of the breakpoints, and (ii) the *D/C-score*, the probability to observe the "Discoradant/Concordant" PEM ratio if there was a real inversion. Moreover, in well covered regions, the *D/C-score* can be used to predict the heterozygosity of genomic inversions and estimate the genotype of each individual.

GRIAL scores are the main tool to eliminate false positive predictions. The comparison takingin to account the GRIAL+ predictions (the most reliable result from the GRIAL standard version) shows that GRIAL+ result is a high quality prediction of inversions. The result of the benchmarking using validated inversions and error-prone identified locations, shows the important contribution of the classification of the prediction by their reliability score and the reporting of the most credible predictions dataset. GRIAL+ outperforms all the other compared programs in all measured parameters of reliability. Attending to the efficiency of detection of validated inversions (the relation between the number of predictions made and the number of inversions detected), GRIAL+ with a 0.95 of efficiency give the best result. The accuracy of each algorithm was also tested through the fraction of well-located breakpoints (overlap between the predicted and the validated breakpoints) from the total of predictions successfully detecting an inversion. In this aspect, the best performance also was from GRIAL+, with 0.89, meaning that it is more likely to find real breakpoints within GRIAL+ predictions than in the other program predictions. Finally, GRIAL also achieves much higher accuracy than the other algorithms in refining the breakpoints, because it has the minor median error distance of the predicted breakpoint boundaries to the real ones.

Furthermore, the important contribution of the GRIAL strategy was especially illustrated by the good performance of the quality filter. GRIAL+ results are the least erroneous data used, both in the number of predictions using wrong PEM signal and the quantity of those PEM used as support for predictions. The advantage of the scoring process is shown by the drastic reduction of the number of predictions in this unreliable regions almost by half in the GRIAL+ dataset, where the 89% of the wrong PEM that points to incorrect inversions is filtered out. This data allow us to estimate the sensitivity of the scoring and filtering process for GRIAL+ compared to the default result from GRIAL. The true positive rate (TPR) of 0.96 and the false discovery rate (FDR) of 0.51 of GRIAL+, confirms the acceptable performance of the filtering process and proves that our high quality result keeps the sensitivity of prediction higher than 90% and reduces by half the number of false predictions.

Finally, it is important mention that although the NGS data simulations suggest that GRIAL performs well also with this data, in this case still some work remains to be done to make reliable predictions from repetitive sequences.

5.2 Obtaining a reliable prediction of inversions in the human genome

In the second objective of this thesis, we aimed to generate reliable polymorphic inversion predictions from available PEM data in the human genome by using GRIAL. The goal was to create a data set of accurately predicted inversions with the greatest certainty that is possible.

Despite that NGS technology has generated a great amount of available PEM data, and even that GRIAL perform well in NGS simulated data [Lucas Lledó and Cáceres, 2013], this thesis mainly considered genomic data from the Human Genome Structural Variation Project [Eichler *et al.*, 2006] of fosmid PEM libraries of 9 HapMap individuals that was previously described in [Kidd *et al.*, 2008]. This data was generated by Sanger sequencing and the increased average read length, sequence quality and paired-end correspondence increases mapping power. This type of data is the most suitable for the hard-clustering algorithms, because the uniquely mapping of reads is favored and the challenge of managing NGS shorter reads data is surpassed [Lucas Lledó and Cáceres, 2013; Onishi-Seebacher and Korbel, 2011; Bashir *et al.*, 2010].

Notwithstanding that this strategy of sequencing offers a low coverage of the genome, the main value of this data resides also in the relative large insert-size of the template fosmids (~40 kb) used for PEM. Although some small inversions should be more difficult to detect because these events could be completely encompassed by the insert-size, the capacity of detection of the bigger and most interesting events is greatly enhanced because the (∼ 40 kb) fragments could span big inverted repeats at the breakpoint loci [Lucas Lledó and Cáceres, 2013]. It is widely recognized that most of the relatively large polymorphic inversions in the human genome are flanked by highly identical inverted segmental duplications [Feuk, 2010; Pang *et al.*, 2013; Kidd *et al.*, 2010]. Another reason for this choice is the availability of, beside the PEM data, the information of full-length sequenced fosmids that possibilitate to validate some predictions by sequence comparison at nucleotide resolution [Kidd *et al.*, 2008, 2010; Eichler *et al.*, 2006] The prediction of GRIAL resulted in a total of 636 inversions, among which there are 220 34.6% predictions in which both breakpoints were detected and 416 65.4% supported by the detection of only one breakpoint. The using of the 9 PEM libraries as a unique merged dataset, allowed us to predict 201 inversions that are supported by just one fosmid PEM in each breakpoint or supported by two PEM from different individuals. These inversions were not detected in the previous analysis of this data set and several of them have been experimental validated. Thus, this work is the most complete analysis of inversions from this data.

The total inversions were located just within 306 regions of overlap. In the GRIAL analysis it was computed the complexity of the inversion regions and around ∼48% of the predictions where classified as simple regions in which only one putative inversion might be present, ∼ 13% as low-complexity regions, in which it may exist a maximum of two putative inversions, and ∼39% as high-complexity regions in which there appears to be above three putative inversions. So far, this complexity has not an explanation. However, our analysis leads us to suspect that most of them are problematic regions. For example, one interesting observation is that the high complex predictions are mostly related with gaps in the human reference assembly.

The most reliable dataset was obtained mainly based on the two scores implemented in the GRIAL pipeline. After applying the filtering to disregard the statistically unreliable results, 65.09% of the predictions were classified as false discoveries and were filtered out. The final GRIAL+ (high quality) dataset that can be used for more confidence analysis is composed just by 222 (34.91%) reliable predictions. This is an important result because it is pointing out that the number of false positives on inversions predicted from PEM data is extremely high. A good example of the value of this work is the collaboration in the recent study [Aguado *et al.* 2013 submitted (appendix C)], where these data have already been used for the experimental analysis of inversion candidates and for the exhaustive characterization of inversions at the population level.

5.3 The first human polymorphic inversion database

In the third objective of this thesis we have considered the task of designing the first database of polymorphic inversions in the human genome that combines all the freely available published information from different studies that have human inversions as a subject. The goal is to obtain a comprehensive catalogue of non-redundant or potentially different predictions of inversion, and for each particular event, collect all the associated information derived from their detailed study.

The result achieved is InvFEST, a database created by integrating data from multiple sources that has been totally implemented as a MySQL multidimensional database with its internal functions and stored procedures to manage the information. InvFEST has a web interface implemented both in PHP and HTML+Ajax that make the database readily accessible online at http://invfestdb.uab.cat through a user-friendly query engine and a complete report for each inversion.

As a database for compilation and analysis purposes, the data model follows a particular snowflake schema, that centralizes all the information in the inversion entity table, and this table is multiply-connected to all dimensions of information of interest. The other important feature of InvFEST is the implementation of an automatic merging engine for the input data (Online Analytical Processing (OLAP)). Both characteristics are the main technical value of the project. The whole process is completely implemented as MySQL stored procedures within the InvFEST database, and thus the database is easily scalable by adding new studies into the existing set of inversions.

The merging strategy used in InvFEST for incorporating a new prediction is based on the overlapping of both corresponding breakpoints to the existing information within the database, always taking into account the resolution (error) of the methodology by which each prediction was obtained. This merging process identifies whether the new prediction represents additional supporting evidence of an already existing inversion, or if it corresponds to an evidence of a completely new inversion to be inserted as an independent entry into the database. The strategy of merging predictions also improves the non-redundant inversion dataset. Particularly, some complex regions in which big inversions encompass small ones, or regions where several segmental duplications are pointing to different putative inversion breakpoints could be clarified.

The main contribution of the InvFEST database is the comprehensive catalogue and the high-quality dataset of human inversions. At this moment, InvFEST combines information from 34 different studies that contribute all type of data of interest, including inversion predictions, validations, and other relevant information. The database reports as current non-redundant dataset, 1092 candidate inversions, of which only 85 have been validated experimentally. This result shows that it remains much work to do in terms of the experimental validation of inversions. However, if false and unreliable predictions are excluded, the total number of inversions is reduced to 617. There are 51 false inversions representing genome assembly errors, PEM errors, or other types of SVs, which are maintained in the database to make possible the tracking of these incorrect predictions in past or future studies. The large-scale detection methods contribute to 98% of the total number of the current catalogued inversions in InvFEST. However, they show a small overlap among their predictions, with the majority of inversions 82% being predicted only by one study, and almost half of them are either unreliable or false. This exemplifies the high false-positive discovery rate of these large-scale detection methods and suggests that there may be diverse biases in each prediction strategy.

It is expected that this database will become a central repository of human inversion information and will be a useful tool for researchers of many diverse fields interested in inversions. Currently, it is already a very valuable resource for the experimental work carried out in the laboratory and will increase exponentially as all the new information generated is incorporated.

5.4 Description of human inversion polymorphisms

In the last objective of this thesis we aimed to summarize the descriptive analysis of the genetic pattern of the current information about inversion polymorphisms in human genomes. This work updates both the size and the chromosomic distribution of polymorphic inversions in the human genome. And also it suggests the answer to the question of to what extent inversions are potentially associated with genomic functional elements.

An important information extracted from the current polymorphic inversions catalogue is that the association between the number of inversion and the size of the chromosome and the number of SDs in the chromosome is more complex than a simple positive correlation. Therefore, it is possible that some other feature, for example recombination hotspot or chromosome architecture could be another factors explaining this distribution. The current information of inversions shows other interesting trends such that the bigger inversions have low minor allele frequency, and that the distribution of lengths shows a size average considerably below the theoretically expected. This preliminary description also shows that real inversions are not likely to affect genes, but there are several cases of inversions breaking genes. Thus the current inversions around a gene regions that have not experimentally validated yet is a very interesting research target.

CHAPTER 6

GENERAL CONCLUSIONS

The main goals of the thesis were successfully achieved through three independent projects, from which the general conclusions are:

1. The study of inversion predictions by paired-end mapping indicates that there is a high degree of false positives and that more rigorous analysis are needed for the reliable detection of this particular type of structural variant.

2. New geometrical rules based on inversion specific characteristic have been deduced to identify precisely the paired-end mapping that support a given inversion and refine the region of the breakpoints more accurately.

3. We have implemented a new algorithm named GRIAL, based on the previous geometrical rules, that is designed to predict inversions from paired-end mapping data. In addition, by using a combination of scores it has been possible to avoid most of the error signals, reducing drastically the false positive rate for inversion predictions.

4. By comparison with other available programs, we have shown that GRIAL has a higher specificity and breakpoint precision in inversion prediction.

5. We have obtained an accurate and reliable set of predicted inversions, from fosmid data of 9 individuals, which has already proven its value as useful data source for subsequent experimental studies. In addition, we have shown that the use of a unique merged dataset of paired-end mapping data from different individuals enhances the power of detection of inversions at low coverage.

6. We have successfully created InvFEST, the first database specific for human polymorphic inversions, through the automatic and scalable integration of data from multiple sources, ranging from large-scale detection studies to targeted validations. This database represents the most reliable catalogue of polymorphic inversions in the human genome.

7. Our analysis of the current list of human inversions shows that there is a low overlap between the inversion predictions from different studies. This suggests that the map of inversions in the human genome is still incomplete and many of the current predictions are low reliable.

8. Genomic distribution of inversions is correlated with the size of the chromosome and its content in inverted segmental duplications, although there are interesting exceptions. Inversion length distribution is clearly smaller than expected by chance, and bigger inversions tend to have a lower frequency of the minor allele.

REFERENCES

1000 Genomes Project C., Abecasis G., Auton A., *et al.* (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

1000 Genomes Project C., Abecasis G. R., Altshuler D., *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

Ackermann M., Weronika S., and Beyer A. (2013). Impact of natural genetic variation on gene expression dynamics. *PLoS genetics*, 9(6).

Adi F., Leffler E. M., Guan Y., *et al.* (2011). Variation in human recombination rates and its genetic determinants. *PloS one*, 6(6).

Agùndez J. A. G., Gallardo L., Ledesma M. C., *et al.* (2001). Functionally active duplications of the *CYP2D6* gene are more prevalent among larynx and lung cancer patients. *Oncology*, 61(1):59–63.

Ahn S. M., Kim T. H., Lee S., *et al.* (2009). The first korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome research*, 19(9):1622–1629.

Aitman T. J., Dong R., Vyse T. J., *et al.* (2006). Copy number polymorphism in fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, 439(7078):851–855.

Alkan C., Coe B. P., and Eichler E. E. (2011). Genome structural variation discovery and genotyping. *Nature reviews. Genetics*, 12(5):363–376.

Alkan C., Kidd J. M., Marques-Bonet T., *et al.* (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*, 41(10):1061–1067.

Alves J. M., Lopes A. M., Chikhi L., and Amorim A. (2012). On the structural plasticity of the human genome: chromosomal inversions revisited. *Current genomics*, 13(8):623–632.

Anderson A. R., Hoffmann A. A., McKechnie S. W., Umina P. A., and Weeks A. R. (2005). The latitudinal cline in the In(3R)Payne inversion polymorphism has shifted in the last 20 years in australian drosophila melanogaster populations. *Molecular ecology*, 14(3):851–858.

Andolfatto P., Depaulis F., and Navarro A. (2001). Inversion polymorphisms and nucleotide variability in drosophila. *Genetics Research*, 77.

Antonacci F., Kidd J. M., Marques-Bonet T., *et al.* (2009). Characterization of six human disease-associated inversion polymorphisms. *Human molecular genetics*, 18(14):2555–2566.

Antonacci F., Kidd J. M., Tomas M., *et al.* (2010). A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nature genetics*, 42(9):745–750.

Antonarakis S. E., Rossiter J. P., Young M., *et al.* (1995). Factor VIII gene inversions in severe hemophilia a: results of an international consortium study. *Blood*, 86(6):2206–2212.

Ayala D., Fontaine M. C., Cohuet A., *et al.* (2011). Chromosomal inversions, natural selection and adaptation in the malaria vector anopheles funestus. *Molecular biology and evolution*, 28(1):745–758.

REFERENCES

Azim M. K., Yang C., Yan Z., *et al.* (2013). Complete genome sequencing and variant analysis of a pakistani individual. *Journal of human genetics*.

Bailey J. A., Yavor A. M., Massa H. F., Trask B. J., and Eichler E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome research*, 11(6):1005–1017.

Bashir A., Bansal V., and Bafna V. (2010). Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance. *BMC genomics*, 11.

Bassaganyas L., Eva R., Manel G., *et al.* (2013). Worldwide population distribution of the common LCE3C-LCE3B deletion associated with psoriasis and other autoimmune disorders. *BMC genomics*, 14(1).

Bell D. A., Taylor J. A., Paulson D. F., *et al.* (1993). Genetic risk and carcinogen exposure: a common inherited defect of the carcinogen-metabolism gene glutathione s-transferase m1 (gstm1) that increases susceptibility to bladder cancer. *Journal of the National Cancer Institute*, 85(14):1159–1164.

Bennardo N., Cheng A., Huang N., and Stark J. M. (2008). Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. *PLoS genetics*, 4(6).

Bentley D. R., Balasubramanian S., Swerdlow H. P., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Bi W., Park S., Shaw C. J., *et al.* (2003). Reciprocal crossovers and a positional preference for strand exchange in recombination events resulting in deletion or duplication of chromosome 17p11.2. *American journal of human genetics*, 73(6):1302–1315.

Bobrow M., Joness L., and Clarke G. (1971). A complex chromosomal rearrangement with formation of a ring 4. *Journal of medical genetics*, 8(2):235–239.

Boettger L. M., Handsaker R. E., Zody M. C., and A M. S. (2012). Structural haplotypes and recent evolution of the human 17q21.31 region. *Nature genetics*, 44(8):881–885.

Bondeson M. L., Dahl N., Malmgren H., *et al.* (1995). Inversion of the ids gene resulting from recombination with ids-related sequences in a common cause of the hunter syndrome. *Human Molecular Genetics*, 4(4):615–621.

Bosch N., Morell M., Ponsa I., *et al.* (2009). Nucleotide, cytogenetic and expression impact of the human chromosome 8p23.1 inversion polymorphism. *PloS one*, 4(12).

Bossé Y. (2013). Genome-wide expression quantitative trait loci analysis in asthma. *Current opinion in allergy and clinical immunology*, 13(5):487–494.

Branzei D. and Foiani M. (2007). Template switching: from replication fork repair to genome rearrangements. *Cell*, 131(7):1228–1230.

Cáceres M., Barbadilla A., and Ruiz A. (1999). Recombination rate predicts inversion size in diptera. *Genetics*, 153(1):251–259.

Cáceres M., of Health Intramural Sequencing Center Comparative Sequencing Program N. I., Sullivan R., and Thomas J. (2007). A recurrent inversion on the eutherian x chromosome. *Proceedings of the National Academy of Sciences of the United States of America*, 104(47):18571–18576.

Camacho C., Coulouris G., Avagyan V., *et al.* (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10.

Carvalho C. M. B., Zhang F., and Lupski J. R. (2011). Structural variation of the human genome: mechanisms, assays, and role in male infertility. *Systems biology in reproductive medicine*, 57(1-2):3–16.

Chen K., Wallis J. W., McLellan M. D., *et al.* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–681.

Chen R., Ren S., and Sun Y. (2013). Genome-wide association studies on prostate cancer: the end or the beginning? *Protein & cell*.

Chung C. C., Magalhaes W. C. S., Gonzalez-Bosquet J., and Chanock S. J. (2010). Genome-wide association studies in cancer–current and future directions. *Carcinogenesis*, 31(1):111–120.

Chung S. J., Jung Y., Hong M., *et al.* (2013). Alzheimer's disease and parkinson's disease genome-wide association study top hits and risk of parkinson's disease in korean population. *Neurobiology of aging*, 34(11):2695.e1–2695.e7.

Church D. M., Lappalainen I., Sneddon T. P., *et al.* (2010). Public data archives for genomic structural variation. *Nature genetics*, 42(10):813–814.

Church D. M., Schneider V. A., Graves T., *et al.* (2011). Modernizing reference genome assemblies. *PLoS biology*, 9(7).

Conrad D. F., Pinto D., Redon R., *et al.* (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712.

Coyne J. A., Aulard S., and Berry A. (1991). Lack of underdominance in a naturally occurring pericentric inversion in drosophila melanogaster and its implications for chromosome evolution. *Genetics*, 129(3):791–802.

Davis A. J. and Chen D. J. (2013). DNA double strand break repair via non-homologous end-joining. *Translational cancer research*, 2(3):130–143.

de Jong S., Chepelev I., Janson E., *et al.* (2012). Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. *BMC genomics*, 13.

de Ravel T. J. L., Devriendt K., Fryns J. P., and Vermeesch J. R. (2007). What's new in karyotyping? the move towards array comparative genomic hybridisation (CGH). *European journal of pediatrics*, 166(7):637–643.

Deng L., Zhang Y., Kang J., *et al.* (2008). An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. *Human mutation*, 29(10):1209–1216.

Dhami P., Coffey A. J., Abbs S., *et al.* (2005). Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *American journal of human genetics*, 76(5):750–762.

Donnelly M. P., Paschou P., Grigorenko E., *et al.* (2010). The distribution and most recent common ancestor of the 17q21 inversion in humans. *American journal of human genetics*, 86(2):161–171.

Edgar R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5.

Edwards J. H., Harnden D. G., Cameron A. H., Crosse V. M., and Wolf O. H. (1960). A new trisomic syndrome. *The Lancet*, 275(7128):787 – 790. Originally published as Volume 1, Issue 7128.

Eichler E., Altshuler D., and Nickerson D. (2006). Human genome structural variation project: January 23, 2006.

Emanuel B. S. and Shaikh T. H. (2001). Segmental duplications: an 'expanding' role in genomic instability and disease. *Nature reviews. Genetics*, 2(10):791–800.

ENCODE Project C., Bernstein B., Birney E., *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

ENCODE Project C., Birney E., Stamatoyannopoulos J. A., *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Entesarian M., Carlsson B., Mansouri M. R., *et al.* (2009). A chromosome 10 variant with a 12 mb inversion [inv(10)(q11.22q21.1)] identical by descent and frequent in the swedish population. *American journal of medical genetics. Part A*, 149A(3):380–386.

Escaramís G., Tornador C., Bassaganyas L., *et al.* (2013). PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data. *PloS one*, 8(5).

Estivill X. and Armengol L. (2007). Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS genetics*, 3(10):1787–1799.

Eva R., He S., Escaramís G., *et al.* (2011). Meta-analysis confirms the LCE3C_LCE3B deletion as a risk factor for psoriasis in several ethnic groups and finds interaction with HLA-Cw6. *The Journal of investigative dermatology*, 131(5):1105–1109.

Feuk L. (2010). Inversion variants in the human genome: role in disease and genome architecture. *Genome medicine*, 2(2).

Feuk L., Carson A., and Scherer S. W. (2006). Structural variation in the human genome. *Nature reviews. Genetics*, 7(2):85–97.

Feuk L., MacDonald J. M., Tang T., *et al.* (2005). Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS genetics*, 1(4).

Fujimoto A., Nakagawa H., Hosono N., *et al.* (2010). Whole-genome sequencing and comprehensive variant analysis of a japanese individual using massively parallel sequencing. *Nature genetics*, 42(11):931–936.

Giglio S., Broman K. W., Matsumoto N., *et al.* (2001). Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *American journal of human genetics*, 68(4):874–883.

Giglio S., Calvari V., Gregato G., *et al.* (2002). Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *American journal of human genetics*, 71(2):276–285.

Gilad Y., Rifkin S. A., and Pritchard J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in genetics : TIG*, 24(8):408–415.

Gilling M., Dullinger J., Gesk S., *et al.* (2006). Breakpoint cloning and haplotype analysis indicate a single origin of the common inv(10)(p11.2q21.2) mutation among northern europeans. *American journal of human genetics*, 78(5):878–883.

Gimelli G., Pujana M. A., Patricelli M. G., *et al.* (2003). Genomic inversions of human chromosome 15q11-q13 in mothers of angelman syndrome patients with class II (BP2/3) deletions. *Human molecular genetics*, 12(8):849–858.

Gnerre S., Maccallum I., Przybylski D., *et al.* (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4):1513–1518.

Gonzalez E., Kulkarni H., Bolivar H., *et al.* (2005a). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science (New York, N.Y.)*, 307(5714):1434–1440.

Gonzalez E., Kulkarni H., Bolivar H., *et al.* (2005b). The influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility. *Science*, 307(5714):1434–1440.

González J., Casals F., and Ruiz A. (2007). Testing chromosomal phylogenies and inversion breakpoint reuse in drosophila. *Genetics*, 175(1):167–177.

Graur D., Zheng Y., Price N., *et al.* (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution*, 5(3):578–590.

Gu W., Zhang F., and Lupski J. R. (2008). Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(1).

Gupta R., Ratan A., Rajesh C., *et al.* (2012). Sequencing and analysis of a south Asian-Indian personal genome. *BMC genomics*, 13.

Hajirasouliha I., Hormozdiari F., Alkan C., *et al.* (2010). Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics (Oxford, England)*, 26(10):1277–1283.

Handsaker R. E., Korn J. M., Nemesh J., and A M. S. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature genetics*, 43(3):269–276.

Haraksingh R. and Snyder M. P. (2013). Impacts of variation in the human genome on gene regulation. *Journal of molecular biology*.

Hastings P. J., Ira G., and Lupski J. R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS genetics*, 5(1).

Hinds D. A., Stuve L. L., Nilsen G. B., *et al.* (2005). Whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–1079.

Hoffmann A. A. and Rieseberg L. H. (2008). Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annual review of ecology, evolution, and systematics*, 39:21–42.

Hoffmann A. A., Sgrò C. M., and Weeks A. R. (2004). Chromosomal inversion polymorphisms and adaptation. *Trends in ecology & evolution*, 19(9):482–488.

Hormozdiari F., Alkan C., Eichler E. E., and Sahinalp S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome research*, 19(7):1270–1278.

Hormozdiari F., Hajirasouliha I., Dao P., *et al.* (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, 26(12):i350–i357.

Human Genome Structural Variation Working Group P., Eichler E. E., Nickerson D., *et al.* (2007). Completing the map of human genetic variation. *Nature*, 447(7141):161–165.

Hurles M. E., Dermitzakis E. T., and Tyler-Smith C. (2008). The functional impact of structural variation in humans. *Trends in genetics : TIG*, 24(5):238–245.

Iafrate A. J., Feuk L., Rivera M. N., *et al.* (2004). Detection of large-scale variation in the human genome. *Nature genetics*, 36(9):949–951.

International Cancer Genome C., Hudson T., Anderson W., *et al.* (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998.

International HapMap I Project C. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

International HapMap II Project C., Frazer K. A., Ballinger D. G., *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861.

International HapMap III Project C., Altshuler D., Gibbs R., *et al.* (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58.

International HapMap Project C. (2003). The international HapMap project. *Nature*, 426(6968):789–796.

International Human Genome Sequencing C. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

Jacobs P. A., Matsuura J. S., Mayer M., and Newlands I. M. (1978). A cytogenetic survey of an institution for the mentally retarded: I. chromosome abnormalities. *Clinical Genetics*, 13(1):37–60.

Jakobsson M., Scholz S. W., Scheet P., *et al.* (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–991003.

Joron M., Frezal L., Jones R. T., *et al.* (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477(7363):203–206.

Kapitonov V. V. and Jurka J. (2008). A universal classification of eukaryotic transposable elements implemented in repbase. *Nature reviews. Genetics*, 9(5):411–2; author reply 414.

Karakoc E., Alkan C., O'Roak B. J., *et al.* (2012). Detection of structural variants and indels within exome data. *Nature methods*, 9(2):176–178.

Kidd J. M., Cooper G. M., Donahue W. F., *et al.* (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64.

Kidd J. M., Graves T., Newman T., *et al.* (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, 143(5):837–847.

Kingsmore S. F., Lindquist I. E., Mudge J., Gessler D. D., and Beavis W. (2008). Genome-wide association studies: progress and potential for drug discovery and development. *Nature reviews. Drug discovery*, 7(3):221–230.

Kirkpatrick M. (2010). How and why chromosome inversions evolve. *PLoS biology*, 8(9).

Kirkpatrick M. and Barton N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, 173.

Kleinjan D. J. and van Heyningen V. (1998). Position effect in human genetic disease. *Human molecular genetics*, 7(10):1611–1618.

Kong A., Barnard J., Gudbjartsson D., *et al.* (2004). Recombination rate and reproductive success in humans. *Nature genetics*, 36(11):1203–1206.

Korbel J. O., Abyzov A., Mu X. J., *et al.* (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome biology*, 10(2).

Korbel J. O., Urban A. E., Affourtit J. P., *et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, 318(5849):420–426.

Koumbaris G., Hariklia H., Alexandrou A., *et al.* (2011). FoSTeS, MMBIR and NAHR at the human proximal xp region and the mechanisms of human xq isochromosome formation. *Human molecular genetics*, 20(10):1925–1936.

Krimbas C. B. and Powell J. R. (1992). *Drosophila Inversion Polymorphism.* Boca Raton, FL.: CRC Press.

Lakich D., Kazazian H. H., Antonarakis S. E., and Gitschier J. (1993). Inversions disrupting the factor VIII gene are a common cause of severe haemophilia a. *Nature genetics*, 5(3):236–241.

Lam H. Y. K., Mu X. J., Stütz A. M., *et al.* (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature biotechnology*, 28(1):47–55.

Lander E. S., Linton L. M., Birren B., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lappalainen I., Lopez J., Skipper L., *et al.* (2013). DbVar and DGVa: public archives for genomic structural variation. *Nucleic acids research*, 41(Database issue):D936–D941.

Lee J. A., Carvalho C. M. B., and Lupski J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, 131(7):1235–1247.

Lee S., Cheran E., and Brudno M. (2008). A robust framework for detecting structural variations in a genome. *Bioinformatics (Oxford, England)*, 24(13):i59–i67.

Levy S., Sutton G., Ng P. C., *et al.* (2007). The diploid genome sequence of an individual human. *PLoS biology*, 5(10).

Li R., Zhu H., Ruan J., *et al.* (2010a). De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20(2):265–272.

Li Y., Hu Y., Bolund L., and Wang J. (2010b). State of the art de novo assembly of human genomes from massively parallel sequencing data. *Human genomics*, 4(4):271–277.

Li Y., Shaw C. A., Sheffer I., *et al.* (2012). Integrated copy number and gene expression analysis detects a CREB1 association with alzheimer's disease. *Translational psychiatry*, 2.

Li Y., Zheng H., Luo R., *et al.* (2011). Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nature biotechnology*, 29(8):723–730.

Lieber M. R. (2008). The mechanism of human nonhomologous DNA end joining. *The Journal of biological chemistry*, 283(1):1–5.

Lieber M. R., Ma Y., Pannicke U., and Schwarz K. (2003). Mechanism and regulation of human nonhomologous DNA end-joining. *Nature reviews. Molecular cell biology*, 4(9):712–720.

Lilleoja R., Sarapik A., Reimann E., *et al.* (2012). Sequencing and annotated analysis of an estonian human genome. *Gene*, 493(1):69–76.

Liu P., Carvalho C. M., Hastings P. J., and Lupski J. R. (2012). Mechanisms for recurrent and complex human genomic rearrangements. *Current opinion in genetics & development*, 22(3):211–220.

López-Correa C., Dorschner M., Brems H., *et al.* (2001). Recombination hotspot in NF1 microdeletion patients. *Human molecular genetics*, 10(13):1387–1392.

Lowry D. B. and Willis J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS biology*, 8(9).

Lubs H. (1969). A marker x chromosome. *American journal of human genetics*, 21(3):231–244.

Lucas Lledó J. I. and Cáceres M. (2013). On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. *PloS one*, 8(4).

Lupski J. R. (1998). Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in genetics : TIG*, 14(10):417–422.

Lupski J. R. and Stankiewicz P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS genetics*, 1(6).

Maegenis R., Donlon T., and Wyandt H. (1978). Giemsa-11 staining of chromosome 1: a newly described heteromorphism. *Science*, 202(4363):64–65.

Mardis E. R. (2006). The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133 – 141.

Marques-Bonet T., Girirajan S., and Eichler E. E. (2009). The origins and impact of primate segmental duplications. *Trends in Genetics*, 25(10):443 – 454.

Marshall C. R., Noor A., Vincent J. B., *et al.* (2008). Structural variation of chromosomes in autism spectrum disorder. *American journal of human genetics*, 82(2):477–488.

Martin J., Han C., Gordon L. A., *et al.* (2004). The sequence and analysis of duplication-rich human chromosome 16. *Nature*, 432(7020):988–994.

McCarroll S. A. and Altshuler D. M. (2007). Copy-number variation and association studies of human disease. *Nature genetics*, 39(7 Suppl):S37–S42.

McKernan K. J., Peckham H. E., Costa G. L., *et al.* (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research*, 19(9):1527–1541.

McVey M. and Lee S. E. (2008). MMEJ repair of double-strand breaks (directorâĂŹs cut): deleted sequences and alternative endings. *Trends in Genetics*, 24(11):529 – 538.

Medvedev P., Stanciu M., and Brudno M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature methods*, 6(11 Suppl):S13–S20.

Meyer L. R., Zweig A. S., Hinrichs A. S., *et al.* (2013). The UCSC genome browser database: extensions and updates 2013. *Nucleic acids research*, 41(Database issue):D64–D69.

Mills R. E., Luttig C. T., Larkins C. E., *et al.* (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research*, 16(9):1182–1190.

Myers A. J., Gibbs J. R., Webster J. A., *et al.* (2007). A survey of genetic human cortical gene expression. *Nature genetics*, 39(12):1494–1499.

Navarro A. and Barton N. H. (2003). Chromosomal speciation and molecular divergence–accelerated evolution in rearranged chromosomes. *Science (New York, N.Y.)*, 300(5617):321–324.

Navarro A. and Ruiz A. (1997). On the fertility effects of pericentric inversions. *Genetics*, 147(2):931–933.

NCBI Resource C. (2013). Database resources of the national center for biotechnology information. *Nucleic acids research*, 41(Database issue):D8–DD20.

Nica A. C. and Dermitzakis E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1620).

Onishi-Seebacher M. and Korbel J. O. (2011). Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 33(11):840–850.

Osborne L. R., Li M., Pober B., *et al.* (2001). A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nature genetics*, 29(3):321–325.

Pang A. W., R M. J., Pinto D., *et al.* (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome biology*, 11(5).

Pang A. W. C., Migita O., Macdonald J. R., Feuk L., and Scherer S. W. (2013). Mechanisms of formation of structural variation in a fully sequenced human genome. *Human mutation*, 34(2):345–354.

Pennisi E. (2007a). Breakthrough of the year. human genetic variation. *Science (New York, N.Y.)*, 318(5858):1842–1843.

Pennisi E. (2007b). Human genetic variation. *Science*, 318(5858):1842–1843.

Perry G. H., Dominy N. J., Claw K. G., *et al.* (2007). Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, 39(10):1256–1260.

Pevzner P. and Tesler G. (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7672–7677.

Pinto D., Darvishi K., Shi X., *et al.* (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature biotechnology*, 29(6):512–520.

Qi L., Qi Q., Prudente S., *et al.* (2013). Association between a genetic variant related to glutamic acid metabolism and coronary heart disease in individuals with type 2 diabetes. *JAMA : the journal of the American Medical Association*, 310(8):821–828.

Quinlan A. R., Clark R. A., Sokolova S., *et al.* (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome research*, 20(5):623–635.

Raffaele Di Barletta M., Ricci E., Galluzzi G., *et al.* (2000). Different mutations in the LMNA gene cause autosomal dominant and autosomal recessive Emery-Dreifuss muscular dystrophy. *American journal of human genetics*, 66(4):1407–1412.

Rausch T., Zichner T., Schlattl A., *et al.* (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)*, 28(18):i333–i339.

Redon R., Ishikawa S., Fitch K. R., *et al.* (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454.

Reinhardt J. A., Baltrus D. A., Nishimura M. T., *et al.* (2009). De novo assembly using low-coverage short read sequence data from the rice pathogen pseudomonas syringae pv. oryzae. *Genome research*, 19(2):294–305.

Rozen S. and Skaletsky H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology (Clifton, N.J.)*, 132:365–386.

Rozen S., Skaletsky H., Marszalek J. D., *et al.* (2003). Abundant gene conversion between arms of palindromes in human and ape y chromosomes. *Nature*, 423(6942):873–876.

Salm M. P. A., Horswell S. D., Hutchison C. E., *et al.* (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome research*, 22(6):1144–1153.

Sankaranarayanan K., Taleei R., Rahmanian S., and Nikjoo H. (2013). Ionizing radiation and genetic risks. XVII. formation mechanisms underlying naturally occurring DNA deletions in the human genome and their potential relevance for bridging the gap between induced DNA double-strand breaks and deletions in irradiated germ cells. *Mutation research*.

Sebat J., Lakshmi B., Troge J., *et al.* (2004). Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y.)*, 305(5683):525–528.

Sharp A. J., Cheng Z., and Eichler E. E. (2006). Structural variation of the human genome. *Annual review of genomics and human genetics*, 7:407–442.

Sharp A. J., Locke D. P., McGrath S. D., *et al.* (2005). Segmental duplications and copy-number variation in the human genome. *American journal of human genetics*, 77(1):78–88.

Shaw C. J. and Lupski J. R. (2004). Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Human molecular genetics*, 13 Spec No 1:R57–R64.

Shen H., Li J., Zhang J., *et al.* (2013). Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four caucasians. *PloS one*, 8(4).

Sindi S. S., Helman E., Bashir A., and Raphael B. J. (2009). A geometric approach for classification and comparison of structural variants. *Bioinformatics (Oxford, England)*, 25(12):i222–i230.

Sindi S. S., Onal S., Peng L. C., Wu H., and Raphael B. J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome biology*, 13(3).

Slack A., Thornton P. C., Magner D. B., Rosenberg S. M., and Hastings P. J. (2006). On the mechanism of gene amplification induced under stress in escherichia coli. *PLoS genetics*, 2(4).

Small K., Iber J., and Warren S. T. (1997). Emerin deletion reveals a common x-chromosome inversion mediated by inverted repeats. *Nature genetics*, 16(1):96–99.

Small K. and Warren S. T. (1998). Emerin deletions occurring on both xq28 inversion backgrounds. *Human molecular genetics*, 7(1):135–139.

Smith D. W., Patau K., and Therman E. (1961). Autosomal trisomy syndromes. *The Lancet*, 278(7195):211 – 212. Originally published as Volume 2, Issue 7195.

Spitz F., Herkenne C., Morris M. A., and Duboule D. (2005). Inversion-induced disruption of the hoxd cluster leads to the partition of regulatory landscapes. *Nature genetics*, 37(8):889–893.

Stajich J. E., Block D., Boulez K., *et al.* (2002). The bioperl toolkit: Perl modules for the life sciences. *Genome research*, 12(10):1611–1618.

Stankiewicz P. and Lupski J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends in genetics : TIG*, 18(2):74–82.

Stankiewicz P. and Lupski J. R. (2010). Structural variation in the human genome and its role in disease. *Annual review of medicine*, 61:437–455.

Starke H., Seidel J., Henn W., *et al.* (2002). Homologous sequences at human chromosome 9 bands p12 and q13-21.1 are involved in different patterns of pericentric rearrangements. *European journal of human genetics : EJHG*, 10(12):790–800.

Stefansson H., Helgason A., Thorleifsson G., *et al.* (2005). A common inversion under selection in europeans. *Nature genetics*, 37(2):129–137.

Steinberg K. M., Antonacci F., Sudmant P. H., *et al.* (2012). Structural diversity and african origin of the 17q21.31 inversion polymorphism. *Nature genetics*, 44(8):872–880.

Stenson P., Ball E., Howells K., *et al.* (2009). The human gene mutation database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Human genomics*, 4(2):69–72.

Stevison L. S., Hoehn K. B., and Noor M. A. F. (2011). Effects of inversions on within- and between-species recombination and divergence. *Genome biology and evolution*, 3:830–841.

Stranger B. E., Forrest M. S., Dunning M., *et al.* (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y.)*, 315(5813):848–853.

Sudmant P. H., Huddleston J., Catacchio C. R., *et al.* (2013). Evolution and diversity of copy number variation in the great ape lineage. *Genome research*, 23(9):1373–1382.

Swaminathan G. J., Bragin E., Chatzimichali E. A., *et al.* (2012). DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Human molecular genetics*, 21(R1):R37–R44.

Tam E., Young E. J., Morris C. A., *et al.* (2008). The common inversion of the Williams-Beuren syndrome region at 7q11.23 does not cause clinical symptoms. *American journal of medical genetics. Part A*, 146A(14):1797–1806.

Tantisira K. G., Lazarus R., Litonjua A. A., Klanderman B., and Weiss S. T. (2008). Chromosome 17: association of a large inversion polymorphism with corticosteroid response in asthma. *Pharmacogenetics and genomics*, 18(8):733–737.

Teague B., Waterman M. S., Goldstein S., *et al.* (2010). High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(24):10848–10853.

Theisen A. (2008). Microarray-based comparative genomic hybridization (acgh). *Nature Education*, 1(1).

Trapnell C., Pachter L., and Salzberg S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–1111.

Turner D. J., Shendure J., Porreca G., *et al.* (2006). Assaying chromosomal inversions by single-molecule haplotyping. *Nature methods*, 3(6):439–445.

Tuzun E., Sharp A. J., Bailey J. A., *et al.* (2005). Fine-scale structural variation of the human genome. *Nature genetics*, 37(7):727–732.

Uddin M., Sturge M., Peddle L., O'Rielly D. D., and Rahman P. (2011). Genome-wide signatures of 'rearrangement hotspots' within segmental duplications in humans. *PloS one*, 6(12).

Venter J. C., Adams M. D., Myers E. W., *et al.* (2001). The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–1351.

Verma R. S., Dosik H., Lubs H. A., and Francke U. (1978). Size and pericentric inversion heteromorphisms of secondary constriction regions (h) of chromosomes 1, 9, and 16 as detected by cbg technique in caucasians: Classification, frequencies, and incidence. *American Journal of Medical Genetics*, 2(4):331–339.

Visser R., Shimokawa O., Harada N., *et al.* (2005). Identification of a 3.0-kb major recombination hotspot in patients with sotos syndrome who carry a common 1.9-Mb microdeletion. *American journal of human genetics*, 76(1):52–67.

Volik S., Zhao S., Chin K., *et al.* (2003). End-sequence profiling: sequence-based analysis of aberrant genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7696–7701.

Wang J., Wang W., Li R., *et al.* (2008). The diploid genome sequence of an asian individual. *Nature*, 456(7218):60–65.

Weischenfeldt J., Symmons O., Spitz F., and Korbel J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature reviews. Genetics*, 14(2):125–138.

Weterings E. and van Gent D. C. (2004). The mechanism of non-homologous end-joining: a synopsis of synapsis. *DNA Repair*, 3(11):1425 – 1435.

Wheeler D. A., Srinivasan M., Egholm M., *et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876.

Xi R., Kim T., and Park P. J. (2010). Detecting structural variations in the human genome using next generation sequencing. *Briefings in functional genomics*, 9(5-6):405–415.

Yalcin B., Wong K., Bhomra A., *et al.* (2012). The fine-scale architecture of structural variants in 17 mouse genomes. *Genome biology*, 13(3).

Ye K., Schulz M. H., Long Q., Apweiler R., and Ning Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, 25(21):2865–2871.

Yorukoglu D., Hach F., Swanson L., *et al.* (2012). Dissect: detection and characterization of novel structural alterations in transcribed sequences. *Bioinformatics (Oxford, England)*, 28(12):i179–i187.

Zeitouni B., Boeva V., Isabelle J., *et al.* (2010). SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics (Oxford, England)*, 26(15):1895–1896.

Zhang J., Wang X., and Podlaha O. (2004). Testing the chromosomal speciation hypothesis for humans and chimpanzees. *Genome research*, 14(5):845–851.

Zhong H., Beaulaurier J., Lum P. Y., *et al.* (2010). Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS genetics*, 6(5).

Zody M. C., Jiang Z., Fung H., *et al.* (2008). Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nature genetics*, 40(9):1076–1083.

Appendixes

APPENDIX A

TABLE OF GENES POTENTIALLY AFFECTED BY THE INVERSIONS

This relation of genes are potentially affected by putative inversions, and represent interesting regions for experimental validation.

Table A.1: Genes potentially affected by inversions.

Gene symbol	Inversion effect
CTRB1	breaks and exchange the gene sequence with another gene affected
CTRB2	breaks the gene and exchange their sequence with another gene affected
KLK7	breaks and reoders the coding sequence within the gene
CRTAP	breaks and reoders the coding sequence within the gene
ARHGEF34P	breaks the gene and exchange their sequence with another gene affected
LOC154761	breaks the gene and exchange their sequence with another gene affected
MARCH1	breaks the gene and exchange their sequence with another gene affected
SORCS2	breaks the gene and exchange their sequence with another gene affected
VIPR2	breaks the gene
KIF27	breaks the gene
MTDH	breaks the gene and exchange their sequence with another gene affected
IKBKB	breaks the gene and exchange their sequence with another gene affected
MOB2	breaks the gene
CES1P1	breaks the gene and exchange their sequence with another gene affected
COL24A1	breaks the gene and exchange their sequence with another gene affected
LOC100506023	breaks the gene and exchange their sequence with another gene affected
ZNF385B	breaks and reoder the coding sequence within the gene
CHN1	breaks the gene
COG7	breaks the gene
CACNA1C	breaks the gene
DDX11-AS1	breaks the gene
CYP4F12	breaks the gene and exchange their sequence with another gene affected
CYP4F24P	breaks the gene and exchange their sequence with another gene affected
LINC00910	breaks the gene

Continued on next page

Table A.1 – continued from previous page

Gene symbol	Inversion effect
LINC00854	breaks the gene
CES1P2	breaks the gene and exchange their sequence with another gene affected
TSPEAR	breaks and reoders the coding sequence within the gene
ANTXRL	breaks the gene
CARD8	breaks the gene
APOL4	breaks the gene and exchange their sequence with another gene affected
APOL1	breaks the gene and exchange their sequence with another gene affected
AKR1C1	breaks the gene and exchange their sequence with another gene affected
AKR1C2	breaks the gene and exchange their sequence with another gene affected
STK31	breaks and reoders the coding sequence within the gene
LOC441242	breaks the gene
C9orf129	breaks the gene
AQPEP	breaks and reoders the coding sequence within the gene
LINC00395	breaks the gene
HERC2	breaks the gene
ZNF257	breaks the gene
CD177	breaks the gene
FAAH2	breaks and reoders the coding sequence within the gene
NBPF3	breaks the gene
HERC2P3	breaks the gene
RNU6-81P	breaks the gene
SMA4	breaks the gene
FAM21C	breaks and reoders the coding sequence within the gene
NOMO1	breaks the gene
SPANXA2-OT1	breaks the gene
CLEC1B	breaks the gene and exchange their sequence with another gene affected
CLEC9A	breaks the gene and exchange their sequence with another gene affected
CST9L	breaks the gene
ZNF100	breaks the gene
VPS13A	breaks the gene and exchange their sequence with another gene affected

Continued on next page

Table A.1 – continued from previous page

Gene symbol	Inversion effect
GNA14	breaks the gene and exchange their sequence with another gene affected
CES1	breaks the gene
LOC399815	breaks the gene and exchange their sequence with another gene affected
C10orf88	breaks the gene and exchange their sequence with another gene affected
GPIHBP1	breaks the gene
TNFRSF10D	breaks the gene and exchange their sequence with another gene affected
TNFRSF10C	breaks the gene and exchange their sequence with another gene affected
SLC41A3	breaks the gene and exchange their sequence with another gene affected
ALDH1L1	breaks the gene and exchange their sequence with another gene affected
CMYA5	breaks and reoders the coding sequence within the gene
ANTXRLP1	breaks the gene and exchange their sequence with another gene affected
CCDC144B	breaks the gene
SRP54	breaks and reoders the coding sequence within the gene
NR2F2-AS1	breaks and reoders the coding sequence within the gene
BRD7	breaks and reoders the coding sequence within the gene
TAOK1	breaks and reoders the coding sequence within the gene
YES1	breaks and reoders the coding sequence within the gene
EFR3B	breaks and reoders the coding sequence within the gene
TSGA10	breaks and reoders the coding sequence within the gene
PCNT	breaks and reoders the coding sequence within the gene
COL23A1	breaks and reoders the coding sequence within the gene

Geòrgia Escaramís[1,2,3,4], *Cristina Tarnador*[1,2,3,4], *Laia Bassaganyas*[1,2,3,4], *Raquel Rabionet*[1,2,3,4], *Jose M. C. Tubio*[1,2,3,4,5], *Alexander Martínez-Fundichely*[1,2,6], *Mario Cáceres*[1,2,6,7], *Marta Gut*[8], *Stephan Ossowski*[2,4,9], *Xavier Estivill*[1,2,3,4]

PLoS One. 2013 May 21;8(5):e63377. doi: 10.1371/journal.pone.0063377.

[1] *Genetic Causes of Disease Group Center for Genomic Regulation (CRG), Barcelona, Spain.*
[2] *Universitat Pompeu Fabra (UPF), Barcelona, Spain.*
[3] *Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESO), Barcelona, Spain.*
[4] *Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain.*
[5] *Galician Fundation of Genomic Medicine-SERGAS, Complexo Hospitalario Universitario de Santiago (CHUS), Santiago de Compostela, Spain.*
[6] *Institut de Biotecnologia i de Biomedicina (IBB), Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain.*
[7] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*
[8] *National Center of Genomic Analysis (CNAG-CRG), Barcelona, Spain*
[9] *Genomic and Epigenomic Variation in Disease Group, Center for Genomic Regulation (CRG), Barcelona, Spain*

APPENDIX B

PESV-FISHER: IDENTIFICATION OF SOMATIC AND NON-SOMATIC STRUCTURAL VARIANTS USING NEXT GENERATION SEQUENCING DATA

Contribution: The work of this thesis contributed to this paper by the study of the theoretical framework of PEM, and the deduction of rules for accurate prediction of breakpoints of different types of structural variants.

Escaramís G, Tornador C, Bassaganyas L, Rabionet R, Tubio JM, Martínez-Fundichely A, Cáceres M, Gut M, Ossowski S, Estivill X. PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data. PLoS One. 2013 May 21;8(5):e63377. doi: 10.1371/journal.pone.0063377

PLoS Genetics 2013, Submitted

Cristina Aguado[1], Magdalena Gayà-Vidal[1], Sergi Villatoro[1], Meritxell Oliva[1], David Izquierdo[1], Carla Giner-Delgado[1], Víctor Montalvo[1], Judit García-González[1], Alexander Martínez-Fundichely[1], Laia Capilla[1], Aurora Ruiz-Herrera[1,2], Xavier Estivill[3,4], Marta Puig[1] and Mario Cáceres[1,5]

[1] *Institut de Biotecnologia i de Biomedicina (IBB), Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain.*
[2] *Departament de Biologia Celular, Fisiologia i Immunologia. Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain.*
[3] *Center for Genomic Regulation (CRG), Barcelona, Spain.*
[4] *Universitat Pompeu Fabra (UPF), Barcelona, Spain.*
[5] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*

APPENDIX C

VALIDATION AND GENOTYPING OF MULTIPLE HUMAN POLYMORPHIC INVERSIONS MEDIATED BY INVERTED REPEATS REVEALS A HIGH DEGREE OF RECURRENCE

Contribution: The work of this thesis contributed to this paper by the creation of a highly reliable dataset of inversions with their breakpoints accurately refined, that were used as candidates to experimental validation. In addition, for each of the validated inversions the PEM support for the standard and inverted orientation was calculated and compared to the genotypes observed by inverse PCR.

Aguado C, Gayà-Vidal M, Villatoro S, Oliva M, Izquierdo D, Giner-Delgado C, Montalvo V, García-González J, Martínez-Fundichely A, Capilla L, Ruiz-Herrera A, Estivill X, Puig M, Cáceres M. Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. PLoS Genet. 2014 Mar 20;10(3):e1004208. doi:10.1371/journal.pgen.1004208