UNIVERSITAT POMPEU FABRA

TESI DOCTORAL UPF 2013

Characterization of Simple and Complex Genomic Structural Variation: a Study of Human Populations and Leukaemia

Laia Bassaganyas Bars

Department of Experimental and Health Science

Genetic Causes of Diseases Group, Bioinformatics and Genomics Program,
Centre for Genomic Regulation – Universitat Pompeu Fabra (CRG-UPF)

Thesis Director

Dr. Xavier Estivill Pallejà

Barcelona, 2013

A la memòria d'en Guillem

Acknowledgments

Durant aquests anys de tesi, són moltes les persones que he tingut la sort de conèixer, que m'han fet aprendre i amb qui he compartit bons moments. Per això m'agradaria deixar constància del meu més sincer agraïment:

A en **Xavier**, no només per haver-me donat l'oportunitat de formar part del teu laboratori, sinó també per tota la teva ajuda, tant a nivell professional com a nivell personal. Moltíssimes gràcies per tot!

A la **Geòrgia**, per ser una gran professional de qui he après moltíssim i per tota la teva inestimable ajuda fins a l'últim dia d'aquesta tesi. Però també, i sobretot, per tot el suport que m'has donat durant tot aquest temps i per la teva amistat. Moltíssimes gràcies Geòrgia, aquesta tesi també és teva.

A tota la gent del laboratori: la **Marta**, l'**Elisa**, l'**Anna P**., l'**Eli**, la **Johanna**, en **Dani**, la **Kelly**, l'**Eulàlia**, la **Mariona**, la **Nàdia**, en **Marc**, l'**Ester**, en **Hyun**, l'**Anna H.**, i a les noves incorporacions **Aparna**, **Tere**, **Laura** i **Joan**, per la vostra ajuda en tots els moments en què ho he necessitat, i perquè compartir el dia a dia és compartir molts petits moments que al final són els que queden i es troben a faltar. També m'agradaria incloure als ex-CRGs que han participat d'alguna manera o altre en aquesta tesi: a en **Cristian** i en **Jose**, per tot el que hem viscut amb la CLL, i a la **Mònica G.**, l'**Eva**, la **Bruni**, en **Sergi**, en **Lluís** i en **Mario**, per la seva col·laboració i suport en el temps que han estat al lab. I a la **Sílvia Carbonell** i a en **Marc Güell**, per la vostra amistat, i per totes les cerveses i converses compartides. Moltes gràcies a tots!!

A tota la gent del consorci de leucèmia, especialment la **Sílvia Bea** i a tota la gent del **Clínic** que m'ha ajudat a tirar endavant l'últim projecte d'aquesta tesi. Thanks as well to the people from **Stephan**'s group, especially **Stephan**, **Oliver** and **Luis**. It was really nice to work with you! Thank you very much!!

Als meus amics amb qui vaig començar tota aquesta història de la biologia i que han sigut un dels recolzaments més importants que he tingut durant aquests últims temps. Perquè ja són deu anys, noies i noi, perquè hem viscut moltíssimes coses i perquè sou ja una part molt important de mi: moltíssimes gràcies per tot i més **Anna**, **Montse**, **Núria** i **Xevi**. I moltes gràcies **Inma**, perquè malgrat haver-te conegut posteriorment, el teu suport i la teva amistat d'aquests últims anys també han estat molt importants. Moltíssimes gràcies!!

A tota la gent que ha format part de la meva altra vida, la que queda fora dels "murs" de la tesi i del PRBB. A les meves cosines **Ingrid** i **Marta**, per la vostra amistat i per la comprensió durant tot aquest procés; a l'**Helena**, per tot el que hem compartit i per ser tan gran; a la **Marina G.**, per ser tan tu i, òbviament, per ser una gran companya de caminades; a la **Debora**, per ajudar a esbargir-me tantes vegades per Barcelona, de dia o de nit; a en **Ribas** i l'**Endika**, per haver-me acollit a casa vostra i per tots els sopars, cerveses i partits del Barça; a la **Marina C.**, la **Laura M.**, la **Paula**, l'**Adrià**, en **Jorge**, la **Mary**, l'**Anna A.**, en **Carles**, l'**Àngels**, i tota la resta de la gent amb qui he compartit sopars, calçotades, cocktelades, barbacoes, esquiades i demés activitats d'oci que han ajudat tant a oxigenar la meva ment. Moltes gràcies tots per aquests moments!!! I a la **Lídia**, per ser-hi sempre, des de fa tants anys. Mil milions de gràcies, nena!!

A tota la meva família, moltíssimes gràcies també! Als meus **pares**, pel vostre suport incondicional, per haver cregut sempre amb mi i per fer-m'hi creure a mi també. A en **Toni**, el meu germà, el meu millor amic. No saps com em sento d'afortunada de tenir-te. I a la **Raquel**, per ser-hi, i per la teva amistat, des del primer dia. Moltíssimes gràcies de tot cor, sense vosaltres aquesta tesi no hauria estat possible.

Finalment, merci beaucoup à toi, **Julien**. Pour ton soutien et la patience infinie au cours de ces derniers mois. Mais, surtout, parce que tout ce que tu donnes es si incroyablement grand.

MOLTES GRÀCIES A TOTS !!

Abstract

Over the last ten years, improvements in molecular techniques and the arrival of the next-generation sequencing technologies have revealed a large amount of structural variation (SV) in the human genome. Consequently, there has been a significant increase in interest from the scientific community to understand the role of the SV in diseases, such as cancer, or in determining phenotypic traits in the general population. The objective of this thesis has been to study in depth the characterization and the functional importance of the SV, through the analysis of different methods for its detection and its biological impact in two different contexts. First, we have analysed the presence of copy-number variants in several human populations using a microarray approach and, by validating one of the detected regions, we have confirmed the reliability of this method for the detection of this type of SV. Second, through the chronic lymphocytic leukaemia (CLL) genome project, we have identified structural variants in patients with CLL by whole-genome sequencing. To obtain a comprehensive analysis of the SV in cancer genomes, we have developed a computational tool with the capacity to characterize and define all forms of the SV using next-generation sequencing data. With this tool we have detected, on one hand, some novel variants in CLL and, on the other hand, a high level of genomic complexity in one of the patients studied. From this last case, we have carried out the evaluation of the phenotypic impact of the complex variants in the progression of the CLL, which has allowed us to determine the importance of analysing cancer as a dynamic process undergoing evolutionary changes over time.

Resum

En els últims deu anys, les millores en tècniques moleculars i l'aparició d'una nova generació de tècniques de seqüenciació han revelat que existeix una gran quantitat de variació estructural (SV, en anglès) en el genoma humà. En conseqüència, hi ha hagut un augment significatiu en l'interès de la comunitat científica per entendre el paper que la SV juga en malalties com el càncer, o en la determinació de trets fenotípics en la població general. L'objectiu d'aquesta tesi ha estat aprofundir en la caracterització i la importància funcional de la SV, analitzant diferents mètodes per a la seva detecció i el seu impacte biològic en dos contextos específics. En primer lloc, hem analitzat la presència de variants en nombre de còpia en diferents poblacions humanes mitjançant una tècnica de *microarray* i, a través de la validació d'una de les regions trobades, hem pogut confirmar la fiabilitat d'aquesta tècnica per detectar aquest tipus de SV. En segon lloc, a través del projecte del genoma de la leucèmia limfàtica crònica (LLC), hem caracteritzat variants estructurals en pacients amb LLC mitjançant la seqüenciació completa del seu genoma. Per tal d'obtenir un anàlisi el màxim d'exhaustiu de la SV en genomes de càncer, hem desenvolupat una eina computacional capaç de caracteritzar i definir totes les formes possibles de SV utilitzant dades de seqüenciació. Amb aquesta eina hem pogut detectar, per una banda, algunes variants noves en LLC i, per l'altra, un alt nivell de complexitat genòmica en un dels pacients. A partir d'aquest últim cas hem dut a terme l'avaluació de l'impacte fenotípic d'un patró de SV complex en la progressió de la LLC, la qual cosa ens ha permès determinar la importància d'analitzar el càncer com a un procés dinàmic sotmès a canvis evolutius al llarg del temps.

Preface

The human genome contains nearly 3 billion base pairs of genomic information organized into chromosomes and small mitochondrial DNA. When the first draft of the human genome sequence was publicized in 2001, it was openly claimed that all the differences among individuals should be attributed only to 0,1% of the genome, which represents a total of ~3 million of the nucleotides. However, with the continuous improvement over the last decade in molecular biology, genomic technologies and bioinformatics skills, our knowledge of human genomic variation has progressed rapidly. We know now that human genomes are highly variable. This implies that the notion of the 99.9% of genome-sequence identity between two individuals might be an erroneous overestimation.

Structural variation (SV) has been recognized as a predominant source of genetic variation among human individuals. Thus, it is likely to make an important contribution to human diversity. However, the SV carries a certain degree of complexity and it still remains difficult to interpret with respect to its functional consequences. The importance of gaining insight into the characterization and interpretation of the SV is for its known active role in human disease and complex traits.

This thesis starts with an **introduction** divided in three main sections. The first two sections show an overview of the SV and the methods for its detection. The last one provides a general outlook about the characteristics of cancer genomes, and the known implications of the SV in tumorigenesis. Since the thesis includes analyses carried out in patients with chronic lymphocytic leukaemia, this last section also contains a general description of the disease. In the **main body**, four articles describing the different studies and the methodology followed in each of them are included. In the **discussion**, a brief summary and an interpretation of the results are provided, as well as an assessment of how this work has contributed to increase the knowledge in the field. Finally, the last section consists of a summarized list of the main **conclusions** of this thesis.

Contents

A. INTRODUCTION

Genomic variation contributes to common phenotypic differences between individuals, populations and species. It is a key element for disease predisposition and is an evident potential substrate for natural selection [1-3]. Variation in the human genome occurs on many different scales, including single nucleotide changes, small insertion-deletion alterations and larger genomic structural rearrangements. First efforts in exploring genetic variation in a wide manner were focused on the identification of single-nucleotide changes (commonly known as single-nucleotide polymorphisms or SNPs). So far millions of SNPs have been described in human populations, mainly through large-scale international projects such as the HapMap [4] and the 1000 Genomes Project (1000GP) [5]. In addition, many genome-wide association studies have allowed the association of numerous SNPs with common diseases and complex traits [6].

The recent discovery of a high number of larger genomic variants in the human genome, called structural variation (SV), has revealed the need of measuring them in comprehensive studies in human genetics research. The availability of high-throughput approaches for the screening of genomes is providing an exceptional opportunity to detect, characterize and analyse SV in an unprecedented detail, and to gain insight into the interpretation of their functional impact in human traits and diseases.

A.1. Structural variation in the human genome

The existence of structural changes in the human genome was first discovered as alterations in the quantity and structure of chromosomes under the microscope. These changes included structural aneuploidies affecting the counts of specific chromosomes that lead to Down syndrome [7], Turner syndrome [8] and Klinefelter syndrome [9], among many others. Generally these aberrations were over 3 megabases (Mb) in size and were detected by cytogenetic methods. However, the high frequency and wide importance of the SV in shaping the variability of the human genome in health and disease, started to be apparent over the last decade with the advent of genome-scanning array technologies. Several analyses of non-diseased genomes uncovered an unexpectedly large extent of submicroscopic SV ranging from 1-kb to 3-Mb. This form of SV involved alterations on the number of copies of genomic fragments (deletions and duplications), which were grouped under the general term of copy-number variants (CNVs) [10-13]. With the observation of the high frequency of CNVs and owing to their comparatively large size, it began to be evident that the SV encompassed millions of bases of DNA and it had to be responsible for larger sequence differences between individuals than SNPs and other smaller classes of variation.

In the past five years, with the arrival of next-generation sequencing (NGS) technologies, the operational spectrum of SV has widened to include smaller events, increasing our capacity to detect all forms of structural alterations with an unprecedented detail [14, 15]. Currently, the classes of SV includes different types of unbalanced (changing the number of copies) and balanced (without altering the copy number) genomic rearrangements of blocks of DNA sequence of at least 50 base pairs (bp) in size [14, 16, 17] (Figure 1).

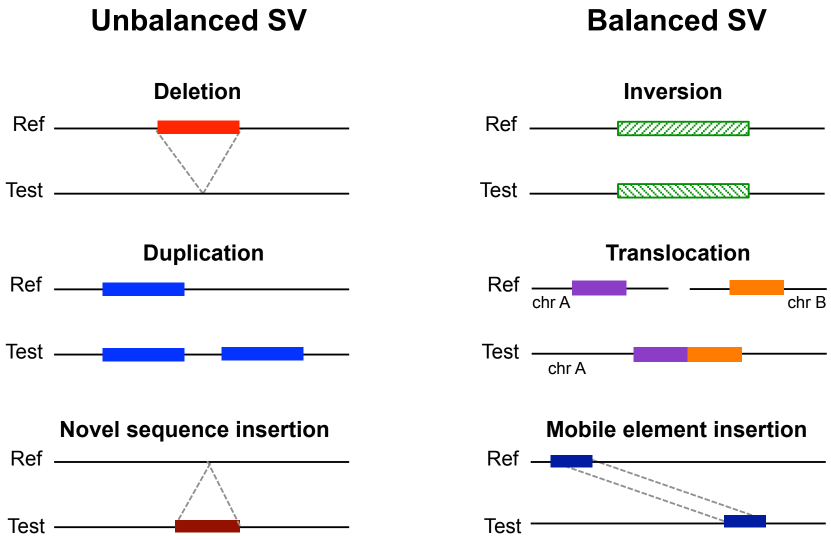


Figure 1. General classes of SV. The schematic illustrate deletions, duplications and novel sequence insertions (unbalanced SV), and inversions, translocations and mobile element sequence insertions (balanced SV) in a test genome (lower line) when compared with the reference genome (upper line)

Thus, SV is considered an independent class of genomic variation excluding SNPs, short insertions and deletions (indels) and variable number of short tandem repeats (VNTRs) (Table 1). SV can be inherited (germline SV) [11-14, 17-19] or can arise during the lifetime of an individual (somatic SV), as it has been observed in the analysis of several cancer genomes [20-22]. Current data [14, 18] suggest that any two humans might differ by 5,000-10,000 inherited structural variants and that both inherited and *de novo* SV can contribute to a variety of normal and disease phenotypes.

Variation type	Class	Size range
Single-nucleotide variant (SNV)	Single base-pair changes	1 bp
Small insertion/deletion (Indel)	Small insertions/deletions of short sequences	1-50 bp
Variable number of short tandem repeats (VNTRs)	Microsatellites, minisatellites and simple repeats	1-500 bp
Structural variation (SV)	Deletions, duplications, low-copy repeats, inversions and translocations	>50 bp
	Retroelement insertions/deletions	300 bp to 10 kb
	Large cytogenetically visible variants (aneuploidies, aneusomies, heteromorphisms, double minutes)	Entire chromosomes

Table 1. The spectrum of variation in the human genome

A.1.1. Simple and complex structural variation

SV can be categorized into two major groups on the basis of the structural features and breakpoint characteristics [23]: (a) **Simple** rearrangements, involving only one type of structural alteration at a specific region (i.e. a single deletion, duplication, inversion or translocation); and (b) **Complex** rearrangements, including a higher level of complexity with multiple clustered breakpoints of more than one simple event at the same region [22, 24-27]. Simple SV is usually recurrent and typical of the germline, and is characterized by the existence of the same breakpoint interval in multiple unrelated individuals. In contrast, complex SV often include those non-recurrent variants more commonly appearing somatically and presenting differences in size, extension and breakpoint positions amongst unrelated subjects.

Simple SV has been largely characterized during the last decade by different types of genome-scanning approaches and using numerous samples of healthy and diseased individuals of diverse geographic origins [10-15]. Instead, the existence of complex chromosomal rearrangements has become more apparent since five years ago through the analysis of NGS data [28, 29]. In addition, the improvement of breakpoint resolution, provided by NGS approaches, has allowed the discovery of an unexpected high extreme form of genomic complexity, generally associated with cancer [22, 24, 30-33]. Such extreme form of complex SV, in which one or a few chromosomes bear dozens to hundreds of clustered rearrangements, was firstly revealed in a case of chronic lymphocytic leukaemia (CLL) and was termed *chromothripsis* [24]. Further studies have confirmed the existence of chromothripsis in several cancers [31, 32, 34, 35] but also in germline and other non-cancer cells resulting in constitutional disorders [36-38]. This suggests that the presence of highly complex SV might be relatively extended in the human diseased genome.

A.1.2. Molecular mechanisms of structural variation formation

Structural alterations generate a change in chromosome structure, joining two formerly separated DNA sequences. Knowledge of the precise junctions of the new structure is essential to estimate its formation mechanism, which is necessary to determine its functional impact and to design targeted genotyping assays. Genome-scanning array technologies and NGS approaches have allowed the characterization of a comprehensive set of SV junctions, providing the definition of five general molecular mechanisms that can generate SV (reviewed in [39, 40]) (Figure 2): (a) recombination between sequences with high homology, as it can occur by non-allelic homologous recombination (NAHR). This mechanism is similar to the normal biological process of homologous recombination during meiosis, but involving sequences that are not the same allele on the homologous chromosomes (Figure 2a); (b) aberrant ligation of double-strand DNA breaks (DSBs), which occurs mostly due to exposure to external DNA damaging agents, through non-homologous end-joining (NHEJ) or microhomology-mediated end-joining (MMEJ) (Figure 2b); (c) DNA replication errors, such as fork-stalling template switching (FoSTeS) or microhomology-mediated break-induced replication (MMBIR) (Figure 2c); (d) mobile element insertions (MEIs) (Figure 2d); and (e) single catastrophic event causing genomic shattering, followed by incorrect re-joining of the fragmented DNA (chromothripsis) (Figure 2e).

Historically, NAHR was considered the prevailing mechanism for the generation of structural rearrangements, involved in both genomic diseases [41-43] and in normal variation [44, 45]. The repeated sequences that recombine might occasionally be retrotransposons (i.e. young Alu and LINE-1 class (L1) elements) that occur widely in the human genome [46, 47], but are usually larger blocks of sequences occurring only twice or a few times (i.e. low-copy repeats, also known as segmental duplications or SDs). Indeed, many studies of copy-number polymorphisms noted a significant enrichment of SDs within intervals that probably contained the breakpoints of deletions, duplications and inversions, suggesting that SDs represent potential hotspots of genomic instability and therefore of the formation of SV [10-14, 17, 44, 48].

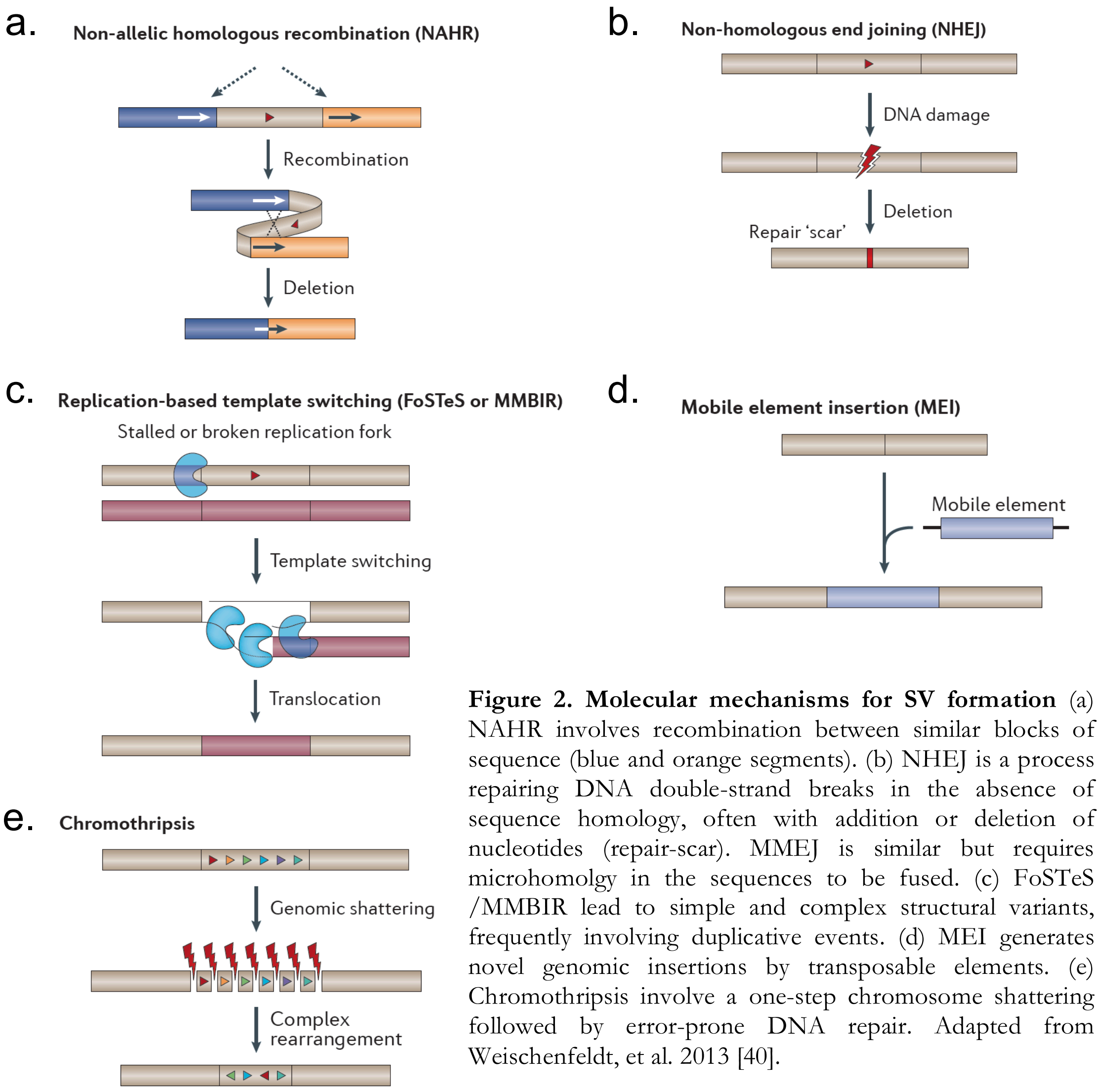

a.
Non-allelic homologous recombination (NAHR)
Recombination
Deletion
b.
Non-homologous end joining (NHEJ)
DNA damage
Deletion
Repair 'scar'
c.
Replication-based template switching (FoSTeS or MMBIR)
Stalled or broken replication fork
Template switching
Translocation
d.
Mobile element insertion (MEI)
Mobile element
e.
Chromothripsis
Genomic shattering
Complex rearrangement

Figure 2. Molecular mechanisms for SV formation (a) NAHR involves recombination between similar blocks of sequence (blue and orange segments). (b) NHEJ is a process repairing DNA double-strand breaks in the absence of sequence homology, often with addition or deletion of nucleotides (repair-scar). MMEJ is similar but requires microhomolgy in the sequences to be fused. (c) FoSTeS /MMBIR lead to simple and complex structural variants, frequently involving duplicative events. (d) MEI generates novel genomic insertions by transposable elements. (e) Chromothripsis involve a one-step chromosome shattering followed by error-prone DNA repair. Adapted from Weischenfeldt, et al. 2013 [40].

However, it has been demonstrated that the frequency of variants formed by NHEJ, MMEJ or replication mechanisms (FoSTeS/MMBIR) is higher than previous estimates [25, 28]. These mechanisms might be responsible for the majority of complex, non-recurrent and small SV (<1 kb), whereas the proportion of homologous recombination is only relevant for simple, larger and recurrent variants. Non-recurrent SV have endpoints in many different positions in unrelated individuals, which implies that they arise at sites that lack extensive homology. Several studies reported the existence of specific properties that are found in the breakpoint junctions of non-recurrent SV [25, 28, 39, 49]: (1) novel junctions are formed at sites of microhomology (2-15 bp) too short to support homologous recombination; (2) the novel structure is usually complex, showing duplication and deletion interspersed with non-duplicated or triplicated lengths; and (3) breakpoints are localized in close proximity to SD.

In germline complex SV, features including multiple copy-number changes, evidence for long-distance template switching, insertion of short sequences at breakpoints, apparently "templated" from nearby genomic intervals, and microhomology at the breakpoint junctions, are consistent with such rearrangements being generated by a DNA replication mechanism (FoSTeS and MMBIR) [37, 39]. In contrast, in somatic complex SV, including those highly complex rearrangements derived from chromothripsis, features like the presence of abundant microhomology sequences at breakpoints, the lack of template-derived insertions and the existence of sequence losses rather than duplications, are consistent with the repair of shattered DNA fragments by NHEJ or MMEJ [24, 36, 38, 50].

On the other hand, it has been reported that a small fraction of mobile elements (mostly Alu, and L1 retrotansposons) remain still active in the human genome, with the capacity to generate SV not only by NAHR. The mechanism of SV formation by MEI is based in their own retrotransposable nature, which allows them to duplicate through RNA intermediates that are reverse transcribed and then inserted at new genomic locations ("jumping" events) [51]. The insertion of L1 and Alu elements at new genomic loci sometimes results in the concomitant deletion of an adjacent genomic sequence, although these L1 and Alu insertion-mediated deletions tend to occur at much lower frequency than single insertions [52, 53]. MEI is though to be the dominating insertion mechanism in the human genome and is responsible for the large number of insertions events of around 300-bp and 6-kb reported in a recent analysis of the 1000GP by NGS [14, 54].

A.1.3. Implications of structural variation for human disease and traits

SV encompass millions of bases of DNA containing entire genes and regulatory regions [11-13, 17]. Therefore, they can influence the biochemical, physiological, morphological and pathological variability among individuals. Some structural changes can have deleterious effect on the reproductive fitness, being in some cases lethal mutations. In these circumstances, the structural variants would eventually be predestined to disappear or to remain in a heterozygous form [55]. In other cases, changes can entail some advantage and therefore they might be under positive selection [56]. However, most of SV are neutral variants, without apparently implication in phenotype or disease. In this case, they can be recurrent or disappear in human populations due to neutral selective forces (such as balancing selection or genetic drift) [57].

As mentioned before, the phenotypic relevance of SV in the human genomes arose through the observation by cytogeneticists of the role of large chromosomal rearrangements in some genetic disorders [7-9]. However, given the emerging picture of genomes being rich in SV, it has become evident that many of variants might contribute to a wide range of common diseases and phenotypes (reviewed in [40, 58]). Currently, we know that the SV can influence on gene expression, phenotypic variation, adaptation and susceptibility to disease, as a result of altering gene dosage, disrupting coding sequences, generating fusion genes, perturbing long-range gene regulation, or unmasking of recessive mutations or functional SNPs on the remaining allele [1, 14, 59, 60] (Figure 3). Briefly, copy-number changes can add or remove entire copies or partial parts of genes, leading generally to a concomitant change in gene dosage. Insertions, deletions and inversions involving only part of a gene can potentially result in the formation of variant proteins through exon shuffling, the creation of splice variants, novel fusion genes or simply creating a truncated protein. SV outside of coding regions can lead to changes in gene expression through positional effects that might alter the location or affect essential regulatory elements. Finally, deletions may also act indirectly revealing recessive mutations on the single remaining haplotype.

Several studies have reported that many forms of SV affect coding sequences, but they show a bias in the types of genes found within structurally variable regions [13, 14, 61, 62]. Genes involved in inflammation, immune response and response to biotic stimuli appear to be significantly enriched. This indicates that genes affected by SV might have roles in the adaptability and fitness of an organism in response to external pressures more than a direct effect to disease. Therefore, SV entailing genes related to these enriched categories might be subject to positive selection and might play a role in human evolutionary adaptation. Otherwise, genes involved in intracellular processes, such as cell signalling, cell proliferation, and biosynthetic and metabolic pathways, are underrepresented in structurally dynamic regions. This probably reflects the sensitive effect that many genes have due to their fundamental role in transcriptional regulation and cellular development. Hence, the impoverishment of these gene functions within SV indicates a possible purifying selection acting against alterations of genes that could be directly associated to disease.

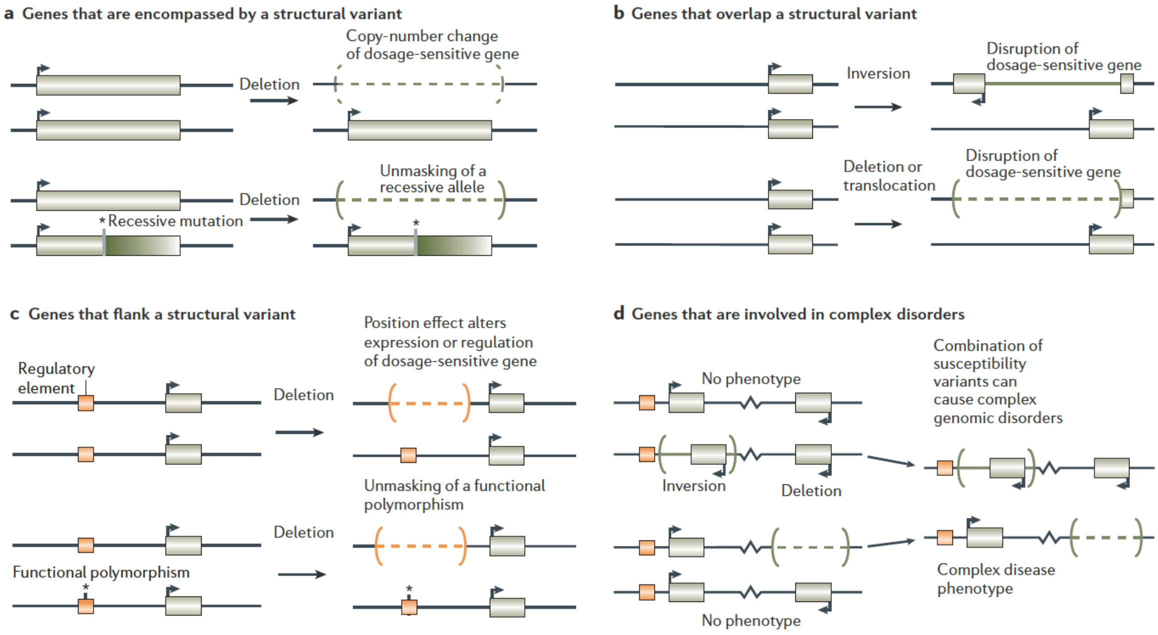

a Genes that are encompassed by a structural variant
Copy-number change of dosage-sensitive gene
Deletion
Unmasking of a recessive allele
Deletion
*Recessive mutation
b Genes that overlap a structural variant
Inversion
Disruption of dosage-sensitive gene
Deletion or translocation
Disruption of dosage-sensitive gene
c Genes that flank a structural variant
Position effect alters expression or regulation of dosage-sensitive gene
Regulatory element
Deletion
Unmasking of a functional polymorphism
Deletion
Functional polymorphism
d Genes that are involved in complex disorders
Combination of susceptibility variants can cause complex genomic disorders
No phenotype
Inversion
Deletion
Complex disease phenotype
No phenotype

Figure 3. **Influence of SV on phenotype.** (a) Deletion (or duplication) of dosage-sensitive gene and deletion unmasking a recessive mutation on the homologous chromosome; (b) Disruption of a gene through an inversion, deletion or translocation; (c) A deletion (or inversion or translocation) of a regulatory element can alter gene expression by positional effect. Alternatively, a deletion (or inversion or translocation) of a functional element could unmask a functional polymorphism within an effector, which could have consequences for gene function; (d) Combination of SV contributing to a complex disease state. Those SV individually do not produce phenotype. Taken from Feuk, L. et al. 2006 [16].

A complete understanding of the implications of SV in human disease and traits requires genome-wide studies that fully examine the frequency of SV among individuals. In this sense, samples provided by the HapMap [4] and the Human Genome Diversity Panel (HGDP) [63] have led the possibility to characterize a vast amount of SV (mainly CNVs and inversions) in thousands of individuals from worldwide populations [11-15, 17-19, 48]. This data has been catalogued in public databases such as the Database of Genomic Variants (DGV, http://projects.tcag.ca/variation/), the Human Structural Variation Database (http://humanparology.gs.washington.edu/structuralvariation/) and the 1000GP Browser (http://1000genomes.org/ensembl-browser/). They provide information on structural variants that are generally not known to directly cause disease (i.e. polymorphic SV). Similarly, databases cataloguing large-scale genotype-phenotype correlations have been developed to provide a source of information about the frequency of SV (mainly CNVs) and the likelihood of causing phenotypic outcome. For example, clinical findings associating directly the presence of submicroscopic chromosomal imbalances with a specific disease phenotype have been archived in DECIPHER (Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources, http://www.sanger.ac.uk/PostGenomics/decipher/) [64]. Furthermore, some disease-genome sequencing studies as the International Cancer Genome Consortium (ICGC, http://icgc.org) [65] have also begun to generate and classify SV genetic variation maps of somatic rearrangements associated with different type of cancers.

A.1.3.1. Structural variation in human disease

Diseases that are associated with specific forms of SV have been called *genomic disorders* [66]. Unbalanced SV (CNVs) can be responsible for sporadic birth defects, other sporadic traits and Mendelian or complex diseases, including a large list of different disorders like Down Syndrome, DiGeorge syndrome, Prader-Willi and Angelman syndromes, Williams-Beuren syndrome, attention-deficit hyperactive disorder, Crohn's disease, rheumatoid arthritis, diabetes, Alzheimer disease, Parkinson disease, systemic autoimmunity and psoriasis (reviewed in [2, 40, 67]. Most genomic disorders associated with CNVs are caused by deletions, whereas only a lower number of cases are the consequence of duplications [64]. In addition, duplications are mostly correlated with less-severe phenotype grading. This can be explained by more-pronounced dosage alterations in the context of deletions and by the possibility that deletions unmask recessive alleles present in the remaining copy of the respective region. Less is known about the functional implication of balanced SV (inversions and balanced translocations), but some of them have also been identified because of their involvement in human disease. For example, a recurrent 400-kb inversion has been found in 40% of patients with haemophilia A [68], as well as smaller inversions affecting the *IDS* gene in Hunter syndrome [69]. Other inversions, mainly involving SDs, have been identified with no detectable effects but conferring a predisposition to further disease-associated deletions in subsequent generation through unequal NAHR. Examples include a 1.5-Mb inversion at 7q11.23 found in some parents of patients with Williams-Beuren syndrome, which present a deletion of this same region [70], or an inversion of 4-Mb at 15q11-q13 that carry about half of the parents of patients with Angelman syndrome, which present microdeletions in such loci [71]. On the other hand, balanced chromosomal translocations between different genes or their regulatory sequences can generate a gain-of-function mutation. This mechanism is prominent in some cancers, mainly haematological malignancies [72]. For example, the fusion of *BCR* and *ABL1* genes, which leads to the formation of Philadelphia chromosome has been directly implicated in the development of chronic myeloid leukaemia (CML) [73]. This phenomenon has also been found in other diseases like the glucocorticoid-remediable aldosteronism (GRA), an autosomal dominant disorder characterized by hypertension [74]. Some inter-chromosomal rearrangements might have also positional effects, such as a translocation that disrupts the *HDCA9* gene at 7p21.1 that has it reciprocal breakpoint on chromosome 1, at 500-kb from the *TGFB2* gene. The patient carrying this translocation has Peter's anomaly, a defect of the anterior chamber of the eye that is more likely consequence of a positional effect at *TGFB2* rather than the *HDCA9* disruption [75].

From a general point of view, non-recurrent SV (complex or not) are more likely to contribute directly to disease phenotype, often independently of external factors. Instead, recurrent (and mainly simple) variants are generally considered polymorphisms or part of the "normal" variation between individuals. Their relation to high risk of disease is often conditioned by other factors that interact with each variant. These factors include SNPs, other SV and specific environmental conditions. Indeed, several studies of individual disease-related loci have identified SV with remarkably high levels of variability between populations [76-78]. This variability might be due to specific local circumstances influencing the presence of such variants, highlighting the relevance of the ethnic background in the susceptibility to disease. For example, population differentiation has been noted for the *CCL3L1* polymorphism, which influences human susceptibility to HIV infection, with median values of 3 and 6 copies per diploid genome in non-Africans and Africans, respectively [77]. A deletion on *UGT2B17* has been associated to osteoporosis and is more common in Eastern-Asian individuals [78]. In addition, a deletion involving the *CFHR3* and *CFHR1* genes, which is related to a decreased risk of age-related macular degeneration, has been found with high frequencies in African individuals and low frequencies in South American and Japanese populations [79]. Thus, investigating the medical impact of the recurrent SV requires that we understand the distribution of such variation within human population, as well as the factors shaping and influencing this variation. The characterization of the variability of the SV in different ethnic groups should provide important clues about adaptation to different geographic environments that could shape phenotype characteristics in humans.

A.1.3.2. Population genetics of structural variation

As it is previously commented, our understanding about the frequency, evolution and population genetics of SV began with the apparition of the HapMap Project [4] and the Human Genome Diversity Panel (HGDP) [63]. Analyses of SNPs and SV using samples from these two large-scale projects have revealed that genetic clusters closely correspond with human groups defined by ethnicity or continental ancestry [13-15, 80, 81]. Furthermore, these studies also have highlighted a gradual decrease of intra-population genetic variability as a function of the distance from sub-Saharan Africa, as expected under the model of out-of-Africa spread of human populations [82]. Finally, intra-population differences among individuals account for the majority of variation, while differences among major population groups represent a minimal fraction [3, 83].

Generally, there are four main evolutionary mechanisms influencing the distribution of variation within and between different populations. They are common to all classes of variant and include balancing selection, local adaptation, mutation-selection balance and founder effects or recent bottlenecks (reviewed in [55]).

Balancing selection refers to a selective process by which multiple alleles are actively maintained in a population at high frequencies (minor allele frequency or MAF >10%), leading to the conservation of genetic polymorphisms. The general idea is that diversity itself is the optimized product of the evolution: the preservation of genetic heterogeneity within a population would be the most favourable situation because it allows punctual adaptations in specific circumstances. Balancing selection might be the result of long-time adaptation and the affected variants probably became common before the arising of the contemporary continental-scale structure of human populations [15]. Otherwise, the same process can occur by chance in small populations, even if the new allele is mildly deleterious. A result of this procedure is illustrated in the unusually high frequency of sickle-cell heterozygotes in regions where the malaria is endemic [84]. Heterozygosis for the allele is selectively advantageous in these areas because it lowers malarial mortality at the cost of only mildly deleterious haematological effects. Thus, in balancing selection, any variant is never driven to fixation regardless of the intensity of selection because the heterogeneity (that is, heterozygosity) have a higher adaptive value than homogeneity (homozygosity).

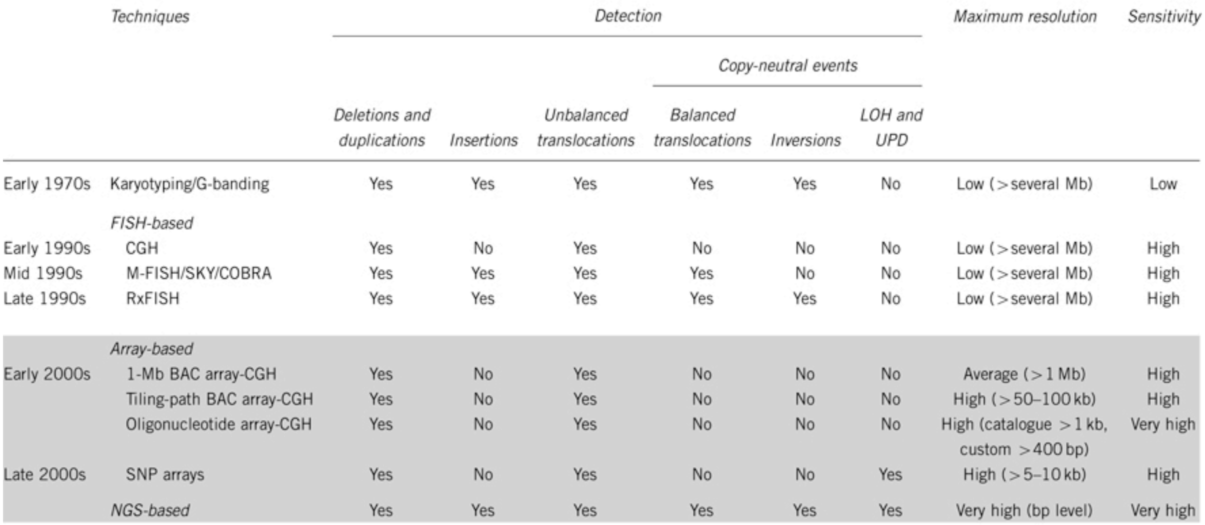
Some traits differ more dramatically between populations than within them. Variation of this type might reflect *local adaptation* to the diverse environments humans encountered following the out-of-Africa migrations 50,000-100,000 years ago [82, 85]. In this case, variants show lower global frequencies (MAF = 0.5-5%) and markedly patterns of population differences. Several CNVs have been reported showing patterns of diversity due to local adaptations. For example, at the *AMY1* locus, between 2 and 10 copies of the gene encoding salivary amylase have been detected, which correlate proportionally with population-specific differences in starch consumption, a prominent characteristic of agricultural societies and hunter-gatherers in arid environments [86]. Other examples include CNV in genes involved in xenobiotic detoxification (*GSTT1*), cellular immunity (*APOBEC3B*) and hormone metabolism (*UGT2B17*) [78, 87-89].

However, most variation in the genome is the consequence of recently appeared variants, particularly in the case of rapidly growing populations. New variants are found mainly in a low global frequency (MAF <0.5%) and are mostly observed in only one or few populations from a continental group [14, 15]. The *mutation-selection balance process* regulates their frequency. If new variants are functionally consequential, they will not become common unless they are under genetic drift, short-term balancing selection or contribute to local adaptation [90]. In most cases, functionally relevant new variants are deleterious and they would be under purifying selection. Many studies of CNVs have showed the relationship between a significant burden of human disease and low-frequency variants, particularly those exhibiting psychiatric and developmental phenotypes, including schizophrenia, autism, intellectual disability and craniofacial anomalies [91].

Finally, recent results from the 1000GP using NGS data have revealed the presence of variants with a very low global frequency (MAF <0.15) [15]. These variants are tightly population specific at the continental scale [92]. It means that most of variability of such variant is found in a very specific population. This is the consequence of *founder effects* or results from quite *recent bottlenecks*, which in fact are the result of an intense amount of genetic drift occurring at the leading edge of a population expansion [93]. As one set of population founders is further sub-sampled to produce a new group, alleles that are at low frequency in the ancestral population, or new mutations occurring during the expansion, can rapidly rise to high frequency in the newly colonized populations. This phenomenon has also been called "*allele surfing*", and it has been associated with general population expansions [94]. Examples of founder effect has been found in the Ashkenazi Jewish population, in which a set of genetic diseases are particularly common [95]. Likewise, recent bottlenecks have been also observed in isolated populations such as Finish and Balearic individuals, which carry excesses of rare variants [15].

A.2. Methods for the detection of structural variation

The detection of aneuploidies and SV began more than 50 years ago with the development of the first cytogenetic techniques, which led to the observation of large chromosomal aberrations under the microscope. With the posterior advent of microarray technologies and NGS the number, resolution and sensitivity of SV identified in any individual genome has increased dramatically (Table 2) (reviewed in [96]).



	Techniques	*Detection*						*Maximum resolution*	*Sensitivity*
					Copy-neutral events				
		Deletions and duplications	*Insertions*	*Unbalanced translocations*	*Balanced translocations*	*Inversions*	*LOH and UPD*		
Early 1970s	Karyotyping/G-banding	Yes	Yes	Yes	Yes	Yes	No	Low (>several Mb)	Low
	FISH-based								
Early 1990s	CGH	Yes	No	Yes	No	No	No	Low (>several Mb)	High
Mid 1990s	M-FISH/SKY/COBRA	Yes	Yes	Yes	Yes	No	No	Low (>several Mb)	High
Late 1990s	RxFISH	Yes	Yes	Yes	Yes	Yes	No	Low (>several Mb)	High
	Array-based								
Early 2000s	1-Mb BAC array-CGH	Yes	No	Yes	No	No	No	Average (>1 Mb)	High
	Tiling-path BAC array-CGH	Yes	No	Yes	No	No	No	High (>50–100 kb)	High
	Oligonucleotide array-CGH	Yes	No	Yes	No	No	No	High (catalogue >1 kb, custom >400 bp)	Very high
Late 2000s	SNP arrays	Yes	No	Yes	No	No	Yes	High (>5–10 kb)	High
	NGS-based	Yes	Yes	Yes	Yes	Yes	Yes	Very high (bp level)	Very high

Table 2. Evolution of genome-wide methods for identifying different types of SV. Abbreviations: BAC, bacterial artificial chromosome; CGH, comparative genomic hybridization; COBRA, combined binary ratio labeling; FISH, fluorescence *in situ* hybridization; LOH, loss of heterozygosity; M-FISH, multiplex-FISH; NGS, next-generation sequencing; RxFISH, Rainbow cross-species FISH or cross-species color banding; SNP, single-nucleotide polymorphism; SKY, spectral karyotyping; UPD, uniparental disomy. Taken from Le Scouarnec, S and Gribble, SM 2011 [96].

However, systematic and comprehensive assessment of SV in a genome has been problematic owing to their complexity and multifaceted features and the nature of genomic sequence, often containing repetitive regions. Hence, despite the great technological improvement in this field, there are still different challenges to face. For the moment, to achieve a nearly complete map of simple and complex SV in a genome is necessary to use a combination of cytogenetic, microarrays and/or NGS approaches.

A.2.1. Cytogenetics

Human cytogenetics was born in the late 1950s with the fundamental discovery that normal human cells contain 46 chromosomes using the technique of *karyotyping* [97] (Figure 4a1). With this method, chromosomes are treated to spread apart from each other, which makes possible to discriminate and arrange them into pairs. Karyotyping allowed for the first time the observation of abnormal number of chromosomes in human disorders, such as Down syndrome [7] (Figure 4a2), Turner syndrome [8] and Klinefelter syndrome [9]. In the early 1970s, cytogenetic analysis became more powerful with the development of chromosomal *banding*. Metaphase chromosomes are stained using a special dye, producing patterns of dark and light bands along the length of each chromosome [98]. These banding patterns become the barcodes by which cytogeneticists can easily identify chromosomes, detect subtle deletions, inversions, insertions and translocations, and refine relatively the breakpoints (Figure 4b1). For example, chromosome banding allowed the identification of a translocation involving chromosomes 7 and 11 in patients with acute myeloid leukaemia (AML) [99] (Figure 4b2). However, the resolution of first chromosome banding was relatively limited because the total number of bands produced on metaphase chromosomes was low due to excessive condensation. The situation improved with the *high-resolution banding* [100], which allows the analysis of pro-metaphase or prophase cells, providing much longer chromosomes containing many sub-bands and therefore increasing the resolution. By this way, several chromosomal abnormalities, both numerical and structural, can be studied more easily and with higher precision, as is shown in the detection of a balanced translocation between chromosomes 11 and 22 represented in Figure 4c. By applying this technique, several already well-known clinical syndromes could be linked to small chromosome aberrations such as micro-deletions, like those found in Prader-Willi and Angelman syndromes, Smith-Magenis and Miller Dieker syndromes, among many others [101].

However, despite the development of high-resolution banding no aberrations were visible at the cytogenetic level in numerous patients showing clear clinical signs. In the early 1980s *fluorescence in situ hybridization* (FISH) was introduced to the cytogenetics field [102], leading to the apparition of molecular cytogenetics. FISH is based on the use of chromosome region-specific fluorescent-labelled DNA probes. Fluorescence microscopy reveals the presence and localisation of probes that bind to targeted complementary sequences in the chromosome, which have been traditionally metaphase chromosome spreads. FISH was rapidly applicable for various situations such as solving complex aberrations, detection of small submicroscopic deletions (Figure 4d), and even in studies of interphase cells. To facilitate the detection of inversions, translocations or unmasked aberrations, whole chromosome specific fluorescent-labelled probes or *paints* were used in a derivative technique called *chromosome painting* [103, 104]. This technique allowed to go one step further in the analysis of more complex genomic aberrations providing, for example, the detection of a deletion on chromosome 5 in patients with myelodysplastic syndrome (Figure 4e) [105]. This deletion was previously masked by a translocation at the same locus between chromosome 5 and chromosome 22. On the other hand, to increase resolution, a new methodology termed *Fiber-FISH* has been developed replacing condensed chromosomes with extended chromatin fibres [106, 107]. Fiber-FISH is used to resolve ambiguities in the order of genes in a chromosomal region, to analyse inversions, the organization of tandem duplications and to detect small-scale rearrangements in chromosomes [108] (Figure 4f).

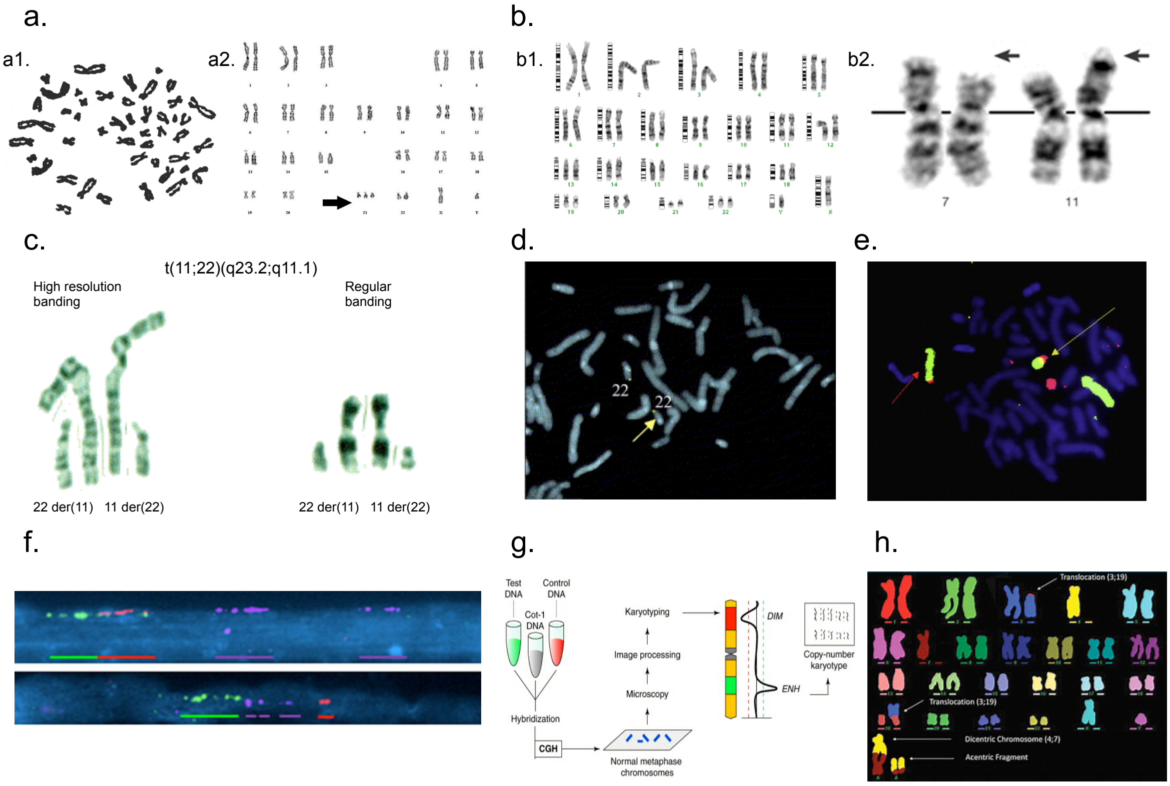


Figure 4. Evolution of SV detection by cytogenetics. (a) a1. First *karyotype*. From Trask, B 2002; a2. Trisomy 21 (Down Syndrome). Source: http://sahha.gov.mt/. (b) b1. Chromosomes by *chromosome banding*. Source: http://www.monctonhigh.ca/mcGBiology/cellular_division.htm; b2. Translocation involving the terminal bands of 7p and 11p (acute myeloid leukaemia). From Trask, B 2002 (c) Resolution of *high-resolution banding* (right) and chromosome banding (left): balanced translocation between 11 and 22. From Smeets, D. 2004. (d) Detection by *FISH* of submicroscopic deletions on 22q (DiGeorge syndrome). From Smeets, D. 2004. (e) Metaphase cell hybridized with whole *chromosome painting*, probes on 5 (green) and 22 (red). Red arrow pointing to the abnormal 5, and yellow arrow to the abnormal 22. From Hoffman, M et al. 2009. (f). Inversion by *fiber-FISH*. Top panel: co-hybridisation of three signals (green, purple and red) in human genome assembly; Bottom panel: co-hybridisation of the signals corresponding to an inversion regarding the genome assembly. From Molina, O et al. 2012. (g) Schematic view of the *comparative-genomic hybridisation* technique. From Baak, J et al. 2003. (h) Different aberrations by *spectral karyotyping* (SKY), where all chromosome pairs have their own color. Source: http://hps.org/hpspublications/journalarchive/155-2012.html.

However, these techniques are very time-consuming, can only be used for the analysis of a limited number of chromosomal loci at one time, and only provide information of loci whose status has been queried. More recently, in order to detect genomic aberrations at the whole-genome level without prior knowledge, other FISH methods based on hybridisation have been developed. For example, copy-number differences between two genomes can be detected using *comparative genomic hybridization* (CGH) [109], which will become the basis of more recent high-throughput hybridization-based approaches. In this approach, the genomic DNA of test and reference are labelled in two different colours and then hybridised to normal human metaphase chromosomes, where DNA sequences from both sources compete for their targets. Subsequently, ratios of the test and reference hybridization signals are quantified along the length of each chromosome, resulting in ratio values of 1, <1 or >1. Depending on whether the chromosomal regions are equally represented, a deletion, or a duplication, are detected [110, 111] (Figure 4g). Furthermore, *multiplex-FISH* (M-FISH) [112] and *spectral karyotyping* (SKY) [113], where all chromosomes are differentially coloured in a single experiment (Figure 4h)*,* have been developed to successfully detect prior unknown translocations and more complex rearrangements. All these methods have still limited resolution by the use of chromosomes as targets, are unable to well detect mosaicism and are also experimentally demanding and labour intensive. However, as a mature enterprise, cytogenetics is still largely used for the diagnoses of chromosomal disorders [114].

Other molecular techniques with higher resolution, mainly PCR-based, have been also developed to detect copy number changes. They are not cytogenetic approaches but, like them, they cannot be defined as high-throughput technologies for SV detection and therefore they are mentioned in this section. Perhaps the best established is the real-time quantitative PCR (qPCR) [115], which works well for scoring individual deletions and duplications but is not suitable for multiplexing. Alternative PCR-based methods for the simultaneous interrogation of multiple regions include *multiplex amplifiable probe hybridization* (MAPH) [116] and *multiplex ligation-dependent probe amplification* (MLPA) [117], in which copy-number differences for up to 40 regions can be scored in one experiment. Both methods rely on the hybridization of multiple locus-specific probes to their target sequences, which are then simultaneously amplified using fluorescently tagged universal primers, and the amount of each resulting product is quantified by capillary electrophoresis. However, these methods suffer from low information content, limited throughput, and the requirement of choosing candidate targets before the test.

A.2.2. Hybridization-based microarray approaches

Hybridization-based microarray approaches are commonly used to unmask CNVs (array comparative genomic hybridization or aCGH) and for SNP genotyping (single-nucleotide polymorphism microarrays or SNP arrays, also able to detect CNVs), as well as to study DNA methylation, alternative splicing, miRNAs and protein-DNA interactions (array-based ChIP –chromatin immunoprecipitation). Each array is composed of thousands of locus-specific probes immobilized onto a solid substrate, typically a glass slide. Labelled DNA or RNA fragments are applied to the array surface, allowing the hybridisation of complementary sequences between probes and targets. These methods allow the comparison of entire genomes at high resolution. In terms of SV, microarray-based methods have been the experimental workhorse of CNV discovery and genotyping during the last decade [10-13, 118]. They offer several advantages respect to cytogenetics for the screening of copy number changes, including higher resolution mapping, directly to genome sequence and higher throughput due to the massive parallelism of the assay. The main types of microarray used for the detection of CNVs are aCGH (Agilent and Roche Nimblegen) and SNP arrays (Affymetrix and Illumina) [119, 120]. Although SNP arrays have initially been designed for large-scale SNP genotyping, they can also be used for the detection of CNVs. Thus, both technologies infer copy-number gains or losses, but differ in methodological details and in the measures that they use for the characterization of such rearrangements.

Array-comparative genomic hybridization

Based on the same principle of comparative hybridization of two fluorescently labelled samples (test and reference) of conventional chromosome CGH approaches [110] (Figure 5). The main difference is that metaphase chromosomes are replaced by a set of cloned DNA fragments (probes) distributed along the genome. The distance and the sizes of the probes will determine the detection sensitivity and the resolution of the microarray. Probes now consist generally in oligonucleotides of 25-75 bp [121, 122], although historically they were performed using bacterial artificial chromosome (BAC) clones (80-200 kb in length) selected throughout the genome at 1-Mb intervals [123-125], or PCR fragments [126]. Each probe generates a fluorescent signal ratio between the test and reference samples, which is then normalized and converted to a $\log_2$ ratio (LRR), the measure used to infer copy-number [120, 127, 128]: an increased LRR represents a gain in the test compared with the reference; conversely, a decrease indicates a loss (Figure 6a).

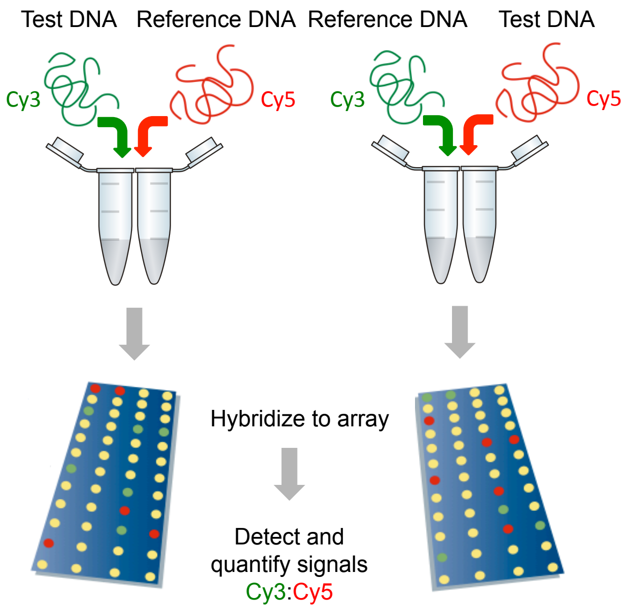


Figure 5. aCGH approach for the identification of copy-number variants. Reference and test DNA samples are differentially labelled with fluorescent tags (Cy5 and Cy3, respectively), and are hybridized to genomic arrays. The array can be spotted with BAC clones, PCR fragments or oligonucleotides. After hybridization, the fluorescence ratio (Cy3:Cy5) is determined, which reveals copy-number differences between the two DNA samples. Typically, aCGH is carried out using a "dye-swap" method, in which the initial labelling of the reference and test DNA samples is reversed for a second hybridization. This detects spurious signals for which the reciprocal ratio is not observed.

Single-nucleotide polymorphisms arrays

Based on the comparative hybridization of a single sample (test) against a collection of reference probes. In this case, the probes are specific to single-nucleotide differences (SNPs) between DNA sequences. SNP arrays allow the possibility to obtain not only a LRR metric, based on the sum of SNP signal intensities, but also an additional one called *B Allele Frequency* (BAF), based on the ratio between SNP signal intensities. Hybridization intensities are transformed to LRR and are compared with the average values derived from references, such that deviations from these averages indicate a change in copy-number [129] (Figure 6b). BAF metric enables a more comprehensive assignment of copy number, allowing the distinction of alleles, the identification of copy-neutral events such as segmental uniparental disomy (segmental UPD) or whole-chromosome UPD and identity by descent (IBD), extended regions of loss of heterozygosity, and the detection of mosaic occurrence of gains and losses [130-132] (Figure 6b). Hence, for example in cancer genomics, in which it is expected to find different clonality –or mosaicism- within tumour cells, or human diseases linked to uniparental disomy (reviewed in [133]), SNP arrays provide a valuable additional information.

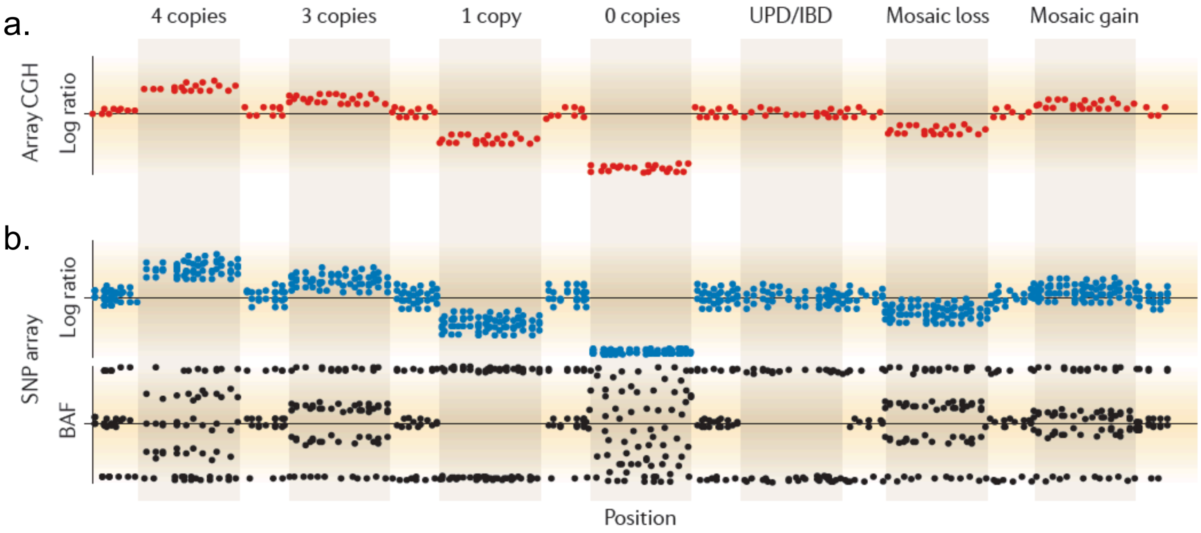


Figure 6. Log ratio (aCGH and SNP array) and B-allele frequency (SNP array). (a) Log_2 ratio (LRR) metric obtained by aCGH between a test and reference sample. An increased LRR represents a gain in copy number in the test compared with the reference; conversely, a decrease indicates a loss. (b) SNP arrays generate a similar LRR metric by comparing the signal intensities for the sample being analyzed to a collection of reference hybridizations, but with a lower per-probe signal-to-noise ratio than aCGH. Additional metric of B-allele frequency (BAF) has a significantly higher per-probe signal-to-noise than the LRR, may be used to more accurately assign copy-numbers, and allows the detection of copy-neutral events such as segmental uniparental disomy (UPD) or whole-chromosome UPD and identity by descent (IBD). In addition, it can be used to reliably detect and type low-level mosaic gains and losses. Adapted from Alkan, C. et al. 2011 [135].

Early studies of CNVs were based on aCGH BAC arrays or low-resolution oligonucleotide platforms. Although they highlighted the incredible number of CNVs in normal samples, they only allowed detection of CNVs greater than 50-100 kb and with very poor breakpoint resolution [10, 12, 13]. Early SNP arrays also demonstrated poor coverage of CNV regions and were used as complements to aCGH platforms to offer higher confidence in CNV detection [13]. With the advent of aCGH comprising longer oligonucleotides, which was first implemented in an assay format known as representational oligonucleotide microarray analysis (ROMA) and after by Agilent and Nimblegen companies, the resolution of CNV detection improved up to 10-30 kb depending on the platform [126]. Currently, an ultra-high-resolution oligonucleotide aCGH is capable of detecting changes to a resolution of 500 pb with breakpoints definition being precise enough to allow the identification of, for example, sequence motifs [61, 134]. Recent SNP arrays have also improved their efficiency of CNV discovery, incorporating better SNP selection criteria for complex regions of the genome and specific non-polymorphic copy-number probes [118, 129], although their resolution is not as great as ultra-high-resolution aCGH (reviewed in [135]).

The most important advantages of microarrays are in terms of throughput and cost. They enable data for thousands of relevant genomic regions of interest with any prior knowledge to be generated rapidly for a large number of samples in a cost-effective manner. In addition, the amount of input sample material required is generally low, usually <1μg of DNA. These have led to their widespread adoption in clinical diagnosis, essentially replacing karyotype analysis for the diagnosis of some type of diseases such as developmental disabilities or congenital anomalies [136].

On the other hand, the general limitations of microarrays are that they can not detect balanced rearrangements such as inversions or reciprocal translocations, they only identify copy number differences of sequences used to design probes, they do not provide the location of duplicated copies and are generally unable to revolve breakpoints at the single base-pair level. In addition, they tend to undergo a reduced sensitivity in the detection of single copy gains (3 to 2 copy-number ratio) compared with deletions (1 to 2 copy-number ratio) [48, 118, 120]. Nonetheless, the most important limitation in terms of CNV detection is the use of hybridization probes in repeat-rich and duplicated regions. Microarray platforms assume each location to be diploid in the reference genome, which is not valid in duplicated sequence. This is particularly challenging because CNVs are enriched in SDs regions and many breakpoints lie within the duplicated blocks of sequences [10-12, 14, 44, 48].

A.2.3. Next-generation sequencing technologies

Conventional Sanger sequencing [137], considered as a "first generation" technology, had dominated the industry for four decades, leading to a number of monumental accomplishments and allowing the completion of the first sequencing of a human genome [138, 139]. However, sequencing one individual took more than a decade of international effort and approximately US$3 billion (around US$300 million in sequencing reagents), showing therefore a need for new and improved technologies for sequencing large numbers of human genomes. With the arrival of next-generation sequencing (NGS) technologies around 2005, sequencing of a whole human genome can now be achieved in a few days and for around 3,000 euros. In addition to the capacity to identify single nucleotide changes and small deletions and insertions (indels), NGS can provide an accurate detection of the SV. In comparison with microarray approaches, the main advantage of NGS is that, in a single experiment, it is possible to identify all types of SV, both balanced and unbalanced events, and with an unprecedented high resolution. The use of NGS has expanded the spectrum of SV to include tens of thousands of smaller events (>50-bp) [5, 14]. At present, the most commonly used platforms have been developed by Illumina (Genome Analyzer/HiSeq), Roche (454 Genome Sequencers) and Applied Biosystems/Life Technologies (SOLiD). They are quite diverse in sequencing biochemistry but their workflows are conceptually similar (reviewed in [140, 141]).

NGS technologies allow the sequencing of millions of DNA or RNA molecules simultaneously after library preparation of fragments to produce sequence reads of 30-400 bp covering all the genome [140, 141]. These reads can be generated through the mate-pairs or paired-end strategies, which differ in the methodological procedure but are similar from the perspective of the final computational analysis. Briefly, in both strategies two paired reads are generated at an approximately known distance in the donor genome. Mate-pairs are created in circularized fragments joined with an adaptor and then the two reads are generated sequencing around the adaptor (Figure 7a). In contrast, paired-end reads are generated by the fragmentation of genomic DNA into short segments followed by sequencing of both ends of the segment (Figure 7b). Sequence reads are then aligned to the reference genome or assembled *de novo*, and single nucleotide changes, small indels and all types of SV can be detected. The paired-end sequencing strategy is generally used to create small insert-sizes libraries (300-400 bp), which have the advantage of a tight size selection of DNA fragments and therefore greater sensitivity for small intra-chromosomal events and breakpoint characterization. In contrast, mate-pair strategy is usually used for larger insert-sizes libraries (2-5 kb), which allows greater genomic coverage per sequenced fragment and to jump repetitive sequences.

Although NGS technologies have reduced considerably the cost of a whole-genome sequencing analysis, they are still relatively expensive when several genomes have to be analysed. In this sense, an alternative strategy has been developed by Roche Nimblegen (SeqCap EZ) and Agilent (SureSelect Target Enrichment) companies, in which specific genomic regions of interest can be selected and isolated from the genome by *sequence capture*. This strategy can be used to examine all of the exons in the genome (exome-sequencing [142]) or sequencing of specific targeted regions [143], such as specific gene families or Mb-size regions that are implicated in disease. Sequence capture allows sequencing a specific subset of the genome across many individuals rather than the whole genome of fewer samples, with the advantage to increase the sequence coverage of each targeted region. However, this strategy limits the capacity of SV detection since only those variants with breakpoints falling in targeted regions are possible to be discovered. Thus, sequence capture is basically useful when the objective is to characterize or genotype at a high resolution the breakpoints of previously known structural variants in multiple individuals.

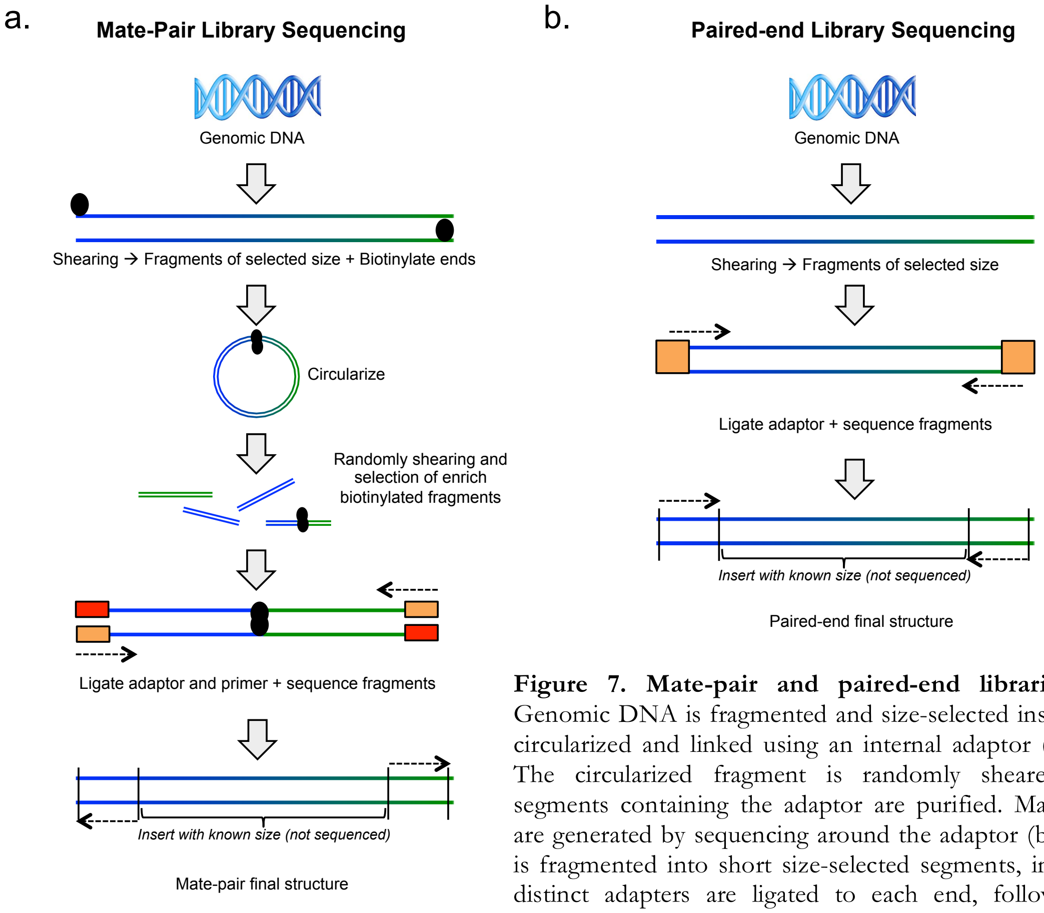

a.
Mate-Pair Library Sequencing
Genomic DNA
Shearing → Fragments of selected size + Biotinylate ends
Circularize
Randomly shearing and selection of enrich biotinylated fragments
Ligate adaptor and primer + sequence fragments
Insert with known size (not sequenced)
Mate-pair final structure
b.
Paired-end Library Sequencing
Genomic DNA
Shearing → Fragments of selected size
Ligate adaptor + sequence fragments
Insert with known size (not sequenced)
Paired-end final structure

Figure 7. Mate-pair and paired-end libraries. (a) Genomic DNA is fragmented and size-selected inserts are circularized and linked using an internal adaptor (biotin). The circularized fragment is randomly sheared, and segments containing the adaptor are purified. Mate-pairs are generated by sequencing around the adaptor (b) DNA is fragmented into short size-selected segments, in which distinct adapters are ligated to each end, followed by sequencing of the ends.

.

The production of large numbers of low-cost reads makes the NGS platforms useful for many other applications. For example, it is possible to catalogue the full transcriptome by RNA sequencing, where shotgun libraries derived from mRNA or small RNAs are deeply sequenced [144]. In addition, epigenetic marks and chromatin structure can be analysed by ChIP-seq (deep sequencing of DNA fragments pulled down by chromatin immunoprecipitation), and the methylation profiles can be obtained by methylome sequencing (deep sequencing of bisulfite-treated DNA) [145].

A.2.3.1. Detection of structural variants by next-generation sequencing

For SV discovery by NGS there are four general types of strategy [14, 135, 146]. They can be used both for mate-pair and paired-end libraries and all of them focus on mapping sequence reads to the reference genome and the subsequently identification of patterns or signatures that are suggesting different classes of SV:

Pair-read method

Also called paired-end mapping strategy [11, 29] is currently the most powerful and extended method to detect all types of SV (Figure 8). Pair-read (PR) began in 2005 when Tuzun et al. [11] utilized over 1.1 million of paired-end sequences from a high-density library of fosmids (that is, phage cloning vectors with DNA packaging limited to 40-kb) for SV discovery. Later, Eichler and colleagues [19] constructed new fosmid libraries from eight HapMap samples and sequenced both ends of approximately one million clones per genome, finding and validating hundreds of SV. The PR method assesses the span and orientation of paired-end or mate-pair reads and cluster "discordant" pairs in which the mapping span and/or orientation of the read pairs are inconsistent with the reference genome. Thus, the mapping procedure is the basis for understanding how the PR strategy allows the detection of SV. When aligned to the reference genome, read pairs are expected to map at a certain distance corresponding to the average library insert size, which can be from 200-bp up to 5-kb. Then, read pairs that map too far apart from the average insert size define potential deletions (Figure 8a); those found too close together are indicative of novel insertions, which are limited by the insert size (Figure 8b); those in which one of the reads maps with aberrant orientation can delineate inversions (Figure 8c); those found with an incorrect order can define tandem duplications and intra-chromosomal translocations (Figure 8d); and those in which one read maps in one chromosome and the other in another one are indicative of inter-chromosomal translocations (Figure 8e) [11, 19, 27, 29]. Although the widely application of this method and the many computational tools based on it [147-149], there are three important limitations that avoid the perfect detection and characterization of SV: (1) the accurate prediction of breakpoints depend on very tight fragment size distribution (low variability of read pairs insert sizes), which can make the library construction difficult and costly [146]; (2) PR depends on the mapping quality and so far there is not a mapping algorithm that completely resolve ambiguous mapping assignments generated by repetitive genomic regions; and (3) the insert size of the library is a limitation for some types of SV. Whereas it is not possible to detect insertions larger than the library insert size (often around 400-bp), small deletions or insertions can be missed with large-insert libraries because the expected size variance between the pairs will not be significantly altered [150].

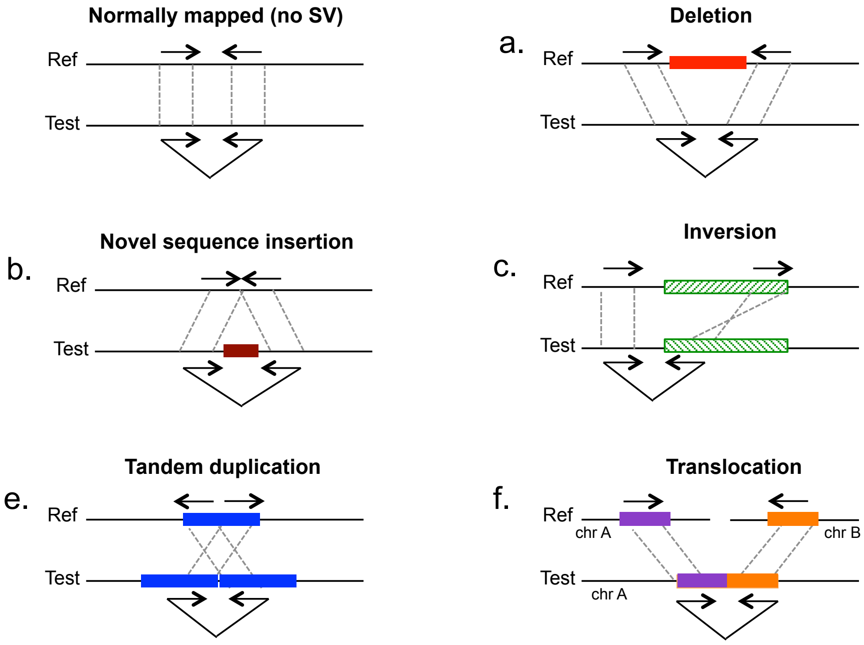


Figure 8. Pair-read method for SV detection. Pairs of sequenced reads (black arrows) are mapped to the reference genome. In deletions (a) and novel insertions (b) the mapped distance is different from the insert size. In inversions (c), the order of the two reads is preserved but one of the changes orientation. In tandem duplications (e) (also applicable for intra-chromosomal events), reads map out of order but with proper orientation. In translocations (f), reads map with correct order and orientation but in different reference chromosomes.

Read-depth method

The read-depth (RD) approach is the only sequencing-based method to accurately predict absolute copy numbers [151], essentially providing similar information to that obtained by aCGH by indicating genomic gains or losses [20, 152-154]. The method assumes that the number of reads mapping at each chromosomal position (sequence read-depth) is random (Poisson distributed). Then, a significant deviation from the mean of the distribution is an indicative of a CNV “position”. The basic idea is that, when compared to diploid regions, duplications will show significantly higher read-depth sequence, heterozygous deletions will show a reduced read-depth and absence of reads will be suggestive of homozygous deletions (Figure 9). However, factors such as GC content, homopolymeric stretches of DNA or preferential PCR amplification at the library preparation stage can introduce biases. In addition, repetitive genomic regions are also problematic as reads are aligned with low confidence, providing poor information on copy-number status, and it is not possible to discriminate between tandem and interspersed duplications. Finally, RD does not allow the detection of balanced rearrangements.

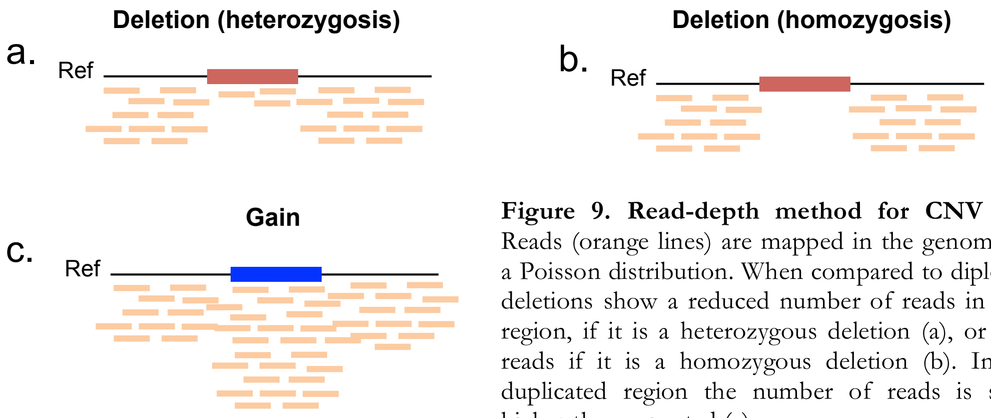


Figure 9. Read-depth method for CNV detection. Reads (orange lines) are mapped in the genome following a Poisson distribution. When compared to diploid regions, deletions show a reduced number of reads in the deleted region, if it is a heterozygous deletion (a), or absence of reads if it is a homozygous deletion (b). Instead, in a duplicated region the number of reads is significantly higher than expected (c).

Split-read method

The split-read (SR) strategy was initially developed for Sanger sequencing and is capable of detecting SV breakpoints with a high resolution. The basis of SR is to select those read pairs that have been aligned to the reference genome responding to the following criteria: one read maps perfectly (no mismatches) and uniquely, and the other read of the pair is classified as unmapped because presents a "split" sequence signature (that is, the alignment of the read to the genome is broken because it falls in a rearrangement breakpoint) [155-157]. Then, following similar criteria than in the PR strategy, when the unmapped read is split into two fragments that map separately, this is suggestive of a deletion breakpoint if the alignment is in the same chromosome (Figure 10) or an inter-chromosomal translocation breakpoint if is in different chromosomes.

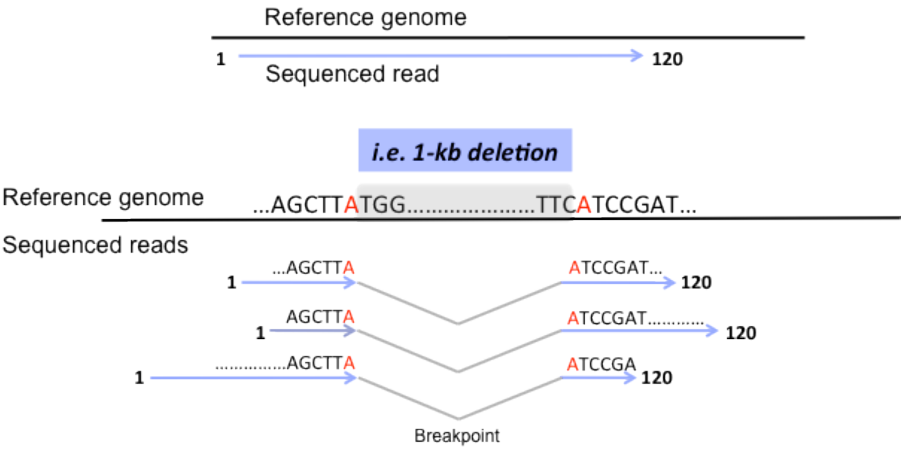


Figure 10. Example of a deletion breakpoint detected by split-reads method. Sequenced reads pointing to the same breakpoint are splitted at the nucleotide where the breakpoint occurs, and a half of a read is mapped correctly and the other in a greater distance than expected for the insert size. This situation is the same for inversions and translocation breakpoints, with the only difference that one part of the read map with inverted orientation or in another chromosome, respectively.

On the other hand, inversion breakpoints will be found when one of the fragments maps in different orientation respect to the other. The SR approach, in contrast to the other methods allows the detection of very small SV and indels, but is currently reliable only in the unique regions of the genome. Therefore the application of this method to current NGS data sets is limited owing to the difficulty in aligning uniquely short reads.

Sequence assembly method

By the sequence assembly method (ASM) all forms of SV could be, *a priori*, accurately typed for copy, content and structure. In fact, the most complete identification and characterization of SV should be possible with the sequencing of a whole genome followed by ASM and comparison to high-quality reference. However, ASM tend to be heavily biased against repeats and duplications owing to assembly collapse over such regions, and the enormous amount of data creates an inevitable barrier to assembly process in terms of memory usage [158]. Several assemblers have specifically been developed with some success to assemble large genomes [159-161], but its application has been used mainly to discover the content and location of novel sequence insertions [162]. One of the long-term goals of NGS should be the *de novo* assembly of human genomes to a standard comparable to or better than that of the current human reference genome (GRCh37 or hg19).

None of these methods for discovering SV using NGS data is comprehensive. In addition, when many algorithms using different types of strategies and experimental techniques are applied to the same samples, a significant fraction of the validated variants remains unique to a particular approach [14] (Figure 11). In order to improve the sensitivity and specificity of SV detection, recent algorithms try to incorporate at least two types of methodologies [163-166], but there is currently no suite of algorithms that could be applied to systematically resolve all types of SV. Biases remain in terms of content, size and class of SV. So far, most discovery efforts have been focused on deletions in unique sequences [5, 14] because they are the easier and the more confident type of structural variant to detect.

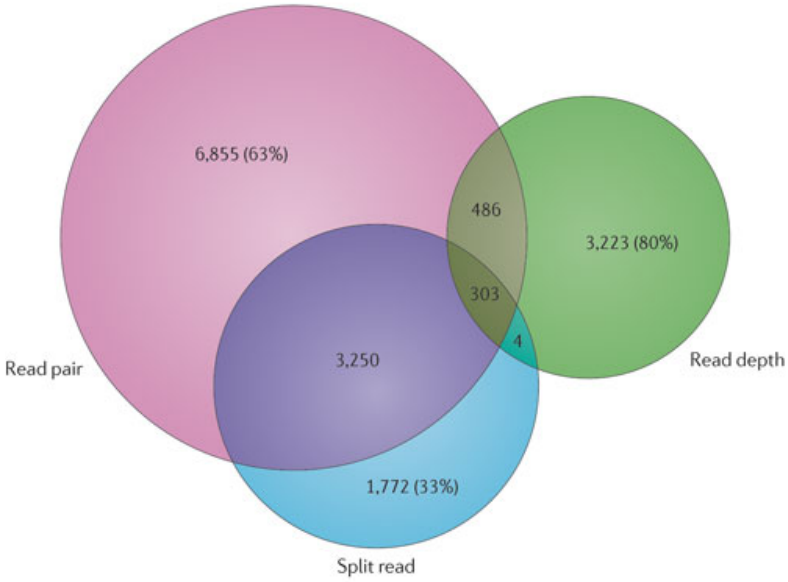


Figure 11. Comparison of SV detection by three different strategies. This Venn diagram shows the number of unique and shared SV found by read-pair, read-depth and split-reads methods that have been used in the 1000 Genomes Project. Read-pair and split-reads strategies show the greatest extend of overlap, whereas read-depth and split-read are the most discordant approaches.. Adapted from Alkan, C. et al. 2009 [135].

The greatest problem of NGS to discover SV is probably the *short reads* generated by the sequencing technologies (from 30-bp to 400-bp depending on the platform [140, 141]), which are considerably shorter than those produced by conventional Sanger sequencing. Considering the nature of the human genome, containing widespread common repeats and SDs, there is considerable read-mapping ambiguity that affects substantially these analytical methods that rely on sequence alignment. Generation of longer reads (of few kb) would increase mapping specificity, facilitating the accurate discovery of SV by PR in more complex regions of the genome and increasing the “capacity” of SR and ASM methods. Another great concern is *sequence coverage*, defined as the average number of times that each base pair in the genome is represented in an aligned read. The success of sequencing approaches in the detection and characterization of SV depend on obtaining sufficient sequence coverage because of the relatively high level of sequencing error in NGS. In addition, the sensitivity of the four strategies for SV detection explained above is also interdependent on the coverage. For example, with increasing coverage, the number of discordant read-pairs pointing to the same SV also increases, achieving a better sensitivity in the breakpoint resolution. However, when the objective is to analyse hundreds of individuals, the cost of high-coverage whole-genome sequencing is still prohibitive for most individual laboratories. Therefore, for the moment the general strategy in the scientific community in the characterization of genome is exome or specific target sequencing. Although this represents a limitation for a comprehensive characterization of the genomes, it allows the simultaneous analyses of many samples.

A.3. The Cancer Genome

Cancer is a major cause of death throughout the world [167] and, although we have a much greater understanding of cancer biology and genetics than ten years ago [168], the eradication or control of advanced disease has not yet been achieved. The translation of the molecular and genetic findings into the clinical practice is still difficult to accomplish due to the genomic complexity of cancer cells and their dynamics and evolutionary characteristics. These features provide both barriers to, and opportunities for, successful early diagnosis and treatment of cancer.

The genomes of all cancer cells carry somatic mutations, including base substitutions, small indels and all classes of SV [169]. The number of these somatic mutations in one cancer case can vary from a handful to hundreds or thousands (more usual). In addition, cancer genomes may acquire epigenetic changes, such as alterations in methylation of cytosine residues and other epigenetic modifications [170]. In some types of cancers it has also been found the presence of external DNA sequences in the tumour cell (for example, human papilloma virus, Epstein Barr virus, hepatitis B virus or human herpes virus), which has been acquired from exogenous sources and may contribute to the genesis of the cancer [171].

The traditional view of cancer evolution and progress is that of a gradual process analogous to Darwinian evolution, involving a continuous acquisition of genetic changes in individual cells by random mutation (Figure 12a), and natural selection acting on the resultant phenotypic diversity [172-174]. Selection favours those cells carrying alterations that confer the capacity to proliferate and survive more effectively than neighbour cells. Subsequently, these selected cells are driven to waves of clonal expansions, with the fittest clone coming to dominate the cellular compartment. However, recent analyses have revealed that cancers are actually mixtures of competing subclones, implying the existence of genetic heterogeneity within a tumour [174-177]. Indeed, tumours likely evolve through the competition and interactions between genetically diverse clones. The accumulation of mutations in cancerous and pre-cancerous cells over time has been demonstrated as a complex and dynamic process undergoing a continuous evolutionary selection [178]. Intrinsic factors, such as error-prone repair processes during normal cell division, cellular components as hormones and growth factors, local cell regulators or cell architectural constrains, are modulators of the cancer-cell environment. Furthermore, the effect of exogenous mutagens such as cigarettes carcinogens, ultraviolet light, and chemotherapeutic drugs [168], can also lead to sustained elevation of mutation rate and therefore to an evolutionary dynamism. However, the course of mutation acquisition might not be smooth, and predecessors of the cancer cells may suddenly acquire a large number of mutations as a consequence of the attrition of the telomeres [179], or due to the single catastrophic mitotic events suggested in some cancer genomes with chromothripsis [24, 31, 37]. These two last observations indicate that cancer evolution may also be accelerated or initiated by punctuated changes in the genome architecture that generate in cell alterations with advantageous capacity of proliferation, which will be subsequently clonally expanded (Figure 12b).

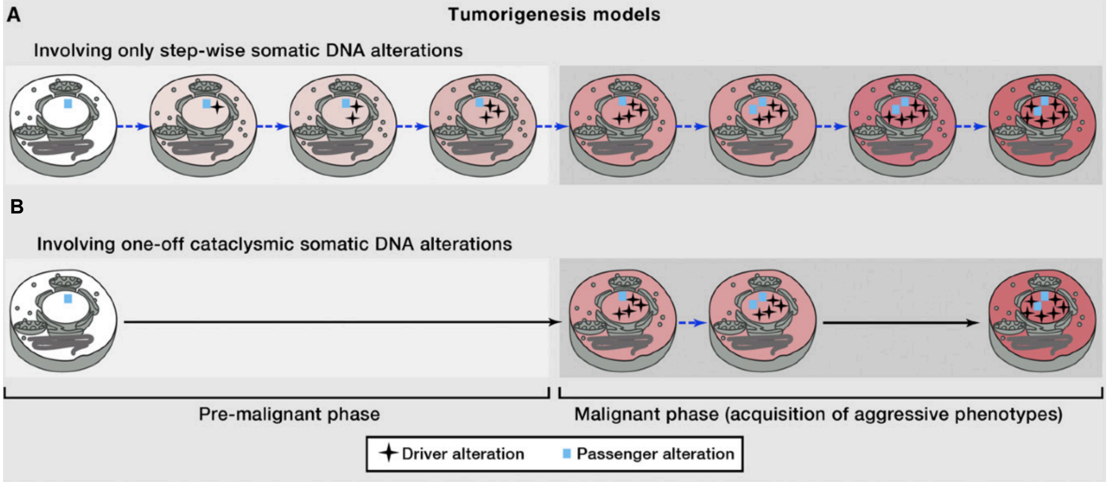

A
Tumorigenesis models
Involving only step-wise somatic DNA alterations
B
Involving one-off cataclysmic somatic DNA alterations
Pre-malignant phase
Malignant phase (acquisition of aggressive phenotypes)
Driver alteration
Passenger alteration

Figure 12. Models for cancer development. (a) Tumorigenesis is classically though to involve the stepwise acquisition of somatic DNA driver alterations (dashed blue arrows) (b) Cellular "crises", such as chromothripsis, may accelerate this process by resulting in several DNA alterations at once (solid black arrows). The red color symbolizes the acquisition of malignant phenotypes in the cell. Adapted from Korbel, J. et al. 2013 [206].

Not all the somatic abnormalities present in a tumour cell have been involved in the development of the cancer. Each somatic mutation may be classified according to its consequences for cancer initiation [169]: (a) *driver* mutations, which are those somatic alterations that confer selective clonal growth advantage and are causally implicated in oncogenesis; and (b) *passenger* mutations, which are the remainder alterations that do not confer growth advantage and therefore, not contribute to cancer development. Driver mutations can be identified by the observation that they occur more frequently in multiple cancers than would be expected in the normal mutation rate and that they are associated with clonal expansions [180, 181]. Driver mutations often alter specific classes of genes, which have been termed *cancer-related* genes. The proteins encoded by cancer-related genes normally regulate cell proliferation, cell differentiation and cell death, but mutations underlying oncogenesis might also occur in genes that mediate DNA-repair processes [182]. All these type of genes corresponds roughly 2% of the protein-coding genes and they have been catalogued in some databases such as COSMIC or IntOGen [183, 184]. In contrast, passenger mutations are randomly distributed throughout the genome and they do not affect any specific type of genes.

Mutations in around 10% of cancer-related genes have been also found in the germline, where they can influence cancer susceptibility [169, 182]. Germline mutations in cancer-related genes might represent inherited cancer susceptibility, especially when there is a familial history of cancer. Hereditary cancers are often characterized by the presence of mutations in DNA repair genes, which leads to genomic instability [185]. Thus, the presence of genomic instability probably precedes the acquisition of mutations in oncogenes and tumour suppressor genes and, therefore, leads to the acquisition of other cancer hallmarks [186]. Well-documented examples of hereditary cancer with mutations in DNA repair genes are hereditary non-polyposis colon cancer (HNPCC, also known as Lynch syndrome) [187], involving mutations in *MLH1, MSH2* and *MSH6* genes*,* and hereditary breast-ovarian cancer, involving mutations mainly in *BRCA1* and *BRCA2* genes [188, 189].

The definition of the type of mutations affecting cancer cells is important because they might influence how tumours will respond to therapy. Currently, there are several examples of specific chemotherapy regiments used in cancers carrying mutations in genes involved in response to DNA damage (DDR) [190]. For example, in familial and sporadic forms of breast-ovarian cancers associated with alterations in *BRCA1* and *BRCA2*, *platinum salts* are frequently given to patients due to their mechanism causing DNA cross-link strand breaks. In tumour cells that lack homologous recombination repair, these agents may be particularly effective [191]. Otherwise, an alternative recent approach to chemotherapy is the design of anticancer drugs that act specifically against DDR components. For example, genes with kinase domains or topoisomerases are considered "druggables" and therefore specifically targeted in therapy [192]. A paradigm of such strategy is the development of *imatinib* and subsequent generations of small-molecule inhibitors of the constitutively activated *BCR-ABL* kinase generated by the Philadelphia chromosome in CML [193].

However, the main problem in cancer therapy is the appearance of therapy-resistance tumour cells. As it occurs at the organisms level, the genetic heterogeneity within tumour cells might be considered an optimized product of the evolution of the cancer process. This genetic variation leads to the opportunity for some cells to become resistance to therapy. For example, mutations that were previously passengers in minor subclones in absence of therapy may be converted into driver mutations when the selective environment is changed by the initiation of treatment [176-178]. Indeed, very few advanced or metastatic malignancies can be effectively controlled or eradicated. In this sense, great expectation has been placed in cancer-genome sequencing. The successive sequencing of cancer genomes should allow to determine the genomic profile for each patient and to eventually influence treatment strategies as a tool to individualise and direct cancer treatment.

To make the increasing large amount of data available to the entire research community, three main international networks of cancer genome projects (The Cancer Genome Project, The Cancer Genome Atlas, and The International Cancer Genomic Consortium) [65, 194, 195] have generated public databases to host the data of the studies that they coordinate. The principal goal is to maximize the efficiency among the scientist working to understand, treat, and prevent cancer.

A.3.1. Structural variation in cancer genome

Decades of cytogenetic studies have shown that somatic chromosome alterations are a feature of many cancer genomes. These early studies, particularly in leukaemia and lymphoma, identified recurrent chromosomal rearrangements that were present in many patients with the same type of cancer [72]. As it has been mentioned before, a well-known example is the formation of Philadelphia chromosome by a translocation between chromosomes 9 and 22 in a significant fraction of patients with CML [73]. Later, genome-wide scanning of tumour samples by microarray approaches allowed the discovering of hundreds/thousands of CNVs related to cancer (reviewed in [196]). Currently, NGS enables the comprehensive and precise identification of many forms of somatic SV, increasingly rapidly the knowledge about their prevalence and distribution in cancer genomes. Since 2008, several studies have reported whole-genome sequencing of various types of clinical samples and cell lines [21, 22, 30, 34, 35]. These studies have revealed that some cancers exhibit few rearrangements whereas others show many [197]. In addition, the distribution of the SV also differs between different types of cancer, suggesting distinctive pattern of genomic instability. For example, tandem duplications are particularly common in breast and ovarian cancers [30], whereas deletions and inversions would be more frequent in pancreatic cancer [21], and inter-chromosomal rearrangements are mainly found in haematological malignancies [72].

Like in the analyses of single nucleotide changes, a critical challenge in whole-genome analyses of cancer is distinguishing driver structural alterations from the numerous passenger variants that accumulate during tumorigenesis. SV can lead to cancer development by the disruption of tumour suppressor genes, activation of oncogenes, or generation of fusion genes that individually or in combination can promote tumour progression. Occasionally, breakpoints directly generate an oncogenic element, such as a driver fusion transcript (e.g. *BCR-ABL1*, *EML4-ALK*, *PML-RARA*), or a deletion of a critical tumour suppressor gene (e.g. *TP53, CDKN2A, RB1, WRN, FBXW7*) (reviewed in [49]). In this case, the structural event can be catalogued as a direct "driver" for the oncogenesis process (Figure 13a).In other cases, the SV have been found that do not directly interrupt genes or change regulatory regions, but, by virtue of the rearrangement, seem to set up subsequent critical oncogenic events [198], acting therefore as a "conductor" event (Figure 13b). Finally, it has been found that SV can also be "indicator" events, with no oncogenic driver function but correlating with expression of nearby oncogenic factors (Figure 13c).

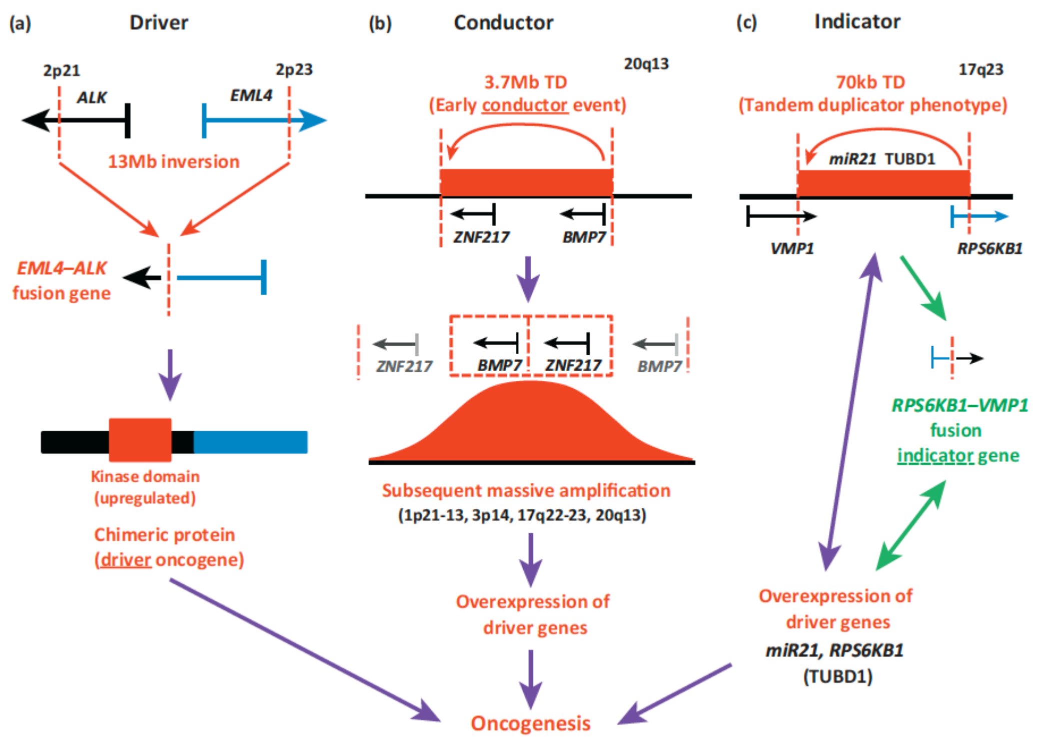


Figure 13. Driver, Conductor and Indicator SV. (a) Example of a driver SV in lung cancer. A 13-Mb inversion between 2p21 and 2p23 generates the oncogenic *EML4–ALK* fusion gene, which has upregulated kinase activity. (b) Example of a conductor SV in breast cancer. A tandem duplication (TD) with a span size of 3.7 Mb on 20q13 juxtaposes the *BMP7* and *ZNF217* genes, which originally resided at the ends of the duplicated segment. The two newly juxtaposed genes and the fusion point of the TD have undergone massive amplification, causing overexpression of driver genes (i.e., *BMP7* and *ZNF217*). Thus, the break point itself is not driver but the SV 'conducts' subsequent genomic amplification of the two drivers. (c) Example of an indicator SV in breast cancer. A TD with span size of 70 kb on 17q23 generates the *RPS6KB1–VMP1* fusion gene, although this has no discernable oncogenic function. However, the expression of the fusion transcript is correlated with high oncogene expression inside and around the SV, such as miR21. Thus, the SM is an indicator of adjacent oncogenic drivers. Adapted from Inaki, K. and Liu, E.T 2012.

Several specific genomic regions or gene families are prone to be affected by SV in cancer genomes. For example, some genomic regions are commonly amplified (e.g., 1q, 8q, 17q and 20q) across cancers [199, 200]. In such cases, key oncogenic drivers within the amplicons activated by overexpression have been identified (for example, *MYC, MYCN, ERBB2, GLI1*). Genomic regions prone to amplification are enriched for genes with ontological terms such as kinase growth, *MYC* and apoptosis. Instead, deleted regions are mainly enriched for genes with a role as tumour suppressor or involved in methylation or cellular adhesion [201].

A high number of SV in cancer genomes generates global genomic instability and increases the mutation rate [185]. Analyses of cancer samples have revealed the existence of unusual complex genomic SV in some tumours, such as those seen in *chromothripsis* [24] or those observed in another particular phenotype called *tandem duplicator* [30, 198, 202] (Figure 14).

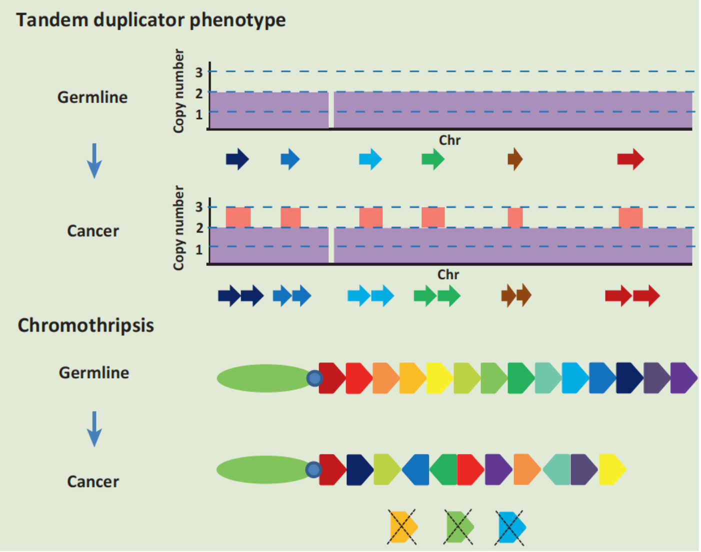


Figure 14. Complex genome-wide SV. (a) In Tandem duplicator phenotype, a few hundred TDs with a range of spans (50-300 kb) are detected across entire chromosomes and so distributed segments with one copy number increase are observed. (b) Chromothripsis is probably caused by a single catastrophic fragmentation and rearrangement through random joining within a few chromosomes. As the figure depicts, some of the fragments are also lost in the reshuffling. Adapted from Inaki, K. and Liu, E.T. 2012

In the tandem duplicator phenotype, a large number of different tandem duplications are distributed throughout the cancer genome, a state that is distinct from most other tumours. This phenomenon was first noted during studies of breast cancer cell lines [203], in which the investigators observed several clusters exhibiting multifocal amplifications (*amplisomes*) interspersed with segments of unaltered copy number at a single locus. A more recent study has demonstrated that complex patterns of amplifications, with rearrangements linking amplified segments, may be common in other cancers, such as lung squamous cell carcinoma [33]. It seems that this phenomenon might presumably result from stepwise mutations, especially if the copy-number state of the amplifications is greater than three copies. The presence of multiple amplified CNV states in amplisomes denotes that an initial amplification might precede rearrangement, followed by subsequent coamplification of the rearranged genomic segments.

In contrast, the chromothripsis phenotype is presumably consequence of a single catastrophic event [24]. In this case, a massive number of SV involving one or several chromosomes is accompanied by copy number changes oscillating between a few, low-order copy states [24]. The identification of chromothripsis in 2011 resulted in much interest by biologist oriented researchers, mainly due to the intrinsic fascination about how a single event can cause the pattern of complex SV and how a cell may be able to live with it. In addition, chromothripsis is also of wide interest to clinical researchers, who are now asking how might the chromothripsis relate to prognosis and treatment responses. During the two last years, many papers have been published about this topic gaining insights into the prevalence of the phenomenon in different cancers, the mutagenic mechanisms underlying its appearance, and how it might contribute to tumorigenesis [204-206].

A.3.1.1. The Chromothripsis phenomenon in cancer

The pattern of the SV observed in the chromothripsis might be considered the maximum expression of genomic structural complexity discovered so far. Samples with chromothripsis share six main genomic features (Figure 15) (reviewed in [207]): (a) the occurrence of remarkable numbers of rearrangements in localized chromosomal regions, encompassing a broad spectrum of complexity and exhibiting vast architectural diversity; (b) regularity of low copy-number states (between one or two) across the rearranged region; (c) an alternation in the rearranged areas of regions where heterozygosity is preserved with regions presenting loss of heterozygosity (LOH); (d) rearrangements biased toward a single parental chromosome (or haplotype); (e) shattered fragments stitched together in random order and orientation (inverted or non-inverted sense respect to the original chromosome); and (f) each DNA segment retained in the derivative chromosome must be demarcated at either end by genomic rearrangement breakpoints detected by RP strategy.

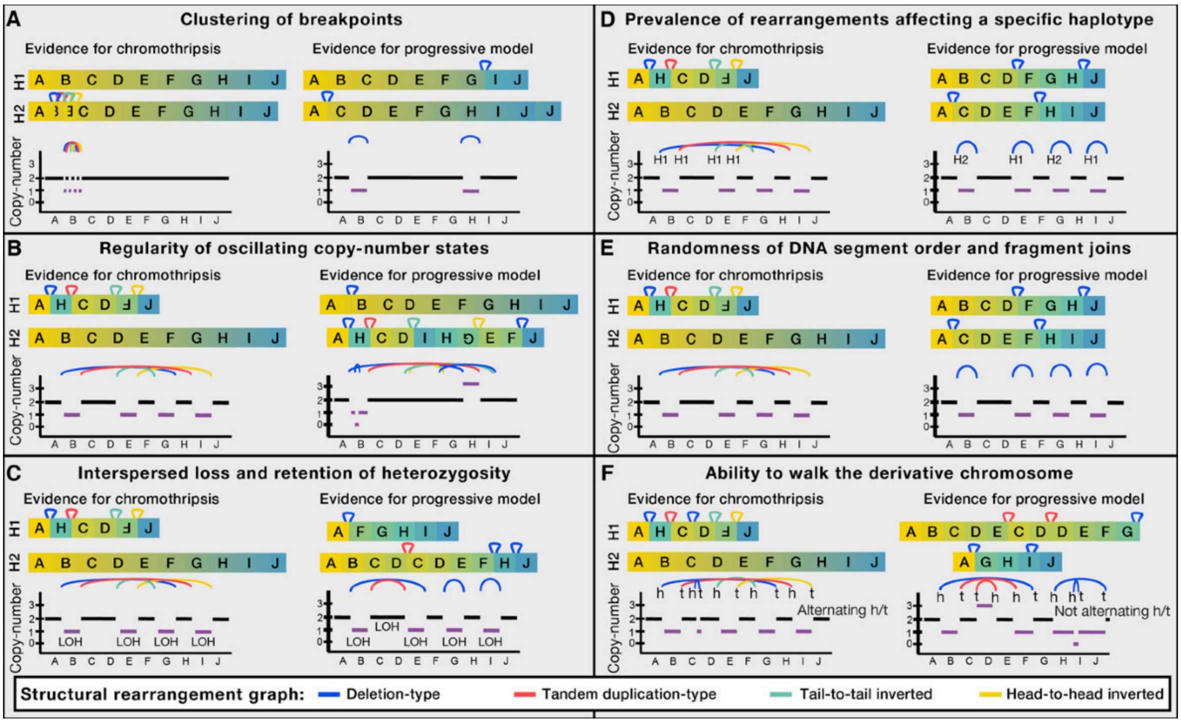

A
Clustering of breakpoints
Evidence for chromothripsis
Evidence for progressive model
B
Regularity of oscillating copy-number states
C
Interspersed loss and retention of heterozygosity
D
Prevalence of rearrangements affecting a specific haplotype
E
Randomness of DNA segment order and fragment joins
F
Ability to walk the derivative chromosome
H1
H2
Copy-number
A B C D E F G H I J
LOH
Alternating h/t
Not alternating h/t
Structural rearrangement graph:
Deletion-type
Tandem duplication-type
Tail-to-tail inverted
Head-to-head inverted

Figure 15. Features defining chromothripsis. Schematic representation of the six principal features shared by samples with chromothripsis. Adapted from Korbel, J. et al. 2013 [206].

These features have important implications for understanding how and when chromothripsis arises. First, the concentration of breakpoints of different forms of SV in a single or several chromosomes suggests that the phenomenon is most likely to occur when chromosomes are largely condensed, such as during mitosis. Second, the number of copy-number states oscillating between one or two copies, and the considerably clustering of the breakpoints in focal regions, imply that such rearrangements probably occurred within a relatively short period. If the accumulation of the SV occurred progressively, the number of states would invariably increase with the number of detected breakpoints, and these would be found widespread in the genome as it is expected in a progressive accumulation of alterations [24]. Third, the alternation of segments retaining heterozygosis with other presenting LOH suggests that the rearrangements took place in early cancer cell development, at the time when both parental copies of the chromosome were present before LOH. Retention of heterozygosis in patches throughout a chromothriptic region is difficult to explain by progressive rearrangement mechanisms, especially because once the heterozygosis is lost, it cannot be regained.

The simplest model to explain chromothripsis is that one or more chromosomal regions were simultaneously broken into many fragments, some of which were then joined together in a random manner. Then, cells that can survive such catastrophic event emerge with a highly mutated genomic landscape that confers a significant selective advantage to the clone (acquiring several simultaneous tumorigenic alterations), thereby promoting cancer progression [24, 34]. Therefore, the chromothripsis phenomenon suggests the possibility of a punctuated equilibrium mechanism for tumour development, opposed to the generally accepted concept of cancer biogenesis based on the idea of gradualism. However, the main question is how the phenomenon is initiated. So far, several possibilities have been proposed and discussed, hypothesizing mechanisms that might lead to the rise of chromothripsis: ionizing radiation acting upon condensed chromosomes [24]; critical telomere shortening followed by chromosome end-to-end fusions and subsequent massive DNA breakage [24]; abortive apoptosis events [208]; and "premature chromosome compaction," in which chromosomes condense before completing DNA replication and may consequently be shatter [209]. However, the most appealing current model has arisen through the recent demonstration that chromosomes or partial chromosomes can be contained in nucleus-external structures called "micronuclei" [210]. DNA segments within micronuclei can suffer aberrant DNA replication and can be pulverized during mitosis, with subsequent re-joining of fragments by one or several mechanism of reparation (NHEJ, MMBIR, FoSTeS). This leads to a derivative chromosome or chromosomes that can be reincorporated into the main nucleus [210]. Nevertheless, it is important to stress that additional processes are likely to contribute to the chromothriptic landscapes, and further in-depth analyses of more events with chromothripsis need to be carried out for a more complete understanding of the relative importance of each mechanism in defining these highly complex pattern of SV.

Chromothripsis was initially observed in a sample of CLL [24], but additional screening revealed similar patterns in around 3% of other type of cancers, with a presumably higher percentage (25%) in bone cancers. Subsequent publications have confirmed that chromothripsis occurs in various human cancers, including paediatric medulloblastoma (13%), neuroblastoma (11%), colorectal cancer (unknown percentage), melanoma (7-

8%), and haematological malignancies (2-10%) [24, 31, 32, 34, 35, 211]. Quite recently it has been reported a significantly higher incidence in glioblastoma (39%) [33]. However, the true incidence of chromothripsis in cancer cannot be well established. Some of these studies have defined the presence of the complex patterns of SV by microarray-based approaches and others by NGS. By microarrays the attention has been focused on the observation of abnormal copy-number profiles in specific chromosomes, and it has been considered that a chromothripsis sample may present at least 10 regions with altered copy-number pattern. Instead, by NGS the emphasis has been put in the accumulation of breakpoints of several types of SV in focal regions, and a sample with chromothripsis not necessarily has to carry 10 regions with abnormal copy-number (reviewed in [207]). Thus, additional work is required for a better evaluation of the real prevalence and impact of this phenomenon in cancer.

Chromothripsis has been associated with poor clinical outcome in several cancers [32, 34, 35, 211], indicating its potential relevance as a prognostic marker and suggesting that it is a feature of some particularly aggressive forms of cancer. In some subtypes of medulloblastoma and in acute myeloid leukaemia, chromothripsis has been associated with germline and/or somatic *TP53* mutations [34], which links the predisposition to the phenomenon with defects in DNA damage response (DDR) or apoptosis. Furthermore, cell culture work has shown that p53 depletion is necessary for micronucleated cells to efficiently progress into the cell cycle and to generate DNA damage in micronuclei [210], which supports the "micronuclei model" for the chromothripsis appearance. In fact, it is thought that not only defects in p53 but also DDR deficiencies enhance the occurrence of micronuclei, suggesting links between chromothripsis and defects in general repair mechanisms [204]. This is important because it opens the possibility for treating a DDR defective chromothriptic cancer with specific drugs targeting residual DDR pathways that the cancer is more dependent on than are the normal cells of the patient [190, 212]. However, from a clinical perspective much remains to be understood, as there are currently many unknown questions like at what stage does chromothripsis occur, the full prevalence of the phenomenon in the different cancer type and why is it more prevalent in certain tumour types.

A.3.2. Chronic lymphocytic leukaemia

Chronic lymphocytic leukaemia (CLL) is the most common type of adulthood leukaemia. Most people newly diagnosed with CLL are over the age of 60 (median age at diagnosis

65-70 years), and the incidence rates in men are nearly twice as high as in women [213]. There is also a substantial geographic variation in its prevalence, with higher frequencies in Western countries [214]. Advanced age and a family history of leukaemia or lymphoma are additional risk factors [215]. However, CLL is a very heterogeneous malignancy, both in terms of clinical biological features (reviewed in [216]). The major success achieved in the characterization of specific biological markers was the classification of CLL patients in two groups, depending on whether their present somatic hypermutations in the immunoglobulin genes (IGHV), and its relation with the clinical outcome [217]. This broad classification was improved with the identification of additional factors such as mutations inactivating *TP53* [218] and *ATM* [219] genes, and the differences found among patients in the expression of ZAP-70 [220] and CD38 [221]. Finally, some genomic aberrations were also recognized as important determinants of the evolution of the CLL [213, 222] but genomic events that would play a clear role in the initiation of the tumour and in the heterogeneous evolution of the disease remain still unknown. This biological landscape probably reflects either an unknown underlying biochemical mechanism playing a key role in the malignancy, or multiple molecular pathways independently driving the development of the leukemic cells.

Given its clinical and societal impact and the gaps in the knowledge of its genomic determinants, CLL is one of the cancers included in the ICGC [65] through the Chronic Lymphocytic Leukaemia Genome Project (CLL-GP) (BOX 1).

Box 1

Chronic Lymphocytic Leukaemia Genome Project

The Chronic Lymphocytic Leukaemia Genome Project (CLL-GP, www.cllgenome.es), a contributing member of ICGC, started in 2009. The project is funded by the Ministry of Science and Innovation, with the participation of a multidisciplinary team of researchers from the Hospital Clínic of Barcelona, the University of Barcelona, the University of Oviedo, the Centre for Genomic Regulation in Barcelona, the Catalan Institute of Oncology, the Spanish National Cancer Research Centre, the Cancer Research Centre of Salamanca, the National Centre for Genomic Analysis in Barcelona, and the University of Deusto.

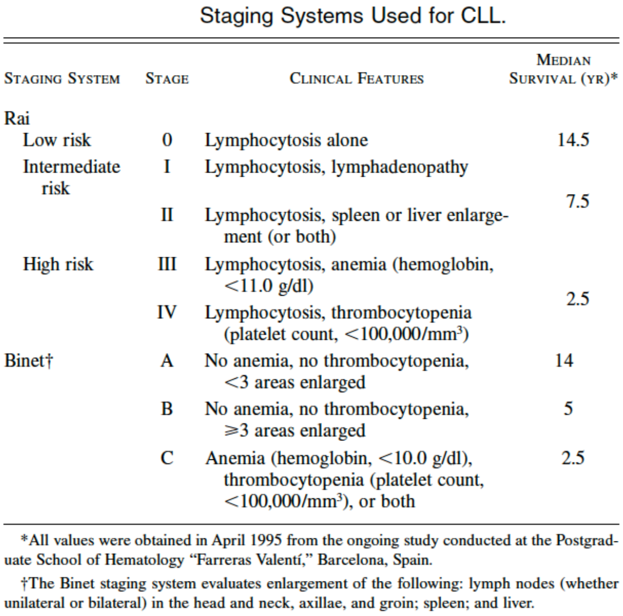
The main objective of the project is to generate a comprehensive catalogue of genomic alterations involved in the development and progression of the disease using 500 independent CLL tumours, including clinical, biological and epidemiological information. The ultimate aspiration is to discern the enormous diversity and complexity of the changes in the tumour genome that may be responsible for the initiation and progression of the CLL. The project has the purposes of creating diagnostic tools, discovering therapeutic targets to improve CLL prevention and diagnosis, and developing new strategies that will allow a customized therapy for CLL in order to make it more precise and effective.

The specific goals designed in order to fulfill the requirements of the project are: (a) to identify the different types of somatic genomic alterations (SNPs and SV) occurring in CLL; (b) to characterize the epigenomic alterations that may be involved in the disease; (c) to establish the transcriptome profiles of the same tumours; (d) to promote functional and clinical studies to validate the usefulness of the identified alterations; and (e) to share the data available with the scientific community as quickly as possible.

Because the Genes and Disease's Group is a member of this project, this thesis has been focused in part on the analysis of this type of cancer. Thus, the next subsections provide a general background about the clinical, molecular and genomic features of the CLL.

A.3.2.1. Clinical and molecular features in chronic lymphocytic leukaemia

CLL is a malignant clonal lymphoproliferative disorder of B-lymphocytes, which slowly accumulate with mature appearance in blood, bone marrow, lymph nodes or other lymphoid tissues [213, 216]. This progressive accumulation of B-lymphocytes can lead to leucocytosis, lymphadenopathy, hepatosplenomegaly, bone marrow failure, recurrent infection and sometimes is associated with autoimmune disease [216]. However, CLL has a highly variable clinical course, with some patients dying from the disease within a few months of the diagnosis and others having a more or less stable disease with normal life span, usually without the need of treatment [213]. The clinical staging systems devised by Rai *et al.* [223] and Binet *et al.* [224], based on the clinical features observed in patients, are the most common used method for predicting the clinical status and the expected survival in CLL (Table 3). The problem of these staging systems is that most patients are asymptomatic at diagnosis [223-225], and they cannot be used to predict the individual risk of disease progression and survival in early stages of the disease.



Staging Systems Used for CLL.

Staging System	Stage	Clinical Features	Median Survival (yr)*
Rai			
Low risk	0	Lymphocytosis alone	14.5
Intermediate risk	I	Lymphocytosis, lymphadenopathy	7.5
	II	Lymphocytosis, spleen or liver enlargement (or both)	
High risk	III	Lymphocytosis, anemia (hemoglobin, <11.0 g/dl)	2.5
	IV	Lymphocytosis, thrombocytopenia (platelet count, <100,000/mm^3)	
Binet†	A	No anemia, no thrombocytopenia, <3 areas enlarged	14
	B	No anemia, no thrombocytopenia, ⩾3 areas enlarged	5
	C	Anemia (hemoglobin, <10.0 g/dl), thrombocytopenia (platelet count, <100,000/mm^3), or both	2.5

*All values were obtained in April 1995 from the ongoing study conducted at the Postgraduate School of Hematology "Farreras Valentí," Barcelona, Spain.

†The Binet staging system evaluates enlargement of the following: lymph nodes (whether unilateral or bilateral) in the head and neck, axillae, and groin; spleen; and liver.

Table 3. Staging systems of Rai and Binet. Taken from Rozman, C and Montserrat, E. 1995 [212].

This and the substantial heterogeneity within clinical stages prompted searches for additional prognosis factors, firstly on the basis of the presence or not of mutations in IGHV genes. This marker distinguishes between leukaemia originating from B cells that have or have not yet undergone the process of somatic hypermutation that occurs as part of normal B cell development [226]. IGHV-mutated CLL patients frequently show mild clinical characteristics, and high overall survival (OS) and progression-free survival (PFS). Instead, IGHV-unmutated cases often suffer an aggressive form of the disease that may be refractory to treatment [217] (Figure 16).

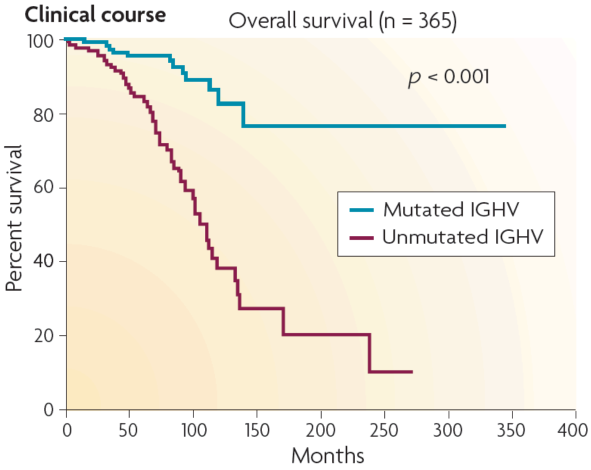


Figure 16. Clinical courses associated with IGVH-unmutated and mutated CLLs. The overall survival of unmutated-IGHV CLL cases is clearly lower respect to the overall survival of mutated-IGHV patients. Adapted from Zenz, T et al. 2010 [215]

Later, other important prognosis factors were identified, such as some cell markers (CD38 and ZAP-70) [220, 221], and several serum markers (for example, thymidine kinase, β2-microglobulin and soluble CD23) [227]. For example, high expression of CD38 (a molecule involved in signalling and activation of immune cells) and ZAP-70 (intracellular protein that promotes activation signals delivered to T cells) are related to patients who would have a more aggressive disease course and low OS, whereas patients with little expression of these markers have an indolent CLL course [220, 221]. All these factors are complemented by recurrent genetic alterations found in several patients, which relevance to prognosis has also been described [213, 222].

A.3.2.2. Genomic aberrations and prognosis markers in chronic lymphocytic leukaemia

The main recurrent genetic aberrations, affecting approximately 70% of CLL cases, are currently among the most important factors in predicting survival [222, 228-231] (Figure 17). They include trisomy 12 and monoallelic or biallelic deletion of chromosomal regions 17p, 11q and 13q [216, 222]. Deletions on 13q (~60% of patients) and 11q (~15-20%) show very heterogeneous sizes among patients, ranging from few kilobases to several megabases [230, 231]. Nevertheless, common critical regions have been recently defined and thus potential targets have been postulated as causal genes in CLL. The minimal deleted region on 13q14 contains *miR15a, miR16*-1 and *DLEU1* genes, whereas on 11q22.3 includes the *ATM* gene [230, 231]. The microRNA cluster involved in the deletion 13q14 seems to negatively regulate the expression of *BCL2* and other genes involved in proliferation and apoptosis [232]. Meanwhile, the *ATM* gene, which point mutations have been found in around 12% of CLL patients [219], plays a role in DNA damage response and is characterized by extreme sensitivity to irradiation, genomic instability and predisposition to lymphoid malignancies [233]. Instead, deletion on 17p (~10% of patients) and trisomy 12 (~20%) are more homogeneous aberrations. Deletions on 17p usually encompass the whole short arm of the chromosome with the breakpoint in the centromere or close to it [231]. In this case the target crucial gene is the *TP53,* which is also found with mutations on the remaining allele in most of cases carrying the deletion [218]. Otherwise, no clear potential target genes on chromosome 12 have been reported in association with CLL FISH panels, containing probes targeting the altered regions [222].

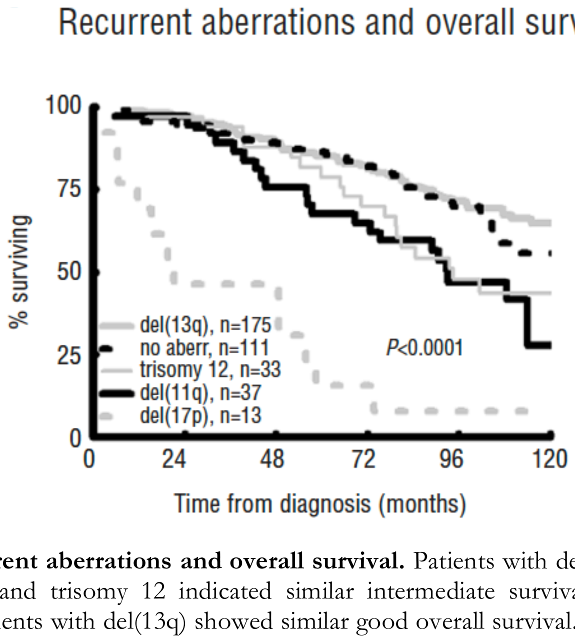


Figure 17. Recurrent aberrations and overall survival. Patients with del(17p) had the worst outcome, whereas del(11q) and trisomy 12 indicated similar intermediate survival. Patients with no recurrent aberration and patients with del(13q) showed similar good overall survival. Adapted from Gunnarsson, R et al. 2011 [229].

Around 90% of CLL patients present SV [229-231]. But the majority carry only one or two aberrations and just a small fraction carry three or more, which are considered CLL cases with "genomic complexity" [228, 230]. This suggests that the genomic instability might not be a prominent feature of CLL. However, the main recurrent aberrations explained above fail to explain the heterogeneity of the disease in its full extent. Indeed, with the arrival of microarray technologies multiple other forms of less-recurrent SV have been described in CLL. Interestingly, these additional structural variants have also been correlated with clinical outcome measures and patients carrying more than three SV (patients with "genomic complexity") have been further classified as "high-risk CLL" [228, 230]. High-risk CLLs involve the subset of patients with an aggressive disease and low OS and PFS [228, 230]. Additional aberrations have been also linked with the presence of deletions in 11q, 17p and/or *TP53* mutations [228-231]. This is probably due to secondary events resulting from defects in the DNA damage response associated with alterations of *ATM* and *TP53*.

Examples of relatively recurrent additional CNAs include deletions on 6q (especially on 6q21), found in approximately 7% of patients and considered an intermediate-risk feature [222, 234], trisomy of chromosomes 18 and 19, a partial trisomy 3, deletions on chromosomes 8p, 18q, 10q24 and 15q15 and gains on 2p15-p16, 3q26 and 8q24 [229, 231, 235]. Although genes involved in these additional SV are unknown, for example a minimally gained region on 8q24 has been recently defined close to the *MYC* locus, and deletions on 15q15 have been located within *MGA* locus. *MYC* has an effect mostly in transcriptional activation promoting cell-cycle progression, apoptosis and cellular transformation, and has been associated also with other B-cell malignancies [236, 237]. *MGA* is involved in regulatory mechanisms for cell proliferation, differentiation and apoptosis, probably acting as a transcriptional repressor [238], and its relation with CLL has been established only recently with high-throughput genome-scanning arrays [231]. On the other hand, although inter-chromosomal alterations are rare in CLL [239], translocations on chromosome 14q32 involving immunoglobulin heavy chain (IGH) locus have been identified in 4-9% of patients [222, 240]. Translocations of this locus, which have been also linked to an unfavourable outcome, are recurrently associated with the *BCL2* gene (IGH/*BCL2*) and sporadically with other genes, including *BCL-11A*, *CCND3* and *CDK6* [240].

Genomic complexity appears to be a marker of progressive disease and inferior survival in newly diagnosed CLL cases [230]. Indeed, the presence of a high number SV in CLL indicates the capacity for clonal plasticity and evolution and the possibility of the generation of successive clones with more aggressive clinical characteristics. Many patients that have received therapy underwent clonal evolution, which means that the treatment *per se* induces genomic instability [230]. Simultaneously, the presence of several subclonal populations carrying different alterations, leads to the possibility of developing therapy resistance with the expansion of previously minor subclones [177]. In fact, most of CLL patients with detected clonal evolution exhibit shorter failure-free survival due to the presence of driver mutations in different subclones [177]. This is because cytotoxic therapy removes the incumbent clones and shifts the evolutionary landscape in favour of one or more aggressive subclones [177, 241]. The presence of pre-treatment subclonal drivers anticipate the dominant genetic composition of the relapsing tumour [177]. Therefore, the estimation of subclonal populations is particularly important, as CLL is an incurable disease characterized by relapses. The knowledge about the tumour composition facilitates the development of new therapeutic paradigms, targeting not only specific drivers but also their evolutionary landscape.

B. OBJECTIVES

During the last ten years, with the advent of microarray-based approaches and next-generation sequencing (NGS) technologies, a large number of genomic structural variation (SV) has been identified in healthy and diseased individuals. Genome-wide studies performed in many individuals and populations have allowed the association of different types of SV with specific phenotypic traits and diseases, highlighting the evidence that the SV might have a significant biological impact on the human genome. However, molecular and mechanistic links between most SV and many phenotypes remain still not completely understood, mainly because the large number of variants found in each genome hinders the correct interpretation of their specific effects in the individual, either by acting alone or in combination with other changes. Furthermore, the variability in the architecture of the different types of SV and the complexity of some structural rearrangements represent a major obstacle for their complete characterization.

When this thesis began microarray-based approaches were the main strategies for the genome scanning of SV and therefore most efforts were focused in the discovery of copy-number gains or losses (CNVs). The association of CNVs with several diseases was identified, and several reports showed that this type of genomic variation could explain phenotypic differences among human populations. Nevertheless, the knowledge about the functional effect of CNVs was still in its infancy, and much remained to be explored. With the advent of NGS technologies, all forms of SV could be analysed at large-scale. The "first generation" of tools developed for the analysis of SV using whole-genome sequencing (WGS) data allowed the detection of all forms of SV, but presented limitations derived from the characteristics of NGS technologies and the existence of repetitive sequences in the genome. This made difficult, and still does, a comprehensive exploration of SV in the different genomes, despite that it has facilitated the identification of high complex rearrangements, as it is the case of somatic changes in cancer genomes. Indeed, the field of SV in cancer genomics is probably the one that has profited the most with the advent of NGS technologies because they have provided the capacity to detect complex rearrangements.

Based on all this, and with the aim to gain insight on the characterization of the SV, using both microarray-based and NGS approaches, and on the evaluation of the SV phenotypic impact in the human genome, the following objectives were established:

1. Detection of copy-number variable regions in different human populations.

1.1. Characterization of CNVs that could show variable copy-number profiles among human populations, which might be influenced by environmental factors and might have a phenotypic impact.

1.2. Evaluation of the reliability of CNV detection by array-comparative genomic hybridization (aCGH), validating a genomic region that shows population copy-number differences, and that is involved in susceptibility to psoriasis and other autoimmune diseases.

2. Development of a computational tool for a comprehensive detection and characterization of all forms of SV using whole-genome sequencing (WGS) data. The following challenges were pursued to bring novel features respect to other available pipelines of SV analysis:

2.1. Integration of read-pair and read-depth methodologies for the correct interpretation of all forms of SV.

2.2. Automatic identification of potential false positive calls, which mainly includes repetitive elements.

2.3. Direct detection of somatic and/or non-somatic SV in cancer genomes.

3. Detection of somatic SV involved in cancer using WGS.

3.1 Characterization of somatic SV in patients with chronic lymphocytic leukaemia (CLL) by WGS.

3.2. Evaluation of the phenotypic impact of highly complex pattern of SV, consistent with chromothripsis, in the progression of cancer in a CLL case.

3.3. Determination of tumour heterogeneity through the quantification of cancer cell fraction using WGS data.

C. RESULTS

C.1. A common deletion in human populations by array-comparative genomic hybridization

The first part of this thesis was focused in the population-genetic analysis of a 32-kb deletion on the *PSORS4* locus, involving the *LCE3C* and *LCE3B* genes. This deletion was previously found to be associated with psoriasis and other autoimmune diseases. We analysed, by array-comparative genomic hybridization (aCGH) and PCR-based genotyping assay, the frequency of the deletion in worldwide populations and the linkage disequilibrium (LD) pattern between the variant allele and the tag SNP rs4112788. Our results show that most ethnic groups tend to have a higher frequency of the deleted allele when comparing with Sub-Saharan Africans. Furthermore, we found strong LD between the rs4112788 and the deletion in most of non-African populations, in contrast to the low concordance between the loci in the Africans. These observations suggest that differences between ethnic groups might not be due to natural selection but the consequence of genetic drift caused by the strong bottleneck during the "out of Africa" expansion. This study provides another example of population variability in terms of biomedical interesting CNV and allows new insight in the evolutionary mechanisms that influence genomic variability between human populations. Furthermore, the study demonstrates the reliability to predict CNVs by aCGH, although it also highlights the lack of precision of this technique in the definition of breakpoints.

The results of this study led to the publication of the following article:

Worldwide population distribution of the common LCE3C-LCE3B deletion associated with psoriasis and other autoimmune disorders

Bassaganyas, L, Riveira-Muñoz E, García-Aragonés M, González JR, Cáceres M, Armengol L, Estivill X.

BMC Genomics. 2013 Apr 17;14(1):261

Bassaganyas, L, Riveira-Muñoz E, García-Aragonés M, González JR, Cáceres M, Armengol L, Estivill X. **Worldwide population distribution of the common LCE3C-LCE3B deletion associated with psoriasis and other autoimmune disorders. Supplementary information.** BMC Genomics. 2013 Apr 17;14(1):261.

C.2. A computational tool for the characterization of structural variation using whole-genome sequencing data

This work was focused on the development of a tool, named *PeSV-Fisher*, for the detection and characterization of all forms of SV using WGS data. *PeSV-Fisher* is the first pipeline designed specifically to obtain directly somatic SV analysing simultaneously normal and tumour DNA from a given cancer patient. It also provides comprehensive information on co-localization of the SV in the genome, a crucial aspect for studying their biological consequences. The algorithm uses a combination of methods based on paired-reads and read-depth strategies. In order to avoid the need of large-scale computing clusters with high number of cores imposed by NGS data analysis, the pipeline has been build dividing the process in multiple threads and with low coupling between classes. Programming languages used are Python and C, and external resources needed for the compilation are MySQL and Samtools [242]. The reliability of the tool was tested using public data generated from the 1000 Genome Project and the Chronic Lymphocytic Leukemia Genome Project.

PeSV-Fisher is available at http://gd.crg.eu/tools.

The results of this study led to the publication of the following article:

PeSV-Fisher: Identification of somatic and non-somatic structural variants using next-generation sequencing data.

Geòrgia Escaramís*, Cristian Tornador*, Laia Bassaganyas*, Raquel Rabionet, Jose M. C. Tubio, Alexander Martínez-Fundichely, Mario Cáceres, Marta Gut, Stephan Ossowski and Xavier Estivill

*equal contribution

PLoS ONE 2013 May 21;8(5). Doi:10.1371/journal.pone.0063377

Escaramís G, Tornador C, Bassaganyas L, Rabionet R, Tubio J.M.C, 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. Supplementary information.** PLoS ON. 2013 May 21; 8(5).

C.3. Structural variation in chronic lymphocytic leukaemia detected by whole-genome sequencing

This study was the first one performed by the chronic lymphocytic leukaemia genome project. Four CLL cases were analysed by WGS identifying 46 somatic point mutations, and 10 structural variants (four of them as novel variants). Due to more promising preliminary results, only the point mutations, but not the SV, were further explored in 363 CLL patients. Four genes (*NOTCH1, XPO1, KLHL6* and *MYD88*) were identified as recurrently mutated in CLL, probably representing the activating events for the tumour development.

Our contribution in this study was in the detection of SV using WGS data. Combining our results with the information obtained by different microarray platforms, we established the 10 structural rearrangements as the more confident ones. We also provided the breakpoint positions at a nucleotide-resolution level.

The results of this study led to the publication of the following article:

Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia.

Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N, Escaramis G, Jares P, Beà S, González-Díaz M, Bassaganyas L, Baumann T, Juan M, López-Guerra M, Colomer D, Tubío JM, López C, Navarro A, Tornador C, Aymerich M, Rozman M, Hernández JM, Puente DA, Freije JM, Velasco G, Gutiérrez-Fernández A, Costa D, Carrió A, Guijarro S, Enjuanes A, Hernández L, Yagüe J, Nicolás P, Romeo-Casabona CM, Himmelbauer H, Castillo E, Dohm JC, de Sanjosé S, Piris MA, de Alava E, San Miguel J, Royo R, Gelpí JL, Torrents D, Orozco M, Pisano DG, Valencia A, Guigó R, Bayés M, Heath S, Gut M, Klatt P, Marshall J, Raine K, Stebbings LA, Futreal PA, Stratton MR, Campbell PJ, Gut I, López-Guillermo A, Estivill X, Montserrat E, López-Otín C, Campo E.

Nature. 2011 Jun 5;475(7354):101-5. doi: 10.1038/nature10113.

Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N, Escaramis G, Jares P, Beà S, González-Díaz M, Bassaganyas L, Baumann T, Juan M, López-Guerra M, Colomer D, Tubío JM, López C, Navarro A, Tornador C, Aymerich M, Rozman M, Hernández JM, Puente DA, Freije JM, Velasco G, Gutiérrez-Fernández A, Costa D, Carrió A, Guijarro S, Enjuanes A, Hernández L, Yagüe J, Nicolás P, Romeo-Casabona CM, Himmelbauer H, Castillo E, Dohm JC, de Sanjosé S, Piris MA, de Alava E, San Miguel J, Royo R, Gelpí JL, Torrents D, Orozco M, Pisano DG, Valencia A, Guigó R, Bayés M, Heath S, Gut M, Klatt P, Marshall J, Raine K, Stebbings LA, Futreal PA, Stratton MR, Campbell PJ, Gut I, López-Guillermo A, Estivill X, Montserrat E, López- Otín C, Campo E. **Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Supplementary information.** Nature. 2011 Jun 5; 475(7354):101-5.

C.4. Characterization of complex structural variation in a chronic lymphocytic leukaemia patient.

The last section of this thesis was focused in the deep characterization of highly complex genomic rearrangements and their functional impact in the progression of cancer. Applying *PeSV-Fisher* we identified a pattern of chromothripsis in one CLL case. Using different types of NGS and microarray approaches we performed a longitudinal analysis of this case over a period of eleven years of disease evolution. We observed by WGS the phenomenon of chromothripsis concurrently with a change of the disease to a malignant state. Targeted re-sequencing and microarray analysis of SV at different time-points revealed that chromothripsis was a sporadic phenomenon, not involved in the tumour development and that it disappeared after treatment. This is the first description of the rise and fall of chromothripsis and we demonstrate that this process did not play a role in the initial tumour development and disease progression of CLL.

The results of this study led to the publication of the following article:

Sporadic and reversible chromothripsis in chronic lymphocytic leukaemia revealed by longitudinal genomic analysis.

Bassaganyas, L., Beà S, Escaramís G, Tornador C, Salaverria I, Zapata L, Drechsel O, Ferreira PG, Rodriguez-Santiago B, Tubio JM, Navarro A, Martín-García D, López C, Martínez-Trillos A, López-Guillermo A, Gut M, Ossowski S, López-Otín C, Campo E, Estivill X.

Leukemia. 2013 Apr 24. doi: 10.1038/leu.2013.127.

Bassaganyas, L., Beà S, Escaramís G, Tornador C, Salaverria I, Zapata L, Drechsel O, Ferreira PG, Rodriguez-Santiago B, Tubio JM, Navarro A, Martín-García D, López C, Martínez-Trillos A, López-Guillermo A, Gut M, Ossowski S, López-Otín C, Campo E, Estivill X. **Sporadic and reversible chromothripsis in chronic lymphocytic leukaemia revealed by longitudinal genomic analysis. Supplementary information.** Leukemia. 2013 Apr 24. doi: 10.1038/leu.2013.127.

D. DISCUSSION

The existence of large human genomic structural variation (SV) is known since the middle of the last century when cytogenetic techniques were discovered. But the real importance of a more intermediate-scale of SV in human genomes began to be apparent at the beginning of the last decade, with the detection of large number of copy-number variants (CNVs) in healthy and diseased individuals by the use of microarray-based approaches [10, 12, 13]. The application of SNP arrays and array-comparative genomic hybridization (aCGH) led to the identification of many genomic regions containing CNV, but these methods lacked precision in the definition of breakpoints and failed to identify complex rearrangements and inversion genomic changes. Later on, with the arrival of next-generation sequencing (NGS) technologies, it has been confirmed and even extended the huge amount of structural rearrangements present in individual genomes [14, 15]. Indeed, NGS technologies have had a revolutionary impact on the field of genomic variation, as they have provided the capacity to detect all forms of SV with an unprecedented resolution and they have increased the operational spectrum of SV to include small sequences variants of at least 50 base pairs. Since ten years ago, genome-wide scanning approaches are performed in many samples, replacing the traditional approaches of individual gene analyses by large data sets generated by high-throughput technologies. Therefore, research on "next-generation genomics" is increasingly drifting towards statistical and computational approaches and using less individual experiments, which has implied a change and consequently an adaptation in the design of the studies.

Despite the considerable improvements in our understanding about the quantity and functional impact of SV, the molecular and mechanistic links between SV and phenotypes remain difficult to ascertain. The large number of discovered variants hampers in many cases the interpretation of their specific effects, whereas the variable architecture of SV and the complexity of the rearranged-induced sequence alterations represent a major obstacle to their correct characterization. Thus, the motivations of this thesis have been, on the one hand, to gain new insight on the methodology for the characterization of SV using both microarray-based approaches and NGS technologies, and on the other hand to evaluate its biological impact from specific examples

Characterization of a common deletion in different worldwide human populations

The first step of the thesis was to characterize, using a microarray-based approach, regions showing variable copy-number profiles among worldwide human populations. The aim was to identify regions that could present inter-population structural genomic differences, as well as their functional-related elements that might be influenced by environmental factors and have a phenotypic impact. To achieve this, we analysed samples from thirteen worldwide populations selected from Human Genome Diversity Panel (HGDP) and the HapMap Phase I Project using the Agilent H244K aCGH platform. By admixing individual samples from each population in pools, we diluted intra-populations differences, and consequently we enhanced the inter-population variability and the capacity to detect population-specific CNVs. Thus, we expected detect CNV regions present at a relatively high frequency within each studied population.

We decided to compare each population-pool against the pool of Yoruban (YRI, from Nigeria) samples considering valid the premise of the serial-founder model of human expansion out of Sub-Saharan Africa [80, 243]. This model defines that there is a high degree of genomic heterozygosis in Sub-Saharan Africans with an increasing degree of genomic homozygosis as we get further away from Africa. Therefore, we expected to find more population-specific CNVs in population samples from regions far from Africa and that these populations could be under relatively recent evolutionary forces, thus reflecting the influence of geography and the environment on human genetics. Indeed, we found 54 loci that showed variable copy-number among worldwide populations, most specifically in populations from America and Oceania followed by Eastern and Western Asian populations. In agreement with previous reports [13, 61], we also observed a significant enrichment of genes involved in sensory perception, immune system and distinct metabolic pathways overlapping these CNVs. This clearly supports the idea that population-specific CNV profiles could explain adaptations to environmental pressure and differences in the prevalence of some diseases among populations.

From all detected regions with copy-number differences between populations, we focused our attention for a detailed characterization on a variable locus encompassing the previously characterized deletion of 32 kb involving the *LCE3C* and *LCE3B* genes (*LCE3C_LCE3B-del*). This deletion was found to be associated with susceptibility to psoriasis [244-246], which has a high prevalence in developed countries and very low prevalence in Africans, as well as with other autoimmune diseases [247-249] which leads to an interesting example for studying the relation between a specific genomic variant, environment and disease. We hypothesized that the variability observed in the prevalence of psoriasis might be explained in part by population differences in the frequency of the *LCE3C_LCE3B-del* CNV. We also suggested that the pattern of the distribution of the deletion among populations might provide insights in the evolutionary mechanism influencing such population differences. The results of this survey have been published in the article entitled "Worldwide population distribution of the common LCE3C-LCE3B deletion associated with psoriasis and other autoimmune disorders".

Evaluation of the reliability of pools of samples in aCGH for the identification of CNV

The aCGH experiments showed that all populations, with the exception of pygmies (PYG), tend to have higher frequency of the deletion respect to the YRI group. The Database of Genomic Variants (DGV) listed more than 25 distinct large CNVs spanning either the *LCE3C* and *LCE3B* genes, or both, which *a priori* could suggest the existence of different breakpoint positions among individuals, such as the case of non-recurrent (and complex) SV. However, the same specific deletion of 32 kb, with the same breakpoint interval, had been found in individuals from clearly distant populations such as Spanish, Mongolian and Chinese [246, 250, 251]. This indicated that the *LCE3C_LCE3B-del* might probably be an example of a recurrent deletion occurring at the same specific genomic region. Thus, it should be possible to genotype the deletion in all population samples using the same genotyping assay. The information provided in DGV probably reflects a problem still persistent in the documentation of SV in most databases, which leads to confusion by considering complex SV that is actually simple: annotated observations came from several studies, each using a different technology platform and data processing algorithms, with different degrees of pre- and post-experimental standardization and validation.

A deeper analysis of the aCGH data on the *LCE3C_LCE3B-del* region showed population differences in signal intensities in a region generally smaller than the 6 probes covering this locus and population variability on the probe extension. This could support the information provided by DGV and suggested population differences in the CNV breakpoints. However, the probe density of the Agilent H244k aCGH platform (around 244,000 probes covering the entire genome at ~10 kb of average resolution) means that, following our strategy of three consecutive probes to determine a CNV, we were searching for regions >30 kb. Therefore, this platform is not probe-dense enough to define specific breakpoints at the needed resolution. Thus, a higher resolution microarray platform could provide more precise detection of breakpoints. In addition, we considered the possibility that some of the probes in the array might not be absolutely specific and might hybridize to similar sequences, like other *late cornified envelope* (LCE) genes, masking the signal from similar regions. This is in agreement with the probe failures, poor power and non-trivial rates of false positives derived from aCGH to identify CNVs [135].

We confirmed the limited power of the aCGH H244K platform for breakpoint characterization amplifying the deleted and non-deleted alleles in all populations and samples using a multiplex PCR-based assay. Although we cannot exclude the presence of other CNV in the region in some individuals, our analysis indicated that the 32 kb *LCE3C_LCE3B-del* is, at least, the predominant one in the studied populations. Therefore, we subsequently extended the PCR genotyping assay to several other worldwide populations in order to gain insight into the evolutionary mechanism influencing the copy-number variability of the region.

Biological impact of population difference in the frequency of the LCE3C_LCE3B-del

Our results demonstrated the tendency of higher frequencies of deletion in non-Sub-Saharan African populations, with the exception of some isolated groups that might be subject to local adaptation or a particular genetic drift. The fact that a deletion involving two genes is found in high frequency suggests that positive or negative selective forces are not acting, since deletions are mostly exposed to purifying selection and therefore expected to be in low frequencies or at least not expanded worldwide. Indeed, this geographic distribution pattern of allele frequencies had been described for other variants [252, 253], making it a likely result of a neutral demographic process called “allele surfing”. This phenomenon would be the result of the intense amount of genetic drift produced by strong bottlenecks that occurred during the exit “out of Africa”, which were followed by a spatial expansion leading to a geographic spread of an allele and increasing its frequency in newly colonized areas [254]. Moreover, we found a similar geographical pattern of linkage disequilibrium (LD) between the CNV and a previously identified tag SNP rs4112788 [244], associated with the deletion, with higher extent of LD in non-Sub-Saharan African populations. This pattern revealed a single origin for the CNV and the

SNP and also matches the prediction from a model of sequential founder effects during the expansion from Africa [80, 82]. Furthermore, it also demonstrated that although a strong LD is particularly useful to optimize genotyping in association studies for complex disorders, the association might not be useful in studies applied to all populations, since the mutation could have occurred recurrently on different haplotypes in different populations.

It is clear that the deleted allele has been established in most world populations, which probably has some kind of functional consequences. Expression of *LCE3C* and *LCE3B* genes is induced upon epidermal activation as a consequence of inflammation or skin disease [244]. However, the high frequency of the deletion worldwide suggests some redundancy in the function of LCE genes present in the same locus. It is possible that other genes fulfil the role of *LCE3C* and *LCE3B*, although imperfectly, contributing to the abnormal differentiation and epidermal hyperproliferation characteristics of psoriatic lesions. Thus, as it happens in most complex diseases, when other susceptibility components do not exist, the *LCE3C_LCE3B-del* is insufficient to produce the abnormal phenotype, needing the concurrence of several susceptibility components for disease development.

In this study we demonstrated that the reliability to predict a region containing a CNV by aCGH was almost excellent and that the main problem of the platform used in our study was its lack of capacity for breakpoint resolution. This limitation is now improved by the use of current ultra-high-resolution microarray platforms, containing 24 to 42 million probes and are able to discover CNVs down to 500 bp [135]. However, the general drawbacks of microarray approaches are still not solved. They are unable to provide information on the location of duplicated copies, they cannot resolve breakpoints at a single-base-pair level, they suffer reduced sensitivity in the detection of single-copy gains (3 to 2 copy-number ratio) compared with deletions (1 to 2 ratio) and they are unable to detect balanced SV. At present, the only possibility to go one step further in the characterization of SV is to use NGS technologies. Although they also have limitations in duplicated or repetitive regions, these technologies can resolve breakpoints with an unprecedented resolution, and allow the characterization of all forms of SV, including highly complex rearrangements. This last point is important since complex SV generally appear at the somatic level, which means that they are predominantly found in cancer genomes. Indeed, the field of SV in cancer genome is probably the one that has gained the most from the advent of NGS, as several publications have described the pattern of SV found in different tumour genomes in the last five years [21, 22, 24, 30, 34, 35].

Development of a computational tool for the detection of structural variation

The next objective of this thesis came with the arrival of NGS technologies and our group joining the Chronic Lymphocytic Leukaemia Genome Project (CLL-GP) in 2009, through the SV analysis team. In the framework of the International Cancer Genome Consortium (ICGC), the design of the CLL-GP was based on the idea that sequencing the CLL genomes could reshape our understanding in the biology of this type of cancer, with direct implications for clinical translation. Specifically, our aim was to identify, using whole-genome sequencing (WGS) data, the different types of SV occurring in CLL that could be responsible for the development and progression of the malignancy.

Nowadays there are over 20 computational tools to analyse SV from NGS data, but when we initiated the project, in 2010, only few were available [147, 148, 153, 255, 256]. These pipelines used only pair-reads (PR) or read-depth (RD) strategies and their SV analysis required high-computational costs. This made them difficult to use in laboratories lacking high computational resources. In addition, these strategies have limitations regarding the type and size of SV that they are able to detect, as each one has different strengths and weaknesses [135, 146]. On the other hand, none of these SV callers reported experimental evidence of their capacity to detect the full range of structural genomic changes. The short size of sequenced reads, the huge amount of data generated by NGS, and the nature of the genome with large number of repetitive sequences [138] hampered, and still hampers, SV discovery and the correct understanding of their biological consequences. Hence, we decided to develop a new tool, called *PeSV-Fisher*, in order to contribute to the improvement of SV analysis by WGS. We constructed the pipeline with the aim to provide capacity to: (a) detect and correctly interpret the results for the characterization of complex SV; (b) discard the maximum number of false positive calls due to the presence of repetitive elements; and (c) directly identify somatic variation in cancer genomes, analysing normal and tumour paired samples simultaneously and producing a list of non-shared variants.

The result of this effort has been published under the title of “PeSV-Fisher: Identification of somatic and non-somatic structural variants using next-generation sequencing data”. The tool is able to work on the two different types of sequence data based on pairs of reads, namely paired-end or mate-pair libraries, and starts from sequence alignment data in the BAM format, which is a widely accepted standard in the sequencing community. To achieve each of our objectives, several strategies have been used. Briefly, to correctly interpret the large amount of results obtained using WGS data, we developed a method combining PR and RD strategies. Although more recent methods have already begun to consider both PR and RD signals [163, 165, 166], our pipeline goes one step further and does not only use RD to confirm CNV breakpoint as predicted by PR, but also to define complex SV. Our strategy is based on the combination of different types of aberrantly aligned read-pairs, which define PR predictions, and results obtained by RD. Thus, *PeSV-Fisher* is capable to define simple deletions, inversions or translocations and also more complex events, such us *copy-paste* events (DNA fragments copied and inserted elsewhere in the genome) and *cut-paste* events (DNA fragments that have been moved to another locus), which are detected by the co-localization of different types of breakpoints identified by PR and RD signals.

Another strategy for the detection of SV by NGS is the identification of split-reads (SR). SR is probably the most sensitive procedure for the detection of breakpoints, as it looks for reads that are broken because they fall just in the breakpoint position. Some recent algorithms like Pindel [155] have introduced the SR approach in their analysis, but its application to NGS data sets is currently limited owing to the difficulty in aligning short reads. The strategy is based on the identification of broken reads and the probability for them to be mapped uniquely in the genome is really lower than in the normal reads.

When comparing our method against validated deletions obtained from a high-coverage whole-genome sample (NA19240) from the 1000 Genomes Project (1000GP), we confirmed the importance of the combination of PR and RD signals. For example, we found that around 30% of *distance*-aberrant clusters (potential deletions) detected by *PeSV-Fisher* and overlapping the list of validated deletions given by the 1000GP, did not show a significant decrease of RD, which may suggest false deletion calls and probably intra-chromosomal transpositions of small fragments of DNA. More interestingly, in six cases the variant definition of *PeSV-Fisher* classified them as other than deletions (two copy-paste and two cut-paste events). Additionally, *PeSV-Fisher* keeps information concerning the orientation of aberrantly aligned read-pairs, which can be important for the functional evaluation of the SV. For example, the positional effect of a gene copy in any region of the genome could be different if this copy is inserted in one orientation or another.

On the other hand, we included in the pipeline, as a user-option, a module to filter out those breakpoints detected by PR and falling in repetitive regions. This module works for rearrangements that have not been defined by any overlapping combination in the previous step, searching for and filtering out those where at least one of the breakpoints falls within repetitive elements, such as segmental duplications, simple repeat sequences or low-divergent transposable elements. However, since there is evidence that segmental duplications or shorter common repeat sequences show a high frequency of overlap with SV breakpoints [19], the removed rearrangements could still indicate valid SV. Therefore, rearrangements involving repetitive sequences are not removed but archived as "putative" SV. Because >50% of the genome involves repetitive sequences [138] and NGS technologies produce reads not large enough to be aligned in a single locus, most studies have been focused on non-repetitive regions and have been blind towards the phenotypic contribution of the SV in complex, repeat-rich, highly duplicated areas of the genome. *PeSV-Fisher* tries to improve the analysis of these regions, but the difficulties will be better faced with futures advances in DNA sequencing technology [257], including longer DNA reads that will increase the accessible genome and will enable the assessment of the SV embedded in long repeat structures, such as most of balanced inversions.

PeSV-Fisher has the capacity to simultaneously analyse two samples and produce a list of non-shared variants between them, but also allowing the possibility of analysing whole-genome samples individually. The simultaneous analysis of two genomes is especially useful in cancer genomics. In addition, the method does not take into account the diploid nature of the genome, although this aspect has recently been considered to solve the co-localization of more than two variants at the same locus [149]. For example, the cancer genome is characterized by the presence of clonal mosaicism [258], which is defined as the coexistence of cells with two or more distinct genotypes [259]. Therefore, the identification of more than two SV at the same locus in a cancer genome is conceivable.

Two of the major challenges that NGS data must face are the computer memory limitation and the runtime needed. Thus, the current tendency is to parallelize processes and to use large-scale computing cluster with high numbers of cores. Being aware that many laboratories do not have the computational infrastructure needed for this high-dimensional data analysis, we invested a great effort in these issues by: (a) splitting the data by chromosomes and making classes with low coupling between them, hence the process can be easily parallelized; and (b) using sorted strategies that directly manipulate files, thus the memory is better balanced. Consequently, *PeSV-Fisher* can be launched on either a cluster or workstation. For example, using a workstation with 12 cores and 48GB of memory under a Linux environment, we spent 805 minutes for the whole process of analysing in multithreading mode two paired genomes simultaneously (to find somatic variants), with a coverage mean of 45x each. The memory by thread was around 0.1 to 0.4GB during the longest part of the execution with isolated peaks no higher than 2.4GB.

To test the performance of *PeSV-Fisher* we compared it with that of *BreakDancer* (PR strategy) and the more recent version of *VariationHunter* (PR and the recently included RD for CNVs) tools. Results showed that the overlap between *PeSV-Fisher* and each tool was similar to the overlap between the two others. However, the main problem when comparing SV callers, as reported previously [14], is that each algorithm makes a large number of unique calls. A fraction of these calls might be due to different breakpoint accuracies, different size distribution patterns of deletions reported by each algorithm, or the presence of false positives calls. Of note, the reliability of SV discovery methods depends on mapping reads onto their genomic locus of origin, being of extreme importance the alignment algorithm used for WGS data. The three tools compared in the analysis employed different alignment algorithms, as *PeSV-Fisher* used BWA, *BreakDancer* MAQ and *VariationHunter* mrFAST [14]. It is well known that mapping programs, even though they can differ in some parameters, generate a large amount of ambiguous alignments, leading to a similarly large amount of false positive breakpoint calls and thus false positive SV predictions by the PR approach.

The strategy designed to evaluate the *PeSV-Fisher* in terms of true positive rates, was based on the construction of different scenarios using a set of confidence parameters that have an influence on the consistency of breakpoint predictions. We used the SV results from one of the four CLL cases, previously analysed by the CLL-GP, and we validated the detected breakpoints using targeted re-sequencing. Results were obtained applying the PR strategy of *PeSV-Fisher* and an independent split-read (SR) analysis for the confirmation of breakpoint predictions using the GEM aligner [260]. Results from both validation approaches showed the same significant considerations: (a) the importance of the phred-scale quality score, *Q*, value, which evaluate the mappability of the read and the base call accuracy, for the reliability of a SV even in those rearrangements that do not contain repetitive sequences in the breakpoints. That is, the higher the value of *Q* (*Q*>35, >99.9% of the probability of a read to be mapped uniquely) in at least one read-pair forming of the total list of read-pairs pointing to the same breakpoint (cluster), the higher the probability of this rearrangement to be real; and (b) the expected importance of the number of read-pairs in a cluster: the larger the number of read-pairs, the more confident the results. However, if the number of read-pairs exceeds significantly the average coverage rate, this is an indication that breakpoints might fall into specific types of highly repetitive or highly polymorphic regions and, therefore, that it might be a potential false positive call.

Further, we searched for the performance of *PeSV-Fisher* in terms of false negative rates. Using another CLL case from the CLL-GP in which we detected a high number of SV by WGS, we extracted breakpoints detected only in the tumour sample and then we asked whether the breakpoints were also found in the normal sample using targeted re-sequencing. Since genotyping errors are a common source of false-positive somatic SV calls in cancer sequencing studies, a germline breakpoint may be misclassified as somatic due to a false negative in the matched normal sample. Following this strategy, we eventually established a reasonably low false negative rate (11.9%).

Characterization of somatic structural variation in chronic lymphocytic leukaemia

The study of the genomics of CLL is an ambitious challenge because leukemic cells present an enormous heterogeneity and there is not a major specific genomic alteration that explains the origin of the disease. Thus, the overall goal of the CLL-GP was to dissect the genome of cancer and normal cells of 500 cases of CLL using different types of NGS and microarray approaches, to generate a comprehensive catalogue of genomic alterations involved in the development and progression of the CLL phenotype, and integrating clinical, biological and epidemiological information.

Around 90% of CLL patients present SV, but only four rearrangements (13q14, 11q22-q23 and 17p13 deletions, and trisomy 12) have been described with a frequency higher than 10% [222, 228-231]. The majority of the CLL cases carry only one type of SV, with a small subset of cases carrying more than three (considered CLL cases with genomic complexity). Moreover, the four main aberrations mentioned above affect ~70% of patients with SV [222, 228-231]. These observations, obtained before the NGS period, might suggest *a priori* that the SV would unlikely explain the heterogeneity of CLL and that they would not constitute the most important type of alteration involved in its tumorigenesis.

It is important to note that most SV in CLL had been identified by cytogenetic methods and only recently, with the use of high-resolution microarrays, several copy-number alterations of intermediate size have begun to emerge. The presence or absence of the mentioned recurrent SV represents an important factor in predicting survival [222, 230] and their identification is still one of the essential routine steps at the diagnosis of the disease. These facts led us to hypothesize that (a) critical SV in CLL might have not yet been detected due to the low resolution of cytogenetic and microarrays approaches; and (b) the clearly recognized role of the known SV in the prognosis of CLL might be consequence of the important implication of this type of variations in the pathogenesis of the disease.

WGS is a comprehensive NGS approach that enables the full genomic examination of the normal and tumour genomes at the level of both single-nucleotide alterations (SNAs) and SV. It allows the detection of all forms of SV and the characterization of their breakpoints with high resolution. Hence, CLL-GP took this opportunity to explore in depth the leukemic and normal genomes, in order to perform a comprehensive study of CLL cases, combining WGS with clinical characteristics and clinical outcomes. The article entitled "Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia" contains the results of the first effort of the CLL-GP. In this study, four CLL cases representative of different forms of the disease (two *IGHV* mutated and two *IGHV* unmutated cases) were analysed by WGS. From the sequencing of the four cases, the evaluation of single-nucleotides reported 46 SNAs potentially affecting gene function and not previously linked to CLL. These SNAs were further explored in 363 CLL patients, revealing that *NOTCH1, XPO1, KLHL6* and *MYD88* mutations are recurrent, that their profiles depend on the *IGHV* mutation status, and that they probably represent activating events for tumour development, which imply that they might also be potential therapeutic targets.

The examination of the SV in these four WGS cases reported 6 known copy-number alterations (CNAs) (deletion of 13q14 in three cases and a deletion on 6q in one case) and 4 novel variants (two complex gains on 3q26 and 6q15-q16, one tandem duplication on 2p16-p15, and one fold-back inversion on 1q21). Interestingly, three of these SV corresponded to complex SV (two complex gains and the fold-back inversion), highlighting the capacity of WGS to identify higher patterns of genomic complexity in a given tumour. The fold-back inversion, which is defined as an inversion co-occurring with duplication, was described for the first time in a previous study of pancreatic cancer

[21]. Our results further demonstrate the presence of complex somatic SV in the cancer genome, with the particularity that CLL is a cancer type typically characterized by a much lower number of structural rearrangements than other tumours.

Characterization of complex somatic structural variants: a chromothripsis case

PeSV-Fisher allowed us to identify a special case of CLL carrying a significantly high number of somatic SV. This case (CLL16) had been analysed by cytogenetic, different microarrays, WGS, whole-exome sequencing (WES), and RNA sequencing (RNA-seq), in the systematic genomic examination of CLL samples included in the CLL-GP. However, only the SNAs of this case found by WES were reported together with other 105 CLL cases [261]. *PeSV-Fisher* identified a large number of somatic SV showing a specific chromosomal geographic localization. This particularity could be consistent with a pattern described for other cancer genomes and coined with the term of chromothripsis [24]. To discard that this unexpected large number of rearrangements was not derived from sequencing or mapping errors, we decided to compare WGS results with microarray data for the confirmation of particular CNA profiles, and to perform a targeted re-sequencing approach for the validation of predicted breakpoints. This targeted re-sequencing was also eventually used for the evaluation of false negative rates of *PeSV-Fisher.*

We clearly confirmed the singular pattern of complex SV with chromosomal geographic localization in specific regions of the genome, defining a new case of chromothripsis in CLL. Since our understanding of the chromothripsis is still at early stages, we decided to go further in the analysis of case CLL16 and we designed a study to evaluate the phenotypic impact of the phenomenon on the progression of the patient disease.

Genome profiles obtained so far by cancer-genome sequencing projects have underestimated the complexity of cancer cells because they have provided snapshots from cases at a single time-point. Following the idea of natural selection acting on cancer, an initial cell becomes tumour under the effect of a specific driver mutation(s). However, during cancer progression certain subgroups of heterogeneous cell populations formed over time might also evolve and adapt. This process explains that in some cases, after treatment, some tumour cells can survive and regenerate the cancer with potential further malignity [178]. Leukaemia is a type of cancer in which is possible to perform a good monitoring of the clonal evolution of tumour cells. Indeed, some recent studies have monitored SNAs and CNAs in CLL samples taken from the same patient, identifying both founder events (eventually potential drivers) and subclonal compositions becoming dominant after therapy or even just before [176, 177]. The dynamics of clonal diversification and selection are critical to distinguish driver from passenger mutations and to understand and prevent therapeutic resistance.

Although the mechanism by which chromothripsis arises is uncertain, the molecular characteristics of the breakpoints suggest that the multiple focal lesions occurred by a single catastrophic oncogenic event, in contrast of the basic principle of Darwinian evolution [24, 204, 207]. Subsequently, it is thought that after the oncogenic transformation of the affected cell due to chromothripsis appearance, the complex pattern of SV is expanded to almost all of derived tumour cells. Since the presence of a high number of structural rearrangements might entail important functional consequences, it could explain why chromothripsis has been associated with more aggressive malignancies [32, 34, 35, 211]. In the first sample in which chromothripsis was identified (a case of CLL), the apparition of the phenomenon resulted in a rapid deterioration of the clinical course, and a re-analysis done after chemotherapy showed that the complex pattern of SV persisted [24]. It suggested the idea that chromothripsis was determined before the patient was first diagnosed. However, an extended evaluation of the genetic evolution of a cancer case with chromothripsis had never been reported, and from a clinical perspective much remains still to be understood.

Thus, in order to characterize in depth the complexity and evolution of some structural alterations in a cancer genome, we carried out a longitudinal analysis of case CLL16 over a period of eleven years of disease progression using NGS and microarray approaches. From this study we have published the article entitled "Sporadic and reversible chromothripsis in chronic lymphocytic leukaemia revealed by longitudinal genomic analysis". Patient CLL16 showed an increase of the aggressiveness of the disease and a requirement of treatment for the first time at the moment of chromothripsis detection, so that he evolved from a CLL clinical stage A0 (low risk) to stage BII (intermediate risk). A longitudinal analysis using samples obtained before and after chromothripsis provided us with a comprehensive view of the evolution of this phenomenon in CLL, gaining insight into the understanding the potential driver role of complex SV in tumorigenesis.

Biological impact of chormothripsis in a chronic lymphocytic leukaemia patient

From a biological point of view, we were aware that conclusions extracted from the analysis of a single case could be limited. However, we consider that the study has provided some new insights that could contribute to a better understanding of the evolution of CLL over time, specially linked to chromothripsis: (a) chromothripsis appeared in a single catastrophic event (all rearrangements are found in a same clonal fraction) but did not play a role in the tumour initiation (it has not been found in any additional sample); (b) chromothripsis might have been the consequence of previous alterations expanded from an early leukaemia cell, which would have led to chromosome instability associated with defects in DNA repair; (c) a subset of non-complex SV were present in all time-points of disease progression and in the same high clonal fraction (~90%), suggesting their potential role in the CLL development, in contrast of the complex SV associated with chromothripsis; (d) the complex SV added on a previous tumour cell carrying expanded alterations might have conferred a selective advantage to the subclonal population increasing its malignancy and therefore providing a more aggressive CLL phenotype; (e) chromothripsis subclones did not survive chemotherapy and did not reappear for a period of 10 years, having therefore no apparent implications for patient diagnosis; and (f) in contrast of SV, almost all point mutations remained fixed during the disease progression, suggesting a more important role in the CLL development.

Of these six evidences obtained in our study, only the demonstration that chromothripsis appeared in a single catastrophic event was consistent with previous knowledge. However, the notion that chromothripsis should be the oncogenic event, present in all tumour cells and directly linked to the tumorigenesis, has only been inferred from the observation of the phenomenon in single snapshots. Only in the first description of a chromothripsis case the researchers went one step further in the understanding of the role of this phenomenon in tumour progression, re-analysing the same sample after therapy. The results were of the same chromothripsis pattern in both samples [24] and no other study has tried to evaluate another possibility. The hypothesis is that, with the presence of chromothripsis, the probability of recovering clones with one or more beneficial driver alterations, and a minimal load of deleterious passenger mutations, might be close to 100%. This is perhaps the more logical assumption, but the situation observed in our case would be also feasible: a subset of driver alterations (probably a combination between SV and point mutations) leads to CLL development, and one or some of these alterations affects DNA damage repair genes, providing a high genomic instability that can facilitate the occurrence of chromothripsis within the tumour cells. The subclones carrying chromothripsis are then expanding and increasing the aggressiveness of the disease, but by the fortunate effect of the chemotherapy this sub-population has ended up being eliminated. In fact, this situation fits well with the relationship between alterations in *TP53* gene and the cases of chromothripsis in medulloblastoma and acute myeloid leukaemia [34].

The example of case CLL16 reinforces the importance of continuously monitoring the genomic evolution of cancer cells during disease progression. Because different tumour cell shape the malignant tissue, cancer can evolve through the competition and interaction among its cells, following the changes on the tissue environment. Therefore, we need to examine how tumour cells adapt to a changing tissue environment during disease progression, especially after chemotherapy rounds, because the heterogeneity leads to the possibility of developing therapy resistance and further aggressiveness in the disease. Patient monitoring by genomic sequencing approaches at different time-points, should result in a better understanding of the complete architecture of tumour cells, that should help in therapeutic decisions along the clinical course of the disease. In addition, a recent analysis demonstrated that clonal evolution patterns are different in individual patients, revealing that each cancer in each patient has an individually unique genomic profile variable over time [176]. The potential to personalize therapeutic choices for patients on the basis of the genomic architecture of their tumour cells, is the long-term aspiration for cancer-genome studies.

Concluding remarks

NGS has become a true revolution in the field of human genomics, with an enormous impact in the characterization of SV. However, in addition to the robustness, flexibility and low input material required, nowadays microarray-based technologies are still widely used. Microarray-based assays have replaced karyotyping for the analysis of developmental disabilities and congenital abnormalities [136], and they will remain the gold standard method until sequencing cost drops more and downstream analyses are facilitated. It is expected that the apparition of the "third-generation" sequencing technologies will improve the ability to analyse all types of variants using whole-genomes at lower costs and high-throughput. Therefore, the number of new discovered changes will also probably increase. A good discrimination of functional important variants from general polymorphisms is one of the more essential challenges to face, mainly in terms of disease-related studies. A good public catalogue of polymorphisms found in the general population will benefit from statistical approaches for rejecting non-functional variants and enhancing the potential damaging ones. In this sense, the 1000GP is already providing valuable information, mainly regarding single-nucleotide polymorphisms (SNPs) found in many worldwide populations. This allows us to discard high frequent variants in studies of diseases, and to better understand the population genetics of such variants. However, more resources are needed to guide the interpretation of the impact of SV. Since the growing of interest for personalized medicine, not only in cancer, the comprehensive characterization of all types of genomic variation is the most essential challenge to face.

.

E. CONCLUSIONS

The analysis of pools of samples by aCGH allows the detection of common CNVs in groups of human populations but lacks precision in the definition of breakpoints even for a simple rearrangement.

1. Most of the population-specific CNVs involved genes related to environmental responses, suggesting that the SV could be the consequence of specific environmental adaptations, which could also influence the different prevalence of some diseases among populations.
2. The worldwide distribution of the 32-kb deletion of the *LCE3C* and *LCE3B* genes, associated with psoriasis and other autoimmune diseases, revealed that neutral selective forces derived from demographic histories might also explain specific patterns of SV between populations.
3. The reliability to predict a region containing a common CNV by aCGH was almost excellent, but the probe density of the platform limited the precise definition of the breakpoints.

PeSV-Fisher provides a novel strategy to correctly interpret the large amount of results obtained by WGS, classifies the data to discard the maximum number of false positive calls, and allows obtaining directly somatic SV in cancer genomes in a relatively low computational cost.

1. The co-localization of breakpoints of SV predicted by PR, in combination with RD signals, allows the interpretation of SV within highly polymorphic genomic regions, SV showing a complex genomic pattern, and the correct definition of SV involving changes in copy-number.
2. The automatic classification of aberrations not involved in any category of SV, and falling into repetitive regions, is particularly interesting to discard potentially false calls.
3. PeSV-Fisher is the first pipeline with the capacity to produce a list of non-shared variants analysing two samples simultaneously, which is especially useful in cancer genomics.

WGS allows the precise detection of SV and reveals that somatic rearrangements can have high levels of genomic complexity, but some features of the NGS approach and nature of the genome hinder the complete characterization of SV.

1. The novel SV detected in four CLL patients by WGS involved balanced rearrangements or was defined as "complex", which explains their previous non-characterization and highlights the importance of WGS for the identification of all forms of SV.

2. WGS further revealed an unusual complex pattern of SV in a CLL case that was consistent with the recently described phenomenon of chromothripsis.

3. The short size of sequenced reads hampers the discovery of SV in repetitive regions and generates a huge amount of false positive calls, revealing a need to improve the strategy for the detection of the SV by WGS.

The longitudinal analysis performed in a CLL case reveals the importance of continuous monitoring the genomic evolution of cancers cells in order to achieve a comprehensive view of the dynamism of driver alterations, which might help in therapeutic decisions along the clinical course of the disease.

1. The longitudinal analysis of the case of choromothripsis demonstrated that this phenomenon did not play a role in the tumour initiation, refusing the previous notion that the appearance of highly complex SV might be an oncogenic event.
2. The study allowed the identification of a subset of SV and SNA present in almost all tumour cells over the time, suggesting their potential driver role in the development of the CLL, and their potential capacity to resistance to therapy.
3. Conclusions extracted from a single case could be limited, but our study demonstrate the importance of the elucidation of clonality in tumour cells for a better understanding of the progression of the disease, and to correctly design of therapeutic intervention.

F. BIBLIOGRAPHY

1. Stranger, B.E., et al., Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science, 2007. **315**(5813): p. 848-53.

2. Zhang, F., et al., *Copy number variation in human health, disease, and evolution.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 451-81.

3. Altshuler, D.M., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.

4. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

5. Abecasis, G.R., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

6. Stefansson, H., et al., *A common inversion under selection in Europeans.* Nat Genet, 2005. **37**(2): p. 129-37.

7. Lejeune, J., M. Gautier, and R. Turpin, *[Study of somatic chromosomes from 9 mongoloid children].* C R Hebd Seances Acad Sci, 1959. **248**(11): p. 1721-2.

8. Ford, C.E., et al., *A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome).* Lancet, 1959. **1**(7075): p. 711-3.

9. Jacobs, P.A. and J.A. Strong, *A case of human intersexuality having a possible XXY sex-determining mechanism.* Nature, 1959. **183**(4657): p. 302-3.

10. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome.* Nat Genet, 2004. **36**(9): p. 949-51.

11. Tuzun, E., et al., *Fine-scale structural variation of the human genome.* Nat Genet, 2005. **37**(7): p. 727-32.

12. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome.* Science, 2004. **305**(5683): p. 525-8.

13. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-54.

14. Mills, R.E., et al., *Mapping copy number variation by population-scale genome sequencing.* Nature, 2011. **470**(7332): p. 59-65.

15. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

16. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.

17. Sharp, A.J., Z. Cheng, and E.E. Eichler, *Structural variation of the human genome.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 407-42.

18. Pang, A.W., et al., *Towards a comprehensive structural variation map of an individual human genome.* Genome Biol, 2010. **11**(5): p. R52.

19. Kidd, J.M., et al., *Mapping and sequencing of structural variation from eight human genomes.* Nature, 2008. **453**(7191): p. 56-64.

20. Campbell, P.J., et al., *Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.* Nat Genet, 2008. **40**(6): p. 722-9.

21. Campbell, P.J., et al., *The patterns and dynamics of genomic instability in metastatic pancreatic cancer.* Nature, 2010. **467**(7319): p. 1109-13.

22. Berger, M.F., et al., *The genomic complexity of primary human prostate cancer.* Nature, 2011. **470**(7333): p. 214-20.

23. Gu, W., F. Zhang, and J.R. Lupski, *Mechanisms for human genomic rearrangements.* Pathogenetics, 2008. **1**(1): p. 4.

24. Stephens, P.J., et al., *Massive genomic rearrangement acquired in a single catastrophic event during cancer development.* Cell, 2011. **144**(1): p. 27-40.

25. Zhang, F., C.M. Carvalho, and J.R. Lupski, *Complex human chromosomal and genomic rearrangements.* Trends Genet, 2009. **25**(7): p. 298-307.

26. Perry, G.H., et al., *The fine-scale and complex architecture of human copy-number variation.* Am J Hum Genet, 2008. **82**(3): p. 685-95.

27. Kidd, J.M., et al., *A human genome structural variation sequencing resource reveals insights into mutational mechanisms.* Cell, 2010. **143**(5): p. 837-47.

28. Pang, A.W., et al., *Mechanisms of formation of structural variation in a fully sequenced human genome.* Hum Mutat, 2013. **34**(2): p. 345-54.

29. Korbel, J.O., et al., *Paired-end mapping reveals extensive structural variation in the human genome.* Science, 2007. **318**(5849): p. 420-6.

30. Stephens, P.J., et al., *Complex landscapes of somatic rearrangement in human breast cancer genomes.* Nature, 2009. **462**(7276): p. 1005-10.

31. Kloosterman, W.P., et al., *Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer.* Genome Biol, 2011. **12**(10): p. R103.

32. Magrangeas, F., et al., *Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients.* Blood, 2011. **118**(3): p. 675-8.

33. Malhotra, A., et al., *Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms.* Genome Res.

34. Rausch, T., et al., *Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations.* Cell, 2012. **148**(1-2): p. 59-71.

35. Molenaar, J.J., et al., *Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes.* Nature, 2012. **483**(7391): p. 589-93.

36. Kloosterman, W.P., et al., *Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline.* Hum Mol Genet, 2011. **20**(10): p. 1916-24.

37. Liu, P., et al., *Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements.* Cell, 2011. **146**(6): p. 889-903.

38. Chiang, C., et al., *Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration.* Nat Genet. **44**(4): p. 390-7, S1.

39. Hastings, P.J., et al., *Mechanisms of change in gene copy number.* Nat Rev Genet, 2009. **10**(8): p. 551-64.

40. Weischenfeldt, J., et al., *Phenotypic impact of genomic structural variation: insights from and for human disease.* Nat Rev Genet, 2013. **14**(2): p. 125-38.

41. Stankiewicz, P. and J.R. Lupski, *Genome architecture, rearrangements and genomic disorders.* Trends Genet, 2002. **18**(2): p. 74-82.

42. Inoue, K. and J.R. Lupski, *Molecular mechanisms for genomic disorders.* Annu Rev Genomics Hum Genet, 2002. **3**: p. 199-242.

43. Lee, J.A., C.M. Carvalho, and J.R. Lupski, *A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders.* Cell, 2007. **131**(7): p. 1235-47.

44. Sharp, A.J., et al., *Segmental duplications and copy-number variation in the human genome.* Am J Hum Genet, 2005. **77**(1): p. 78-88.

45. Eichler, E.E., *Recent duplication, domain accretion and the dynamic mutation of the human genome.* Trends Genet, 2001. **17**(11): p. 661-9.

46. Han, K., et al., *L1 recombination-associated deletions generate human genomic variation.* Proc Natl Acad Sci U S A, 2008. **105**(49): p. 19366-71.

47. Sen, S.K., et al., *Human genomic deletions mediated by recombination between Alu elements.* Am J Hum Genet, 2006. **79**(1): p. 41-53.

48. Itsara, A., et al., *Population analysis of large copy number variants and hotspots of human genetic disease.* Am J Hum Genet, 2009. **84**(2): p. 148-61.

49. Inaki, K. and E.T. Liu, *Structural mutations in cancer: mechanistic and functional insights.* Trends Genet. **28**(11): p. 550-9.

50. Kloosterman, W.P., et al., *Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms.* Cell Rep. **1**(6): p. 648-55.

51. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.

52. Callinan, P.A., et al., *Alu retrotransposition-mediated deletion.* J Mol Biol, 2005. **348**(4): p. 791-800.

53. Han, K., et al., *Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages.* Nucleic Acids Res, 2005. **33**(13): p. 4040-52.

54. Stewart, C., et al., *A comprehensive map of mobile element insertion polymorphisms in humans.* PLoS Genet. **7**(8): p. e1002236.

55. Olson, M.V., *Human genetic individuality.* Annu Rev Genomics Hum Genet. **13**: p. 1-27.

56. Sabeti, P.C., et al., *Positive natural selection in the human lineage.* Science, 2006. **312**(5780): p. 1614-20.

57. Brcic-Kostic, K., *Neutral mutation as the source of genetic variation in life history traits.* Genet Res, 2005. **86**(1): p. 53-63.

58. Hurles, M.E., E.T. Dermitzakis, and C. Tyler-Smith, *The functional impact of structural variation in humans.* Trends Genet, 2008. **24**(5): p. 238-45.

59. Kleinjan, D.A. and V. van Heyningen, *Long-range control of gene expression: emerging mechanisms and disruption in disease.* Am J Hum Genet, 2005. **76**(1): p. 8-32.

60. Schlattl, A., et al., *Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions.* Genome Res, 2011. **21**(12): p. 2004-13.

61. Conrad, D.F., et al., *Origins and functional impact of copy number variation in the human genome.* Nature, 2010. **464**(7289): p. 704-12.

62. Nguyen, D.Q., C. Webber, and C.P. Ponting, *Bias of selection on human copy-number variants.* PLoS Genet, 2006. **2**(2): p. e20.

63. Cann, H.M., et al., *A human genome diversity cell line panel.* Science, 2002. **296**(5566): p. 261-2.

64. Firth, H.V., et al., *DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources.* Am J Hum Genet, 2009. **84**(4): p. 524-33.

65. Hudson, T.J., et al., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993-8.

66. Lupski, J.R., *Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits.* Trends Genet, 1998. **14**(10): p. 417-22.

67. Stankiewicz, P. and J.R. Lupski, *Structural variation in the human genome and its role in disease.* Annu Rev Med, 2010. **61**: p. 437-55.

68. Lakich, D., et al., *Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A.* Nat Genet, 1993. **5**(3): p. 236-41.

69. Bondeson, M.L., et al., *Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome.* Hum Mol Genet, 1995. **4**(4): p. 615-21.

70. Osborne, L.R., et al., *A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome.* Nat Genet, 2001. **29**(3): p. 321-5.

71. Gimelli, G., et al., *Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions.* Hum Mol Genet, 2003. **12**(8): p. 849-58.

72. Kuppers, R., *Mechanisms of B-cell lymphoma pathogenesis.* Nat Rev Cancer, 2005. **5**(4): p. 251-62.

73. Rowley, J.D., *Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining.* Nature, 1973. **243**(5405): p. 290-3.

74. Lifton, R.P., et al., *A chimaeric 11 beta-hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension.* Nature, 1992. **355**(6357): p. 262-5.

75. David, D., et al., *Molecular characterization of a familial translocation implicates disruption of HDAC9 and possible position effect on TGFbeta2 in the pathogenesis of Peters' anomaly.* Genomics, 2003. **81**(5): p. 489-503.

76. Flint, J., et al., *High frequencies of alpha-thalassaemia are the result of natural selection by malaria.* Nature, 1986. **321**(6072): p. 744-50.

77. Gonzalez, E., et al., *The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility.* Science, 2005. **307**(5714): p. 1434-40.

78. Xue, Y., et al., *Adaptive evolution of UGT2B17 copy-number variation.* Am J Hum Genet, 2008. **83**(3): p. 337-46.

79. Holmes, L.V., et al., *Determining the Population Frequency of the CFHR3/CFHR1 Deletion at 1q32.* PLoS One, 2013. **8**(4): p. e60352.

80. Jakobsson, M., et al., *Genotype, haplotype and copy-number variation in worldwide human populations.* Nature, 2008. **451**(7181): p. 998-1003.

81. Rosenberg, N.A., et al., *Clines, clusters, and the effect of study design on the inference of human population structure.* PLoS Genet, 2005. **1**(6): p. e70.

82. Ramachandran, S., et al., *Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa.* Proc Natl Acad Sci U S A, 2005. **102**(44): p. 15942-7.

83. Barbujani, G., et al., *An apportionment of human DNA diversity.* Proc Natl Acad Sci U S A, 1997. **94**(9): p. 4516-9.

84. Aidoo, M., et al., *Protective effects of the sickle cell gene against malaria morbidity and mortality.* Lancet, 2002. **359**(9314): p. 1311-2.

85. Li, H. and R. Durbin, *Inference of human population history from individual whole-genome sequences.* Nature. **475**(7357): p. 493-6.

86. Perry, G.H., et al., *Diet and the evolution of human amylase gene copy number variation.* Nat Genet, 2007. **39**(10): p. 1256-60.

87. Armengol, L., et al., *Identification of copy number variants defining genomic differences among major human groups.* PLoS One, 2009. **4**(9): p. e7230.

88. Polimanti, R., et al., *Genetic variability of glutathione S-transferase enzymes in human populations: functional inter-ethnic differences in detoxification systems.* Gene. **512**(1): p. 102-7.

89. Kidd, J.M., et al., *Population stratification of a common APOBEC gene deletion polymorphism.* PLoS Genet, 2007. **3**(4): p. e63.

90. Pritchard, J.K., *Whole-genome sequencing data offer insights into human demography.* Nat Genet. **43**(10): p. 923-5.

91. Girirajan, S., C.D. Campbell, and E.E. Eichler, *Human copy number variation and complex genetic disease.* Annu Rev Genet. **45**: p. 203-26.

92. Gravel, S., et al., *Demographic history and rare allele sharing among human populations.* Proc Natl Acad Sci U S A. **108**(29): p. 11983-8.

93. Novembre, J. and A. Di Rienzo, *Spatial patterns of variation due to natural selection in humans.* Nat Rev Genet, 2009. **10**(11): p. 745-55.

94. Excoffier, L. and N. Ray, *Surfing during population expansions promotes genetic revolutions and structuration.* Trends Ecol Evol, 2008. **23**(7): p. 347-51.

95. Motulsky, A.G., *Jewish diseases and origins.* Nat Genet, 1995. **9**(2): p. 99-101.

96. Le Scouarnec, S. and S.M. Gribble, *Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics.* Heredity (Edinb), 2012. **108**(1): p. 75-85.

97. Ford, C.E. and J.L. Hamerton, *The chromosomes of man.* Nature, 1956. **178**(4541): p. 1020-3.

98. Caspersson, T., et al., *Chemical differentiation along metaphase chromosomes.* Exp Cell Res, 1968. **49**(1): p. 219-22.

99. Sato, Y., et al., *Reciprocal translocation involving the short arms of chromosomes 7 and 11, t(7p-;11p+), associated with myeloid leukemia with maturation.* Blood, 1987. **70**(5): p. 1654-8.

100. Yunis, J.J., *High resolution of human chromosomes.* Science, 1976. **191**(4233): p. 1268-70.

101. Smeets, D.F., *Historical prospective of human cytogenetics: from microscope to microarray.* Clin Biochem, 2004. **37**(6): p. 439-46.

102. Bauman, J.G., et al., *A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA.* Exp Cell Res, 1980. **128**(2): p. 485-90.

103. Cremer, T., et al., *Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes.* Hum Genet, 1988. **80**(3): p. 235-46.

104. Pinkel, D., et al., *Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4.* Proc Natl Acad Sci U S A, 1988. **85**(23): p. 9138-42.

105. Hoffman, M.W., S. Janney, and J.R. Batanian, *Cryptic deletion of EGR1 in association with a novel balanced t(5;22)(q31;q11.2) in a patient with myelodysplastic syndrome.* Cancer Genet Cytogenet, 2009. **191**(2): p. 106-8.

106. Heng, H.H., J. Squire, and L.C. Tsui, *High-resolution mapping of mammalian genes by in situ hybridization to free chromatin.* Proc Natl Acad Sci U S A, 1992. **89**(20): p. 9509-13.

107. Parra, I. and B. Windle, *High resolution visual mapping of stretched DNA by fluorescent hybridization.* Nat Genet, 1993. **5**(1): p. 17-21.

108. Molina, O., et al., *High-resolution fish on DNA fibers for low-copy repeats genome architecture studies.* Genomics. **100**(6): p. 380-6.

109. Kallioniemi, A., et al., *Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.* Science, 1992. **258**(5083): p. 818-21.

110. Weiss, M.M., et al., *Comparative genomic hybridisation.* Mol Pathol, 1999. **52**(5): p. 243-51.

111. Baak, J.P., et al., *Genomics and proteomics in cancer.* Eur J Cancer, 2003. **39**(9): p. 1199-215.

112. Speicher, M.R., S. Gwyn Ballard, and D.C. Ward, *Karyotyping human chromosomes by combinatorial multi-fluor FISH.* Nat Genet, 1996. **12**(4): p. 368-75.

113. Schrock, E., et al., *Multicolor spectral karyotyping of human chromosomes.* Science, 1996. **273**(5274): p. 494-7.

114. Kannan, T.P. and B.A. Zilfalil, *Cytogenetics: past, present and future.* Malays J Med Sci, 2009. **16**(2): p. 4-9.

115. Bieche, I., et al., *Novel approach to quantitative polymerase chain reaction using real-time detection: application to the detection of gene amplification in breast cancer.* Int J Cancer, 1998. **78**(5): p. 661-6.

116. Hollox, E.J., S.M. Akrami, and J.A. Armour, *DNA copy number analysis by MAPH: molecular diagnostic applications.* Expert Rev Mol Diagn, 2002. **2**(4): p. 370-8.

117. Schouten, J.P., et al., *Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification.* Nucleic Acids Res, 2002. **30**(12): p. e57.

118. Cooper, G.M., et al., *Systematic assessment of copy number variant detection via genome-wide SNP genotyping.* Nat Genet, 2008. **40**(10): p. 1199-203.

119. Curtis, C., et al., *The pitfalls of platform comparison: DNA copy number array technologies assessed.* BMC Genomics, 2009. **10**: p. 588.

120. Coe, B.P., et al., *Resolving the resolution of array CGH.* Genomics, 2007. **89**(5): p. 647-53.

121. Brennan, C., et al., *High-resolution global profiling of genomic alterations with long oligonucleotide microarray.* Cancer Res, 2004. **64**(14): p. 4744-8.

122. Carvalho, B., et al., *High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides.* J Clin Pathol, 2004. **57**(6): p. 644-6.

123. Snijders, A.M., et al., *Assembly of microarrays for genome-wide measurement of DNA copy number.* Nat Genet, 2001. **29**(3): p. 263-4.

124. Fiegler, H., et al., *DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones.* Genes Chromosomes Cancer, 2003. **36**(4): p. 361-74.

125. Chung, Y.J., et al., *A whole-genome mouse BAC microarray with 1-Mb resolution for analysis of DNA copy number changes by array comparative genomic hybridization.* Genome Res, 2004. **14**(1): p. 188-96.

126. Lucito, R., et al., *Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation.* Genome Res, 2003. **13**(10): p. 2291-305.

127. Pinkel, D., et al., *High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.* Nat Genet, 1998. **20**(2): p. 207-11.

128. Oostlander, A.E., G.A. Meijer, and B. Ylstra, *Microarray-based comparative genomic hybridization and its applications in human genetics.* Clin Genet, 2004. **66**(6): p. 488-95.

129. McCarroll, S.A., et al., *Integrated detection and population-genetic analysis of SNPs and copy number variation.* Nat Genet, 2008. **40**(10): p. 1166-74.

130. Conlin, L.K., et al., *Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis.* Hum Mol Genet, 2010. **19**(7): p. 1263-75.

131. Rodriguez-Santiago, B., et al., *Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome.* Am J Hum Genet, 2010. **87**(1): p. 129-38.

132. Gonzalez, J.R., et al., *A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data.* BMC Bioinformatics, 2011. **12**: p. 166.

133. Yamazawa, K., T. Ogata, and A.C. Ferguson-Smith, *Uniparental disomy and human disease: an overview.* Am J Med Genet C Semin Med Genet, 2010. **154C**(3): p. 329-34.

134. Park, H., et al., *Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing.* Nat Genet, 2010. **42**(5): p. 400-5.

135. Alkan, C., B.P. Coe, and E.E. Eichler, *Genome structural variation discovery and genotyping.* Nat Rev Genet, 2011. **12**(5): p. 363-76.

136. Miller, D.T., et al., *Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies.* Am J Hum Genet, 2010. **86**(5): p. 749-64.

137. Swerdlow, H., et al., *Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette.* J Chromatogr, 1990. **516**(1): p. 61-7.

138. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

139. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.

140. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.

141. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

142. Coffey, A.J., et al., *The GENCODE exome: sequencing the complete human exome.* Eur J Hum Genet, 2011. **19**(7): p. 827-31.

143. Hedges, D.J., et al., *Comparison of three targeted enrichment strategies on the SOLiD sequencing platform.* PLoS One, 2011. **6**(4): p. e18595.

144. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

145. Wold, B. and R.M. Myers, *Sequence census methods for functional genomics.* Nat Methods, 2008. **5**(1): p. 19-21.

146. Medvedev, P., M. Stanciu, and M. Brudno, *Computational methods for discovering structural variation with next-generation sequencing.* Nat Methods, 2009. **6**(11 Suppl): p. S13-20.

147. Chen, K., et al., *BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.* Nat Methods, 2009. **6**(9): p. 677-81.

148. Korbel, J.O., et al., *PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data.* Genome Biol, 2009. **10**(2): p. R23.

149. Hormozdiari, F., et al., *Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery.* Bioinformatics. **26**(12): p. i350-7.

150. Bashir, A., et al., *Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer.* PLoS Comput Biol, 2008. **4**(4): p. e1000051.

151. Sudmant, P.H., et al., *Diversity of human copy number variation and multicopy genes.* Science, 2010. **330**(6004): p. 641-6.

152. Chiang, D.Y., et al., *High-resolution mapping of copy-number alterations with massively parallel sequencing.* Nat Methods, 2009. **6**(1): p. 99-103.

153. Yoon, S., et al., *Sensitive and accurate detection of copy number variants using read depth of coverage.* Genome Res, 2009. **19**(9): p. 1586-92.

154. Abyzov, A., et al., *CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.* Genome Res, 2011. **21**(6): p. 974-84.

155. Ye, K., et al., *Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.* Bioinformatics, 2009. **25**(21): p. 2865-71.

156. Zhang, Z.D., et al., *Identification of genomic indels and structural variations using split reads.* BMC Genomics, 2011. **12**: p. 375.

157. Abel, H.J., et al., *SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data.* Bioinformatics, 2010. **26**(21): p. 2684-8.

158. Alkan, C., S. Sajjadian, and E.E. Eichler, *Limitations of next-generation genome sequence assembly.* Nat Methods, 2011. **8**(1): p. 61-5.

159. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing.* Genome Res, 2010. **20**(2): p. 265-72.

160. Gnerre, S., et al., *High-quality draft assemblies of mammalian genomes from massively parallel sequence data.* Proc Natl Acad Sci U S A, 2011. **108**(4): p. 1513-8.

161. Wu, X.L., et al., *TIGER: tiled iterative genome assembler.* BMC Bioinformatics, 2012. **13 Suppl 19**: p. S18.

162. Hajirasouliha, I., et al., *Detection and characterization of novel sequence insertions using paired-end next-generation sequencing.* Bioinformatics, 2010. **26**(10): p. 1277-83.

163. Sindi, S.S., et al., *An integrative probabilistic model for identification of structural variation in sequencing data.* Genome Biol, 2012. **13**(3): p. R22.

164. Rausch, T., et al., *DELLY: structural variant discovery by integrated paired-end and split-read analysis.* Bioinformatics, 2012. **28**(18): p. i333-i339.

165. Qi, J. and F. Zhao, *inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W567-75.

166. Medvedev, P., et al., *Detecting copy number variation with mated short reads.* Genome Res, 2010. **20**(11): p. 1613-22.

167. Jemal, A., et al., *Cancer statistics, 2008.* CA Cancer J Clin, 2008. **58**(2): p. 71-96.

168. Stratton, M.R., *Exploring the genomes of cancer cells: progress and promise.* Science, 2011. **331**(6024): p. 1553-8.

169. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.

170. Laird, P.W., *Cancer epigenetics.* Hum Mol Genet, 2005. **14 Spec No 1**: p. R65-76.

171. Talbot, S.J. and D.H. Crawford, *Viruses and tumours--an update.* Eur J Cancer, 2004. **40**(13): p. 1998-2005.

172. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-8.

173. Merlo, L.M., et al., *Cancer as an evolutionary and ecological process.* Nat Rev Cancer, 2006. **6**(12): p. 924-35.

174. Nik-Zainal, S., et al., *The life history of 21 breast cancers.* Cell. **149**(5): p. 994-1007.

175. Ding, L., et al., *Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing.* Nature. **481**(7382): p. 506-10.

176. Schuh, A., et al., *Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns.* Blood. **120**(20): p. 4191-6.

177. Landau, D.A., et al., *Evolution and impact of subclonal mutations in chronic lymphocytic leukemia.* Cell. **152**(4): p. 714-26.

178. Greaves, M. and C.C. Maley, *Clonal evolution in cancer.* Nature. **481**(7381): p. 306-13.

179. Stewart, S.A. and R.A. Weinberg, *Telomeres: cancer to human aging.* Annu Rev Cell Dev Biol, 2006. **22**: p. 531-57.

180. Maley, C.C., et al., *Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus.* Cancer Res, 2004. **64**(10): p. 3414-27.

181. Tao, Y., et al., *Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data.* Proc Natl Acad Sci U S A, 2011. **108**(29): p. 12042-7.

182. Futreal, P.A., et al., *A census of human cancer genes.* Nat Rev Cancer, 2004. **4**(3): p. 177-83.

183. Pleasance, E.D., et al., *A comprehensive catalogue of somatic mutations from a human cancer genome.* Nature, 2010. **463**(7278): p. 191-6.

184. Gundem, G., et al., *IntOGen: integration and data mining of multidimensional oncogenomic data.* Nat Methods, 2010. **7**(2): p. 92-3.

185. Negrini, S., V.G. Gorgoulis, and T.D. Halazonetis, *Genomic instability--an evolving hallmark of cancer.* Nat Rev Mol Cell Biol, 2010. **11**(3): p. 220-8.

186. Loeb, L.A., *Mutator phenotype may be required for multistage carcinogenesis.* Cancer Res, 1991. **51**(12): p. 3075-9.

187. Lynch, H.T. and A.J. Krush, *Cancer family "G" revisited: 1895-1970.* Cancer, 1971. **27**(6): p. 1505-11.

188. Miki, Y., et al., *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.* Science, 1994. **266**(5182): p. 66-71.

189. Wooster, R., et al., *Identification of the breast cancer susceptibility gene BRCA2.* Nature, 1995. **378**(6559): p. 789-92.

190. Lord, C.J. and A. Ashworth, *The DNA damage response and cancer therapy.* Nature. **481**(7381): p. 287-94.

191. Banerjee, S., S.B. Kaye, and A. Ashworth, *Making the best of PARP inhibitors in ovarian cancer.* Nat Rev Clin Oncol, 2010. **7**(9): p. 508-19.

192. Pommier, Y., et al., *DNA topoisomerases and their poisoning by anticancer and antibacterial drugs.* Chem Biol. **17**(5): p. 421-33.

193. Druker, B.J., *Translation of the Philadelphia chromosome into therapy for CML.* Blood, 2008. **112**(13): p. 4808-17.

194. Dickson, D., *Wellcome funds cancer database.* Nature, 1999. **401**(6755): p. 729.

195. Collins, F.S. and A.D. Barker, *Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies.* Sci Am, 2007. **296**(3): p. 50-7.

196. Pinkel, D. and D.G. Albertson, *Array comparative genomic hybridization and its applications in cancer.* Nat Genet, 2005. **37 Suppl**: p. S11-7.

197. Yates, L.R. and P.J. Campbell, *Evolution of the cancer genome.* Nat Rev Genet, 2012. **13**(11): p. 795-806.

198. Hillmer, A.M., et al., *Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes.* Genome Res, 2011. **21**(5): p. 665-75.

199. Beroukhim, R., et al., *The landscape of somatic copy-number alteration across human cancers.* Nature. **463**(7283): p. 899-905.

200. Bignell, G.R., et al., *Signatures of mutation and selection in the cancer genome.* Nature, 2010. **463**(7283): p. 893-8.

201. Beroukhim, R., et al., *The landscape of somatic copy-number alteration across human cancers.* Nature, 2010. **463**(7283): p. 899-905.

202. Ng, C.K., et al., *The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer.* J Pathol, 2012. **226**(5): p. 703-12.

203. Raphael, B.J. and P.A. Pevzner, *Reconstructing tumor amplisomes.* Bioinformatics, 2004. **20 Suppl 1**: p. i265-73.

204. Forment, J.V., A. Kaidi, and S.P. Jackson, *Chromothripsis and cancer: causes and consequences of chromosome shattering.* Nat Rev Cancer, 2012. **12**(10): p. 663-70.

205. Korbel, J.O. and P.J. Campbell, *Criteria for inference of chromothripsis in cancer genomes.* Cell, 2013. **152**(6): p. 1226-36.

206. Jones, M.J. and P.V. Jallepalli, *Chromothripsis: chromosomes in crisis.* Dev Cell, 2012. **23**(5): p. 908-17.

207. Korbel, J.O. and P.J. Campbell, *Criteria for inference of chromothripsis in cancer genomes.* Cell. **152**(6): p. 1226-36.

208. Tubio, J.M. and X. Estivill, *Cancer: When catastrophe strikes a cell.* Nature. **470**(7335): p. 476-7.

209. Meyerson, M. and D. Pellman, *Cancer genomes evolve by pulverizing single chromosomes.* Cell. **144**(1): p. 9-10.

210. Crasta, K., et al., *DNA breaks and chromosome pulverization from errors in mitosis.* Nature, 2012. **482**(7383): p. 53-8.

211. Hirsch, D., et al., *Chromothripsis and focal copy number alterations determine poor outcome in malignant melanoma.* Cancer Res, 2013. **73**(5): p. 1454-60.

212. Jackson, S.P. and J. Bartek, *The DNA-damage response in human biology and disease.* Nature, 2009. **461**(7267): p. 1071-8.

213. Rozman, C. and E. Montserrat, *Chronic lymphocytic leukemia.* N Engl J Med, 1995. **333**(16): p. 1052-7.

214. Dores, G.M., et al., *Chronic lymphocytic leukaemia and small lymphocytic lymphoma: overview of the descriptive epidemiology.* Br J Haematol, 2007. **139**(5): p. 809-19.

215. Goldin, L.R., et al., *Familial risk of lymphoproliferative tumors in families of patients with chronic lymphocytic leukemia: results from the Swedish Family-Cancer Database.* Blood, 2004. **104**(6): p. 1850-4.

216. Zenz, T., et al., *From pathogenesis to treatment of chronic lymphocytic leukaemia.* Nat Rev Cancer, 2010. **10**(1): p. 37-50.

217. Hamblin, T.J., et al., *Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia.* Blood, 1999. **94**(6): p. 1848-54.

218. Zenz, T., et al., *Monoallelic TP53 inactivation is associated with poor prognosis in chronic lymphocytic leukemia: results from a detailed genetic characterization with long-term follow-up.* Blood, 2008. **112**(8): p. 3322-9.

219. Austen, B., et al., *Mutations in the ATM gene lead to impaired overall and treatment-free survival that is independent of IGVH mutation status in patients with B-CLL.* Blood, 2005. **106**(9): p. 3175-82.

220. Crespo, M., et al., *ZAP-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia.* N Engl J Med, 2003. **348**(18): p. 1764-75.

221. Damle, R.N., et al., *CD38 expression labels an activated subset within chronic lymphocytic leukemia clones enriched in proliferating B cells.* Blood, 2007. **110**(9): p. 3352-9.

222. Dohner, H., et al., *Genomic aberrations and survival in chronic lymphocytic leukemia.* N Engl J Med, 2000. **343**(26): p. 1910-6.

223. Rai, K.R., et al., *Clinical staging of chronic lymphocytic leukemia.* Blood, 1975. **46**(2): p. 219-34.

224. Binet, J.L., et al., *A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis.* Cancer, 1981. **48**(1): p. 198-206.

225. Chiorazzi, N., K.R. Rai, and M. Ferrarini, *Chronic lymphocytic leukemia.* N Engl J Med, 2005. **352**(8): p. 804-15.

226. Klein, U. and R. Dalla-Favera, *Germinal centres: role in B-cell physiology and malignancy.* Nat Rev Immunol, 2008. **8**(1): p. 22-33.

227. Wierda, W.G., et al., *Characteristics associated with important clinical end points in patients with chronic lymphocytic leukemia at initial treatment.* J Clin Oncol, 2009. **27**(10): p. 1637-43.

228. Stilgenbauer, S., et al., *Clonal evolution in chronic lymphocytic leukemia: acquisition of high-risk genomic aberrations associated with unmutated VH, resistance to therapy, and short survival.* Haematologica, 2007. **92**(9): p. 1242-5.

229. Ouillette, P., et al., *Acquired genomic copy number aberrations and survival in chronic lymphocytic leukemia.* Blood. **118**(11): p. 3051-61.

230. Gunnarsson, R., et al., *Array-based genomic screening at diagnosis and during follow-up in chronic lymphocytic leukemia.* Haematologica, 2011. **96**(8): p. 1161-9.

231. Edelmann, J., et al., *High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations.* Blood. **120**(24): p. 4783-94.

232. Cimmino, A., et al., *miR-15 and miR-16 induce apoptosis by targeting BCL2.* Proc Natl Acad Sci U S A, 2005. **102**(39): p. 13944-9.

233. Gumy-Pause, F., P. Wacker, and A.P. Sappino, *ATM gene and lymphoid malignancies.* Leukemia, 2004. **18**(2): p. 238-42.

234. Stilgenbauer, S., et al., *Incidence and clinical significance of 6q deletions in B cell chronic lymphocytic leukemia.* Leukemia, 1999. **13**(9): p. 1331-4.

235. Gunnarsson, R., et al., *Large but not small copy-number alterations correlate to high-risk genomic aberrations and survival in chronic lymphocytic leukemia: a high-resolution genomic screening of newly diagnosed patients.* Leukemia. **24**(1): p. 211-5.

236. Nagy, B., et al., *Abnormal expression of apoptosis-related genes in haematological malignancies: overexpression of MYC is poor prognostic sign in mantle cell lymphoma.* Br J Haematol, 2003. **120**(3): p. 434-41.

237. Aref, S., et al., *c-Myc oncogene and Cdc25A cell activating phosphatase expression in non-Hodgkin's lymphoma.* Hematology, 2003. **8**(3): p. 183-90.

238. Grandori, C., et al., *The Myc/Max/Mad network and the transcriptional control of cell behavior.* Annu Rev Cell Dev Biol, 2000. **16**: p. 653-99.

239. Novak, U., et al., *A high-resolution allelotype of B-cell chronic lymphocytic leukemia (B-CLL).* Blood, 2002. **100**(5): p. 1787-94.

240. Cavazzini, F., et al., *Chromosome 14q32 translocations involving the immunoglobulin heavy chain locus in chronic lymphocytic leukaemia identify a disease subset with poor prognosis.* Br J Haematol, 2008. **142**(4): p. 529-37.

241. Maley, C.C., et al., *Genetic clonal diversity predicts progression to esophageal adenocarcinoma.* Nat Genet, 2006. **38**(4): p. 468-73.

242. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

243. Rosenberg, N.A., et al., *Genetic structure of human populations.* Science, 2002. **298**(5602): p. 2381-5.

244. de Cid, R., et al., *Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis.* Nat Genet, 2009. **41**(2): p. 211-5.

245. Huffmeier, U., et al., *Replication of LCE3C-LCE3B CNV as a risk factor for psoriasis and analysis of interaction with other genetic risk factors.* J Invest Dermatol, 2010. **130**(4): p. 979-84.

246. Riveira-Munoz, E., et al., *Meta-analysis confirms the LCE3C_LCE3B deletion as a risk factor for psoriasis in several ethnic groups and finds interaction with HLA-Cw6.* J Invest Dermatol, 2011. **131**(5): p. 1105-9.

247. Docampo, E., et al., *Deletion of the late cornified envelope genes, LCE3C and LCE3B, is associated with rheumatoid arthritis.* Arthritis Rheum, 2010. **62**(5): p. 1246-51.

248. Lu, X., et al., *Deletion of LCE3C_LCE3B is associated with rheumatoid arthritis and systemic lupus erythematosus in the Chinese Han population.* Ann Rheum Dis, 2011. **70**(9): p. 1648-51.

249. Docampo, E., et al., *Deletion of LCE3C and LCE3B is a susceptibility factor for psoriatic arthritis: a study in Spanish and Italian populations and meta-analysis.* Arthritis Rheum, 2011. **63**(7): p. 1860-5.

250. Xu, L., et al., *Deletion of LCE3C and LCE3B genes is associated with psoriasis in a northern Chinese population.* Br J Dermatol, 2011. **165**(4): p. 882-7.

251. Li, M., et al., *Deletion of the late cornified envelope genes LCE3C and LCE3B is associated with psoriasis in a Chinese population.* J Invest Dermatol, 2011. **131**(8): p. 1639-43.

252. Evans, P.D., et al., *Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size.* Hum Mol Genet, 2004. **13**(11): p. 1139-45.

253. Mekel-Bobrov, N., et al., *Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens.* Science, 2005. **309**(5741): p. 1720-2.

254. Hofer, T., et al., *Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection.* Ann Hum Genet, 2009. **73**(1): p. 95-108.

255. Hormozdiari, F., et al., *Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes.* Genome Res, 2009. **19**(7): p. 1270-8.

256. Sindi, S., et al., *A geometric approach for classification and comparison of structural variants.* Bioinformatics, 2009. **25**(12): p. i222-30.

257. Onishi-Seebacher, M. and J.O. Korbel, *Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond.* Bioessays, 2011. **33**(11): p. 840-50.

258. Jacobs, K.B., et al., *Detectable clonal mosaicism and its relationship to aging and cancer.* Nat Genet, 2012. **44**(6): p. 651-8.

259. Youssoufian, H. and R.E. Pyeritz, *Mechanisms and consequences of somatic mosaicism in humans.* Nat Rev Genet, 2002. **3**(10): p. 748-58.

260. Marco-Sola, S., et al., *The GEM mapper: fast, accurate and versatile alignment by filtration.* Nat Methods, 2012. **9**(12): p. 1185-8.

261. Quesada, V., et al., *Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia.* Nat Genet, 2012. **44**(1): p. 47-52.

G. ABBREVIATIONS

1000GP: 1000 Genome Project

aCGH: Array-comparative genomic hybridization

ASM: Assembly method

BAC: Bacterial artificial chromosome

BAF: B-allele frequency

CGH: Comparative genomic hybridization

CLL: Chronic lymphocytic leukaemia

CLL-GP: Chronic lymphocytic leukaemia genome project

CML: Chronic myeloid leukaemia

CNA: Copy-number alteration

CNV: Copy-number variant

DDR: DNA-damage response

DGV: Database of genomic variants

FISH: Fluorescence in-situ hybridization

FoSTeS: Fork-stalling template switching

HGDP: Human genome diversity panel

ICGC: International cancer genomic consortium

IGHV: Immunoglobulin genes

LD: Linkage disequilibrium

LOH: Loss of heterozygosity

LRR: Log2 ratio

MAF: Minor allele frequency

MEI: Mobile element insertion

MMBIR: microhomology-mediated break-induced replication

MMEJ: Microhomology-mediated end-joining

NAHR: Non-allelic homologous recombination

NHEJ: Non-homologous end-joining

NGS: Next-generation sequencing

OS: Overall survival

PR: Pair-read

PFS: Progression-free survival

RD: Read-depth

SD: Segmental duplication

SR: Split-read

SNA: Single-nucleotide alteration

SNP: Single-nucleotide polymorphism

SV: Structural variation

WES: Whole-exome sequencing

WGS: Whole-genome sequencing

YRI: Yoruba population

G. ANNEX

List of publications

Geòrgia Escaramís*, Cristian Tornador*, Laia Bassaganyas*, Raquel Rabionet, Jose M. C. Tubio, Alexander Martínez-Fundichely, Mario Cáceres, Marta Gut, Stephan Ossowski and Xavier Estivill
PeSV-Fisher: Identification of somatic and non-somatic structural variants using next-generation sequencing data. *PLoS ONE 2013 May 21;8(5).*

Bassaganyas, L., Beà S, Escaramís G, Tornador C, Salaverria I, Zapata L, Drechsel O, Ferreira PG, Rodriguez-Santiago B, Tubio JM, Navarro A, Martín-García D, López C, Martínez-Trillos A, López-Guillermo A, Gut M, Ossowski S, López-Otín C, Campo E, Estivill X.
Sporadic and reversible chromothripsis in chronic lymphocytic leukaemia revealed by longitudinal genomic analysis. *Leukemia. 2013 Apr 24.*

Bassaganyas, L, Riveira-Muñoz E, García-Aragonés M, González JR, Cáceres M, Armengol L, Estivill X.
Worldwide population distribution of the common LCE3C-LCE3B deletion associated with psoriasis and other autoimmune disorders *BMC Genomics. 2013 Apr 17;14(1):261*

Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, Ramsay AJ, Beà S, Pinyol M, Martínez-Trillos A, López-Guerra M, Colomer D, Navarro A, Baumann T, Aymerich M, Rozman M, Delgado J, Giné E, Hernández JM, González-Díaz M, Puente DA, Velasco G, Freije JM, Tubío JM, Royo R, Gelpí JL, Orozco M, Pisano DG, Zamora J, Vázquez M, Valencia A, Himmelbauer H, Bayés M, Heath S, Gut M, Gut I, Estivill X, López-Guillermo A, Puente XS, Campo E, López-Otín C.
Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet 2011 Dec 11:44(1):47-52*

Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N, Escaramis G, Jares P, Beà S, González-Díaz M, Bassaganyas L, Baumann T, Juan M, López-Guerra M, Colomer D, Tubío JM, López C, Navarro A, Tornador C, Aymerich M, Rozman M, Hernández JM, Puente DA, Freije JM, Velasco G, Gutiérrez-Fernández A, Costa D, Carrió A, Guijarro S, Enjuanes A, Hernández L, Yagüe J, Nicolás P, Romeo-Casabona CM, Himmelbauer H, Castillo E, Dohm JC, de Sanjosé S, Piris MA, de Alava E, San Miguel J, Royo R, Gelpí JL, Torrents D, Orozco M, Pisano DG, Valencia A, Guigó R, Bayés M, Heath S, Gut M, Klatt P, Marshall J, Raine K, Stebbings LA, Futreal PA, Stratton MR, Campbell PJ, Gut I, López-Guillermo A, Estivill X, Montserrat E, López-Otín C, Campo E.
Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature. 2011 Jun 5;475(7354):101-5.*

Tubio JM, Tojo M, Bassaganyas L, Escaramis G, Sharakhov IV, Sharakhova MV, Tornador C, Unger MF, Naveira H, Costas J, Besansky NJ.
Evolutionary dynamics of the Ty3/gypsy LTR retrotransposons in the genome of Anopheles gambiae. *PLoS One. 2011 Jan 24;6(1)*

Communications to scientific meetings

Oral communications

European Human Genetics Conference 2011. May 28-31 Amsterdam (Netherlands)
Oral presentation: *Landscape of somatic structural alterations in chronic lymphocytic leukemia (CLL) detected by whole-genome sequencing.* Bassaganyas L*, Tubío JMC*, Escaramis G, Tornador C, Bea S, Puente XS, Gonzalez-Knowles D, Guigó R, Gut I, Lopez- Otín C, Campo E, Estivill X, on behalf of the CLL-ICGC

ENGAGE Consortium Meeting 2009, Parc de Recerca Biomèdica de Barcelona, January 14-16, 2009, Barcelona
Oral presentation: *Profiles of structural variation between world human populations.* Bassaganyas L, García-Aragonés M, Montfort M, Gonzalez JR, Cáceres M, Armengol L and Estivill X.

European Human Genetics Conference 2008. CCIB Barcelona May 31- June 3, 2008
Oral presentation: *Genomic Structural Variation Profiles of world human populations.* Bassaganyas L, García-Aragones M, Montfort M, Escaramís G, Cáceres M, Armengol L and Estivill X.

Poster presentations

CSHL 2011. Meeting on the Biology of Genomes. Cold Spring Harbor Laboratories, New York (USA) 2011
Poster presentation: *Landscape of somatic structural alterations in chronic lymphocytic leukemia detected by whole-genome sequencing.* Bassaganyas L, Tubio, JMC, Escaramís G, Tornador C, Bea S, Puente XS, Gonzalez-Knowles D, Guiguó R, Gut I, Lopez-Otin C, Campo E and Estivill X. (presented by Estivill, X)

CSHL 2011. Meeting on the Biology of Genomes. Cold Spring Harbor Laboratories, New York (USA) 2011
Poster presentation: *PeSV-Fisher: a pipeline for somatic structural variant identification from high throughput sequencing data.* Escaramís G, Tornador C, Bassaganyas L, Tubio, JMC, Rabionet R and Estivill X. (presented by Rabionet, R)

European Human Genetics Conference 2010. Gothenburg, Sweden June 12-15, 2010
Poster presentation: *Geographic distribution of the LCE3C-LCE3B deletion, a susceptibility factor for psoriasis, across different human ethnic groups.* Bassaganyas L, Riveira-Muñoz E, García-Aragones M, Gonzalez JR, Cáceres M, Armengol L and Estivill X.

American Society of Human Genetics Conference 2008. Philadelphia, Pennsylvania (USA) November 11-15, 2008
Poster presentation: *Profiles of structural variation between ethnic groups*. Bassaganyas L, García-Aragones M, Montfort M, Escaramís G, Cáceres M, Armengol L and Estivill X. (presented by Estivill, X)

CSHL 2008. Meeting on the Biology of Genomes. Cold Spring Harbor Laboratories, New York (USA) May 6-10, 2008
Poster presentation: *Spectrum of Genomic Structural Variation in Human Populations*. Bassaganyas L, García-Aragones M, Montfort M, Escaramís G, Cáceres M, Armengol L and Estivill X. (presented by Estivill, X)