UNIVERSITAT DE BARCELONA

Identification and characterization of non-coding genomic variations associated to cancer diseases

Santiago González Rosado

Programa de doctorat Biomedicina EEES H0101

Facultad de Biologia, Universtitat de Barcelona

Identification and characterization of non-coding genomic variations associated to cancer diseases

Memòria presentada per Santiago González Rosado per optar al grau de doctor per la Universitat de Barcelona

UNIVERSITAT DE BARCELONA

Doctorand

Santiago González Rosado

Director

David Torrents Arenales

Tutor

Josep Lluís Gelpí Buchaca

Tesi realitzada al Barcelona Supercomputing Center

Acknowledgements

Durante el desarrollo de una tesis uno tiene muchas veces el placer, algunas pocas la desgracia, de tener que interactuar con numerosas personas. Todas ellas aportan algo al resultado final de la misma, ya sea por su aportación científica, por su apoyo a mantener el estado anímico o simplemente porque el que se cruzaran en tu camino condicionó el resultado final. Conocedor de que han sido muchas las personas, más de las que recordaría si intentara nombrarlas a todas, me gustaría centrarme más en lo que todas ellas han aportado que en una fría lista con sus nombres.

Empezaré haciendo una excepción y nombrando a David Torrents, básicamente porque el grupo al que pertenece empieza y acaba en él, el de director y jefe durante toda mi tesis. Primero de todo quisiera agradecer su confianza al aceptarme como doctorando suyo en el Barcelona Supercomputing Center. En plena crisis financiera, con el gobierno tanto autonómico como nacional sometiendo a la investigación pública a unos terribles recortes, confiar en un simple estudiante de biología que llega a la puerta de tu despacho puede parecer todo un acto de valentía. Si a lo anteriormente mencionado añadimos un currículum universitario nada destacable y la imposibilidad de solicitar becas, la valentía de depositar la confianza en dicha persona empieza a tornarse peligrosamente en temeridad. Pero si algo ha caracterizado a David durante los años que he cohabitado en el grupo de Computational Genomics con él ha sido su patológico optimismo. Resulta redundante destacar que gran parte de lo aquí escrito es fruto directo de su persona.

Haré una última excepción con Bàrbara Monsterrat, la postdoc a la que he respetado y obedecido como segunda jefa. De ella aprendí que había personas que vivían la investigación pública con absoluta devoción, y que lamentablemente yo no era una de ellas. Ella me guio en mis primeros pasos enseñándome a ser paciente y perseverante.

A todo el grupo de técnicos, doctorandos y postdocs del departamento de Life Science en general y muy particularmente a todos los que han formado a lo largo de estos años el grupo de Computational Genomics. Muchos de ellos han tenido aportaciones científicas de gran relevancia en esta tesis, y los que no ha sido el caso han ayudado a mantener un excelente clima de trabajo. Todos ellos han sabido ser algo más que simples conocidos de trabajo y han hecho que venir al BSC sea un placer que ofenda llamarse "venir a trabajar".

A todos los científicos externos que han ayudado a que esta tesis brillara, como son las colaboraciones con diferentes universidades y centros (UB, IRB, EMBL, Univerisadad de Kiel…). Capítulo especial merece el Hospital Clínic, concretamente el grupo de Oncomorfologia funcional humana i experimental, y la enorme confianza que siempre depositaron en mi persona, algo que a lo largo de estos años he podido constatar como extremadamente excepcional.

A cualquiera que haya sentido directa o indirectamente que han sido responsabilidad mía o que hayan tenido que obedecer alguno de mis mandatos. Pese a considerarme un blando mandando, estoy seguro que en algún momento tuvieron que armarse de paciencia. A todos ellos les deseo mucha suerte y espero que guarden un grato recuerdo.

Finalmente me gustaría agradecerle a toda mi familia su apoyo, desde el que ha tenido que sufrirme poco al que ha tenido que aguantarme en los momentos más difíciles. Muchas veces infravaloramos lo importante que es el tener un buen estado anímico, mucho más importante que la mejor de las ideas para tu estudio. Son todas esas personas más cercanas, aquellas que realmente merecen ser llamadas familia, tengan lazos de sangre o no, las que ofrecen el apoyo y la guía necesaria. Sin ellos, no solo estoy totalmente convencido de que jamás hubiera llegado a depositar esta tesis, sino que posiblemente no hubiera ni finalizado la totalidad de mis estudios que actualmente adornan mi CV.

Muchas gracias a todos.

Table of contents

Prologue

This thesis represents my research trajectory at the Barcelona Supercomputing Center, as part of the computational genomics group. The computational genomics group main goal is the analysis of biological data to understand the genetic and molecular causes and consequences of the most frequent human diseases.

In particular I developed my studies in the analysis of cancer data. I have combined the development of bioinformatic tools to analyze genome information with their application on cancer genome data in order to answer specific questions regarding the genomic basis of the disease. Therefore, the present thesis focuses on the biological aspects in the study of cancer patients, the capability to annotated genomic regions using different sources of data and how all this information can be integrated to help us to understand the basis of the development and progression of the disease.

The study of cancer genomes has grown dramatically with the production of thousands of sequenced samples from thousands of patients. In parallel, many bioinformatic applications have been developed to analyze the different sources of data: whole genome sequencing, exome sequencing, RNAseq, SNP arrays, epigenomic data, and others. Several databases and web portals are also publicly available providing all types of information regarding these samples, from raw data to preliminary results. Due to the vast amount of programs, studies, consortiums and available genomic data the introduction will focus on general technical and strategical aspects of the field, using, as examples those, those activities that are related to this thesis.

Finally, I would like to apologize to all the people and studies that, due to extension constraints, are not cited in the thesis, despite their relevant contribution to the field.

Genetic disorders

The identification of the genetic and molecular basis of disease has been one of the central interests of biology and biomedicine. Uncovering the modifications in the genome associated to specific pathological phenotypes allows the identification of the molecular processes behind each disease. From the generation of specific gene panels within the clinics, to the design of precise drugs targeting specific proteins related to the pathology, this research activity is fundamental to understand the mechanisms of the diseases and to develop better and more precise diagnosis and therapeutic protocols.

Nowadays, the explosion of sequencing technologies has made the analysis of genomic sequences cheap and accessible, expanding the possibilities of finding disease markers in the genome. The availability of complete genomic sequences for a large number of patients complements the traditional genomic analysis of diseases in two major ways: (1) by giving the possibility of generating richer and more precise profiles of polymorphic variation (haplotypes) within the population, which, in turn, increases the statistical power to find disease associated risk variants. It impacts in the study of DNA modifications that are heritable and affect multigenic diseases. They usually confer a given risk or susceptibility of developing the disease and are present in all the cells of the organism. These are commonly named as germline variations or polymorphisms; and (2) by allowing the search and direct analysis of disease mutations through the gathering of significant amount of genomic data from patients. This can be applied to the analysis of heritable mutations giving rise, mostly, to monogenic or rare diseases, and to the analysis of somatic

variants, which occur during the life span of the individual and are involved in several pathologies, including cancer. An important fraction of this thesis is centred in the analysis of somatic mutations (see below) and their potential implication in tumor development and progression.

Somatic variants

Acquired during the lifetime, somatic variants appear de novo in some cells of the organism. Most of them are expected to be harmless but, in a few cases, those changes in the genome can give rise to genetic disorders. If one single variant is enough to trigger the disease, it is classified as a monogenic disorder (Weatherall 2001; Erez and DeBerardinis 2015). When several somatic variants act together altering different cell processes, it is considered that they are involved in complex genetic diseases. Cancer is a challenging example of this type of diseases due to its complexity in the number of somatic variants involved and the different biological processes affected as shown in figure 1(Kan et al. 2010).

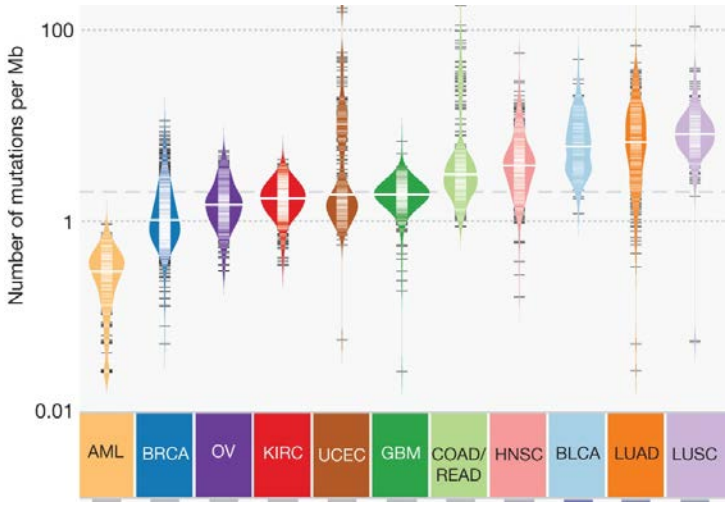


Figure 1. Distribution of mutation frequencies across 12 cancer types. Dashed grey and solid white lines denote average across cancer types and median for each type, respectively (Kandoth et al. 2013).

The somatic mutations arise from different endogenous and exogenous factors. As part of the endogenous causes the replication errors are common in all the different cells under division and in some cases these errors are not repaired or it is done incorrectly. Other inner causes are the DNA damage due to reactive oxygen, malfunction of enzymes involved in DNA repair, retrotransposons, other DNA binding proteins, and many more. The list of external factors is large, including the most common potential mutagens, such as tobacco, UV light and radiation. Clear examples can be found in the substitutions of C>T and C>G produces by over-activity of members of the APOBEC family (Alexandrov et al. 2013). The somatic mutagenesis is the fundamental cornerstone of the molecular basis of several disorders, such as cancer (Friedberg 2003).

The role of sequencing technologies within the field

Sequencing technologies have been essential to help us understanding the human genome and to uncover the genetic variability within and among individuals, as well as its role in human disease (Escaramis et al. 2015). The evolution of DNA sequencing covers a wide range of possibilities and technologies. Each type of sequencing technology has involved a particular range of use in a particular moment on research and has also entailed specific limitations.

Sanger sequencing technologies have contributed to biomedicine for more than 20 years, and still do, by initially introducing molecular genetics techniques into the research lines of nearly every bio-research group (Sanger et al. 1977). For example, thousands of cDNAs and millions of Expressed Sequenced Tags (ESTs) have been sequenced using Sanger technology, which have been essential to build the basis for almost all what we know about the molecular biology of diseases. Mostly used for the sequencing of amplified DNA targeting relatively small regions, Sanger sequencing was also later used for deciphering complete bacterial and eukaryotic genomes. In 2001 the first draft of the human genome was finished by the public consortium, setting up the basis of the new era of biomedical genomics (International Human Genome Sequencing 2004). This constituted a great effort involving more than 3 US$ billion, more than 10 years, and a large number of countries and research groups. Despite late improvements in price and speed, the price and the processing power of the Sanger technology did not allow a massive

sequencing of different individuals of a given population or phenotypic group, which posterior sequencing technologies could.

Next generation sequencing

Sanger sequencing technology was displaced by novel techniques that bring sequencing closer to most of the researches groups, and enabling large-scale sequencing. Pyrosequencing method was first released in 1998 (Ronaghi et al. 1998) and the first commercial product was the 454 Life Science in 2005. The new method had several advantages. The main one was the use of DNA libraries allowing the automatization because it no longer depends on specific primers. The second is the capability to directly detect the read strand without electrophoresis, eliminating the human intervention and permitting the parallelization. 454 pyrosequencing triggered the next generation sequencing (NGS). One year later Solexa platform was commercialized and, in 2007, the SOLID system by Applied Biosystems (Valouev et al. 2008). Nowadays Illumina Hiseq technology can produce more than 3 Billion reads in less than 3 days reducing the cost of analyze a human genome in less than $1000. All NGS platforms have common traits: highly automated and parallelized protocols, short read length (from tens to few hundred nucleotides) and, most importantly, a reduced cost per sample run. The fast evolution of sequencing technologies last 15 years has produced a dramatically growth of sequenced data, as can be observed in figure 2 with the number of new eukaryotic organisms sequenced.

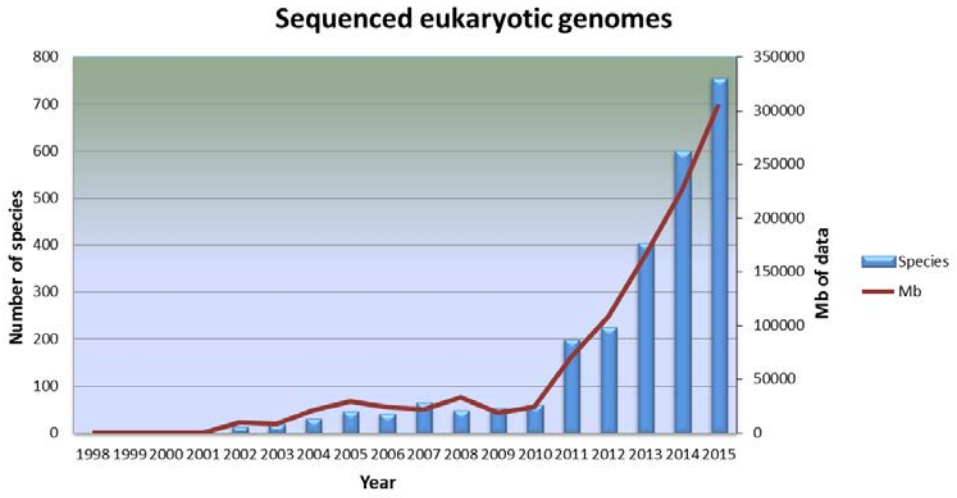


Figure 2. Evolution of new eukaryotic sequenced species since 1998 according to NCBI data. Blue bars represent the total number of new organisms. Red line corresponds to the total amount of storage information regarding the sequenced genomes.

Major applications of Next Generation Sequencing technologies

The low costs combined with the sequencing speed revolutionized genomics allowing medium and small laboratories to include even large-scale sequencing within their projects and research plans. The immediate profit was the access to the genomes of many individuals, phenotypes and conditions, considering large targeted fragments (single or multigenic panels), coding regions (Whole Exome Sequencing, WES; (Ng et al. 2010; Kiezun et al. 2012)) or the entire genome (Whole Genome Sequencing, WGS). Nowadays technologies based on NGS are predominant in genomic scientific studies as can be observed figure 3. With the availability of genome sequences we can directly search for risk or causal disease mutations using massive

computational approaches to later link them with their functional impact (Ciriello et al. 2013).

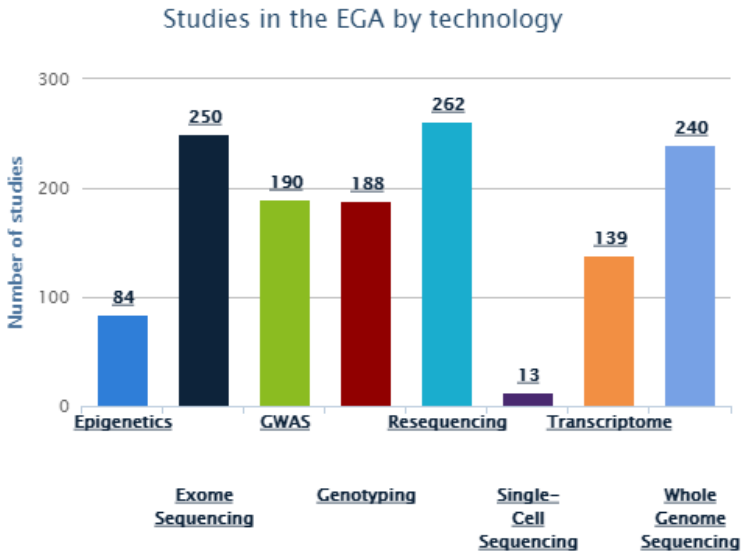


Figure 3. Abundance of different studies group by technology in EGA database. Epigenetics, exome sequencing, resequencing, single-cell sequencing, transcriptome and whole genome sequencing are all techniques partially or complete derived from NGS (Lappalainen et al. 2015).

In addition to DNA sequence, NGS technologies have also given access to the entire transcriptome through RNAseq, also known as Whole Transcriptome Shotgun Sequencing (WTSS), which is probably the best example of novel techniques that use NGS not limited to genomic DNA (Morin et al. 2008). The results are the sequencing of the entire collection of mRNAs of a given sample, which can be then analyzed quantitatively, to detect relative abundances of particular mRNAs isoforms, and qualitatively, to identify new fusion/chimeric genes and splicing aberrations. RNAseq is currently also included in the study of

genetic diseases in combination with the analysis of the genome of the same sample. At the end, this allows to correlate particular changes in the genome with changes in the expression of particular genes (Wang et al. 2009; Ren et al. 2012; Teles Alves et al. 2015).

NGS technologies have also permitted the massive sequencing of short fragments of DNA, which has been key to setup complex experiments involving the isolation and sequencing of particular regions of the genomes that interact internally or with other molecules. For example, methods based on chromatin immunoprecipitation sequencing (ChIP-seq) have largely contributed to the annotation of non-coding part of the genome (Johnson et al. 2007). Through a first step of purification of the protein of interest, which is physically interacting with the DNA (chromatin immunoprecipitation), followed by a massive parallel sequencing of all these DNA fragments. This technique has already provided extremely useful information, for example, thousands of bindings sites for a given transcription factor protein, DNA polymerase binding sites, histones positioning, and others (Gerstein et al. 2012; Wang et al. 2012).

Another relevant example of NGS application is the chromosomal conformation capture (3C) and their variants 4C, 5C and Hi-C (Dekker et al. 2002). In exactly the same direction that previous methods it is based on, they sequence cross-linked DNA regions that are proximal in the space. Digestion of these compounds and forward sequencing allows the detection of proximal DNA regions and the understanding of the interactions and the spatial distribution of the genomic DNA within the nucleus.

Analysis of NGS data

The recent and growing amount of sequencing data generated around diseases using NGS is revolutionizing our understanding of genetic disorders permitting the detection of driver (causal) variants in the genome and, at the same time, the study of their potential impact in the pathology, by combining it with the functional annotation of the genome. The analysis of all this data has been a challenge at different levels, conceptually, but also from the point of view of the methods and technologies needed to process it. Currently, our capability to generate sequencing data is growing faster than our power to analyze and process it. The scientific community has to overcome enormous computational challenges in order to store, manage and analyze all this information (Eisenstein 2015; Marx 2015). The current thesis describes our contribution in solving these limitations by providing novel bioinformatic solutions that connect the generated information with our understanding about the genetic causes and consequences in disease.

All NGS approaches have in common the massive parallel sequencing. As a result, the user obtains millions of short reads containing the targeted information. These sequence reads cannot be directly interpreted and it is necessary to process and analyze them with complex bioinformatic protocols and applications. In contrast to Sanger sequencing technology, NGS generates such amount of small reads that the bioinformatic community had to invent new protocols or adapt existing ones for the analysis of sequence data. For example, the most common initial step for the analysis of NGS data in nearly all its applications is to align all the reads against a reference genome,

which allows the user to study them grouped by regions of interest. This has involved a redesign of the strategy and alignment algorithms, for example BWA (Li and Durbin 2009) and GEM (Marco-Sola et al. 2012).

The mapping step is extremely sensitive to the uniqueness of the target sequence within the genome and to the level of sequence identity. To be able to align NGS reads in a reasonable timeframe, these methods force a highly concordance between the sequenced data and the reference genome (Li and Durbin 2009). While this does not affect many of the applications, it does interfere with the analysis of mutations, as, in these cases, the reads of interest, i.e. those containing changes, are expected to have lower mapping scores. The problem becomes much more complex if we include the thousands of repetitive or low complexity region of the genome. To aid during the mapping process most of the NGS techniques now incorporate what is known as paired-end reads, which consists in the generation of pairs of sequencing reads whose distance in the genome is known (Fullwood et al. 2009). Although, this has not completely solved the problem of aligning complex or mutated regions, it has reduced its impact considerably.

Cracking the genetic code

The next NGS allowed hundreds of different studies to analyze a large number of patients in order to identify the variation associated to a large number of genetic diseases (van Dijk et al. 2014) (Mardis 2008). Promptly the scientific community needed to transform all these information into real knowledge to interpret the genomic code and understand the functional impact of each of the genomic changes identified as associated to disease. This is an essential step, not only to link a specific disease to a given single or group of variants, but also to be able to understand the biological impact of the different variants in the development and progression of the diseases.

Genome annotation

Right after a new variant is associated with a disease, the immediate question to answer is how it is affecting the cell behavior and how relevant might be for the development of the molecular mechanisms suspected or known to be driving the disease. The impact that a genetic variant can have within cell functionality is wide: from the direct modification of a gene producing a malfunction of the encoded RNA or protein, to changes in gene expression regulatory regions, in the overall stability of the chromatin and others. Different large-scale initiatives have been launched to complete the annotation of functional elements within the human genome in order to, among others, have better and more accurate possibilities of correlation between genetic modification and functional impact.

The ENCyclopedia Of Dna Elements (ENCODE) was the first international project with the enormous challenge of elaborate a comprehensive catalog of the structural and functional components encoded in the human genome (Consortium 2004). The catalog included protein-coding genes, non-protein-coding genes, transcriptional regulatory elements and sequences that mediate chromosome structure and dynamics. To elaborate this catalog the ENCODE consortium integrated thousands of different analysis that could have not be done without the NGS technologies: ChIPseq, 3C, DNaseI, DNA sequencing, RNA sequencing, Methylations and others. Figure 4 represents a briefing about some of the most relevant techniques applied to the genome. In its first approach they planned to comprehensibly annotate 1% of the human genome, but that first objective was quickly overcome by the improvement of the analysis techniques and finally most of their results were extended to the annotation of the whole genome. Several international subconsortia and subprojects have taken part within ENCODE annotation initiative. For example, VEGA (Ashurst et al. 2005), GENCODE (Harrow et al. 2006) and EGASP (Guigo et al. 2006). The last, but still uncompleted, frozen set of results were published in 30 different publications in 2012 (http://www.nature.com/encode/). Currently the collaboration between UCSC and ENCODE gives access to 288 human cell and tissue types, 32 different assays and mapping information about more than 300 DNA binding sites (Rosenbloom et al. 2013).

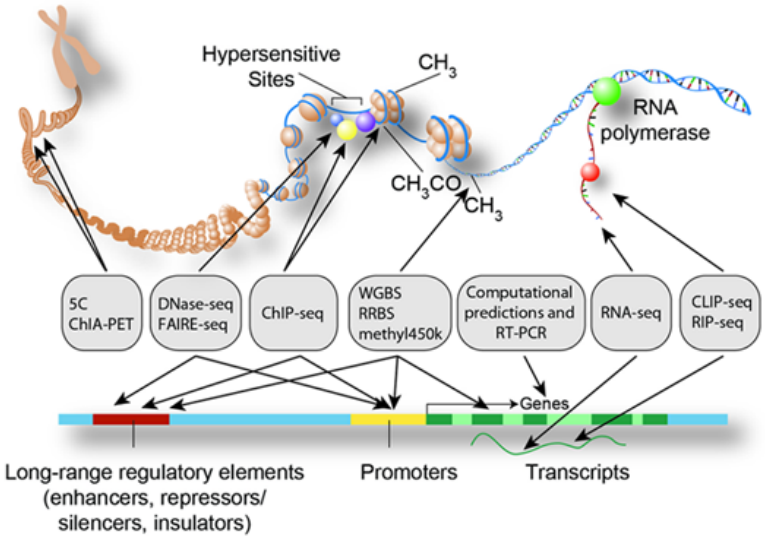


Figure 4. Representation of different experimental techniques applied during ENCODE analysis in order to provide functional information about the different genomic regions.

Despite the new technologies have allowed us to obtain and interpret all this annotation data, underlying methodologies and statistical frames to transform all sequencing data into useful knowledge in relation to a specific application (RNAseq, Chip-seq, for example) are still under development and improvements, and often still generate contradictory or inconsistent results. This is why part of this information should be used as suggestive and supporting evidence only.

Finding and classifying functional elements in the genome

Several functional elements can be found in our genome and they are responsible of gene regulation, signaling, DNA stabilization, etc. Assuming that coding exons are the part of the genome with the assignment to codify proteins, most of our DNA is involved in other tasks. Figure 5 represents a distribution of what nowadays is known about the distribution of functional elements in human genomes.

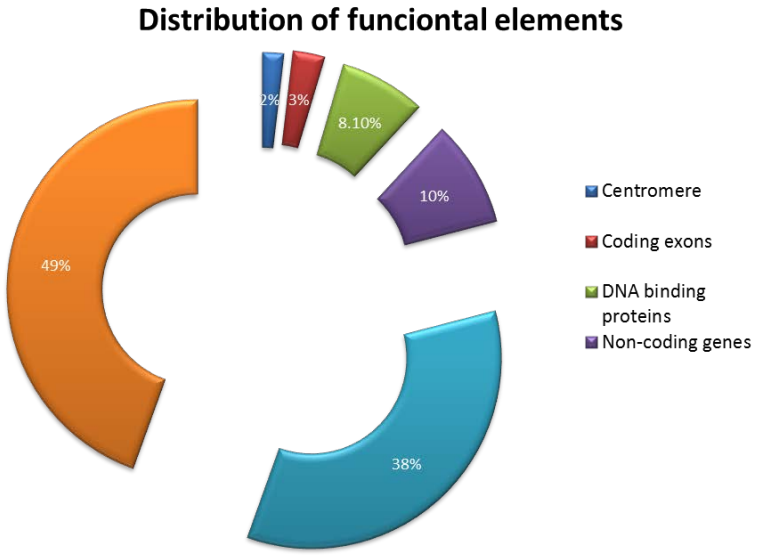


Figure 5. Distribution of different functional elements across human genome according to ENCODE data. Notice that percentages can exceed 100% because some functional elements can overlap between them.

According to GENCODE, initially formed as part of the pilot phase of the ENCODE project to identify and map all protein-coding genes, the annotation of the coding genes is closed to be finished with a bit less than 20.000 genes found in human. However, the RNA high throughput techniques, supported by the new sequencing platforms,

have increased the number of known non-coding genes. Different categories have been created to classify the non-coding genes, for example, into long non-coding RNA or small non-coding RNA. In total, more than 30.000 non-coding genes have been described. Although the biological function of most of them is still unknown, there are several examples where they are involved in gene regulation, RNA inactivation, signaling, and RNA post-processing (Huttenhofer et al. 2005) (Ziats and Rennert 2013; Palazzo and Lee 2015).

Among the functional elements that cannot be transcribed into RNA, transposable elements are, by far, the most abundant. The definition of the category is diffuse and includes all small pieces of DNA that copy and translocate within the genome. In fact, that genomic movement can be observed from an evolutionary point of view, not only between different species, but also within individuals from the same population. These repeats can be particularly important, as they have been associated to several diseases (Xiao-Jie et al. 2015) and other functionality of the genome, such as with retroviruses, DNA stability and gene regulation (Callinan and Batzer 2006) and gene duplication during evolution enabling in some cases speciation processes (Wicker et al. 2007) (Kazazian and Moran 1998) (Kim et al. 1998)..These moving regions can include other functional elements, even genes, dynamizing over time the combination of exons, complete genes and regulatory elements.

Regulatory regions constitute more than 8% of the genome (Consortium 2012). Their annotation is a huge challenge because their function is dependent of cell and tissue type, as well as of developmental stage. From the traditional view where a gene needs a immediately upstream region, named proximal promoter, to start the

transcription, studies have later demonstrate a much more complex genomic architecture that regulates the expression of our genes. The gene regulation became possible thanks to the interaction between the proximal upstream region to the gene (promoter) with one or several distal regions (enhancers) in collaboration with several trans elements known as transcription factors. Experimental approximations, such as 3C, 5C and Hi-C allow us to widen our understanding on how the genome adopts an structure that favors these interactions, even between regions which are far away in terms of DNA sequence but closely in the 3D structure of the nucleus (Belton et al. 2012).

A large number of bioinformatic applications have been developed to identify and classify gene regulatory regions (Hallikas et al. 2006) (Sun et al. 2009) (Abeel et al. 2009) (Dubchak et al. 2013) (Palin et al. 2006). Due to the intrinsic difficulty in detecting these heterogeneous regions, not a single method or approach seems to be powerful enough to capture and characterize all the different regulatory regions in the human genome. In the last years, novel methods have focused in the combination of different sources of data in order to obtain good balances between sensitivity and specificity (Fu et al. 2014) (Seumois et al. 2014). Currently, a large fraction of the accompanying genomic data can be used to infer regulatory potential: histone modification marks, ChIPseq of TFBSs, DNase I accessibility, evolutionary conservation, sequence motifs, relative distances to known genes and the 3D organization of the DNA.

All these analyses have allowed to observe the high level of plasticity of these regions. Different studies expanded their analysis to different cell lines of specific tissues, development stages and pathological

states. The results confirm how the accuracy in detecting regulatory regions depends on the selection of the proper cells and conditions. Several research consortiums are generating and offering a wide range of different information about the most relevant aspects of chromatin in a large number of cell lines that can be used to support the potential functionality of non-coding regions: BLUEPRINT (Abbott 2011), FANTOM (Carninci et al. 2005) and ENCODE (Consortium 2004).

Finally, the results of nearly all the genome annotation efforts done in the community can be accessed and easily interpreted through powerful web platforms that organize all these results according to their genomic position: UCSC genome browser (Kent et al. 2002), ENSEMBL (Hubbard et al. 2002). Figure 6 shows a screenshot of UCSC genome browser with several tracks associated with gene regulation.

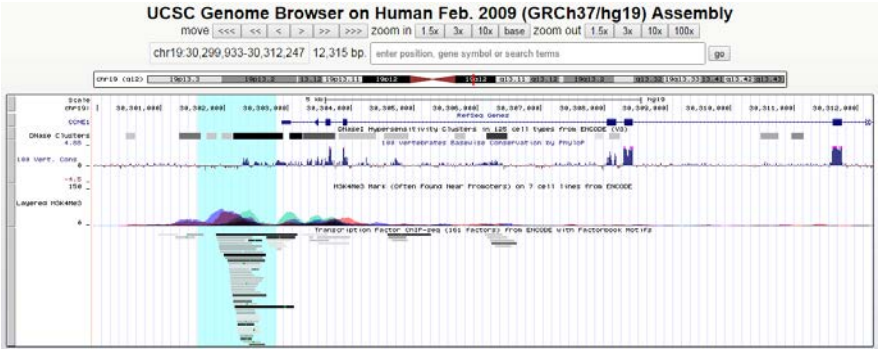


Figure 6. UCSC genome browser representation of promoter region of CCNE1 gene. Highlighted region correspond to the promoter region according to the most common marks that includes: DNase I accessibility, conservation of DNA across species, Histone marks associated with regulation and TFBSs ChIP-seq.

The cancer genome

Cancer can be considered as a model example of the use of NGS on genomes to uncover the mutational spectrum underlying the disease and, through the use of the genome annotation, infer its functional consequences. Its impact in the society, in combination with the complexity of its biology, has drawn important research efforts trying to unveil the specific biology underneath the different types of cancer processes. Figure 7 shows the number of independent studies according the principal diseases where cancer studies are clearly predominant.

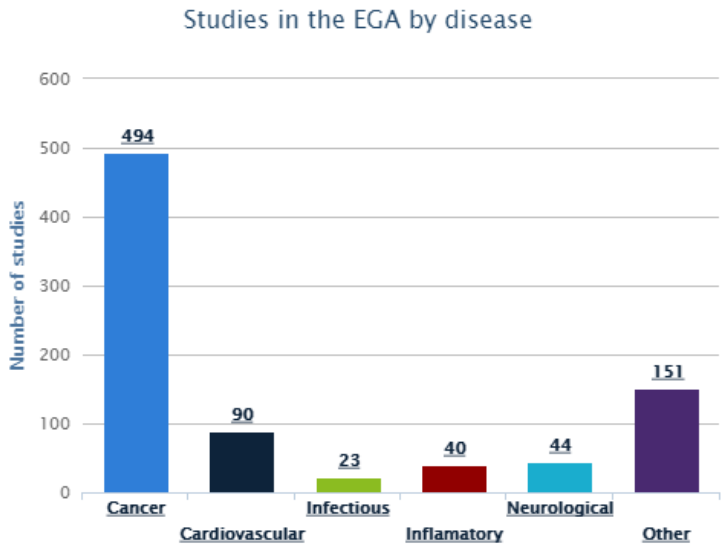


Figure 7. Distribution of studies by disease type on EGA database. (Lappalainen et al. 2015)

All tumor types are characterized by relatively unrestrained proliferation of cells that can invade beyond normal tissue boundaries and metastasize to distant organs (Stratton et al. 2009). These cells escape from both the normal cell behavior and the exogenous restraints of growth. Cancer has been described as an example of positive selective evolution in which a given number of cells acquire mutations that can confer an advantage, i.e. resistance to death and continuous proliferation.

The processes of somatic mutagenesis allow the tumor cell to gradually acquire a set of functional capabilities which are common in most, if not all, of the cancer types (Hanahan and Weinberg 2000). Firstly, the self-sufficiency in growth signals, or the capability to activate proliferation states without the regulation of external stimulus. Secondly, the insensitivity to antigrowth signals that maintain the quiescence and tissue homeostasis in normal cells. Thirdly, in most of the cases the tumor cell can evade the apoptosis programmed in their code. Fourthly, instead of autonomously regulated their replicative potential in tumors this limitations does not exists. Fifthly, the sustained angiogenesis, or the potential to constitute real functional tissues with the formation of new blood vessels. Lastly, the competence to perform tissue invasions and metastasis. Last few years different treatments and drugs have been developed targeting those exclusive capabilities of tumor cells, some examples are represented in figure 8.

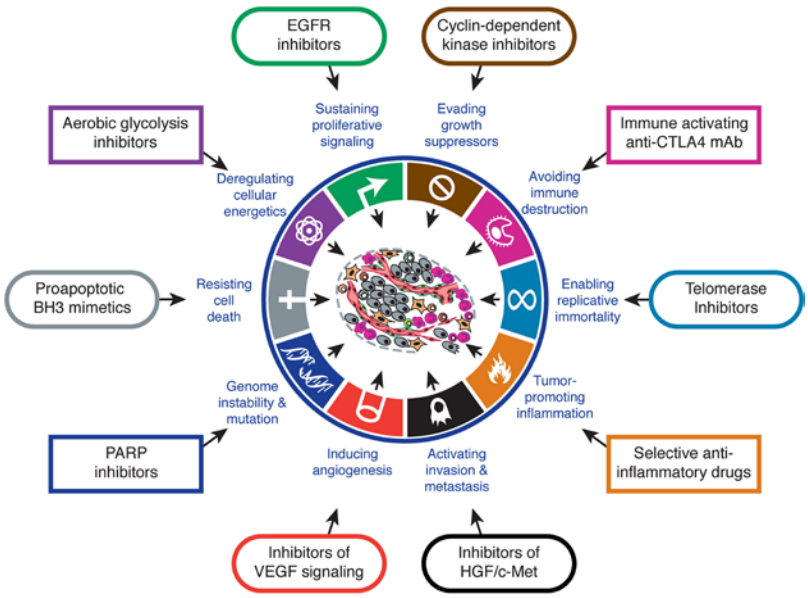


Figure 8. Therapeutic Targeting of the Hallmarks of Cancer. Drugs that interfere with each of the acquired capabilities necessary for tumor growth and progression have been developed and are in clinical trials or in some cases approved for clinical use in treating certain forms of human cancer (Hanahan and Weinberg 2011).

Because cancer is a category of probably hundreds of particular diseases with multifactorial genetic causes and it implies several cell mechanisms our comprehension strongly depends how much we can understand the biology of the cell. At this point, please note (below) a text written in 2000, which has become premonitory and at the same time can be reused nowadays:

> We anticipate otherwise: those researching the cancer problem will be practicing a dramatically different type of science than we have experienced over the past 25 years. Surely much of this change will be apparent at the technical level. But ultimately, the more fundamental change will be conceptual.
>
> (Hanahan and Weinberg 2000)

The overall complexity of tumorigenesis and its forms of progression makes this field of research a challenge that needs, not only the identification and characterization of all the mutations involved, but at the same time, a deeper understanding about the specific functionality of the different regions of the genome affected. It is in this sense, and currently applied to cancer research, where NGS technologies and the improvements in genome annotation allow a clear shift in basic research strategies towards "from genetics to function" approximations.

Previous to the wide accessibility to whole genome sequencing technologies, and still very active, a large number of studies have provided key genetic and molecular information about most commonly affected oncogenes and tumor suppressor genes, mainly by using the "from function to genetics" approach (Hanahan and Weinberg 2000). Complementing these essential molecular studies, large scale genomic analysis are providing to the entire community an unprecedented amount of candidate novel cancer genes, together with information about mechanism of structural genomic variation, often taking place within the tumor cell. However, not all the somatic mutations appearing in a particular cell that becomes immortal are involved in this process of transformation. Only a small percentage of them can be considered as tumor mutations, commonly named as driver mutations. All the other, the passenger mutations, are not directly associated with that selective growth advantage.

The amounts of cancer driver mutations, in combination with all the different potential paths that can lead to the disease, add a new layer of complexity in the study of cancer. Tumors originating in the same organ or tissue can vary substantially in their alterations while

similar patterns can be observed in tumors from different tissues (Ciriello et al. 2013). This intracancer and intercancer heterogeneity puts into question, and highlights the limitations, of the traditional approach of treating all tumors of the same tissue as equal. On the contrary, new diagnosis and therapeutic protocols should take into account the nature of the genomic alterations of each tumor in particular in order to make more precise and effective treatment protocols. This is actually the basis of personalized medicine, where patients will be treated according to specific genetic or molecular markers. Cancer is one of the first diseases that benefits of this molecular and genetic analysis of the patient. Ras mutation is a well-known biomarker that defines different populations within colorectal cancer to determine their treatment (Stintzing et al. 2015).

With this aim, a large international initiative was launched few years ago that included the compromise of most of developed countries of the world to sequence and analyze the genomic and molecular basis of several types of cancer. This consortium, The International Cancer Genome Consortium (ICGC; https://icgc.org), aimed at sequencing the genome of at least 500 patients of particular cancer types, together with generating accompanying functional data, such as gene expression, epigenetic marks, and others. The research environment and the strategies generated by this consortium have become standard in the field of cancer genomics. The general protocol within these analyses is summarized in Figure 9.

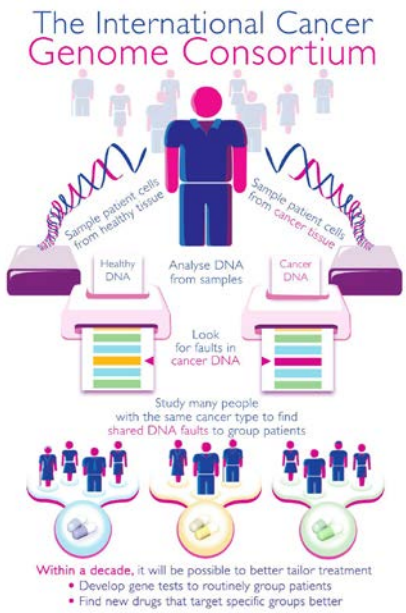


Figure 9. Summary of ICGC protocol to analyze cancer patients. All the involved countries sequence healthy and tumor cells from the same patient and tissue. The objective is to determine groups of patients sharing the same DNA faults to develop gene tests and drugs.

In order to favor the identification of the specific genetic and molecular basis of tumors, normal and tumor cell samples (ideally from the same tissue) are extracted from each patient and analyzed at the level of genome and transcriptome sequencing, and of specific chromatin states. All this data is then analyzed using computational approaches that combine, among others, (i) the identification of all the spectrum of somatic variations in the genome. It includes single nucleotide variants (SNVs) to structural variants (SV) that involve small or medium size insertion, deletions and inversions (commonly known as indels) and large chromosomal rearrangements, viral integration and other structural modifications of the genome. It is in this particular frame that the present thesis had its major contribution. Additionally, copy number variation (CNV) is also

explored within these tumor genomes; (ii) analysis of RNAseq data to explore tumor specific expression profiles; and (iii) analysis of DNA methylation and other chromatin modifications.

All these results are then interpreted and crossed with genome annotation to identify what are the genomic and transcriptomic modifications that are related to the development or progression of the tumor, i.e. which are driver events. The distinction between driver and passenger (non tumorigenic) events is still a challenge. Although a number of methods (Ng and Henikoff 2003; Carter et al. 2009; Reva et al. 2011; Gonzalez-Perez and Lopez-Bigas 2012) have been generated within the community to prioritize, from the list of all mutated genes found in a tumor, which are likely to have an impact in the biology of the tumor. Together with the challenge of finding the mutations within tumor genomes, the distinction between driver and passenger variation events remains unsolved, being the frequency of certain events or mutated genes the most reliable criteria to infer association with the tumor. In other words, if a gene or any other functional region is recurrently found to be mutated in tumor genomes, it is then taken as potentially driving somehow the tumor.

The final goal of this general strategy of analysis of tumor genomes is to translate all this knowledge into effective and specific clinical protocols for treatment.

Identification of somatic mutations in cancer research

An important challenge within cancer genomic studies is the identification of somatic mutations, to ultimately isolate the causal fraction that plays a role in the development or progression of the tumor. A large number of studies are based on incomplete analysis of genomic sequences: exome sequencing, point mutations, mutations affecting coding genes, etc. They have only uncovered the tip of the iceberg, leaving a large mutational space unexplored. A clear example can be observed in the largest international cancer genome consortium ICGC where the exome sequencing data represents one order of magnitude more than that coming from whole genome (Zhang et al. 2011). The easy access to exome sequencing compared with whole genome has limited most of our knowledge to the coding exons and, mostly to point mutations (Kandoth et al. 2013) (Ciriello et al. 2013) (Kan et al. 2010) (Alexandrov et al. 2013). Additionally, several methods to evaluate the biological impact of a somatic mutation affecting coding regions have also been developed (Gonzalez-Perez and Lopez-Bigas 2012) (Ng and Henikoff 2003) (Reva et al. 2011) (Carter et al. 2009). All this leaves an important fraction of causal mutations outside coding regions significantly less studied. Regulatory variation, as well as the variation associated to transposons, viruses, and with large chromosomal rearrangements is still largely unexplored due to the general limitations of methods to detect them, and only a limited number of examples exist (Puente et al. 2015) (Kulis et al. 2012) (Huang et al. 2013) (Akhtar-Zaidi et al. 2012) (Herz et al. 2014).

The process of variant calling requires both, complex algorithms and efficient computational protocols to deal with massive amounts of sequences, making it a "big data" challenge (Puckelwartz et al. 2014; Eisenstein 2015; Marx 2015). Most of the available software to identify somatic mutations emerges from the adaptation of the methods originally developed to detect germline variation (Sudmant et al. 2015). For the past years, we have experienced an explosion of different methods for the identification of somatic mutations by comparing normal and tumor genomes. Before this thesis, all existing methods for somatic variant calling were based on the inspection of the reads aligned to the reference genome, i.e. from a BAM format. Each of these methods is usually restricted to the detection certain types of somatic variation (Cibulskis et al. 2013) (Rausch et al. 2012b) (Chen et al. 2009) (Ye et al. 2009) (Wang et al. 2011) (McKenna et al. 2010). Some methods are designed to detect point mutations and small (of a few nucleotides) deletions or insertions, others are focused on small size indels (less than sequencing read size) and the least of them, on the detection of large structural variants (i.e. chromosomal rearrangements). Each of these tools has been usually developed in a different bioinformatics groups and, often, using different programming models and languages.

All this makes that a comprehensive analysis of tumor genomes require the development of complex computational pipelines, gluing together many of these methods and adding extra filters to minimize the rate of false positive calls. These pipelines, which require the intervention of deep computing expertise, are not distributed within the community, leaving most of the small and medium groups with access to the sequencing of tumor genomes, with no possibilities of analyzing properly the data that they generate. Figure 10 represents

the filtering pipeline used by ICGC that must be applied to the sequencing data previous to the different variant calling methods. Novel approaches are trying to remove all these technical barriers and at the same time improve our capability to detect the most complex variants. Different strategies have been developed in these direction, being the direct comparison of the sequenced reads and the de novo assembly two of the most promising ones (Rimmer et al. 2014).

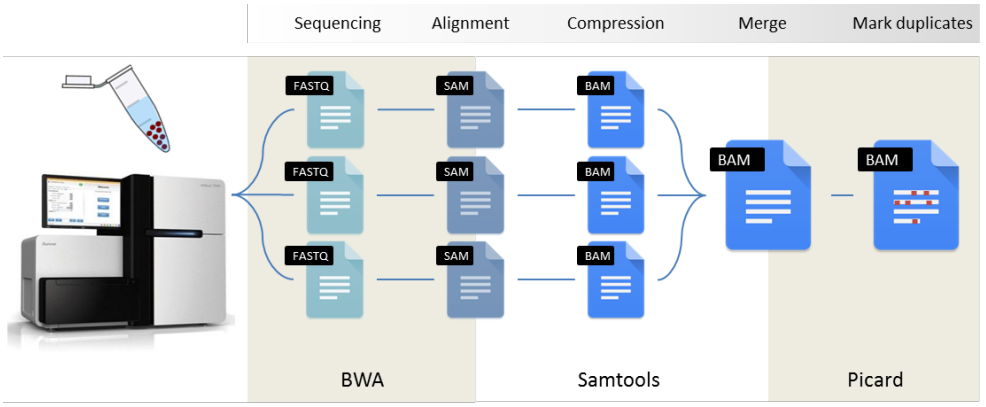


Figure 10. Filtering process applied to sequenced data. Top line represents the different steps, for each file their current format is given (FastQ, SAM and BAM), at the bottom the different programs used for produce each file. This step graph represents the minimal 5 step filtering process currently applied in ICGC-CLL studies previously to variant calling methods.

Large structural rearrangements

From all the different types of somatic sequence variation, those that constitute large chromosomal rearrangements are among the most challenging. The range of large structural variants (LSV) includes chromosomal translocations, but also copy number variants, mobile elements, insertions of non-human DNA, such as viruses, and other types.

The study of LSVs becomes essential in the study of the cancer genome. All the possible functional alterations that these rearrangements can cause are many. For example, (i) the breakage of functional elements such as genes. The disruption of *PTEN* in prostate cancer is an example (Baca et al. 2013); (ii) The complete deletion of large genomic regions that including functional elements, such as the deletion of part of the 13q chromosome arm, identified as recurrent in leukemia, which involves specific microRNA genes. (Liu et al. 1995) (Smonskey et al. 2012) (Klein et al. 2010); (iii) The modification of the genomic context, for example rearrangements that translocate regulatory regions close to other genes, resulting in the deregulation of the expression of specific genes (Affer et al. 2014); (iv) The generation of gene fusions are also the result of genomic translocations. This category includes the inactivation of gene transcription, the production of non-functional RNA that can interfere with the normal allele, or even the production of new genes as a combination of different functional domains that can be translated into a protein with a new and fatal functionality (Mitelman et al. 2007); finally, (v) large structural variants can also produce large reorganizations of the chromosomes affecting the stability of the genome. When these LSVs occur several times within one single

catastrophic event we face, what we call, mainly chromothripsis or chromoplexy. (Korbel and Campbell 2013) (Rode et al. 2015) (Baca et al. 2013) (Shen 2013) (Rausch et al. 2012a). Differences between those processes are still vague and their causes are not completely well understood.

In general, an aberrant event is considered chromothripsis when multiple (sometimes hundreds) of rearrangements occur within a restricted portion of the genome, involving one or two chromosome. Figure 11 represent a chromothripsis event in a paediatric medulloblastoma patient. In contrast, chromoplexy involves fewer LSVs and multiple chromosomes. Other marks such as the level of DNA gain and loss or the mutation of certain genes involved in DNA stability have been proposed as intermediates of these large reorganizations, but their role is not complete clear.

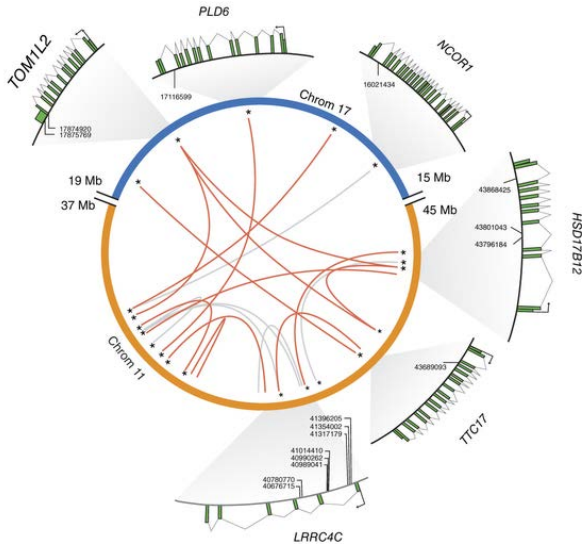


Figure 11. Chromothripsis in paediatric medulloblastoma. Red lines represents the different translocations between chromosome 17 and chromosome 11, when those translocations affect a certain gene, it is represented and the breakpoint highlighted.

Final considerations

Taken altogether, the recent advances in the technologies related to the production of biological data, primarily of DNA and RNA sequences, is complementing the research in biomedicine in an unprecedented way. The possibility of generating sequences from thousands of patients allows the addition of novel approaches into research, expanding the possibilities of finding the genetic and molecular basis of disease. The possibility of going from the genetics to the function using all the generated sequence data and the annotation of the genome, is quickly contributing to widen our understanding of disease in general, and of cancer in particular.

But these advances entail important technical and conceptual challenges to a point that the capacities for the analysis of the genome and the accuracy of the annotation of functional elements in the genome become the bottleneck of this process and are not accessible for most of the biomedical groups. This thesis focuses on overcoming part of these limitations by contributing in three major aspects: (i) the development of novel methods for the identification of somatic mutations from tumor genomes; (ii) the generation of tools for the annotation of gene regulatory regions; and (iii) the application of these tools in order to answer questions related to the biology of the cancer genome.

Objectives

I. To contribute to overcoming the limitations of the analysis of big data in genomics through the development of novel strategies and bioinformatics solutions for the massive analysis of whole genome sequences and the identification of somatic mutations in tumors.

II. To identify and classify the somatic variation landscape in tumor genomes, focusing on the characterization of complex chromosomal rearrangements, as to their underlying mechanisms and potential functional impact.

III. To contribute to the annotation of regulatory regions in genomes through the development of more efficient bioinformatics tools.

IV. To combine the developed tools in order to identify and characterize the somatic variation of tumors with a potential impact in the regulation of gene expression.

List of publications and scientific contributions

Santi has fulfilled his PhD contributing with up to four published studies, and a fifth one that is about to be sent for publication to Nature Communications. In general, the contribution of Santi to all his publications has covered, both the technical and the biological aspects of the studies. In addition, he has also coordinated other peoples work, particularly in the publications of ReLA and SMUFIN. In all the publications, Santi has followed and has contributed to answer the underlying biological questions, either directly or through discussions. The double background of Santi (biological and computational) has given Santi a broad vision of all the technical and biomedical points of each of the studies, allowing him to contribute, one way or another, in nearly all the aspects covered.

First author publications

Title: ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites.

Authors: Santi González*, Bàrbara Montserrat-Sentís*, Friman Sánchez, Montserrat Puiggròs, Enrique Blanco, Alex Ramirez, David Torrents.

Journal: Bioinformatics

Impact factor: 5.323

Citations: 7

Contribution:

This is the first publication, in which Santi took part in his first years in the group. This study was initially pushed by Barbara Montserrat, a postdoc in the group. Even though, Santi started with a secondary role, he soon took the lead of all the work, mostly when Barbara left the group. Santi was responsible of the generation of the ReLA code in collaboration with the department of computer science at the BSC. Santi generated all the examples and the biological information behind this study, as well as coordinated the generation of the web server associated to this publication. He also played a crucial role in the overall design of the study and in the generation of the manuscript.

Title: Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads.

Authors: Valentí Moncunill*, Santi Gonzalez*, Sílvia Beà, Lise O Andrieux, Itziar Salaverria, Cristina Royo, Laura Martinez, Montserrat Puiggròs, Maia Segura-Wang, Adrian M Stütz, Alba Navarro, Romina Royo, Josep L Gelpí, Ivo G Gut, Carlos López-Otín, Modesto Orozco, Jan O Korbel, Elias Campo, Xose S Puente, David Torrents.

Journal: Nature biotechnology

Impact factor: 41.514

Citations: 5

Contribution:

Santi co-authored this publication with Valentí Montcunill, the software engineer responsible of writing the code of the SMUFIN software. Santi's contribution to this work was crucial, as he coordinated, not only the details of the algorithm, but also all that had to do with the application of the program, i.e. the comparison with other methods and the experimental validation of the results obtained using in-silico and real tumor genomes. This task involved the collaborations with the Hospital Clinic and the EMBL, which were also coordinated at daily basis by Santi. Its role in this was not restricted to particular specific tasks, but also involved the coordination of other members of the group that performed particular subtasks. Finally, the contribution of Santi also extended to the general design of the manuscript.

Title: Deciphering the genomic architecture of IGH-ZFP36L1 fusion in mature B-cell lymphomas with del(14)(q24q32) reveals cooperating molecular mechanisms.

Authors: I Nagel*, I Salaverria*, S Gonzalez*, B Rodríguez, G Clot, D Martin-García, I Vater, M Sczepanowski, A Navarro, C Royo, Judit Pinteño, JI Martin-Subero, W Klapper, J Richter, M Kreuz, M Ritgen, E Callet-Bauchu, MJ Calasanz, F Sole, E Schroers, M Kneba, Martin J.S. Dyer, Julio Delgado, A López-Guilllermo, XS Puente, C López-Otín, E Campo, D Torrents, S Beà, R Siebert.

Journal: Not published. (The manuscript is his last stage of corrections and is almost ready to be sent to Nature Communications).

Impact factor: -

Citations: -

Contribution:

Santi co-authored this publication with Inga Nagel and Itziar Salaverria who perform the laboratory analysis, collection of clinical data and sequencing of whole genome and RNA. Santi is the responsible of coordinating all the contribution of the BSC within this publication. Whereas the specific analysis of isoforms was accomplished by Bernardo Rodriguez, Santi pushed and took care of all the other aspects of the BSC activity: interpretation of the rearrangements found in CLL genomes, identification of Translin motive as a potential mechanism of the deletion, and the analysis of gene expression. This last task also involved the mentoring of Judith Pinteño, a visiting undergraduate student in the group, which also contributed to this study.

Collaborations

Title: Unravelling the hidden DNA structural/physical code provides novel insights on promoter location.

Authors: Elisa Durán, Sarah Djebali, Santi González, Oscar Flores, Josep Maria Mercader, Roderic Guigó, David Torrents, Montserrat Soler-López, Modesto Orozco.

Journal: Nucleic acids research.

Impact factor: 8.808

Citations: 3

Contribution:

Santi's contribution to this study involved the analysis of the regulatory potential of candidate regions identified by the ProStar method. This involved in the comparison of thousands of regions with the annotation of regulatory regions found in UCSC.

Title: Non-coding recurrent mutations in chronic lymphocytic leukaemia.

Authors: Xose S Puente, Silvia Beà, Rafael Valdés-Mas, Neus Villamor, Jesús Gutiérrez-Abril, José I Martín-Subero, Marta Munar, Carlota Rubio-Pérez, Pedro Jares, Marta Aymerich, Tycho Baumann, Renée Beekman, Laura Belver, Anna Carrio, Giancarlo Castellano, Guillem Clot, Enrique Colado, Dolors Colomer, Dolors Costa, Julio Delgado, Anna Enjuanes, Xavier Estivill, Adolfo A Ferrando, Josep L Gelpí, Blanca González, Santiago González, Marcos González, Marta Gut, Jesús M Hernández-Rivas, Mónica López-Guerra, David Martín-García, Alba Navarro, Pilar Nicolás, Modesto Orozco, Ángel R Payer, Magda Pinyol, David G Pisano, Diana A Puente, Ana C Queirós, Víctor Quesada, Carlos M Romeo-Casabona, Cristina Royo, Romina Royo, María Rozman, Nuria Russiñol, Itziar Salaverría, Kostas Stamatopoulos, Hendrik G Stunnenberg, David Tamborero, María J Terol, Alfonso Valencia, Nuria López-Bigas, David Torrents, Ivo Gut, Armando López-Guillermo, Carlos López-Otín, Elías Campo

Journal: Nature

Impact factor: 41.456

Citations: 5

Contribution:

Santi contributed to this study by coordinating and performing the analysis of 150 whole CLL genomes with SMUFIN. Santi generated and interpreted all the results of structural variation found in these genomes with the help of Marta Munar, which was under the supervision of Santi. A major tasks accomplished by santi was the coordination of the reconstructions of complex karyotypes observed within these tumors. This work led to the identification of chromotriptic and chromoplectic rearrangement events observed in CLL patients that are associated to worst prognosis of the tumor progression.

Publication 1

Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads

Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads

Valentí Moncunill[1,10], Santi Gonzalez[1,10], Sílvia Beà[2], Lise O Andrieux[1], Itziar Salaverria[2], Cristina Royo[2], Laura Martinez[1], Montserrat Puiggròs[1,3], Maia Segura-Wang[4], Adrian M Stütz[4], Alba Navarro[2], Romina Royo[1,3], Josep L Gelpí[1,3,5], Ivo G Gut[6], Carlos López-Otín[7], Modesto Orozco[1,5,8], Jan O Korbel[4], Elias Campo[2,9], Xose S Puente[7] & David Torrents[1,9]

The development of high-throughput sequencing technologies has advanced our understanding of cancer. However, characterizing somatic structural variants in tumor genomes is still challenging because current strategies depend on the initial alignment of reads to a reference genome. Here, we describe SMUFIN (somatic mutation finder), a single program that directly compares sequence reads from normal and tumor genomes to accurately identify and characterize a range of somatic sequence variation, from single-nucleotide variants (SNV) to large structural variants at base pair resolution. Performance tests on modeled tumor genomes showed average sensitivity of 92% and 74% for SNVs and structural variants, with specificities of 95% and 91%, respectively. Analyses of aggressive forms of solid and hematological tumors revealed that SMUFIN identifies breakpoints associated with chromothripsis and chromoplexy with high specificity. SMUFIN provides an integrated solution for the accurate, fast and comprehensive characterization of somatic sequence variation in cancer.

The recent development of high-throughput sequencing technologies has made possible the sequencing of genomes at an unprecedented speed, allowing the identification of the genetic basis of numerous diseases. These advances have been particularly important in the study of cancer, providing information on thousands of tumor genomes and a large catalog of genomic alteration associated with oncogenesis[1]. The characterization of somatic variation in tumor samples is, therefore, rapidly becoming a standard practice in biomedicine[2]. In a large fraction of biomedical studies that rely on high-throughput sequencing, the production of genome sequence data exceeds available computer resources and the capabilities of analytic protocols. This is particularly pertinent in the field of cancer genomics, where the increasing sequencing of tumor genomes calls for faster and more accurate analyses.

The identification of somatic variants associated with cancer typically requires sequencing tumor and normal genome samples from the same patient, followed by multiple sequence comparisons. Normal and pathological reads are aligned to a reference genome, and the alignment is used to identify sequence changes to isolate the somatic fraction of variants (i.e., those detected only in the tumor). In principle, this simple strategy can be used to detect single-nucleotide variants (SNVs) and structural variants. Existing methods for the detection of somatic SNVs show high sensitivity and specificity[3,4], but identifying structural variants is still challenging and remains largely unsolved. The need for a reference sequence is particularly limiting. Reads carrying variations, such as those covering somatic changes in the tumor, are more difficult to align to the reference genome[5], and corresponding variants might become undetectable. Moreover, reference-based methods also must discriminate germline changes from somatic variants. In addition to these limitations at detection level, this alignment step is also time consuming and requires a considerable amount of computing resources.

To define the complete catalog of somatic variation (SNVs and structural variants) for a given tumor still requires complex computational pipelines with combinations of different methods, each of them restricted to the detection of a particular type of variant or structural variants of particular sizes. This restricts the general usage of this methodology to centers and groups with considerable amounts of computing resources and expertise. For example, widely used programs, such as BreakDancer[6] or Delly[7], can only identify structural variants larger than 20 and 150 base pairs, respectively. Each of the methods needed for a complete structural characterization of somatic variation in tumor genomes further require complex scoring and filtering schemes to achieve acceptable levels of specificity, but such procedures drastically lower the sensitivity, leaving a substantial

[1]Joint IRB-BSC Program in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain. [2]Department of Pathology, Hematopathology Unit, Hospital Clinic, Institut d'Investigació Biomèdica August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain. [3]Computational Bioinformatics, National Institute of Bioinformatics, Barcelona, Spain. [4]European Molecular Biology Laboratory, Genome Biology Research Unit, Heidelberg, Germany. [5]Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Spain. [6]Centro Nacional de Analisis Genomico (CNAG), Barcelona, Spain. [7]Departamento de Bioquímica y Biología Molecular, Universidad de Oviedo–IUOPA, Oviedo, Spain. [8]Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain. [9]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [10]These authors contributed equally to this work. Correspondence should be addressed to D.T. (david.torrents@bsc.es).

Received 18 December 2013; accepted 22 August 2014; published online 26 October 2014; doi:10.1038/nbt.3027

fraction of structural variants undetected. Even experimental procedures, such as those that use single-nucleotide polymorphism (SNP) arrays, generate only a partial description of the rearranged tumor, as they detect only the fraction of structural variation that generates sequence imbalance. The fact that the most recent and complete catalog for signatures of somatic mutations in cancer[8] does not yet include structural variants is a clear consequence of all these limitations.

To fill these gaps, we have developed SMUFIN (for somatic mutation finder), a computational approach for the accurate and complete characterization of somatic variation in cancer. SMUFIN searches for SNVs and structural variants of all sizes by directly comparing normal and tumor sequencing reads without the need of their initial mapping onto a reference genome. Here, we evaluated its performance in the context of existing strategies and the application to cancer genomics, as well as its potential to define complex chromosomal rearrangements in aggressive forms of mantle cell lymphomas and medulloblastoma. The implementation of SMUFIN, including latest releases, documentation, example data sets and supplementary information is freely available at http://cg.bsc.es/smufin/. Source code files are also in **Supplementary Source Code**.

RESULTS

The SMUFIN algorithm

The underlying search algorithm of SMUFIN comprises two major steps (**Fig. 1**). First, under the assumption that any somatic variation occurring in the tumor genome will generate a unique sequence, tumor-specific reads are identified and isolated. This is achieved by creating a quaternary sequence tree (implemented as a generalized suffix array) using all tumor and normal reads (**Fig. 1**). In this tree, genomic regions of unaltered sequence will generate identical tumor and normal reads, and these will cluster together in common branches. Reads covering sequence variations in one or both alleles of the tumor are expected to form isolated branches without normal reads. These unique reads are then grouped into read blocks, each expected to cover a single sequence change or break in the tumor. By further interrogating the tree for overlapping regions (of at least 30 bp), each of these blocks is further expanded by adding and aligning the corresponding normal reads.

Next, potential tumor variants are defined and classified on each of the breakpoint blocks in two steps (**Fig. 1**). First, 'small' variants are identified—that is, SNVs and structural variants that can be completely defined within the size of a read. Second, 'large' structural rearrangements, which expand beyond the size of the input read, are defined. We expect that each of these blocks will represent one of the breaks generated by large insertions, inversions or deletions in the tumor genome, or to single translocation points. SMUFIN provides to the user these large structural variants as single breakpoints along with the corresponding surrounding sequence in the tumor. A simple filtering scheme is also used to ensure a minimum of physical coverage of all detectable variants and to correct for potential contamination of tumor cells in normal samples. Although default parameters have been adjusted in SMUFIN for common sequencing scenarios (i.e., ≥30-fold coverage depth in Illumina sequencing platforms), the user can also tune these filters to adapt the method to the particular characteristics of the data.

In summary, distinct features of SMUFIN that are not available in existing strategies for the detection of somatic variants include (i) the direct comparison of normal and tumor reads without the need to generate mapped BAM files; (ii) the detection, in a single execution, of SNVs and structural variants, such as inter- and intrachromosomal translocations, inversions, insertions and deletions of any size; (iii) the identification of variants at base pair resolution; and (iv) the reconstruction of exact changes in the tumor genome, including the sequence at both sides of all breakpoints detected.

Furthermore, we have developed a Message Passing Interface (MPI) implementation of SMUFIN that yields direct improvements of its usability and execution times. Using 16 nodes (2xIntel SandyBridge, 8-core/2.6 GHz) SMUFIN was able to complete the analysis of a tumor-normal, whole-genome pair in 4–8 h for samples with 30× of sequencing coverage, and 9–15 h for 60× samples. These executions showed discrete peaks of RAM usage of 8–10 Gb and 13–17 Gb per node, respectively.

Assessment and comparison of SMUFIN with model genomes

To assess SMUFIN's performance, we measured both the fraction of somatic variants detected (sensitivity) and the precision of this detection (specificity) using simulated and real cancer genome data together with orthogonal experimental techniques.

We generated normal and tumor test genomes by first applying to the human reference genome the sequence variation corresponding to a random human haplotype[9] and to a predesigned catalog of somatic changes, and then simulating whole-genome sequencing at different depths of coverage (Online Methods, **Supplementary Fig. 1** and **Supplementary Table 1**). To assess the applicability of SMUFIN in the current context of cancer genome analysis, we compared its performance with a representative set of somatic variant callers that are common parts of current pipelines for the analysis of tumor genomes: Mutect for SNVs[3], and BreakDancer[6], Pindel[10], Delly[7] and CREST[11]

Figure 1 SMUFIN. (**a**) (Left) As input, SMUFIN takes high-quality read data (FASTQ) of normal and tumor genomes of the same individual. (Middle) Starting and ending nucleotide sequences of representative example reads from tumor and normal samples. Reads containing no somatic mutations are shown in blue. Somatic mutations and downstream sequences are red. Nucleotide positions are indicated at the bottom, where *n* corresponds to the size of the read (**Supplementary Fig. 4**). Reads are numbered on the right side of the boxes. Pairs 1 and 1′, 3 and 3′, and 6 and 6′ would cover the same region in the nonmutated and mutated allele of the cancer genome, respectively. The other reads represent the two nonmutated alleles. (Right) These reads have different properties inside the quaternary tree. Because nonmutated cancer reads are expected to have their counterpart among healthy reads, they are also expected to share the same branches. Cancer reads that carry variations are expected to be unique and, therefore, to be located in isolated branches. These branches become cancer-specific exactly at the point where they differ, that is, in a breakpoint. SV, structural variation. (**b**) SMUFIN collects all the reads expanding on these cancer-specific branches and takes them as reads containing potential somatic variant breakpoints. Because any particular breakpoint is expected to be represented by several reads, we group all detectable reads that are overlapping and complementary and construct breakpoint blocks (**Supplementary Fig. 5**), covered by only one (single orientation) or by two strands (double orientation). This step, which includes filters for minimum overlap and coverage, removes a large fraction of false-positive variations, mostly derived from sequencing errors. (**c**) Each of the accepted blocks is then analyzed, as to the type of change detected. First, small variants, which can be defined within a single block, are identified. These include SNVs and small insertions, deletions and inversions. The remaining unclassified blocks are then passed into the next step where sequence translocations of large structural variants are defined. Here, for each of the breakpoints, we interrogated the tree and retrieved up to 100 bp of overlapping normal and tumor reads at each side of the break. (**d**) Finally, small and large variants are unambiguously positioned onto the reference genome by mapping[18] the normal consensus region covering and flanking each of the variants. BWA, Burrows-Wheeler Aligner.

for structural variants of different sizes (**Supplementary Table 2**). For the present comparison, we ran them as described in their companies' corresponding publication or website.

We first observed that the calling of somatic SNVs was nearly optimal and within the same range in Mutect and SMUFIN, with sensitivities of 97% and 92%, and specificities of 93% and 99%, respectively (**Table 1** and **Supplementary Table 3**). On the other hand, the calling efficiency of somatic structural variants varied greatly between different methods, revealing clear differences when compared to SMUFIN. Some methods reached reasonable levels of sensitivity when the evaluation was restricted to the range of structural variants they were designed to detect (Pindel and Delly), but these dropped drastically when compared against the complete catalog of structural variations in the tumor (**Supplementary Table 4**). By contrast, SMUFIN was

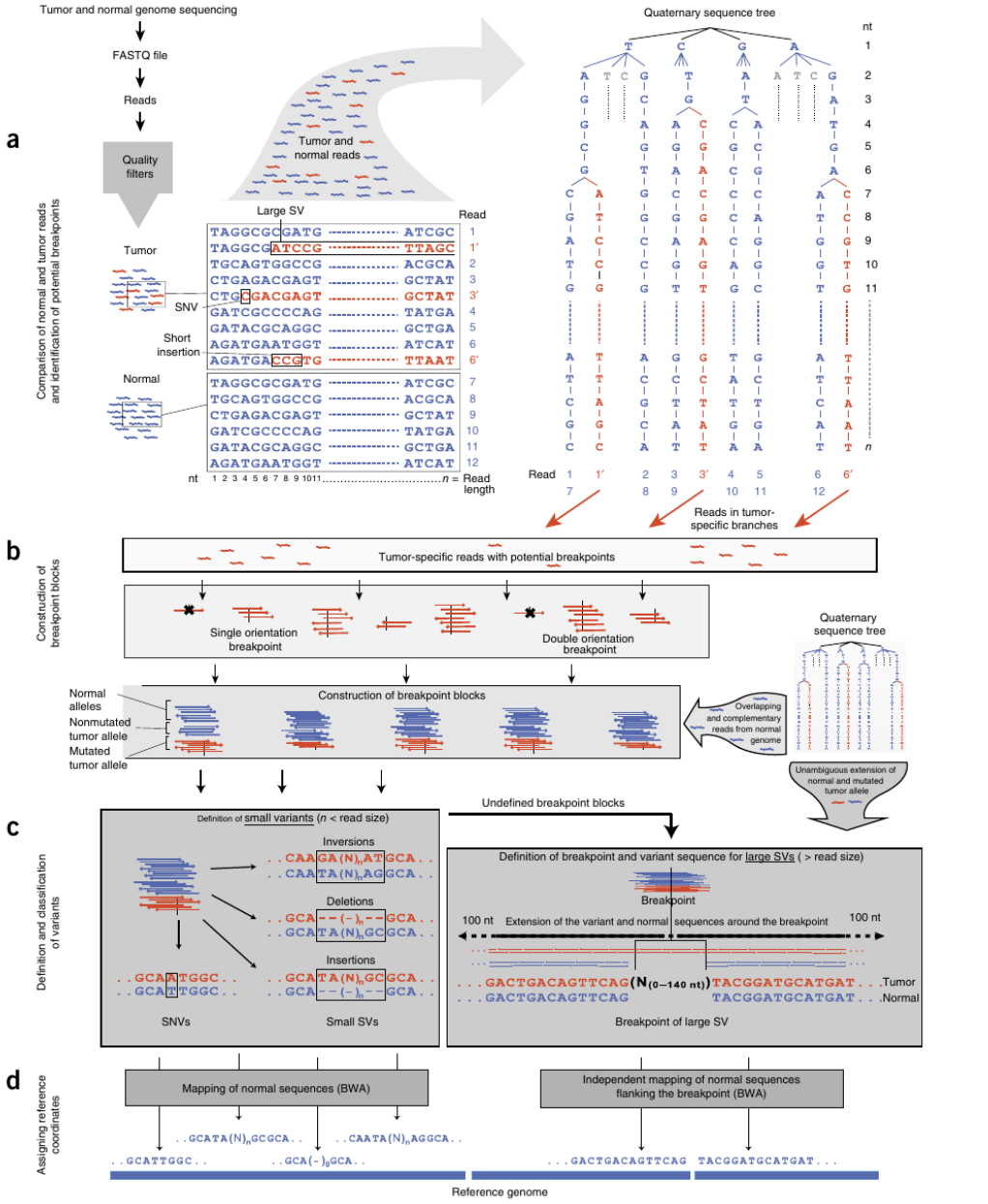


Table 1 ***In silico*** **assessment of variant calling**

	Type of variant[a]	Range of SV detection	Number of detectable variants[b]	Variant calling (sensitivity/specificity)[c]	Deviation from target (nt)[d]
SMUFIN	SNV	–	8,240	92/99	0
	SV	≥1 nt	1,798	74/91	1 ± 1
Mutect	SNV	–	8,240	97/93	0
BreakDancer	SV	≥20 nt	923	63/78	285 ± 145
Pindel	SV	≥1 nt	1,798	74/28	2 ± 26
CREST	SV	≥20 nt	923	42/53	28 ± 111
Delly	SV	≥150 nt	448	89/63	52 ± 77

[a]Variants are distributed as follows: 8,240 SNVs and 1,798 SVs (738 deletions, 715 insertions and 345 inversions). The table shows the number of breakpoints that define SVs. [b]Variants that fall into the range of detection for each of the methods. [c]Performance values obtained counting only variants within the detection range of each of the methods. See **Supplementary Table 4** for a comparison against the complete SV catalog. [d]Expressed as average distance ± s.d. from the breakpoint position. [e]CREST[11] has no size limit at detection level[11]. Nevertheless, among all the predictions obtained, none was below 20 nt.
SV, structural variant; nt, nucleotides.

able to identify somatic structural variants with a sensitivity of 74% independently of the size of the structural variant, reaching >90% sensitivity when only structural variants larger than the read size were taken into account. SMUFIN's sensitivity for somatic SNV and structural variant calling is actually similar to that resulting from the combination of all the methods above: 94% versus 89% for SMUFIN.

The downside of combining these methods as a strategy for variant calling is the low levels of specificity achieved. In fact, in terms of specificity, the values for the external structural variant callers were 29–77%, whereas SMUFIN reached values of 91% across all structural variants. We also tested for consistency at sensitivity level in the identification of medium structural variants (i.e., variant size of 5–500 bp), which constitute a group of variants that have been particularly challenging for structural variant-calling methods that rely on pre-aligned data. This analysis showed that only SMUFIN and Pindel, which has been specifically designed also for small structural variants, kept a similar sensitivity when compared with the identification of the total of structural variants (**Supplementary Table 4**). When further testing SMUFIN, Pindel and CREST using lower levels of *in silico* sequencing coverage, we observed an overall decrease in performance, both at sensitivity and specificity levels, at physical sequencing coverage below 20-fold (**Supplementary Fig. 2**).

Detection of small somatic variants in human tumors

To further investigate the performance of SMUFIN in real data, we calculated and assessed the positive discovery rate of somatic SNVs and structural variants calling using whole-genome sequence (WGS) data from primary tumor and matched nontumor samples. We first tested the detection of small variants by analyzing a previously described sample (M004) of mantle cell lymphoma (MCL)[12], an aggressive subtype of lymphoid neoplasia. SMUFIN identified 4,409 somatic SNVs and 1,094 small structural variants (**Supplementary Table 5**).

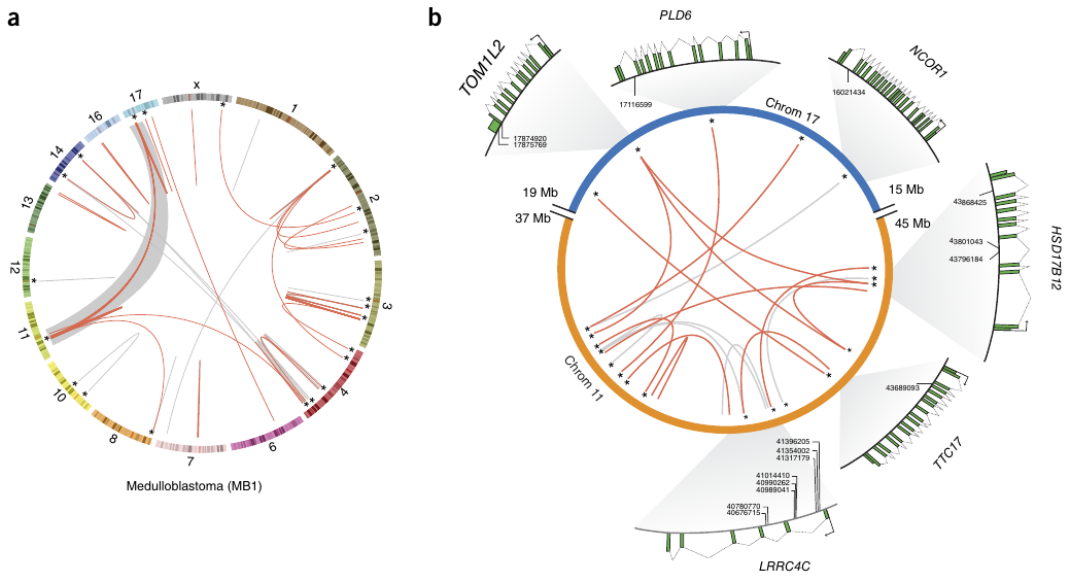


Figure 2 Large structural variation in pediatric medulloblastoma tumor MB1. (**a**) Circos representation of a genome-wide view of all the intra- and interchromosomal translocations identified by SMUFIN in this tumor (chromosomes with no breakpoints are excluded). Novel breakpoints are displayed in red, whereas those already reported are in gray. Breaks marked with "*" correspond to those that were tested and could be confirmed, resulting in a local specificity of 100%. Shadowed area indicates the interconnection between two regions in chromosomes 11 and 17 with high density of DNA breakage and rejoining events. (**b**) Circos map displaying all the breakpoints of chromosome 11 (within the 37–45 Mb region) and the interaction with chromosome 17 (15–19 Mb) in more detail. Genes affected by, at least one previously undescribed breakpoint are drawn, along with the exact position of the break.

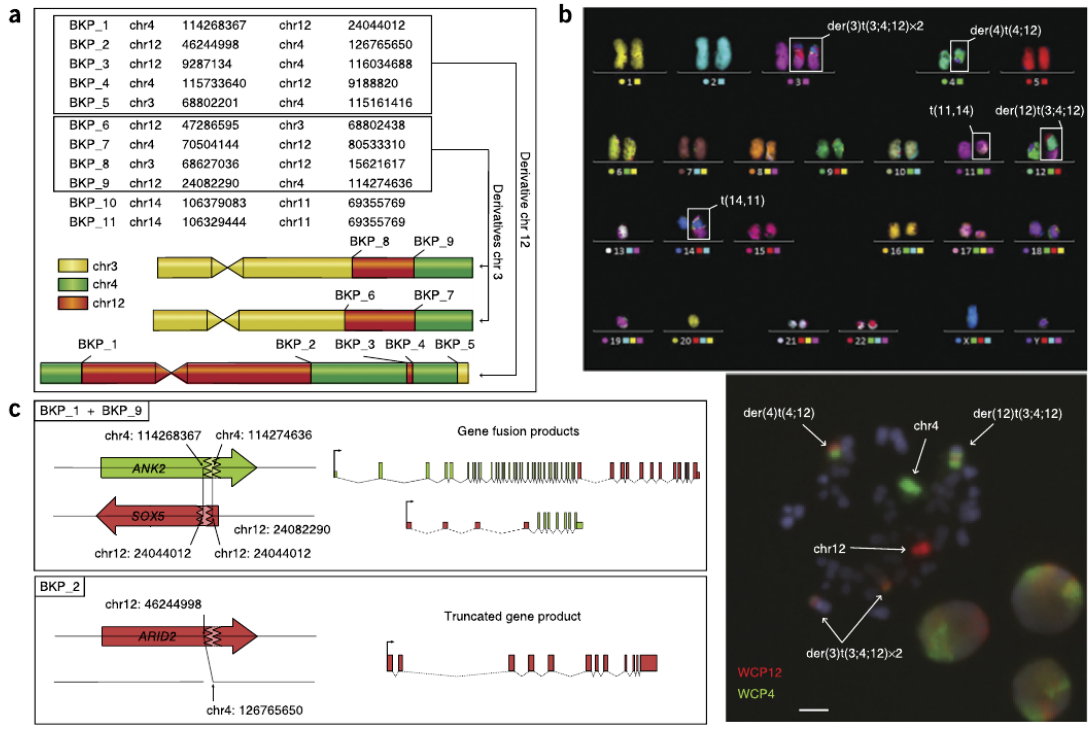


Figure 3 Identification and validation of chromoplexy in mantle cell lymphoma tumor M003. (**a**) Three chimeric chromosomes formed by parts of chromosomes 3, 4 and 12 and the primary hallmark MCL translocation t(11;14). These rearrangements were identified by SMUFIN and all were experimentally verified by PCR. (**b**) A representative 24-color multicolor-FISH (mFISH) karyogram (top) that shows an unbalanced karyotype, with the t(11;14)(q13;q32) (BKP 10 and 11), a centromeric deletion of 17p, and several rearrangements between chromosomes 3, 4 and 12, all of them consistent with the breakpoints identified by SMUFIN. Bottom image shows a metaphase hybridized with whole-chromosome painting (WCP) 4 (green) and 12 (orange) probes showing four derivative chromosomes with material of these two chromosomes. Combination of mFISH and WCP analysis confirmed the presence of two different derivative chromosomes der(3)t(3;4;12), one der(12)t(3;4;12), and identified a fourth, der(4)t(4;12), which is not detectable by SMUFIN owing to the centromeric location of the breakpoint in chromosome 4. Scale bar, 10 μm. (**c**) Genes affected by chromoplexy—a reciprocal fusion of two genes (*ANK2*, in green and *SOX5*, in red) and a truncated chromatin remodeler (ARID2). Coding and noncoding exons are displayed as taller and shorter boxes, respectively.

To evaluate the specificity of SMUFIN, we verified >94% of SNVs (76 of 81) and >80% of structural variants (28 of 35) from a random set of 111 of these somatic calls by Sanger sequencing using the same DNA used for whole genome sequencing (**Supplementary Table 6**). These specificity rates are in agreement with the corresponding values obtained from the *in silico* analysis.

Complex structural variation in aggressive tumors

We next evaluated SMUFIN's accuracy in detecting large structural variants involving the somatic insertion, deletion, inversion or translocation of DNA fragments that are hundreds to millions of base pairs in length. For this test, we analyzed whole-genome sequence data from another mantle cell lymphoma sample (M003) and a sample from a pediatric form of a medulloblastoma (MB1), both known to present complex landscapes of chromosomal rearrangements[12,13]. Because these representative examples corresponded to a hematological and a solid tumor, each sequenced in a different sequencing facility, this analysis also measured SMUFIN's consistency across different types of data.

Identification of chromothripsis

MB1 was previously described as presenting chromothripsis, a complex structural alteration of the genome hypothesized to arise from a single catastrophic event that generates multiple breakpoints, often affecting one single chromosome[14]. In this tumor sample, SMUFIN uncovered a total of 102 breakpoints corresponding to large structural variants (i.e., beyond the read size), covering 85 intra- and 17 interchromosomal translocations (**Supplementary Table 7**). From the assessment of a random set of 39 of these breaks through PCR amplification and Sanger sequencing, we verified 36 (92%). Among all the breakpoints detected, 25 agreed with the intervals of chromosomal translocations that previously led to the definition of chromothripsis in this tumor, including three of the four verified at base-pair resolution.

In addition, we detected 65 previously unidentified breakpoints in the same tumor, covering 53 intra- and 12 interchromosomal translocations (**Supplementary Fig. 3**). From a random subset of 37 of these translocations (16 intra- and 11 interchromosomal), we verified 25 (92.5%) using Sanger sequencing. Together with the clusters

of breakpoints already reported for chromosomes three and four in this tumor, new calls uncovered by SMUFIN enabled us to define a third damaged region in chromosome 11, with a density of six DNA breaks per Mb (between positions 39 and 45 Mb). Notably, many of these breakpoints correspond to translocations with chromosome 17 (**Fig. 2**). Furthermore, and complementary to the previous functional characterization of this tumor, we identified affected genes that were not reported in the previous study (**Supplementary Table 7**), including some that have been identified as possible driver genes, such as *NCOR-1*, *SIN3P*, *WDR52* and *PALLD*, in several types of tumors[15]. Of the 65 breakpoints, 54 were predicted (allowing up to 100-nt deviation in the prediction) by at least one of the methods used above for the comparative assessment of SMUFIN, with 44 found only by Delly. This is not surprising considering the results of the *in silico* analysis, as sensitivity is not the major limitation of the reference alignment-dependent approaches.

Identification of chromoplexy

We also analyzed a sample from an aggressive form of mantle cell lymphoma (M003), previously described to have undergone complex chromosomal rearrangements[12]. We used SMUFIN to identify 30 breakpoints corresponding to large structural variants (**Supplementary Table 8**). Using PCR amplification followed by Sanger sequencing, we verified 19 of the 22 breakpoints tested, involving 7 intra- and 15 interchromosomal translocations (**Supplementary Table 6**). This not only confirms the correct location and the type of translocation identified, but it also shows that SMUFIN was able to reconstruct the correct sequence around the variants, as five of the breakpoints (six inter- and one intrachromosomal; **Supplementary Table 8**) included stretches of a new DNA insertion 5–30 nt long.

We next evaluated whether SMUFIN could be used to define the chromosomal arrangement of this tumor. We compared all 30 breakpoints identified, with 18 noncentromeric and nontelomeric regions of chromosomal imbalances previously detected using Affymetrix SNP6.0 array (Affymetrix, Santa Clara, CA)[12]. SMUFIN could redefine, at base pair resolution, 16 of these 18 regions. By manually assembling the fragments between all the translocations detected, we could model the landscape of this genome, which included three derivative chromosomes formed by combinations of large fragments of chromosomes 3 and 12 with smaller parts of chromosome 4. These chimeric chromosomes were experimentally confirmed in the mantle cell lymphoma cells by a combination of multicolor fluorescence *in situ* hybridization (FISH) and whole-chromosome painting analysis (**Fig. 3**). Furthermore, the resolution provided by SMUFIN allowed the identification of the fragmentation and fusion of genes not previously described in this sample. For example, we found that these translocations caused the fusion of *ANK2* and *SOX5* genes. Notably, these two rearrangement events did not appear to be independent as the corresponding fragments generated after the double-strand break were rejoined again reciprocally—that is, generating both, 12 to 4 and 4 to 12 translocations and two different forms of *ANK2-SOX5* fusions (**Fig. 3**). In fact, 8 out of the 18 breakpoints appeared to be rejoined reciprocally, as recently described in prostate tumors[16,17], suggesting an original organization of the chromatin where these regions were physically proximal and somehow interacting. A third translocation identified in the M003 tumor implies the breakage and putative inactivation of *ARID2*, a gene involved in chromatin remodeling.

By considering the number of rearrangements identified in this tumor, their distribution and the number of chromosomes involved, we classify this scenario as chromoplexy, a recently described phenomenon that, in contrast to chromothripsis[14], is characterized by the presence of tens of unclustered chained rearrangements involving two or more chromosomes[16,17]. The high fraction of reciprocal rejoining events found in this tumor, together with the fusion of genes and the disruption of a chromatin remodeler gene, is also in agreement with the results of the chromoplectic events identified in prostate tumors.

CONCLUSIONS

We describe SMUFIN, a methodology for the identification of somatic variation in tumor genomes from their direct comparison with their corresponding normal samples. SMUFIN also provides an integrated solution for the identification, in a single run, of somatic SNVs and structural variants (insertions, deletions, inversions and translocations of any size), which can currently be partially achieved only by combining several independent programs and in-house filtering schemes into complex computational pipelines. Our method defines, at base pair resolution, complex scenarios of chromosomal rearrangements, such as chromoplexy and chromothripsis. SMUFIN was able to identify the translocations defined before using other computational and experimental methods, as well as novel breakpoints that complete the corresponding landscapes of chromosomal rearrangements. Owing to the underlying mechanism of the algorithm used in SMUFIN, our method is not suitable to quantify copy number variations or detect complete losses of chromosome arms or inversions flanked by palindromic sequences.

Beyond the benefits of the detection capabilities of SMUFIN, the current parallel implementation of the program also shows substantial improvements at the level of usability and execution time compared with available pipelines, as it can currently analyze a pair of whole genome sequences with coverage of 30–60× in 4–15 h, using 50–80 standard cores and requiring less than 17 Gb of RAM memory per computing node. This, together with the scalability of the program, will realistically allow a systematic and parallel analysis of cancer samples, accessible to nonexpert users with standard computing resources.

Taken together, the underlying search mechanism of SMUFIN constitutes an alternative way of processing and analyzing genomic data, which can inspire the development of new tools for other types of genomic analyses. Because SMUFIN actually finds changes in one sequence set relative to another, it could potentially be adjusted to other types of biomedical and evolutionary studies that rely on the comparative analysis of two genomes, even if they are from different species.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. For validation sequences produced in this study (**Supplementary Table 6**), European Genome-phenome Archive: EGAS00001000510.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The ICGC-CLL Genome Consortium is funded by the Spanish Ministry of Economy and Competitiveness (MINECO) through the Instituto de Salud Carlos III (ISCIII), Red Temática de Investigación del Cáncer (RTICC) of the ISCIII (RD12/0036/0036) and National Institute of Bioinformatics (INB). This study was also supported by Ministerio de Economía y Competitividad, Secretaría De Estado De Investigación, Desarrollo e Innovación PLAN NACIONAL de I+D+i 2008-2011, Subprograma de Apoyo a Centros y Unidades de excelencia Severo

Ochoa; and Plan Nacional SAF12/38432; Generalitat de Catalunya AGAUR 2009-SGR-992; Fondo de Investigaciones Sanitarias (PI11/01177); Association for International Cancer Research (12-0142). J.O.K. and M.S.W. were supported by the European Commission (Health-F2-2010-260791). C.L.-O. is an investigator of the Botin Foundation. E.C. and M.O. are ICREA Academia Researchers. We also thank S. Guijarro and C. Gómez for their excellent technical assistance.

AUTHOR CONTRIBUTIONS

V.M., S.G. and D.T. conceived and designed the study. L.O.A., L.M., M.P., J.L.G., R.R. and M.O. performed data analysis. S.B., I.S., C.R., A.N., E.C. and I.G.G. generated and experimentally validated the MCL samples. M.S.-W., A.M.S. and J.O.K. generated and experimentally validated the MB1 sample. C.L.-O., X.S.P., E.C. and D.T. wrote the manuscript; and D.T. supervised the whole study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
2. Frampton, G.M. *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
3. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
4. Puente, X.S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101–105 (2011).
5. Degner, J.F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
6. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
7. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
8. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
9. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
10. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
11. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8**, 652–654 (2011).
12. Beá, S. *et al.* Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc. Natl. Acad. Sci. USA* **110**, 18250–18255 (2013).
13. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**, 59–71 (2012).
14. Korbel, J.O. & Campbell, P.J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226–1236 (2013).
15. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
16. Baca, S.C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
17. Shen, M.M. Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer Cell* **23**, 567–569 (2013).
18. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

ONLINE METHODS

The SMUFIN algorithm. The general structure and the internal mechanism of SMUFIN is displayed in **Figure 1**. The complete variant identification and characterization process comprises the following specific steps:

Input data. As input, SMUFIN takes high-quality sequencing data directly from FASTQ files of tumor and normal samples of the same individual. Alternatively, SMUFIN is also able to accept BAM files, from which it extracts all the sequencing reads. Sequences having over 10% of its bases with a phred quality score < q20 are discarded.

Construction of the quaternary sequence tree. A 'quad-tree'-based structure is first generated using all high-quality normal and tumor reads. All these sequences are sequentially loaded into the tree on the basis of their sequence (**Fig. 1** and in **Supplementary Fig. 4a**). Each node of the tree has, at most, four branches, each one representing one of the four nucleotides. To avoid sequence ambiguity derived from the complexity of the genome, only fragments of at least 30 bp are inserted into the tree. In the case of the presence of undefined base pairs ("N"), these are removed and the original sequence is split forming new shorter reads, which are inserted in the tree only if they are longer than 30 base pairs. Each of sequences accepted is inserted into the tree, from the root, in original form (i.e., starting from nucleotide 1 to the end of the read), together with all derived suffixes larger than 30 bp (recursively starting from nucleotide 2 to the end, 3 to the end, etc…; **Supplementary Fig. 4a**). Because posterior searches through the tree start from the root, the presence of read suffixes allows a rapid identification of particular sequences and reads.

Selecting reads containing candidate variants. Once all the sequences and derived suffixes are loaded into the tree, the next step consists in identifying all tumor-specific reads. Because we expect that variants generate new and distinct sequences in the mutated genome compared with the nonmutated sample, SMUFIN first searches and collects sequences (reads) that are only present in the tumor sample. These sequences are identified from the tree, as nodes and branches with an unbalanced representation of normal (count normal reads; CNR) and tumor (count tumor reads; CTR) reads (**Supplementary Fig. 4b**). We expect that nodes or branches covering a variation in the tumor sequence will theoretically have no representation of normal reads. To favor this condition, we start to search the tree from the level 30 toward the leafs. We accepted only nodes and branches that have a CTR of at least 4. Internal tests suggest that setting CTR ≥ 4 improves specificity in a factor 1.4× with a negligible loss of sensitivity (not shown). Additionally, nodes or branches with a CNR to CTR ratio below a certain threshold (E_CONT) are selected. This threshold can be adjusted by the user to account for expected levels of contamination of tumor cells into the normal sample. Please, be aware that an E_CONT of 0 implies no expected contamination, that is, no acceptance of reads coming from the normal sample (CNR) on that candidate variant node or branch, which implies lower final sensitivity but higher specificity. On the other hand, an E_CONT larger than 0 always results in a higher sensitivity, but at the cost of lower specificity. E_CONT was set to 0 for the *in silico* analysis and to 0.05 for the real tumor samples analyzed here, where we assume a maximum of 5% contamination of tumor reads into the normal sample.

Grouping candidate reads. After all detectable tumor-specific reads have been identified, the next step consists in grouping those that are suspected to cover the same variant. For this, candidate sequences are organized by identity: two sequences belong to the same group if they overlap by at least 30 bp. Reverse complementary sequences are also evaluated during this grouping in order to be able to cover the variant in both orientations. Sequence blocks (groups) with sequences in only one of the orientations or with less than four tumor reads are discarded. Once these groups are generated, we interrogate the tree, also on the 30-bp overlap basis, to extract the normal (nonmutated) reads of the same region and add them to the block. Ideally, each block will represent a region in the genome containing the mutated and the nonmutated version (see a detailed example of a breakpoint block in **Supplementary Fig. 5**). In order to classify and characterize the type of variation identified, we extract the consensus mutated and normal sequences from these blocks. Normal consensus sequences will be also used at the end of the procedure and mapped onto the reference genome to obtain the coordinates of the variant.

Identification and characterization of variants. Once all possible breakpoint blocks are defined, the next step consists in identifying and classifying the variation included there. Normal and tumor consensus sequences derived from these blocks (**Supplementary Fig. 5**) are recursively compared to identify differences. A first evaluation will search for small variants, which consist of those that are completely included within the consensus sequences (SNV and small structural variants: insertions, deletions and inversions). All the blocks that do not match this criterion are then considered candidates for large structural variants, that is, those likely to cover breakpoints of intra- or interchromosomal transitions, part of large deletions, insertions, inversions or translocations. In this case, each tumor consensus sequence is extended on both ends (**Fig. 1**) by interrogating the tree for unambiguous tumor reads that overlap at least 30 bp with the tumor consensus, reconstructing a (maximum) 200-bp region around the break and allowing the detection of newly generated sequence at the point of the break.

After small and large somatic variants are defined, we identify the coordinates of the changes by mapping onto the reference genome the normal consensus sequences corresponding to each of the variants, avoiding potential mapping conflicts derived from the presence of the variant, as usually happens when using reference-based approaches. Sequences mapping (with the same score) to several positions in the genome are discarded.

Calibration and default parameters for SMUFIN were adjusted using a high-quality set of ~1,000 SNVs identified with the Sidrón software in a chronic lymphocytic leukemia sample[4].

SMUFIN's pseudo-code.

```
SeqReader normalReader = openSeqReader(normal_
input_file);
SeqReader tumorReader = openSeqReader(tumor_
input_file);
Tree qtree = initTree();
Foreach read in normalReader:
If quality_check(read):
insertIntoTree(qtree, read, as_normal);
Foreach read in tumorReader:
If quality_check(read):
insertIntoTree(qtree, read, as_tumor);
List candidate_reads = GenerateEmptyList();
Foreach node in qtree:
If depth(node) >= 30 and CTR(node) >= 4 and
CNR(node)/CTR(node) < E_CONT):
reads = GetTumorReadsFromNode(node);
InsertReadsIntoList(candidate_reads, reads);
List breakpoint_blocks = GenerateEmptyList();
Foreach read in candidate_reads:
tumor_reads = GetOverlappingReadsFromCandidateRea
ds(read, candidate_reads);
normal_reads = GetOverlappingReadsFromTree(tumor_
reads, qtree, as_normal);
bp_block = GenerateBPBlock(normal_reads, tumor_
reads);
If Coverage(bp_block) >= 4:
InsertIntoList(breakpoint_blocks, bp_block);
List large_variant_candidates = GenerateEmpty
List();
Foreach bp_block in breakpoint_blocks
normal_consensus_sequence = GetNormalConsensus
SequenceFromBPBlock(bp_block);
tumor_consensus_sequence = GetTumorConsensusSequence
FromBPBlock(bp_block);
If HasSmallVariant(normal_consensus_sequence,
tumor_consensus_sequence)
align_info = MapSequenceToReference(normal_consensus_
sequence)
If (UnambiguousMapping(align_info)
outputSmallSV(align_info, bp_block);
Else
InsertIntoList(large_variant_candidates, bp_block);
Foreach bp_block in large_variant_candidates
```

```
extended_sequence = ExtendTumorSequenceFromBPBlock
(bp_block, qtree);
align_info  =  MapExtendedToReference(extended_
sequence);
If UnambiguousMapping(align_info)
OutputLargeSV(align_info, extended_sequence);
```

Construction of the *in silico* genome. A personalized genome was simulated using the hg19 reference genome downloaded from UCSC (with no repeat-masking), and modifying it to match a randomly chosen human haplotype from the 1000 Genome database. These 7,194,026 variants consist of 4,745,917 SNPs and 2,447,367 deletions. The complete list of these germline events can be found at http://cg.bsc.es/smufin/download. The catalog of somatic variants further added to this personalized genome includes 8,240 SNVs (more than 100 bp apart), 20 known tumor translocations[19,20], 715 random insertions, 738 random deletions and 345 random inversions, all ranging from 1 bp to 100 Mbp (**Supplementary Fig. 4** and **Supplementary Table 1**). *In silico* sequencing was simulated using ART Illumina[21]. For this, we first generated a profile using the M004 sample to extract parameters, like sequence variation or read length. We then run the program at different depths of coverage, using the resulting parameters and a default error rate (0.00009).

Analysis of the *in silico* genome with external methods. Each of the external methods for the comparison with SMUFIN was run on pooled libraries (normal and tumoral) using default settings except for the following parameters: BreakDancer was run with -q 10 (mapping quality) and score cutoff of >80, as described before[6,22]; Pindel's results with less than five supporting reads were not considered as recommended elsewhere to increase specificity[23]; predictions obtained with Delly were rejected if the number of supporting reads were less than three and the mapping quality 20. For BreakDancer, Pindel and Delly, somatic variants were obtained by filtering out all the structural variants found in both normal and tumor libraries: we only kept those structural variants with no unique supporting reads from the normal library. CREST and Mutect already provided somatic variants as direct results. BreakDancer and Pindel were used as complementary methods covering large and small structural variants, respectively, as advised by the developers.

Data sets. M003, M004 and MB1 were obtained with informed consent and an ethical vote (Institutional Review Board) following ICGC guidelines (https://icgc.org). M003, M004 and MB1 were accessed through the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega/) under access numbers EGAS00001000510 and EGAS00001000085.

Identification and analysis of variant genes. Variants genes in tumor samples were identified by analyzing all the changes identified with ANNOVAR[24]. The analysis of the resulting genes potentially modified at coding or splicing level were further analyzed with Intogen[15] in order to infer their potential role in oncogenesis.

Experimental verification of variants. PCR primers were designed on sequence blocks of 2,000 bp around the target variant using Primer 3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/)[25]. PCR reactions were performed for tumor and control samples. Each target locus was amplified using 50 ng of DNA. The amplification was performed using Qiagen Multiplex PCR Kit (Qiagen), and the reaction mix contained 2× QIAGEN Multiplex PCR Master Mix, 10× primer mix (2 µM of each primer) and RNase-free water until a total reaction volume of 25 µl. PCR conditions were as follows: 96 °C, 10 min; 2 cycles of 96 °C, 30 s/60 °C, 30 s/72 °C, 1 min 30 s; 2 cycles of 96 °C, 30 s/58 °C, 30 s/72 °C, 1 min 30 s; 2 cycles of 96 °C, 30 s/56 °C, 30 s/72 °C, 1 min 30 s; 35 cycles of 96 °C, 30 s/54 °C, 30 s/72 °C, 1 min 30 s/70 °C, 10 min. All the PCR products were run in a capillary electrophoresis gel (QIAxcel Advanced System, Qiagen) with the QIAxcel DNA screening kit (Qiagen), and the multiband PCR products were purified using NucleoSpin Gel and PCR Clean-up (Mercherey-Nagel). Regarding the Sanger sequencing, PCR products were cleaned using ExoSAP-IT (USB) and sequenced using ABI Prism BigDye terminator v3.1 (Applied Biosystems) with 5 pmol of each primer. Sequencing reactions were run on an ABI-3730 Sanger sequencing platform (Applied Biosystems). Sequences were examined with the Mutation Surveyor DNA Variant Analysis Software (Softgenetics).

G-banding, FISH and M-FISH analysis. Conventional cytogenetics was performed on Giemsa-banded chromosomes (G-banding) obtained after a 72-h culture and stimulation with tetradecanoyl-phorbol-acetate. Results of the ten metaphases analyzed were described according to the International System for Human Cytogenetic Nomenclature[26]. FISH studies for the presence of the t(11;14) translocation and 17p deletions were performed using Vysis LSI IGH/CCND1 Dual Color Dual Fusion and Vysis LSI TP53 (17p13.1) (Abbott Molecular, Des Plaines, IL) on fixed cells according to the manufacturer's specifications. Two hundred nuclei were examined for each probe. To identify the chromosomes involved in marker chromosomes and to disclose other possible structural balanced abnormalities, we performed 24-color karyotyping using 24XCyte human multicolor FISH (mFISH) probe kit according to manufacturer's instructions (MetaSystems, Altlussheim, Germany) consisting of 24 different chromosome painting probes (combinatorial labeling). Image capture was done with Nikon Eclipse 50i equipped with a CCD-camera (CoolCube1, MetaSystems) and appropriate filters using Isis software. Karyotyping was done using the 24-color mFISH upgrade package. Additionally, whole chromosomal paintings (WCP) of chromosome 4 (spectrum green) and 12 (spectrum orange) were performed simultaneously.

Figure 3 was done using CIRCOS software[27].

19. Richter, J. *et al.* Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.* **44**, 1316–1320 (2012).
20. Teles Alves, I. *et al.* Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer. *Oncogene* doi:10.1038/onc.2013.591 (3 February 2014).
21. Huang, W., Li, L., Myers, J.R. & Marth, G.T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
22. Young, M.A. *et al.* Background mutations in parental cells account for most of the genetic heterogeneity of induced pluripotent stem cells. *Cell Stem Cell* **10**, 570–582 (2012).
23. Jones, D.T. *et al.* Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. *Nat. Genet.* **45**, 927–932 (2013).
24. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
25. Untergasser, A. *et al.* Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
26. Shaffer, L.G., McGowan-Jordan, J. & Schmid, M. (eds.) *ISCN 2013: An International System for Human Cytogenetic Nomenclature (2013)* (Kargar, 2013).
27. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

doi:10.1038/nbt.3027

Publication 2

Non-coding recurrent mutations in chronic lymphocytic leukaemia

ARTICLE

doi:10.1038/nature14666

Non-coding recurrent mutations in chronic lymphocytic leukaemia

Xose S. Puente[1], Silvia Beà[2], Rafael Valdés-Mas[1], Neus Villamor[3], Jesús Gutiérrez-Abril[1], José I. Martín-Subero[4], Marta Munar[5], Carlota Rubio-Pérez[6], Pedro Jares[7], Marta Aymerich[3], Tycho Baumann[8], Renée Beekman[2], Laura Belver[9], Anna Carrio[3], Giancarlo Castellano[2], Guillem Clot[2], Enrique Colado[10], Dolors Colomer[3], Dolors Costa[3], Julio Delgado[8], Anna Enjuanes[7], Xavier Estivill[11], Adolfo A. Ferrando[9], Josep L. Gelpí[5], Blanca González[3], Santiago González[5], Marcos González[12], Marta Gut[13], Jesús M. Hernández-Rivas[12], Mónica López-Guerra[3], David Martín-García[2], Alba Navarro[2], Pilar Nicolás[14], Modesto Orozco[5], Ángel R. Payer[10], Magda Pinyol[7], David G. Pisano[15], Diana A. Puente[1], Ana C. Queirós[4], Víctor Quesada[1], Carlos M. Romeo-Casabona[14], Cristina Royo[2], Romina Royo[5], María Rozman[3], Nuria Russiñol[2], Itziar Salaverría[2], Kostas Stamatopoulos[16], Hendrik G. Stunnenberg[17], David Tamborero[6], María J. Terol[18], Alfonso Valencia[15], Nuria López-Bigas[6], David Torrents[5], Ivo Gut[13], Armando López-Guillermo[8], Carlos López-Otín[1]§ & Elías Campo[3]§

Chronic lymphocytic leukaemia (CLL) is a frequent disease in which the genetic alterations determining the clinicobiological behaviour are not fully understood. Here we describe a comprehensive evaluation of the genomic landscape of 452 CLL cases and 54 patients with monoclonal B-lymphocytosis, a precursor disorder. We extend the number of CLL driver alterations, including changes in *ZNF292*, *ZMYM3*, *ARID1A* and *PTPN11*. We also identify novel recurrent mutations in non-coding regions, including the 3′ region of *NOTCH1*, which cause aberrant splicing events, increase NOTCH1 activity and result in a more aggressive disease. In addition, mutations in an enhancer located on chromosome 9p13 result in reduced expression of the B-cell-specific transcription factor PAX5. The accumulative number of driver alterations (0 to ≥4) discriminated between patients with differences in clinical behaviour. This study provides an integrated portrait of the CLL genomic landscape, identifies new recurrent driver mutations of the disease, and suggests clinical interventions that may improve the management of this neoplasia.

CLL is a B-cell neoplasia that exhibits a very heterogeneous course, with some patients following an indolent disease course, clearly contrasting with others experiencing an aggressive disease[1–3]. Patients have been classically categorized in two groups, depending on whether their tumour B cells express B-cell receptor (BCR) immunoglobulin with immunoglobulin heavy variable (IGHV) genes bearing somatic hypermutation (IGHV-mutated) or not (IGHV-unmutated)[4]. Further studies have led to the identification of additional biological features with prognostic value for CLL patients[5–8]. However, the molecular mechanisms responsible for the initiation and heterogeneous evolution of CLL remain largely unknown.

Whole-genome sequencing (WGS) and whole-exome sequencing (WES) studies in CLL patients have identified recurrently mutated genes such as *NOTCH1*, *SF3B1*, *TP53*, *BIRC3* and *POT1*, and delineated clonal evolution events in this neoplasia[9–15]. Moreover, recent works have profiled the transcriptome and the DNA methylome of many CLL cases[16–18]. Nevertheless, these studies have unveiled a high level of molecular heterogeneity, thus creating the need for integrated analysis of different genomic parameters in a larger number of patients. In this work, and as part of the International Cancer Genome Consortium (ICGC) project[19], we have performed a comprehensive analysis of the genetic alterations driving the oncogenic transformation in 506 patients with monoclonal B-lymphocytosis (MBL) or CLL. We have also carried out additional genomic studies involving single nucleotide polymorphism (SNP) arrays, DNA methylation arrays, RNA sequencing (RNA-seq) analyses and gene expression arrays. Finally, we have performed clinical studies aimed at translating the observed molecular alterations into clinical applications for CLL patients.

Mutational signatures in CLL subtypes

We studied pre-treatment tumour and matched non-tumour samples from 506 patients (452 CLL and 54 MBL): 317 (62%) were IGHV-mutated (IGHV-MUT), 179 (35%) IGHV-unmutated (IGHV-UNMUT), and 10 (2%) undetermined (Extended Data Table 1 and Supplementary Table 1). We performed WGS of 150 tumour/normal pairs, and WES of 440 cases (including 84 with both WGS and WES data). Somatic mutations analysed using the Sidrón pipeline[10] revealed the presence of 359,456 substitutions and small indels in

[1]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología (IUOPA), Universidad de Oviedo, 33006 Oviedo, Spain. [2]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), 08036 Barcelona, Spain. [3]Unitat de Hematología, Hospital Clínic, IDIBAPS, Universitat de Barcelona, 08036 Barcelona, Spain. [4]Departament d'Anatomía Patològica, Microbiología i Farmacología, Universitat de Barcelona, 08036 Barcelona, Spain. [5]Programa Conjunto de Biología Computacional, Barcelona Supercomputing Center (BSC), Institut de Recerca Biomèdica (IRB), Spanish National Bioinformatics Institute, Universitat de Barcelona, 08028 Barcelona, Spain. [6]Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, 08003 Barcelona, Spain. [7]Unidad de Genómica, IDIBAPS, 08036 Barcelona, Spain. [8]Servicio de Hematología, Hospital Clínic, IDIBAPS, 08036 Barcelona, Spain. [9]Institute for Cancer Genetics, Columbia University, New York 10032, USA. [10]Servicio de Hematología, Hospital Universitario Central de Asturias, 33011 Oviedo, Spain. [11]Center for Genomic Regulation (CRG), Pompeu Fabra University (UPF), Hospital del Mar Research Institute (IMIM), 08003 Barcelona, Spain. [12]Servicio de Hematología, IBSAL-Hospital Universitario de Salamanca, Centro de Investigación del Cáncer, Universidad de Salamanca-CSIC, 37007 Salamanca, Spain. [13]Centro Nacional de Análisis Genómico, Parc Científic de Barcelona, 08028 Barcelona, Spain. [14]Cátedra Inter-Universitaria de Derecho y Genoma Humano, Universidad de Deusto, Universidad del País Vasco, 48007 Bilbao, Spain. [15]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Spanish National Bioinformatics Institute, 28029 Madrid, Spain. [16]Institute of Applied Biosciences, Center for Research and Technology Hellas, 57001 Thermi, Thessaloniki, Greece. [17]Department of Molecular Biology, Faculty of Science, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen, 6500 HB Nijmegen, The Netherlands. [18]Servicio de Hematología, Hospital Clínico de Valencia, 46010 Valencia, Spain.

§These authors jointly supervised this work.

WGS analyses (240–5,416 per tumour), and an average mutation burden of 0.87 mutations per megabase (Mb) (Extended Data Fig. 1 and Supplementary Table 2). CLL and MBL samples had a similar mutation burden (0.87 versus 0.89 mutations Mb^{-1}, respectively, $P = 0.8$), and were considered together for WGS analysis. The number of somatic substitutions (excluding *IG* loci) was higher in IGHV-MUT tumours than in IGHV-UNMUT cases (2,847 versus 1,975, $P < 3 \times 10^{-8}$) (Extended Data Fig. 1). Three main mutational signatures were identified (Extended Data Fig. 1): an age-related signature involving C-to-T transitions at CpG sites; signature 2, characterized by T:A > G:C transversions; and an activation-induced cytidine deaminase (AID) signature[20]. This latter pattern was only detected on *IG* loci, although we also confirmed AID-induced mutations in some off-target genes highly expressed in the germinal centre[21,22]. Signature 2 was almost exclusively present in IGHV-MUT tumours, and its presence clearly separated IGHV-MUT from IGHV-UNMUT tumours (Extended Data Fig. 1).

Landscape of somatic mutations

We combined somatic mutations from the 506 tumour/normal pairs detected by either WGS or WES (excluding *IG* genes), resulting in a total of 13,631 somatic mutations affecting protein-coding genes (average 26.9 per tumour) and 951 copy number alterations (CNAs) (average 1.9) (Fig. 1 and Supplementary Table 3). We identified 36 genes (tier 1) as recurrently mutated in CLL (false discovery rate (FDR) < 10%), and 23 additional genes (tier 2) were significantly mutated in one subgroup (IGHV-MUT or IGHV-UNMUT), had recurrent or truncating mutations, or had driver mutations described in other malignancies (Extended Data Table 2). Two genes (*BTG2* and *DTX1*) were excluded as they are known targets of the SHM machinery[21]. The remaining genes included most of the drivers previously described by different WES studies[9,11,13]. The most frequently mutated gene in CLL was *NOTCH1* (57 cases, 12.6%), followed by *ATM* (11%), *SF3B1* (8.6%), *BIRC3* (8.8%), *CHD2* (6%), *TP53* (5.3%) and *MYD88* (4%). Furthermore, we identified 12 novel genes recurrently mutated in CLL and not previously linked to this disease, including *ZNF292*, *ARID1A*, ZMYM3 and *PTPN11*. Most CLL driver genes were preferentially mutated in IGHV-UNMUT tumours and had subclonal mutations[11] (Supplementary Fig. 1). Notably, a similar frequency of mutated drivers was found in CLL and MBL cases of similar IGHV gene SHM status (Extended Data Table 2).

We also identified some genes (tier 3) that probably contain driver mutations but were found in three or less CLL patients. This is the case of activating mutations in the oncogenes *KRAS* and *NRAS*, truncating

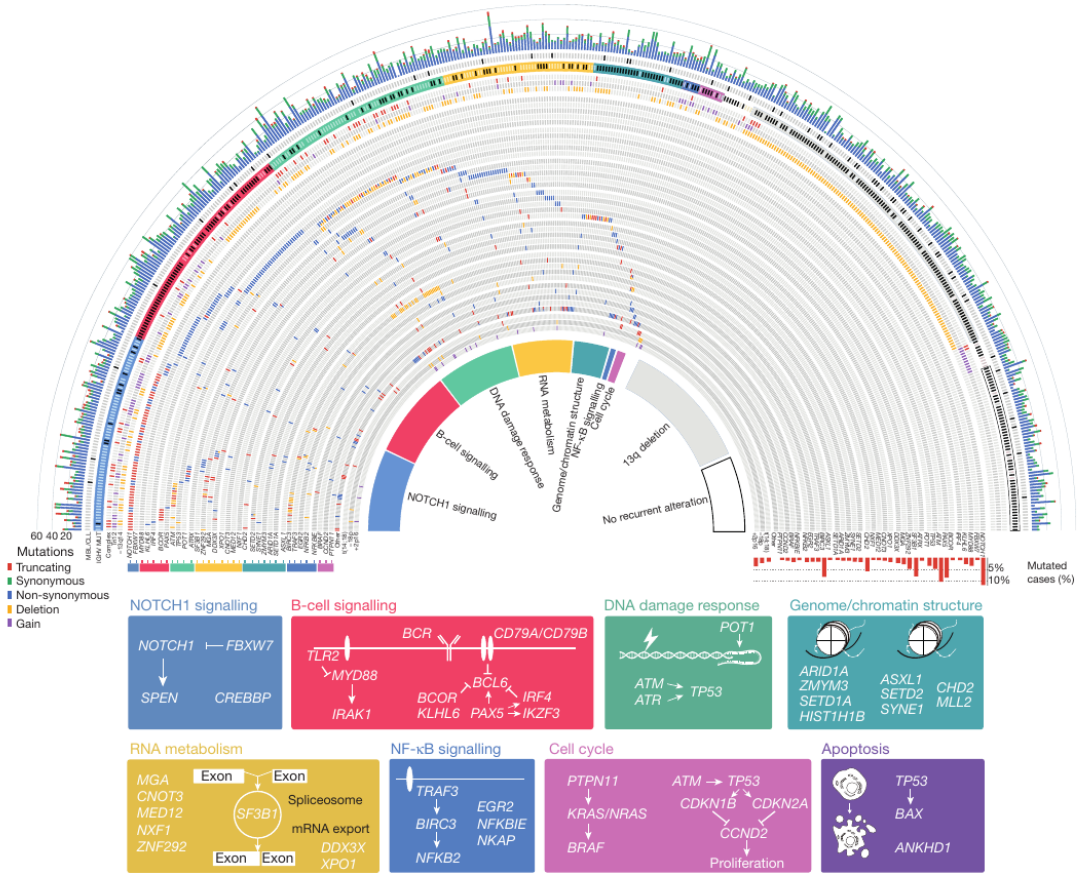


Figure 1 | Significantly mutated genes and pathways. The outer layer represents the number of truncating, non-synonymous and synonymous mutations for 506 CLL (grey) and MBL (black) cases. Clinical classification as well as IGHV-status is shown on the two outermost layers. Inner layers show the most recurrently mutated genes grouped according to the biological pathways involved (bottom). The percentage of cases with mutations in each gene is shown on the right.

mutations in the tumour suppressors *CDKN1B* and *CDKN2A*, and recurrent mutations in the transcription factor *IKZF3*. Mutations in components of the BCR and Toll-like receptor pathway were exclusively present in IGHV-MUT tumours. They included those in *MYD88*, *CD79A*, *CD79B*, *TLR2* and *IRAK1*, detected in 22 of the 278 IGHV-MUT cases, but in none of the 166 IGHV-UNMUT CLL patients ($P = 4.1 \times 10^{-5}$), confirming the importance of the BCR and Toll-like receptor pathways both in CLL pathobiology and as therapeutic targets[23]. Collectively, eight main pathways are frequently altered in CLL, including BCR signalling, cell cycle regulation, apoptosis, DNA damage response, chromatin remodelling, NF-κB signalling, NOTCH1 signalling, and RNA metabolism (Fig. 1).

DNA structural alterations

Analysis of structural variants confirmed the presence of known CNAs such as loss of 13q14, 11q22-q23, 17p, 6q15-q21 and trisomy 12 (Extended Data Fig. 2 and Supplementary Table 4). In addition, we identified novel candidate CLL driver genes in regions of recurrent chromosomal alterations (Fig. 1). They included deletions involving *ZNF292* at 6q15 (2.4%), deletions of 2q37 encompassing *SP140* and *SP110*, loss of 3p21 (2%) affecting *SMARCC1* and *SETD2*, and loss of 10q24 (1.8%) involving *NFKB2* (Supplementary Fig. 2).

Unlike other B-cell malignancies, translocations involving *IG* genes were uncommon in CLL with the exception of *BCL2* rearrangements (10 cases). They occurred exclusively in IGHV-MUT cases, and resulted in overexpression of *BCL2* and recruitment of the SHM machinery (Extended Data Fig. 3). Analysis of WGS data using SMUFIN[24] also revealed the presence of 147 interchromosomal translocations in 43 out of 148 cases (Supplementary Table 5). Recurrent translocations involving chromosome 13q14 with different chromosomal partners and associated with deletion or disruption of the microRNA cluster miR-15a/miR-16 were identified in nine cases ($P < 10^{-8}$). We also detected 15 non-recurrent chromosomal translocations, one of them involving the *IG* locus (*IGH-CBFA2T3*), and 14 predicted to originate in chimaeric genes, five of which could be confirmed by RNA-seq (Supplementary Table 5).

Complex rearrangements (chromothripsis/chromoplexy)[25,26] were identified in 15 out of 452 CLL cases (Extended Data Fig. 3), being more frequent in IGHV-UNMUT than in IGHV-MUT tumours (6% versus 1.8%, $P < 0.05$). Although these complex alterations did not result in any recurrent rearrangement, we observed involvement of chromosome 13 in 4 out of 15 tumours, resulting in *mir-15a/mir-16* loss. Similar to previous studies[27], mutations in *TP53* were more frequent in tumours with chromothripsis (26% versus 4.6%, $P < 0.006$). Furthermore, *SETD2* inactivation was more frequent in CLL cases with chromothripsis than in non-chromothriptic cases (26% versus 1.4%, $P < 2 \times 10^{-4}$).

This analysis revealed significant relationships between several alterations, including co-occurrence of *NOTCH1* mutations and chromosome 12 trisomy[28], trisomy 12 with trisomy 18 ($q < 0.01$), and the mutually exclusive pattern of 13q14 deletion and trisomy 12 ($q < 0.01$). We also observed a higher co-occurrence of mutations in *NOTCH1* with those in *MGA* ($q < 0.01$), *BCOR* ($q < 0.01$) and *BIRC3* ($q < 0.05$), or gain of 2p16 with loss of 18p ($q < 0.01$), among others (Supplementary Fig. 3).

Mutations in non-coding regions

The presence of functional mutations outside of protein-coding regions remains an open question in cancer research[29]. We observed in one CLL case a previously described mutation in the *TERT* promoter (C228T)[29]. Eight mutations in *mir-142* were identified in five cases (Supplementary Fig. 4), with seven of them within AID target consensus (WRCY or WA), reinforcing it as a target of the SHM[30]. We also identified 88 mutations in non-coding regions present in at least two WGS cases (Supplementary Table 6). Most of them were located either within hypermutated late-replication regions[31], or within the 5′-region of *BACH2*, *BCL6*, *BTG2*, *CXCR4* and *TCL1A*, genes known to undergo SHM during the germinal centre reaction[21,22]. Most mutations were within the AID target sequence (WRCY), probably reflecting the passage of the respective progenitor cells through the germinal centre.

Notably, the most frequent recurrent non-coding mutation was detected in the 3′ UTR of *NOTCH1* (chr9: 139390152T > C), present in 4 of the 150 cases with WGS data (Fig. 2a). Sequencing of this region in 356 cases with only WES data revealed seven additional tumours with the same mutation, and two cases with a mutation seven or nine bases downstream of the original one. RNA-seq from six of these 3′ UTR *NOTCH1*-mutated tumours confirmed the presence of a novel splicing event within the last exon of *NOTCH1* (Fig. 2a), which was absent in 290 tumours without these mutations (Extended Data Fig. 4). This splicing event occurred preferentially between a cryptic donor site located in the coding region of the last exon of *NOTCH1* and a newly created acceptor site in the 3′ UTR, resulting in a deletion that includes the last 158 coding bases. Nevertheless, some splicing events occurred between the canonical donor site on exon 33 and the newly created acceptor site in the 3′ UTR of exon 34 (Fig. 2a). Reverse transcription PCR (RT–PCR) analysis confirmed the presence of this aberrant splicing only in cases with mutations in the 3′ UTR (Fig. 2b). This within-exon splicing is predicted to remove a PEST domain of NOTCH1 and to increase protein stability, as in the previously

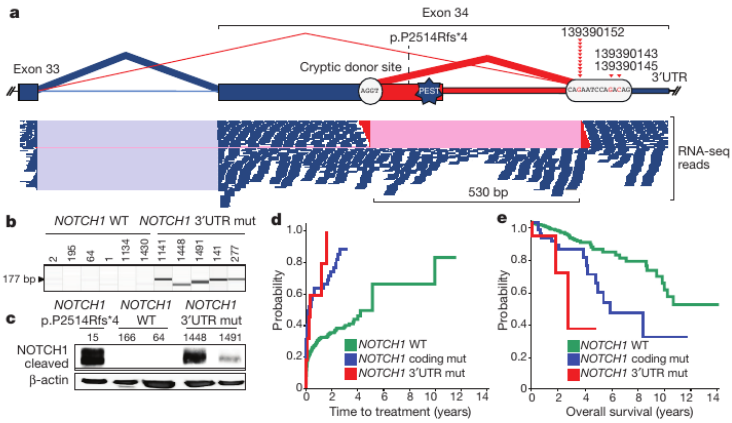


Figure 2 | Activating mutations in the 3′ UTR non-coding region of *NOTCH1*. **a**, Mutant bases are shown in red and the number of cases denoted by arrowheads. Aberrant spliced reads detected by RNA-seq (red) are shown below. **b**, RT–PCR amplification shows the expected 177-base-pair (bp) band in tumours with the recurrent 139390152T > C mutation, and a smaller one in cases with the 139390145 and 139390143 mutations. WT, wild type. **c**, Western blot analysis showing the accumulation of a lower molecular mass NOTCH1 protein in CLL cells with the p.P2514Rfs*4 or the 3′ UTR mutation. β-actin was used as loading control. **d**, **e**, Kaplan–Meier plot of time-to-treatment (**d**) or overall survival (**e**) of CLL patients grouped on the basis of mutations in the 3′ UTR of *NOTCH1*, the presence of *NOTCH1* coding mutations, or *NOTCH1* wild type.

described p.P2514Rfs*4 *NOTCH1* mutation[10]. Western blot analysis confirmed the presence of a smaller molecular mass band in 3′-UTR- and p.P2514Rfs*4-mutated cells, which was absent in cells without mutations in *NOTCH1* (Fig. 2c). Immunohistochemical analysis showed a strong NOTCH1 nuclear signal in tumour cells from patients with 3′ UTR or p.P2514Rfs*4 mutations (Extended Data Fig. 4). All cases with mutations in the 3′ UTR of *NOTCH1* belonged to the IGHV-UNMUT subgroup, accounting for up to 6.7% (12 out of 179) of all IGHV-UNMUT cases. Patients with 3′ UTR *NOTCH1* mutations had features of adverse prognosis (Extended Data Fig. 4) and behaved similarly to patients with coding mutations in *NOTCH1* in terms of the time to first treatment (TTT) and overall survival (Fig. 2d, e).

We further explored the presence of genome regions with high mutational density and found 24 loci enriched in somatic mutations (Fig. 3a). Most of them correspond either to recurrently mutated genes in CLL or to known targets of the SHM process. However, we identified a densely mutated cluster in a small intergenic region of chromosome 9p13, in which 17 different tumours had somatic mutations (Fig. 3b). This region is enriched for both lymphocyte-specific transcription factor binding sites and histone marks related to enhancer elements only in a lymphoblastoid B-cell line (Supplementary Fig. 5). DNase-seq and chromatin immunoprecipitation sequencing (ChIP-seq) analysis in normal B cells and CLL cases revealed that the region contains an active enhancer characterized by a DNase I hypersensitive site and nucleosomes containing histone 3 Lys4 methylation (H3K4me1) and H3K27 acetylation (H3K27ac) (Fig. 3b and Supplementary Fig. 5). Chromosome conformation capture sequencing (4C-seq) analysis[32] in tumour cells from two CLL patients revealed that this potential enhancer shows high three-dimensional contact frequencies extending towards the telomere up to the *PAX5* locus, located 330 kilobases (kb) away (Fig. 3c and Supplementary Fig. 5). Expression analysis of 15 genes located within 1 Mb of this element revealed that the only gene showing a significant expression difference correlated with the presence of mutations within the putative enhancer region was indeed *PAX5* (average expression 87 versus 131, $P = 1.9 \times 10^{-4}$) (Extended Data Fig. 5). *PAX5* encodes a transcription factor that has an essential role in B-cell differentiation[33] and, based on the evidence provided above, is the most likely target of the identified enhancer region. CRISPR/Cas9-based genome editing of this region allowed us to demonstrate that either the introduction of a specific point mutation, or the deletion of this putative enhancer in a lymphoblastoid B-cell line or in RAMOS cells, resulted in a 40% reduction in the expression of *PAX5* (Extended Data Fig. 6).

Sequencing of this region in all CLL cases with WES data identified 25 new cases with somatic mutations. We also found somatic mutations in this enhancer in diffuse large B-cell lymphomas (29%, 26 out of 89), follicular lymphomas (23%, 20 out of 86) and mantle-cell lymphomas (5%, 3 out of 66) (Supplementary Table 7). Interestingly, 84% of CLL cases with mutations in this enhancer belong to the IGHV-MUT subgroup, accounting for up to 13% of IGHV-MUT CLL cases. Mutations in the *PAX5* enhancer were the only recurrent alteration observed in 7 cases, while in 11 tumours this alteration was only combined with 13q14 deletion, raising the possibility that *PAX5* enhancer mutations might constitute driver events contributing to the development of these tumours.

Integrative analysis

We then integrated the standard genetic classification of CLL with a recent patient categorization in three subgroups based on a DNA methylation signature of naive and memory B cells[17,34] (Supplementary Table 1). The three epigenetic subgroups showed a distinct distribution of genetic changes, IGHV gene repertoire and stereotyped B-cell receptors (Extended Data Fig. 7). The intermediate group had moderate IGHV mutation levels, an intermediate contribution of signature 2 mutations, higher frequencies of *SF3B1* and *MYD88* mutations, biased usage of the IGHV-3-21 and IGHV-1-18 genes

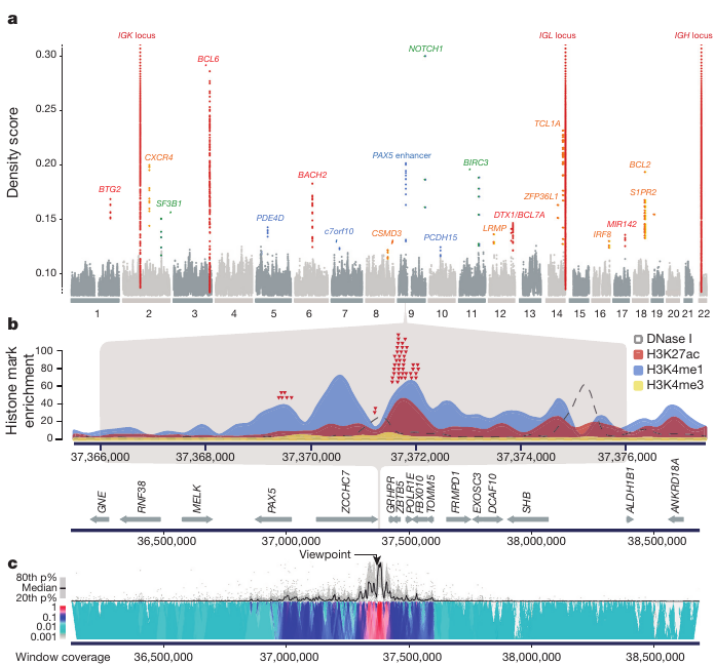


Figure 3 | Identification of somatic mutations in a *PAX5* enhancer. **a**, Regions with a high density of somatic mutations in 150 WGS analyses. Regions correspond to recurrently mutated genes (green), targets of SHM (red/orange), and other regions (blue). **b**, Detailed view of a 9p13 region showing the accumulation of somatic mutations (arrowheads) in CLL tumours as well as DNase I hypersensitivity and histone H3K27ac, H3K4me1 and H3K4me3 enrichment from CLL tumour 110. **c**, 4C-seq analysis in CLL cells showing the interaction frequencies of the enhancer with the surrounding regions. p%, percentile.

and increased frequency of stereotyped subset #2. These results support the hypothesis that this group has a distinct genetic and epigenetic makeup[17,34–36]. We also found a highly significant correlation ($r = 0.64$, $P < 0.001$) between the number of WGS mutations per case and the number of CpGs showing differential methylation as compared to naive B cells (Extended Data Fig. 7). Similarly, the proportion of signature 2 mutations was also correlated with differential methylation in IGHV-MUT cases.

MBL cases were indistinguishable at the genomic, transcriptomic and epigenomic level from CLL cases assigned to the same IGHV subgroup (Extended Data Fig. 7 and Extended Data Table 2), in accordance with the overlapping biological features of both processes. Notably, the burden of driver alterations was significantly lower in patients with MBL than with CLL (1.2 versus 1.7, for IGHV-MUT cases, $P = 8 \times 10^{-4}$), consistent with a model in which MBL/CLL evolution is accomplished by the progressive accumulation of driver alterations.

Clinical implications

Our data support the hypothesis that the observed genomic differences between the two major molecular subgroups of CLL might be in part responsible for their different outcome. The average number of driver mutations in IGHV-UNMUT tumours was higher than in IGHV-MUT cases (3.5 versus 1.7, $P < 10^{-19}$), despite the 44% higher mutational burden of IGHV-MUT tumours. We found that 88% of cases had at least one driver mutation, with almost all IGHV-UNMUT tumours containing at least one driver alteration, while a smaller fraction was found in the IGHV-MUT subgroup (96% versus 83%, $P < 5 \times 10^{-5}$).

We evaluated the influence of the presence of each alteration on the TTT and overall survival from the time of sampling. The mutation of several drivers and CNAs was significantly correlated with an adverse prognosis, in some cases independently from Binet stage and IGHV mutational status (Fig. 4a, Extended Data Fig. 8 and Supplementary Table 8). We confirmed the independent prognostic value of known gene mutations (*SF3B1* and *TP53*), and identified novel independent prognostic drivers for both shorter TTT (*BRAF*, *ZMYM3*, *IRF4*, *NFKB2*, 20p deletion, and 2p16 and 5q34 gains), and overall survival (*ASXL1*, *POT1* and 14q24 deletion). Remarkably, the accumulative number of drivers (0 to ≥4) per tumour had a progressively worse effect on outcome that could discriminate patient subsets differing by more than 10 years in the median TTT, independently of IGHV status and Binet stage. They also showed prognostic value for overall survival, although not independent in the multivariate analysis (Fig. 4b, c). Finally, we examined the potential druggability of the alterations in genes and pathways identified in CLL patients[37], finding candidate drugs for 19 of the 59 driver genes in 42% of the CLL cases (190 out of 452) (Supplementary Fig. 6 and Supplementary Tables 9 and 10).

Discussion

In this work, we have provided a comprehensive and integrated molecular characterization of CLL. We have also unveiled new biological aspects of this disease and identified novel driver genes presumably implicated in its pathogenesis. The large number of different genomic alterations found in our cohort illustrates the enormous biological heterogeneity of CLL. Notably, the use of WGS has allowed us to identify recurrent mutations in non-coding regions, including the 3′ UTR of *NOTCH1* and a *PAX5* enhancer, resulting in marked alterations in the activity of these transcription factors of well-known importance in leukaemia and other malignancies[38,39]. Previous studies have shown the effect of *NOTCH1* mutations in CLL prognosis[10,40]. However, these studies may seriously underestimate the true incidence of *NOTCH1* deregulation in CLL, given our finding that about 20% of *NOTCH1*-mutated tumours contain mutations in the 3′ non-coding region. These findings emphasize the value of large genome-wide studies to discover new molecular alterations that may have a profound effect on cancer development and progression.

The evaluation of putative associations between these molecular alterations and the clinicopathological features of our cohort of CLL patients has been challenging owing to the low frequency of many significantly mutated genes. Patients in which no recurrent alterations

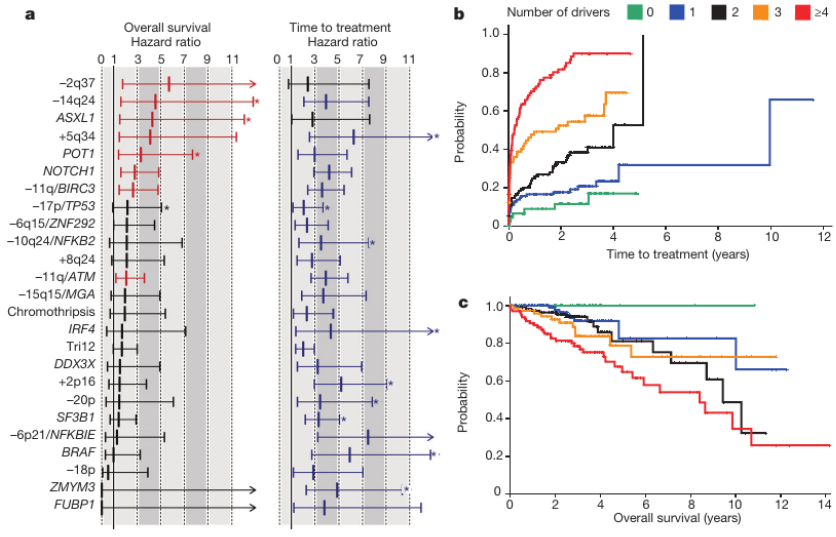


Figure 4 | Prognostic effects of individual alterations and number of drivers. **a**, Effect on overall survival (left) and time-to-treatment (right) for each genomic alteration. Labels including genes and chromosomal regions represent combined analysis of mutations and copy number alterations. Hazard ratios and 95% confidence intervals are shown. Alterations conferring statistically significant (adjusted $P < 0.05$) hazard ratios are shown in colour (red for overall survival and blue for TTT), and those in which the effect was independent of Binet stage and IGHV-status are labelled with an asterisk. **b**, **c**, Kaplan–Meier plots of TTT (**b**) or overall survival (**c**) of CLL patients grouped by the number of driver mutations identified.

were found had the best prognosis and near normal overall survival, suggesting that this study has uncovered most driver alterations involved in CLL evolution, opening new avenues to explore the clinical impact of the heterogeneous molecular composition of the disease in independent cohorts. Hopefully, this work will finally result in new opportunities for improving the clinical management and personalized treatment of CLL patients.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 February; accepted 15 June 2015.
Published online 22 July 2015.

1. Gaidano, G., Foa, R. & Dalla-Favera, R. Molecular pathogenesis of chronic lymphocytic leukemia. *J. Clin. Invest.* **122,** 3432–3438 (2012).
2. Zenz, T., Mertens, D., Kuppers, R., Dohner, H. & Stilgenbauer, S. From pathogenesis to treatment of chronic lymphocytic leukaemia. *Nature Rev. Cancer* **10,** 37–50 (2010).
3. Pekarsky, Y., Zanesi, N. & Croce, C. M. Molecular basis of CLL. *Semin. Cancer Biol.* **20,** 370–376 (2010).
4. Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G. & Stevenson, F. K. Unmutated Ig V_H genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* **94,** 1848–1854 (1999).
5. Damle, R. N. *et al.* Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* **94,** 1840–1847 (1999).
6. Crespo, M. *et al.* ZAP-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia. *N. Engl. J. Med.* **348,** 1764–1775 (2003).
7. Malek, S. N. The biology and clinical significance of acquired genomic copy number aberrations and recurrent gene mutations in chronic lymphocytic leukemia. *Oncogene* **32,** 2805–2817 (2013).
8. Döhner, H. *et al.* Genomic aberrations and survival in chronic lymphocytic leukemia. *N. Engl. J. Med.* **343,** 1910–1916 (2000).
9. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor *SF3B1* gene in chronic lymphocytic leukemia. *Nature Genet.* **44,** 47–52 (2011).
10. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475,** 101–105 (2011).
11. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152,** 714–726 (2013).
12. Fabbri, G. *et al.* Analysis of the chronic lymphocytic leukemia coding genome: role of *NOTCH1* mutational activation. *J. Exp. Med.* **208,** 1389–1401 (2011).
13. Ramsay, A. J. *et al. POT1* mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nature Genet.* **45,** 526–530 (2013).
14. Damm, F. *et al.* Acquired initiating mutations in early hematopoietic cells of CLL patients. *Cancer Discov.* **4,** 1088–1101 (2014).
15. Rossi, D. *et al.* Disruption of *BIRC3* associates with fludarabine chemorefractoriness in *TP53* wild-type chronic lymphocytic leukemia. *Blood* **119,** 2854–2862 (2012).
16. Ferreira, P. G. *et al.* Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res.* **24,** 212–226 (2014).
17. Kulis, M. *et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nature Genet.* **44,** 1236–1242 (2012).
18. Oakes, C. C. *et al.* Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia. *Cancer Discov.* **4,** 348–361 (2014).
19. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
20. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
21. Khodabakhshi, A. H. *et al.* Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* **3,** 1308–1319 (2012).
22. Pasqualucci, L. *et al.* Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* **412,** 341–346 (2001).
23. Byrd, J. C. *et al.* Targeting BTK with ibrutinib in relapsed chronic lymphocytic leukemia. *N. Engl. J. Med.* **369,** 32–42 (2013).
24. Moncunill, V. *et al.* Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nature Biotechnol.* **32,** 1106–1112 (2014).
25. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
26. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153,** 666–677 (2013).
27. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. *Cell* **148,** 59–71 (2012).
28. Balatti, V. *et al. NOTCH1* mutations in CLL associated with trisomy 12. *Blood* **119,** 329–331 (2012).
29. Huang, F. W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339,** 957–959 (2013).
30. Yamane, A. *et al.* Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. *Nature Immunol.* **12,** 62–69 (2011).
31. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
32. Simonis, M., Kooren, J. & de Laat, W. An evaluation of 3C-based methods to capture DNA interactions. *Nature Methods* **4,** 895–901 (2007).
33. Revilla-i-Domingo, R. *et al.* The B-cell identity factor Pax5 regulates distinct transcriptional programmes in early and late B lymphopoiesis. *EMBO J.* **31,** 3130–3146 (2012).
34. Queirós, A. C. *et al.* A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia* **29,** 598–605 (2015).
35. Strefford, J. C. *et al.* Distinct patterns of novel gene mutations in poor-prognostic stereotyped subsets of chronic lymphocytic leukemia: the case of *SF3B1* and subset #2. *Leukemia* **27,** 2196–2199 (2013).
36. Agathangelidis, A. *et al.* Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. *Blood* **119,** 4467–4475 (2012).
37. Rubio-Perez, C. *et al. In silico* prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* **27,** 382–396 (2015).
38. Lobry, C., Oh, P. & Aifantis, I. Oncogenic and tumor suppressor functions of Notch in cancer: it's NOTCH what you think. *J. Exp. Med.* **208,** 1931–1935 (2011).
39. O'Brien, P., Morin, P. Jr, Ouellette, R. J. & Robichaud, G. A. The Pax-5 gene: a pluripotent regulator of B-cell differentiation and cancer disease. *Cancer Res.* **71,** 7345–7350 (2011).
40. Villamor, N. *et al. NOTCH1* mutations identify a genetic subgroup of chronic lymphocytic leukemia patients with high risk of transformation and poor outcome. *Leukemia* **27,** 1100–1106 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was funded by Spanish Ministry of Economy and Competitiveness through the Instituto de Salud Carlos III (ISCIII) and Red Temática de Investigación del Cáncer (RTICC). We are grateful to E. Santos for his continued support to this project, and N. Villahoz and M. C. Muro for their excellent work in the coordination of the CLL Spanish Consortium. C.L.-O. is an Investigator of the Botin Foundation supported by Banco Santander through its Santander Universities Global Division, and E.Ca. and D.T. are Institució Catalana de Recerca i Estudis Avançats-Academia investigators. We acknowledge Partnership for Advanced Computing in Europe (PRACE) for awarding us access to resource Marenostrum based in Spain at the BSC, the Pershing Square Sohn Cancer Research Alliance and European Union's FP7 through the Blueprint Consortium. We are also very grateful to all patients with CLL who have participated in this study.

Author Contributions The Chronic Lymphocytic Leukaemia Genome consortium contributed to this study as part of the International Cancer Genome Consortium. Investigator contributions are as follows: T.B., J.D., A.L.-G., A.R.P., M.G. and J.M.H.-R. contributed to sample collection and clinical annotation; M.R., N.V., E.Ca., E.Co., J.M.H.-R. and M.G. were the pathologists who reviewed and confirmed the diagnoses; P.N., C.M.R.-C. and M.A. prepared and supervised the bioethical requirements; M.P., A.E. and C.R. processed samples and performed validation analysis; M.G., I.G. and D.A.P. were responsible for generating libraries, performing exome capture and sequencing; S.B., D.To., M.M., S.G., I.S., G.C., D.M.-G., A.C., X.E. and D.Cos. analysed copy number alterations and structural variants; X.S.P., R.V.-M., J.G.-A. and V.Q. developed the bioinformatic pipeline for analysis of somatic mutations and performed functional data integration; D.Col., M.L.-G. and B.G. were responsible for downstream validation analysis and functional studies; A.N. and K.S. analysed *IG* gene rearrangements and stereotypes; J.I.M.-S., A.C.Q., G.C., R.B., R.G., N.R., H.G.S. and P.J. performed epigenetic and transcriptomic analysis and 4C-seq experiments; L.B. and A.A.F. performed enhancer analysis and CRISPR experiments; N.V., T.B., A.L.-G. and E.Ca. performed clinical and biological studies; J.L.G., R.R., M.O., D.G.P. and A.V. were in charge of bioinformatics data management; N.L.-B., C.R.-P. and D.Ta. contributed to pathway analysis and *in silico* prescription. X.S.P., C.L.-O. and E.Ca. directed the research, analysed the data and wrote the manuscript.

Author Information Sequencing, expression and genotyping array data have been deposited at the European Genome-Phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted at the European Bioinformatics Institute (EBI), under accession number EGAS00000000092. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.L.-O. (clo@uniovi.es) or E.C. (ecampo@clinic.ub.es).

METHODS

Patients. The clinical and biological characteristics of the 506 patients are shown in Extended Data Table 1. Among these patients, 452 were diagnosed with CLL and 54 with MBL. Cases were defined as IGHV-MUT when the identity of immunoglobulin genes was less than 98%. The tumour samples were obtained before administration of any treatment. All patients gave informed consent for their participation in the study following the International Cancer Genome Consortium (ICGC) guidelines and the ICGC Ethics and Policy committee[19].

Collection and preparation of samples. Tumour samples were obtained from fresh or cryopreserved mononuclear cells. To purify the CLL or MBL fraction, samples were incubated with a cocktail of magnetically labelled antibodies directed against T cells, natural killer cells, monocytes and granulocytes (CD2, CD3, CD11b, CD14, CD15 and CD56), adjusted to the percentage of each contaminating population (AutoMACS, Miltenyi Biotec). The degree of contamination by non-CLL cells in the CLL fraction was assessed by immunophenotype and flow cytometry. DNA was extracted from purified samples by using a Qiagen kit, and the quality of purified DNA was assessed by SYBR-green staining on agarose gels and quantified using a Nanodrop ND-100 spectrophotometer. The tumour DNA and RNA samples for further genomic analysis contained ≥95% neoplastic cells and the contamination by neoplastic cells in normal DNA was <2%.

WGS, WES and RNA-seq. For WGS, 2 μg of genomic DNA from each sample was used for the construction of two short-insert paired-end sequencing libraries. One library was prepared using a standard TruSeqDNA Sample Preparation Kit v2 (Illumina Inc.) with some modifications. In short, following the fragmentation (CovarisE220) the libraries were size-selected on the agarose gel and processed through end-repair, adenylation and indexed adaptor ligation. The gel eluate was directly amplified by 10 PCR cycles. The second library was prepared following the same protocol as above, however, it included a heating step to 72 °C before adaptor ligation and was suddenly cooled down to 4 °C. This resulted in a biased proportion of high GC content reads and counterbalanced some of Illumina's PCR sample preparation methods' GC-bias, thus improving coverage of increased GC-content regions of the genome. Both types of libraries were sequenced in paired-end mode on Illumina GAIIx (2 × 151 bp) using Sequencing kit v4 or Illumina HiSeq2000 (2x101 bp) using TruSeq SBS Kit v3 (Illumina Inc.).

For other samples (Supplementary Table 1), the library preparation procedure was modified to remove the PCR step during short-insert paired-end library preparation. The TruSeq DNA Sample Preparation Kit v2 (Illumina Inc.) and the KAPA Library Preparation kit (Kapa Biosystems) were used. In brief, 2 μg of genomic DNA was sheared on a Covaris E220, size-selected and concentrated using AMPure XP beads (Agencourt, Beckman Coulter) to reach the fragment size of 220–480 bp. Fragmented DNA was end-repaired, adenylated and ligated to Illumina specific indexed paired-end adaptors. All libraries were quantified by Library Quantification Kit (Kapa Biosystems). Each library was sequenced using TruSeq SBS Kit v3-HS (Illumina Inc.), in paired-end mode, 2 × 101-bp, in three sequencing lanes of HiSeq2000 flowcell v3 (Illumina Inc.) according to standard Illumina operation procedures with minimal yield of 85 Gb for each sample. Primary data analysis was carried out with the standard Illumina software Real Time Analysis (RTA 1.13.48) and followed by generation of FASTQ files.

For WES, 3 μg of genomic DNA from each sample were sheared and used for the construction of a paired-end sequencing library as described in the paired-end sequencing sample preparation protocol provided by Illumina[41]. Enrichment of exonic sequences was then performed for each library using either the Sure Select Human All Exon 50 Mb or All Exon+UTRs v4 kits (Supplementary Table 1) following the manufacturer's instructions (Agilent Technologies). Exon-enriched DNA was pulled down by magnetic beads coated with streptavidin (Invitrogen), followed by washing, elution and 18 additional cycles of amplification of the captured library. Enriched libraries were sequenced (2 × 76 bp) in one lane of an Illumina GAIIx sequencer or in two lanes of a HiSeq2000 when using pools of eight samples.

RNA was assayed for quantity and quality using Qubit RNA HS Assay (Life Technologies) and RNA 6000 Nano Assay on a Bioanalyzer 2100. RNA-seq libraries were prepared from total RNA using the TruSeq RNA Sample Prep Kit v2 (Illumina Inc.) with minor modifications. In brief, 0.5 μg of total RNA was used as the input material for poly-A-based messenger RNA enrichment with oligo-dT magnetic beads. Selected mRNA was fragmented (resulting RNA fragment size was 80–250 nucleotides, with the major peak at 130 nucleotides). After first and second strand cDNA synthesis the double-stranded complementary DNA was end-repaired, 3′ adenylated and the 3′ 'T' nucleotide of the adaptor was used for the Illumina indexed adapters ligation. The ligation product was enriched by 10 cycles of PCR. Each library was sequenced using TruSeq SBS Kit v3-HS, in paired-end mode with a read length of 2 × 76 bp. We generated more than 20 million paired-end reads for each sample in a fraction of a sequencing lane on HiSeq2000 (Illumina Inc.) following the manufacturer's protocol. Image analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (RTA 1.13.48) and followed by generation of FASTQ sequence files.

Read mapping and processing. For WGS and WES, reads from each library were mapped to the human reference genome (GRCh37) using BWA[42] with the same option, and a BAM file was generated using SAMtools[43]. Reads from the same paired-end libraries were merged, and optical or PCR duplicates were flagged using Picard (http://picard.sourceforge.net/index.shtml). For the identification of somatic substitutions and indels, we used the Sidrón algorithm[9,44]. This algorithm was adapted to identify subclonal mutations in which the mutant allele fraction is low, but supported by at least three reads. Visual inspection of recurrent mutational hotspots allowed the inclusion of some somatic mutations that were originally discarded owing to the presence of an excess of mutant reads in the non-tumour sample, or owing to low coverage, especially in the case of *NOTCH1*, in which a high GC content on exon 34 usually resulted in very low coverage by WES. In samples in which *NOTCH1* coverage was too low to make a call, mutations were analysed by Sanger sequencing. A comparison of mutation calls by Sidrón and by Sanger sequencing of some of the most frequently mutated genes in CLL (*SF3B1*, *TP53*, *MYD88*) revealed more than 97% specificity and at least 90% sensitivity. Mutational signatures were extracted using the WTSI Mutational Signature Framework[45]. To estimate the presence of subclonal mutations in recurrently mutated genes, the fraction of reads supporting a mutant allele was calculated for those mutations in which the depth of coverage was at least 20 reads. Flow cytometry analysis confirmed that the percentage of tumour cells was at least 98%. A case was considered as having a clonal mutation when at least 80% of cells were estimated to contain the mutation, and the mutant allelic fraction was within the 95% confidence interval.

Analysis of CNAs and structural variants. For the identification of CNAs, tumour and normal DNA from 505 CLL patients were analysed using Affymetrix SNP6.0 microarrays (Affymetrix) as previously described[46]. SNP array experiments were carried out at CeGen (http://www.cegen.org). Additionally, for 230 cases array-comparative genomic hybridization was performed in SurePrint G3 Human aCGH Microarray 1M (Agilent Technologies). Array-comparative genomic hybridizations were performed at qGenomics (http://www.qgenomics.com). Nexus 6.0 Discovery Edition software (Biodiscovery) was used for global analysis and visualization. Copy number neutral loss of heterozygosity was considered when the size of alteration was larger than 5 Mb. Acquired copy number neutral loss of heterozygosity was observed in 28 regions, 16 of them affecting known driver genes that already contained mutations, resulting in homozygous deletion of *mir-15a/mir-16* at 13q14, or inactivation of *ATM* and *TP53* (Supplementary Table 4). According to the literature, the presence of chromothripsis was considered when at least seven switches between two or more copy number states were detected on an individual chromosome in which LOH was retained, and chromoplexy was defined when at least three chained chromosomal rearrangements were detected in a tumour[27,47]. In one case in which genotyping data were not available, we used exome2cnv[48] to identify CNAs from WES data.

For the identification of breakpoints in WGS derived from structural variants, we used SMUFIN[24], a program that directly compares sequence reads from normal and tumour samples, to identify chromosomal breakpoints corresponding to large structural variants at base-pair resolution. We analysed 150 tumour/normal whole-genome pairs setting the cross-sample contamination filter to 5%. Two WGS tumours (019 and 029) showed an abnormal number of breakpoints owing to the presence of sequence lanes with high error rates that interfere with SMUFIN and were not considered for this analysis. All predicted breakpoints that were not confirmed through the BAM file after manual inspection were systematically discarded. A total of 48 out of 53 (91%) selected predicted breakpoints could be verified using PCR amplification followed by Sanger sequencing (Supplementary Table 5). This verification rate is similar to the one observed in our initial description of the method[24]. In addition, custom scripts were used to identify potential translocations involving immunoglobulin genes either in WGS or WES. This resulted in the identification of ten cases (5 WGS and 5 WES) containing putative translocations with the *BCL2* locus (nine with the t(14;18)(q32;q21), and one with the t(2;18)(p11;q21) translocation), all of which were confirmed by either Fluorescence *in situ* hybridization (FISH), cytogenetics or PCR (Extended Data Fig. 3).

G-banding and FISH analysis. Conventional cytogenetics was performed on Giemsa-banded chromosomes (G-banding) obtained after a 72-h culture and stimulation with tetradecanoyl-phorbol-acetate. At least 20 G-banded metaphases per sample were analysed. Results were described according to the International System for Human Cytogenetic Nomenclature. FISH analyses on fixed cells were performed using probes that interrogated for 11q23/*ATM*, 13q14.3 and 17p13/*TP53* deletions and trisomy 12 (Abbott Molecular). Two hundred nuclei were examined for each probe. LSI *IGH/BCL2* dual colour fusion for the t(14;18)(q32;q21) (Abbot Molecular) was used to confirm *BCL2*

rearrangements detected by WGS and WES. Additionally, in case 853, whole chromosomal paintings of chromosomes 8, 11 and X were performed to determine the complex karyotype (with four derivative chromosomes), and rearrangements predicted by SMUFIN algorithm.

Analysis of DNA methylation. DNA methylation was analysed using the 450k Human Methylation Array (Illumina). We used the EZ DNA Methylation Kit (Zymo Research) for bisulphite conversion of 500 ng of genomic DNA, and the Infinium methylation assay was carried out as described by the manufacturer[49,50]. These array experiments were performed at CeGen (http://www.cegen.org). Data from the 450k Human Methylation Array were analysed in R using the minfi package (version: 1.6.0)[51], available through the Bioconductor open source software, applying several custom filters. Unsupervised analyses were performed by principal component analysis and differential methylation between individual CLL/MBL samples and controls was detected using an absolute difference of 0.25.

Gene expression profiling. We studied the gene expression profiling of 468 cases using highly purified leukaemic CLL cells. Total RNA was extracted with the TRIzol reagent following the recommendations of the manufacturer (Invitrogen Life Technologies). RNA integrity was examined with the Agilent 2100 Bioanalyzer (Agilent Technologies) and only high-quality RNA samples were hybridized to Affymetrix Human Genome Array U219 array plates according to Affymetrix standard protocols. Summarized expression values were computed using the robust multichip average approach implemented in the Expression Console Software (Affymetrix Inc.).

RT–PCR. cDNA was synthesized from 500 ng of total RNA using High Capacity RNA-to-cDNA kit (Life Technologies) following the manufacturer's instructions. Amplification was performed using 50 ng of DNA using Qiagen Multiplex PCR Kit (Qiagen), and the reaction mix contained 1× Qiagen Multiplex PCR Master Mix (12.5 µl), primer mix (0.4 µM of each primer) and RNase-free water for a total reaction volume of 25 µl. For *NOTCH1* within-intron splicing, primers used were: forward 5′-CCTAACAGGCAGGTGATGCT-3′ and reverse 5′-TACTC CTCGCCTGTGGACAA-3′. PCR amplification was performed for *NOTCH1* 3′ UTR forward primer 5′-CCTAACAGGCAGGTGATGCT-3′ and reverse primer 5′-ATCTGGCCCCAGGTAGAAAC-3′, *PAX5* enhancer first region forward 5′-TAGATTGTGCCGAATGCTGA-3′ and primer 5′-ACAAGCTCT CCTCCCAGGAA-3′, and *PAX5* enhancer second region forward primer 5′-AGGATGAGAACGGGCAAAC-3′ and reverse primer 5′-GGAGCTTCCA GCTGAACTGA-3′. All PCR products were run on a capillary electrophoresis gel (QIAxcel Advanced System, Qiagen) with the QIAxcel DNA screening kit (Qiagen).

Western blot analysis. For western blot analysis, tumour cells were lysed for 30 min in Triton buffer (1% Triton X-100, 50 mM Tris–HCl, pH 7.6, 150 mM NaCl, 1 mM EDTA) supplemented with protease and phosphatase inhibitors (1 mM PMSF, 2 mM sodium pyrophosphate, 2 mM sodium β-glycerophosphate, 1 mM NaF, 1 mM sodium orthovanadate, 10 µg ml^{-1} leupeptin and 10 µg ml^{-1} aprotinin). Lysates were cleared by centrifugation at 15,000*g* at 4 °C for 15 min, and protein concentrations determined using the Bradford method. Thirty micrograms of protein was separated by SDS–PAGE and transferred onto Immobilon-P membranes. Membranes were blocked with 2.5% phospho-Blocker (Cell Biolabs) in TBS-Tween 20. For protein immunodetection, the specific primary antibodies were used: anti-cleaved NOTCH1 (Val1744) (D3B8; Cell Signaling Technology) and β-actin (Sigma). Anti-rabbit and anti-mouse horseradish peroxidase-labelled IgG (Sigma) were used as secondary antibodies. Chemiluminescence was detected by using ECL substrate (Pierce) on a mini-LAS4000 Fujifilm device (GE Healthcare).

Immunohistochemical analysis. NOTCH1 immunohistochemical staining was performed on a Leica Bond system using formalin-fixed paraffin-embedded tissue sections[52]. Samples were pre-treated using heat-mediated antigen retrieval with EDTA buffer (pH 9.0), epitope retrieval solution 2 (HIER2) for 30 min. Then, sections were incubated with anti-cleaved NOTCH1 rabbit monoclonal antibody (clone D3B8, catalogue number 4147, Cell Signaling Technology) at a final concentration of 8.5 µg ml^{-1}, for 60 min at room temperature and detected using a horseradish peroxidase (HRP)-conjugated compact polymer system. DAB was used as the chromogen. The section was then counterstained with haematoxylin and mounted with DPX.

Sanger sequencing. PCR products were treated using ExoSap IT (USB Corporation) and sequenced with ABI Prism BigDye terminator v3.1 (Applied Biosystems) and 5 pmol of each primer. Sequencing reactions were run on an ABI-3730 automated sequencer (Applied Biosystems). All sequences were examined with the Mutation Surveyor DNA Variant Analysis Software (Softgenetics).

ChIP-seq and DNase-seq. ChIP-seq was performed in normal B-cell subpopulations and in cells (>90% tumour cell content) of a CLL patient with mutated IGHV, and DNase-seq only in the latter following standard protocols generated within the Blueprint Consortium. In brief, cells for ChIP-seq were fixed for 8–16 min in 1% formaldehyde at 4 °C, and chromatin was sonicated for 15 min with a Biorruptor (Diagenode). Chromatin fragments ranging from 50 to 500 bp were selected and immunoprecipitation was carried out with antibodies from Diagenode against H3K4me3 (pAb-003-050 lot:A5051-001P), H3K4me1 (pAb-194-050 lot:A1863-001P) and H3K27ac (pAb-196-050 lot: A1723-0041D) using approximately 500,000 cells per antibody. DNase I digestion was performed using 60 units of the enzyme (Sigma) and 2.5 million cells. ChIP-seq and DNase-seq libraries were constructed using the Kapa Hyper Prep Kit (Kapa Biosystems). For each experiment, from 25 to 50 million reads were sequenced with an Illumina HiSeq2000 sequencer. Detailed protocols can be obtained from the Blueprint Consortium (http://www.blueprint-epigenome.eu/index.cfm?p = 7BF8A4B6-F4FE-861A-2AD57A08D63D0B58).

4C-seq. 4C-seq template generation and amplification was performed as previously described[53,54]. In brief, 1×10^7 cells of two CLL patients were crosslinked with 2% formaldehyde (Merck), chromatin was digested with DpnII (New England Biolabs) followed by ligation with T4 ligase (Roche). Next, chromatin was decrosslinked, DNA was digested with Csp6I (NEB) and re-ligated. PCR amplification of viewpoint regions and their ligated fragments was performed using primers 5′-TGCCACACCTCCTTTTGATC-3′ and 5′-CCTTGTGGAAAGAGTCTC AC-3′ (*PAX5* putative enhancer, viewpoint fragment-end chr9:37,370,916-37,371,635) or 5′-CCGAGCTGGGGTAGCTGATC-3′ and 5′-TTGTGTCCA AAAGTTGTTTG-3′ (*PAX5* promoter, viewpoint fragment-end chr9:37,033, 553-37,034,192). Samples were sequenced using a MiSeq instrument (Illumina) using 50-bp single-end reads, and adding 5% PhiX control DNA. Data analysis was performed using 4Cseqpipe version 0.7 (May 2012) (downloaded from http://compgenomics.weizmann.ac.il/tanay/). Before mapping of the interacting regions to the genome, reads that are a consequence of undigested templates or self-ligation of the viewpoint fragment were removed.

Deletion and mutation of human PAX5 enhancer in B-cell lines using CRISPR/Cas9. Human *PAX5* enhancer was deleted or mutated in RAMOS cells and in an Epstein–Barr virus (EBV)-transformed lymphoblastoid B-cell line using CRISPR/Cas9 genome editing. Guide RNAs (gRNAs) were designed using E-CRISP tool (http://www.e-crisp.org/E-CRISP/index.html)[55]. For the deletions, four gRNAs were designed flanking the *PAX5* enhancer, two at each side (L1/L2 and R1/R2) to be used in combinations (L1+R1, L1+R2, L2+R1, and L2+R2). In addition, two gRNAs were designed to target sites of mutations found in CLL (M1/M2). gRNAs sequences are: L1, 5′-GGGAACCAGGGCGTGGGAGC-3′; L2, 5′-GTGAGGCAGAAACACCACAG-3′; R1, 5′-GGCAGCATGCGGGCG TCATG-3′, R2, 5′-GCCAGGACCTGCTCTCCCAA-3′; M1, 5′-GTGAAAATT TACTCATGCTG-3′; and M2, 5′-GGTGGTACTCAGAGGCTGGG-3′. The gRNA oligonucleotides were cloned in pL-CRISPR.EFS.GFP vector (Addgene plasmid 57818)[56], and lentiviral particles were produced on HEK293T cells by cotransfection with Gag-Pol and vesicular stomatitis virus G (VSV-G)-expressing vectors using the JetPEI transfection reagent (Polyplus). Viral supernatants were collected after 48 h and used for infection by spinoculation of Ramos and EBV-transformed lymphoblastoid B cells. After infection, green fluorescent protein (GFP)-positive cells were sorted (BD Influx, BD Bioscience) and grown for 1 week. Total RNA was extracted with TRIzol (Invitrogen) and converted into cDNA with SuperScript First-Strand Synthesis System (Invitrogen). Then, human *PAX5* expression was determined by quantitative real-time PCR (FastStart Universal SYBR Green Master Mix, Roche) using a 7500 Real-Time PCR system (Applied Biosystems). *GAPDH* was used as normalization control. The following primers were used: *PAX5* forward, 5′-GAGCGGGTGTGT GACAATGA-3′; *PAX5* reverse, 5′-GCACCGGAGACTCCTGAATAC-3′; *GAPDH* forward, 5′-GAAGGT GAAGGTCGGAGT-3′; and *GAPDH* reverse, 5′-GAAGATGGTGATGGGATTTC-3′.

To analyse the efficiency of the CRISPR/Cas9-induced deletions, DNA was extracted and *PAX5* enhancer was PCR-amplified using HotStarTaq DNA Polymerase (Qiagen) and *PAX5* enhancer-flanking oligonucleotides (forward) 5′-GTTGTCTTGGAGGACTTTCAG-3′, and (reverse) 5′- GTGTTATTGTGT ATGTGGCAG-3′. To determine the presence of CRISPR/Cas9-induced mutations we performed heteroduplex cleavage assays using the Guide-it Mutation Detection Kit (Clontech) with primers (forward) 5′-AGGATGAGAACG GGCAAAC-3′ and (reverse) 5′-GGAGCTTCCAGCTGAACTGA-3′.

Statistical analysis. Fisher's test or non-parametric tests were used to correlate clinical and biological variables according to MBL or CLL, and the presence or absence of the different drivers herein analysed. We evaluated the clinical effect (TTT and overall survival) of all driver mutated genes and chromosomal regions with recurrent CNAs in 5 (1%) or more patients. TTT was evaluated only in patients with Binet A and B. TTT and overall survival curves from the date of sampling were plotted by the Kaplan–Meier method and compared by the log-rank test[57]. We examined separately the prognostic impact of point mutations in driver genes (substitutions or small indels) and CNAs. The clinical impact (TTT)

of *TP53*, *ATM* and *BIRC3* mutations was relatively similar to that of the loss of their respective chromosomal region, that is, del(17p) (*TP53*) and del(11q) (*ATM* and *BIRC3*), respectively (Extended Data Fig. 8). Therefore, to evaluate the prognostic impact for each gene/region, both types of alterations were combined. Although the clinical effect of deletions and mutations was somehow different for del(6q15)/*ZNF292* (Extended Data Fig. 8), owing to the fact that most point mutations in *ZNF292* were truncating, we also combined these two alterations to investigate the clinical effect. Finally, the number of cases with mutations or CNAs in the respective chromosomal region of 6p21/*NFKBIE*, 10q24/*NFKB2*, and 15q15/*MGA* was too small to perform a separate analysis and therefore we also combined both types of alterations. Multivariate Cox regression analysis was used to assess the independent prognostic impact from Binet stage and *IGHV* mutational status of each driver in the outcome of the patients. Proportional hazards were checked using Schoenfeld's test. We adjusted all the *P* values for multiple comparisons using the Benjamini–Hochberg correction. All statistical tests were two-sided and statistical significance was considered to be significant with an adjusted $P \leq 0.05$. All the analyses were performed using the SPSS 20 software (http://www.ibm.com) or R software v3.1.3.

Recurrently mutated genes in CLL were defined considering number and type of mutations, gene size and coverage, and local density of mutations derived from the 150 CLL/MBL WGS studies. To test whether a gene was mutated more frequently than expected by chance, we calculated the basal probability for each gene to suffer a non-synonymous mutation (P_{ns}) as: $P_{ns} = \frac{n_{ns} L \delta}{(n_{ns} + n_s) E}$

In this equation, n_{ns} is the total number of possible non-synonymous mutations for this gene, n_s the total number of possible synonymous mutations, L is the effective length of the gene open reading frame (ORF), defined as the sum of the number of bases of the ORF for that gene which are callable at 10× coverage for all exomes or whole genomes analysed, and E is the effective length of all coding regions analysed, defined as the sum of the total lengths of the coding regions that are callable at 10× coverage for all exomes or whole genomes. Finally, δ is the local density of mutations for this locus, which is determined by dividing the number of somatic mutations identified in the 150 WGS studies analysed in a 0.5-Mb region centred on the gene of interest. Thus, the probability P to find M or more non-synonymous mutations in a given gene from a set of N total number of somatic mutations in all patients is:

$$P = 1 - \sum_{j=0}^{M-1} \binom{N}{j} p_{ns}^{j} (1 - p_{ns})^{N-j}$$

A score is computed by taking the base-10 logarithm of this probability (P). Genes for which more than 10% of somatic mutations caused a synonymous change were removed. Finally, 1,000 Monte–Carlo simulations were performed to estimate the FDR based on the total number of mutations observed (N), and the local mutational density for each gene. To identify genes that might be recurrently mutated in an IGHV subgroup, the same analysis was performed only with tumours belonging to the same group (IGHV-MUT or IGHV-UNMUT), and adjusting the local density of mutations for each subgroup according to the mutations obtained from WGS data. Genes were classified in three different tiers (Extended Data Table 2). Tier 1 corresponds to those genes that were identified as statistically mutated in CLL as described above. Tier 2 includes those genes that are not statistically mutated when analysing CLL, but appeared significant when only one subclass (IGHV-MUT or IGHV-UNMUT) was considered. In addition, genes showing either recurrent mutations affecting the same residue, or resulting in mainly loss-of-function mutations, were included in tier 2. Finally, genes classified in tier 3 include those genes that were not in tiers 1 or 2, but containing somatic mutations previously described as driver mutations in the literature.

A sample size of at least 500 tumours was selected during the ICGC study design, as this will give enough power to detect driver genes mutated in at least 3% of tumours[19].

***In silico* prescription.** Drugs with potential therapeutic interactions with driver oncogenic protein products were retrieved as described[37].

41. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
43. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
44. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475,** 101–105 (2011).
45. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3,** 246–259 (2013).
46. Delgado, J. *et al.* Genomic complexity and *IGHV* mutational status are key predictors of outcome of chronic lymphocytic leukemia patients with *TP53* disruption. *Haematologica* **99,** e231–e234 (2014).
47. Edelmann, J. *et al.* High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. *Blood* **120,** 4783–4794 (2012).
48. Valdés-Mas, R., Bea, S., Puente, D. A., Lopez-Otin, C. & Puente, X. S. Estimation of copy number alterations from exome sequencing data. *PLoS ONE* **7,** e51422 (2012).
49. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98,** 288–295 (2011).
50. Bibikova, M. *et al.* Genome-wide DNA methylation profiling using Infinium^{R} assay. *Epigenomics* **1,** 177–200 (2009).
51. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30,** 1363–1369 (2014).
52. Kluk, M. J. *et al.* Gauging NOTCH1 activation in cancer using immunohistochemistry. *PLoS ONE* **8,** e67306 (2013).
53. van de Werken, H. J. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods* **9,** 969–972 (2012).
54. van de Werken, H. J. *et al.* 4C technology: protocols and data analysis. *Methods Enzymol.* **513,** 89–112 (2012).
55. Heckl, D. *et al.* Generation of mouse models of myeloid malignancy with combinatorial genetic lesions using CRISPR-Cas9 genome editing. *Nature Biotechnol.* **32,** 941–946 (2014).
56. Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: fast CRISPR target site identification. *Nature Methods* **11,** 122–123 (2014).
57. Peto, R. & Pike, M. C. Conservatism of the approximation sigma (O-E)2-E in the logrank test for survival data or tumor incidence data. *Biometrics* **29,** 579–584 (1973).

Publication 3

ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites.

Bioinformatics Advance Access published January 16, 2012

ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites

Santi González[1*], Bàrbara Montserrat-Sentís[1*], Friman Sánchez[2], Montserrat Puiggròs[1, 5], Enrique Blanco[3], Alex Ramirez[1,2] & David Torrents [§1,4]

[1] Joint IRB-BSC program on Computational Biology. BSC. c/ Jordi Girona, 29, 08034 Barcelona.

[2] Department of Computer Architecture. Campus Nord UPC D6-117, Jordi Girona 1-3, 08034 Barcelona

[3] Departament de Genètica i Institut de Biomedicina (IBUB), Universitat de Barcelona. Diagonal 645, 08028 Barcelona, Catalonia, Spain.

[4] Institució Catalana de Recerca i Estudis Avançats (ICREA) Pg. Lluís Companys 23, 08010 Barcelona.

[5] Computational Bioinformatics. National Institute of Bioinformatics

* Equal contributors

§ corresponding author: David Torrents: david.torrents@bsc.es

Associate Editor: Prof. John Quackenbush

ABSTRACT

Motivation: The prediction and annotation of the genomic regions involved in gene expression has been largely explored. Most of the energy has been devoted to the development of approaches that detect transcription start sites (TSS), leaving the identification of regulatory regions and their functional transcription factor binding sites (TFBSs) largely unexplored and with important quantitative and qualitative methodological gaps.

Results: We have developed ReLA (for REgulatory region Local Alignment tool), a unique tool optimized with the Smith-Waterman algorithm that allows local searches of conserved TFBS clusters and the detection of regulatory regions proximal to genes and enhancer regions. ReLA's performance shows specificities of 81 and 50% when tested on experimentally validated proximal regulatory regions and enhancers, respectively.

Availability: The source code of ReLA's is freely available and can be remotely used through our web-server under http://www.bsc.es/cg/rela.

Contact: David Torrents (david.torrents@bsc.es)

1 INTRODUCTION

The identification of the genomic regions that control the transcription of genes still remains a challenge despite the recent and continuous development of new experimental and computational methodologies (Tompa, *et al.*, 2005). Multiple automatic approaches have been proposed, ranging from those that search for phylogenetic conservation of sequence or transcription factor binding motifs in non-transcribed DNA regions (Blanchette and Tompa, 2003; Van Loo and Marynen, 2009) to the analysis of DNA physical properties characteristic of regions expected to bind transcription factors (Abeel, *et al.*, 2008; Goni, *et al.*, 2007). However, the incorporation of novel biological knowledge into these programs is not necessarily improving the quality of their predictions, which still contain a substantial fraction of false positives.

Currently, methods that *de novo* detect and characterize proximal regulatory regions show specificity levels below 50% (Van Loo and Marynen, 2009). Even though phylogenetic footprinting using pre-aligned homologous regulatory regions offers promising results in the identification of Conserved Regulatory Modules (CRMs) of TFBSs (Blanchette and Tompa, 2003; Blanco, *et al.*, 2007; Pavesi, *et al.*, 2007; Sebestyen, *et al.*, 2009; Tokovenko, *et al.*, 2009; Tonon, *et al.*, 2010), they cannot define the regulatory region itself in most real scenarios because they are based on global alignment strategies and require that all matching binding sites across all compared sequences are located in the same (or similar) position within each sequence, i.e. they require predefined and pre-aligned regulatory regions. As a result, in spite of the existing methodologies, still the most common and reliable way to identify proximal regulatory regions in genomes is the analysis of a few nucleotides (typically up to 1,000) immediately upstream of annotated TSSs, which likely constitutes the proximal promoter. But the annotation of gene starts is still unsolved, particularly for non-human species. For example, a simple search in the ENSEMBL database (Hubbard, *et al.*, 2009) identified substantial fractions of vertebrate genes without annotated 5'UTR (from 17% in mouse to 91% in opossum, 42% for human). This result is even more dramatic within invertebrates. Other problems that constitute a barrier for the automatic inference of promoters (even in human or mouse) are the abundance and overprediction of alternative transcripts. Taken together, most computational methods that detect or align promoters strongly depend on or are coupled with the annotation of untranslated gene regions, which is generally insufficient for this purpose (Guigo, *et al.*, 2006).

On the other hand, the computational identification of enhancers is even more complex. These regulatory regions that work in cooperation with promoters throughout multiple structural constraints are, apparently, delocalized relative to the genes that are controlling (Arnosti and Kulkarni, 2005). Therefore, their identification through computational methods requires strategies based on local alignments. Some existing

Table 1. Performance results on ABS promoters.

	ReLA	TFM	rVISTA[1]	PromoterExplorer	Eponine	ARTS
Recall	**0.81**	**0.6**	**0.37**	**0.51**	**0.2**	**0.14**
Precision	**0.81**	**0.61**	**0.46**	**0.69**	**0.21**	**0.14**
Prediction Type	**Defined regions with conserved TFBS**	**Conserved TFBS**		**TSS**		

methods, like rVISTA, look for conserved TFBS clusters between regions that have been pre-aligned with local alignment tools, such as BLASTz in rVISTA (Loots and Ovcharenko, 2004), while others use directly local-alignment based search strategies, like the Enhancer Element Locator (EEL) that uses the Smith-Waterman algorithm (Palin, *et al.*, 2006). These tools have shown good prediction rates on enhancers, but also show important limitations regarding the number of species that they can analyse, the required parametization, and the accuracy of the prediction.

To overcome these limitations, and to provide novel and improved solutions to the prediction of regulatory regions, we have developed ReLA, a public local-based alignment tool that is capable of detecting promoters and enhancers by identifying clusters of regulatory motifs conserved in any position within large homologous DNA regions (i.e. independently of gene annotation). Considering the wide range of potential users of this tool, we have also developed a user-friendly web server for remote predictions. The source code of ReLA is distributed also as a standalone program that can be used locally in Unix-based computational platforms.

2 RESULTS

2.1 Rationale and underlying search strategy of ReLA

The goal of this study is to develop a novel methodology that would overcome the current limitations mentioned above by focusing on: (i) the detection of conserved TFBS, (ii) the use of local search strategies; (iii) simplicity of use, and (iv) a low computational cost to perform genome-wide searches. For this, we decided to use the same strategies that have been used for fast local sequence comparisons of protein sequences. In particular, we adapted the Smith-Waterman algorithm (Smith and Waterman, 1981) to make it capable of comparing and detecting the best optimal local alignment of regions with similar sequences of TFBSs. In our procedure, each TFBS is internally transformed into symbols of an arbitrary alphabet, as if they were amino acid or nucleotides in traditional protein-protein and DNA-DNA comparative searches. This search algorithm is the core of a pipeline, referred from now on as ReLA (for REgulatory region Local Alignment tool). The complete procedure can be divided into three major steps. Firstly, input DNA sequences are transformed into sequences of TFBSs by mapping with the MATSCAN software (Blanco, *et al.*, 2006b) all the position frequency matrices (PFMs) provided; Next, the resulting TFBS sequences are compared with each other to identify conserved groups of TFBSs using the modified Smith-Waterman algorithm under different scoring scenarios. Finally, all the resulting preliminary alignments are evaluated to produce the final prediction of the promoter region.

2.2 Evaluation of ReLA's performance

We first applied ReLA to a collection of experimentally validated promoter and enhancer regions, both to define its internal search parameters and to evaluate its performance. Despite ongoing efforts of acquiring experimental data, on functional TFBS, still the vast majority of detailed and reliable data can only be retrieved from the literature. In this direction, the ABS database (Blanco, *et al.*, 2006a) is the result of one of the few initiatives to gather, from the literature, promoters with two or more of their TFBSs experimentally validated. For this reason we used this database for ReLA's evaluation. We selected the subset of 73 (35 human and 38 mouse) promoters from this database that showed, in ENSEMBL (Hubbard, *et al.*, 2009), one-to-one orthologous relationship with at least three out of seven chosen vertebrate species (human or mouse, rat, horse, dog, cow, opossum and chicken). By reproducing a common and realistic search scenario, where the TSS and 5'UTRs of query homologous regions are not known, we collected the putative upstream region of these genes and their corresponding orthologous regions. These upstream regions comprise 5,000 bp upstream DNA, from the first annotated codon according to the encoded ENSEMBL protein. From the measurement of the length of 5'UTRs regions of "known" ENSEMBL genes, we previously had estimated that this selection of 5,000 bp is sufficient to capture the proximal promoters for more than 85% of known vertebrate genes (data not shown). In addition, we have also compared the resulting performance of ReLA with the prediction ratio of other reported TFBS-based search tools: TFM (Tonon, *et al.*, 2010) and rVISTA (Loots and Ovcharenko, 2004), as well as with that of TSS predictors: ARTS (Sonnenburg, *et al.*, 2006); Eponine (Down and Hubbard, 2002) and PromoterExplorer (Xie, *et al.*, 2006), all of them run on the same regions (Table 1).

Table 2. Performance results on EPD TSSs.

	ReLA	TFM	PromoterExplorer	Eponine	ARTS
Recall	**0.56**	**0.49**	**0.78**	**0.23**	**0.17**
Precision	**0.56**	**0.51**	**0.67**	**0.27**	**0.17**

Following the same strategy we alternatively evaluated ReLA using 740 regions derived from the Eukaryotic Promoter Database (EPD; (Schmid, *et al.*, 2006)), which, despite not being ideal for this purpose, sets our tool into the context of previous evaluations of these other existing search strategies: ARTS, Eponine and PromoterExplorer against which we have also compared ReLA's predictions (Table2).

From all resulting predictions, we calculated different performance parameters, such as recall and precision by adapting an evaluation protocol used for the comparison of a large number of TSSs predicting methods (Abeel, *et al.*, 2009). This adaptation is necessary because the different methods we used provide

different type of outputs, e.g. ARTS, Eponine, PromoterExplorer provide single TSS positions, TFM and rVISTA lists of conserved TFBS, and ReLA delimited regions of conserved TFBSs.

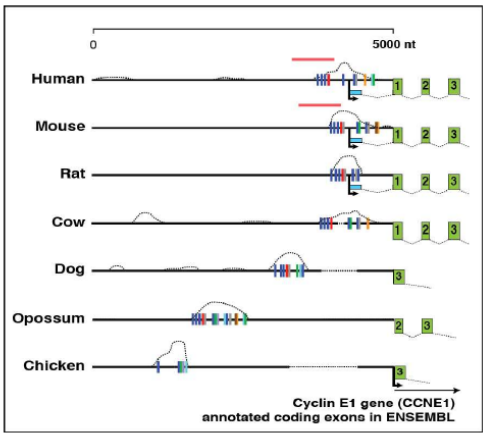


Figure 1. Prediction of the proximal regulatory region of the Cyclin E1 (*CCNE1*) gene in seven vertebrate species. Typical search scenario where the TSS for most of the species compared is not known or missannotated: no TSS information was available for cow, dog and opossum, whereas in chicken is wrongly placed. The predicted promoter for these species appears in different location within the input sequence. Dashed lines show the distribution of all primary predictions along these regions. Consensus predictions are delimited by the first and the last colored box (each box corresponds to a conserved TFBS). Red horizontal lines indicate the experimentally characterized regulatory region. Initial fragments of each transcript are shown on the right: non-coding regions in blue, and coding exons in green. The numbers indicate the position of the coding exons in the human mRNA. ENSEMBL Transcription Start Sites (TSSs) are indicated with arrows.

From the results of this evaluation, we observe that the overall performance is different between the different methods and databases: while ReLA's performance was the best on ABS entries, PromoterExplorer on EPD regions outperformed it. Interestingly, ReLA's precision values for EPD regions are lower than those shown with ABS. To discard a possible bias in the performance of ReLA towards specific promoter types that are more abundant in the ABS database, we divided all EPD and ABS regions in different promoter classes (see Methods) and calculated the same precision values for each promoter type and each prediction method separately (see Supplementary tables 1 and 2). This analysis has shown that, despite ABS appears to be enriched in TATA-box containing promoters (a 42% versus a 20% in EPD), ReLA's performance is not affected by this, as precision values are similar among most of the promoter types present in ABS and EPD. It is also worth noticing that predictors based on TFBSs show a better performance on the ABS, which is also based on TFBS, than with the TSS-based EPD entries, where TSS predictors tent to do better. We did not find any significant difference when comparing performance values for Human or Mouse entries (data not shown).

In order to have a sense of the TFBS conservation levels, upon which ReLA is able to build predictions, we have also analysed the distribution of the number of conserved boxes detected within all ABS and EPD results. This analysis shows that, indeed, there is a wide range of TFBS conservation, both in number and in composition (see Supplementary Figure 1). Similarly, from the analysis of the contribution of each of the species in ReLA's performance, we observe that all the other vertebrates used in this study contribute substantially to the final prediction in human: for example cow and dog contribute to around 60% of the predictions while opossum and chicken to 36 and 32% respectively (see Supplementary Table 3).

A detailed inspection of ReLA's results on ABS entries uncovered some interesting features. Often, the promoters that we identified on each of the species present different locations within the input 5,000 bp region (Figure 1). This typical scenario, which must be necessarily solved with local-based comparative approaches, is observed when the annotation of orthologous gene 5'UTRs and first exons is practically absent, as occurs for most of available genomes. These results highlight the potential of using ReLA for the systematic identification and annotation of regulatory regions in non-model organisms, such as chicken, cow, dog, opossum, and any other that has incomplete gene and cDNA data.

2.3 Prediction of multiple alternative promoters

During the evaluation of ReLA, we also observed that, in some cases, the distribution of preliminary predictions along the reference sequence highlighted two regions with similar scores, which could correspond to alternative promoters. From these two options, ReLA selects the one that generated more preliminary predictions (see Material and Methods). Suboptimal solutions, i.e. potential alternative promoters, can be obtained by simply masking the previously predicted regions in the reference sequence and running ReLA again. For a number of such cases, we confirmed the presence of two TSSs through the analysis of ESTs or known alternatively transcribed full-length mRNAs. For this reason, we have implemented this option in the web server, where the user can launch a second search run to find suboptimal solutions. Figure 2 shows the best two predictions of regulatory regions located upstream from *SLC7A7*, an amino acid transporter gene, which has been experimentally proven to have two alternative promoters (Puomila, *et al.*, 2007).

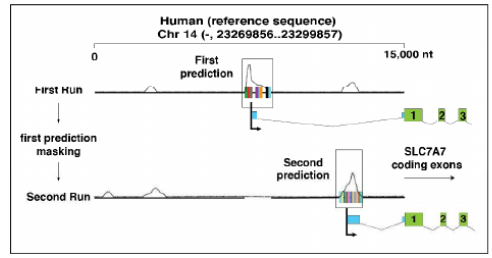


Figure 2. Prediction of alternative known promoter regions of the solute carrier family 7 member 7 (SLC7A7). Searches were done on 15,000 bp region upstream of the first amino acid annotated in the ENSEMBL database for the human SLC7A7 gene. Dashed lines show the distribution of preliminary predictions in the first and second run. Final fist and second predictions are enclosed in a box and delimited by the first and the last conserved TFBS (each designed by a colored box). Initial fragments of each transcript are shown on the right: non-coding regions in blue, and coding exons in green. The numbers indicate the position of the coding exons in the mRNA. ENSEMBL transcription Start Sites (TSSs) are indicated with arrows. Distances are not drawn at real scale.

The finding of two high scoring regions in any ReLA prediction could suggest, instead of the presence of an alternative promoter, the existence of highly conserved coding exons, which would constitute a false positive prediction. Thus, the identification of regulatory regions with ReLA would be based only on the level of sequence conservation and the presence of highly conserved non-

regulatory DNA, like coding exons, could negatively influence the results. To discard this, we studied how this scenario can affect ReLA predictions. The example in Figure 3 shows a positive promoter prediction when all seven orthologous input sequences include the complete region of the *E2F1* gene and the additional upstream untranslated regions (a total of 15,000 nt each). In this case, the distribution of hits along the human sequence shows two high scoring regions that appear to share similar conservation levels of nucleotides. One of these fragments constitutes the third exon of this gene, while the other matches the 5'UTR, the TSS and the core promoter. ReLA is able to successfully discriminate the correct promoter region, including sites that have been experimentally proven (Blanco, *et al.*, 2006a). In particular, the two TFBS that ReLA scores highest in conservation among input sequences are precisely described in the ABS database as two E2F1 factor binding sites necessary for self-regulation during the transition from G1 to S phase in the cell cycle (Johnson, *et al.*, 1994). Interestingly, the third TFBS following the conservation ranking corresponds to ADF1, which was located within the 5'UTR and is known to bind the same motifs recognized by the E2F1 factor in mice (Hsiao, *et al.*, 1994). Despite these results, we cannot discard the possibility that exons are wrongly predicted as promoters in certain situations. Therefore, we recommend performing preliminary evaluations of the coding potential within the input sequences, for example, by comparing them against protein sequence databases with BLASTX (Altschul, *et al.*, 1997). Putative coding regions should be preferentially masked from the input sequences to ensure the correct prediction.

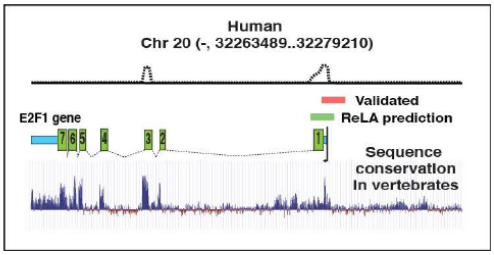


Figure 3. Prediction of the proximal regulatory region of the E2F transcription factor 1 (E2F1) using the sequence of the whole gene. Predicted regulatory region along a highly conserved region of 15,722 bp that includes the E2F1 complete gene transcript and 5,000 bp upstream of the first amino acid annotated in ENSEMBL. Dashed line shows the distribution of all preliminary predictions along this region. A schematic representation of the structure of this gene is shown: non-coding regions are in blue, and coding exons in green. The numbers designate the position of the coding exons in the mRNA, according to human. Data related to DNA conservation from UCSC is also included (http://genome.ucsc.edu; (Kent, et al., 2002)).

2.4 Identification of enhancers

The local nature of the underlying search engine and the capacity to compare large DNA sequences makes ReLA a suitable tool for the identification of enhancers, which are often located distant from other functional elements. In order to test ReLA's capabilities in enhancer detection we have gathered a collection of experimentally validated human enhancers from the VISTA database with activity assessed on transgenic mice (Visel, *et al.*, 2007). To test ReLA, we selected 44 enhancers that are located within the first 50,000 bp upstream of a known gene. In order to search for each of these distal regulatory regions we extracted up

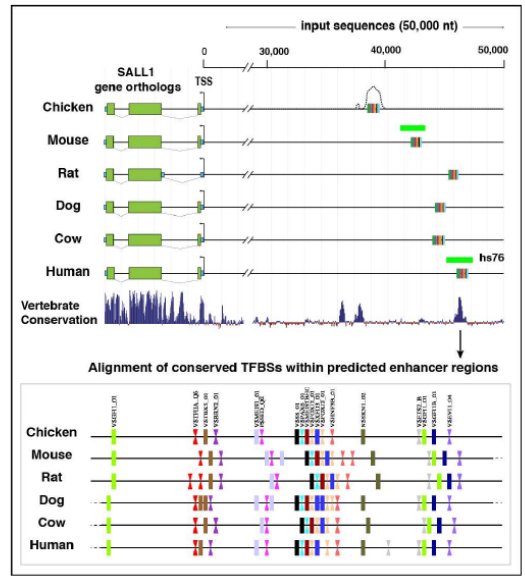


Figure 4. Prediction of the SALL1 enhancer in six vertebrate genomes. The upper panel shows the SALL1 gene and its corresponding 50kbp upstream region for six vertebrate species. Coordinates and strand of these regions are: Chicken (chr11: 6,148,098 - 6,213,605, -), Mouse (chr8: 91,551,143 - 91,618,061, +), Rat (chr19: 19,227,337 - 19,293,298, -), Dog (chr2: 67,080,056 - 67,144,522,-), Cow (chr18: 18,688,671 - 18,752,78, +) and Human (chr16: 51,169,886 - 51,235,181,+). In each line we display the structure of the gene (green boxes are coding exons, while blue are untranslated). Known and predicted TSSs are also shown. ReLA's predictions are shown for each species as groups of colored boxes. Please, note that these regions are not drown to scale. ReLA's predicted enhancer regions expanded from 223 and 233 bp for dog and mouse, 301bp for cow, to 359, 360 and 366 bp for rat, human and chicken respectively. The locations of the experimentally proven regions (as shown in rVISTA db) are displayed as green boxes. The bottom line of this panel shows the sequence conservation profile (according to human coordinates; http://genome.ucsc.edu). In the bottom panel we display the alignment of the conserved TFBS detected within each of the predicted enhancer regions. TFBS are labeled (in TRANSFAC format) and differentiated using arbitrary shapes and colors. Coordinates shown here indicate the position of the predicted enhancer within the 50kbp input sequence.

to 50,000 bp from the most upstream TSS annotated for the closest gene in human and from each of the corresponding one-to-one orthologous genes in mouse, rat, horse, dog, cow, opossum and chicken. The first run of ReLA on these 44 regions generated 40 predictions, from which 11 (28%) overlapped with the annotated enhancer. Considering that the regions selected for the search theoretically contains other unknown regulatory regions (promoters, for instance) that could match with the first prediction, we performed a second run on the remaining 29 cases, which yielded 9 other positive hits. In total, with two iterative runs, ReLA showed a positive predictive value of 50% of the screened subset of annotated human enhancers. A similar prediction rate (49%) is obtained over the same enhancer benchmark set when using a specific enhancer locator tool, EEL (Palin, *et al.*, 2006) that also relies on local-search strategies (EEL searches implied only human and mouse sequences, as it does not accept more than two sequences per search). It is worth mentioning that an important difference between both methods is that ReLA provides more precise results, as the regions predicted are shorter (up to 750 nt long, with an average of 485 nt) than those coming from EEL (up to 11563 nt, with an average of 2644

nt).

These results indicate that ReLA is capable of searching large genomic DNA fragments and identifying multiple proximal and distal regulatory regions, which makes this tool suitable for genome-wide screenings and across several genomes (see an example in Figure 4).

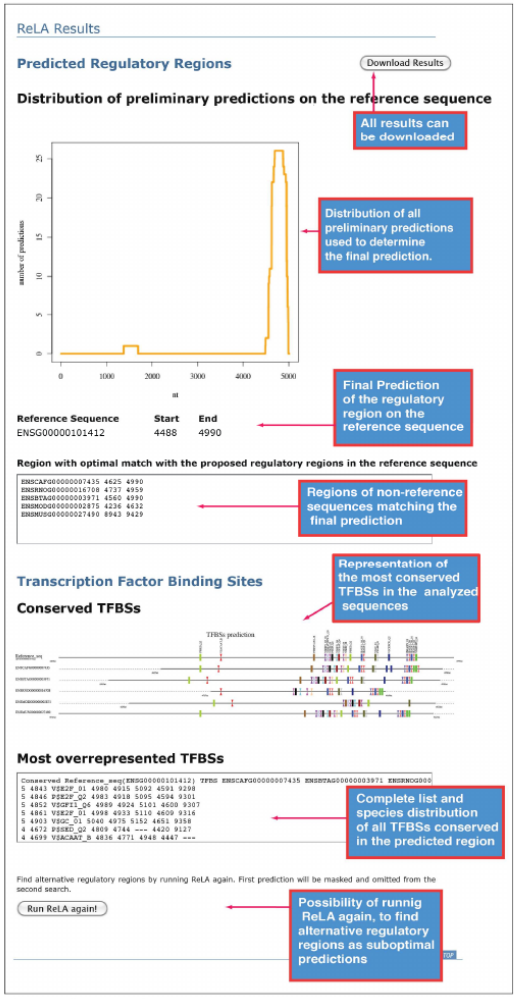


Figure 5. Snapshot of the ReLA web server output. Graphical representation of the putative promoters and alignments of TFBSs of the human E2F1 gene, as well as the lists of predicted regions and conserved binding motifs. See Methods and Results sections for a complete interpretation of each of the results provided.

To further exemplify this feature, we also performed a genome-wide analysis on a 109 kb long ENCODE region (ENm011; chr11:1858751-1968592; (Birney, *et al.*, 2007)) that includes 6 genes coding for, at least, 11 transcripts, with their corresponding intergenic regions. By using SYT8 and MRPL23 flaking genes as anchors, we identified and characterized the corresponding orthologous regions for mouse, rat, dog cow and chicken. The complete analysis consisted in ten iterative ReLA searches and implied the screening of TFBS sequences in more than 600 Kb of genomic DNA. In order to obtain an estimation of the performance on these genome-wide conditions, we have taken as positive predictions those that match ChIP-Seq transcription evidences (Birney, *et al.*, 2007), as well as those falling immediately upstream of annotated gene starts. This count shows that 8 out of 10 predictions have evidence of expression or regulation (see Supplementary Figure 2). Please note that we cannot discard that additional runs would provide other overlooked regulatory regions and, at some point, also false positives.

3 DISCUSSION

Taking into account the available methods to *in silico* recognize gene regulatory regions, a substantial improvement is necessary to accurately annotate genes and promoter sequences in most genomes. Here we report the development of ReLA, a computational tool to identify such regulatory regions using genome-wide comparisons. ReLA is distributed as a standalone program and as a web server. Our approach is mostly based in an adaptation of the popular Smith-Waterman algorithm that is able to rapidly identify coincidences of TFBSs between two sequences (conceptually similar to traditional protein-protein comparisons). ReLA is able to efficiently process long sequences in standard computational platforms (e.g. less than a minute to obtain the results shown in Figure 5). We have evaluated the accuracy of ReLA, first in a dataset of experimentally validated human and mouse promoters, on an extensive collection of validated TSS from the Eukaryotic Promoter Database, as well as on an experimentally validated collection of rVISTA enhancers. We have reached maximums of 0.81 of recall and precision levels on ABS sequences On the other hand, and surprisingly, ReLA's performance results lower when using EPD TSS entries. A possible explanation for this observation could be that ReLA performs better on certain types of promoter regions. But, after we classified all ABS and EPD entries into different promoter types according to their composition and evaluated their associated performance obtained with all the methods used here for the validation, we observed that ReLA's accuracy is similar among most of the identified promoter types. We cannot discard though that other uncontrolled biases present either inside the underlying search methodology of each of the protocols used here, or in the used databases could actually explain the different behaviour observed. It is worth noticing that overall, TFBS-based prediction methods perform better on the TFBS-based ABS database than on the TSS-based EPD, where TSS predictors are doing better. In any case, the levels of precision and recall obtained with ReLA are sufficient to provide reliable predictions that guide posterior experimental validation. This study also demonstrates the benefits of using the Smith-Waterman algorithm to directly search for conserved binding sites, as it outperforms other methods, like rVISTA and TFM, that are based on pre-aligned DNA and global search strategies. See the example in figure 1, which cannot be solved using global alignment approaches. Please note that other methods based on similar strategies could not be included in the comparison, as they did not provide results on our benchmark set because of limitations in the size (MMETA) and on the number of sequences (Conreal; (Berezikov, *et al.*, 2005)).

Furthermore, we also show that ReLA is able of predicting alternative promoters and even enhacer regions, dealing with multiple suboptimal solutions in most cases. Our approach is suitable for integrating a computational annotation pipeline in which other predictive methods such as homology searches (e.g. BLAST against protein databases) can assist in the improvement of the final predictions.

In summary, we believe that the development of ReLA constitutes a significant step forward in the field of the prediction of regulatory regions, as it shows the highest predictive power reported so far. ReLA is able to locally compare multiple large genomic regions and identify non-alignable conservation events across different genomes. This is relevant if we consider that the limited information regarding regulatory regions in eukaryotes is restricted to human and mouse, e.g. from 2540 vertebrate entries in the Eukaryotic Promoter Database, (Schmid, *et al.*, 2006) 2067 (81%) belong to these two species. Thus, with this tool in hand we can now, not only fill missing gaps in the annotation of the genomes of model organisms, mostly with the identification of enhancers and alternative promoters, but also to start a reliable and consistent annotation of conserved promoters throughout the rest of genomes that have little or no information regarding 5'UTRs and often first coding exons (see Figure 1). Beyond the current performance of ReLA and, as we are planning a genome-wide search of regulatory regions across sequenced vertebrate genomes, we are actively searching for ways of improving further its predictive power by, for example, applying more sophisticated scoring systems and accepting even larger DNA regions with low additional computational costs.

4 MATERIALS AND METHODS

4.1 From DNA to TFBS sequences

The first step of our method consists on the mapping of putative TFBSs sequences along the input sequences according to a certain catalog of PFMs (in the documentation included with the program and in the web-server, we provide guidance for selecting a set of homologous sequences). Users locally running ReLA should provide their own PFM files (information about accepted file formats is also provided with the program and in the web-server). It has been shown that the selection of particular collections of matrices yields slightly different results (Blanco, *et al.*, 2006b). After evaluating different options (data not shown), we obtained the optimal results by using PFMs from TRANSFAC (Matys, *et al.*, 2006). We classified this collection of models into three subsets: whole collection of TRANSFAC PFM, the first 600 and the first 400 PFMs ranked by their information content. These three sets are included as the default option in the web server. The identification of potential TFBSs is performed using the MATSCAN software (Blanco, *et al.*, 2006b) at high levels of stringency: 80% of similarity threshold. For this study, we calculate the similarity score using SS= ((current – min)/(max-min)), where, “current” is the actual matching score, “min” is the minimum possible matching score and “max”, the maximum possible matching score of a particular PFM, as described elsewhere (Kel, *et al.*, 2003). Because the next Smith-Waterman step requires a single sequence of TFBSs, and because MATSCAN results usually contain PFM that overlap in all possible ways, we next simplify this output. We remove this overlap by sliding a window of n bp (n = 3 or 5, both conditions are included in the global search, see below). For each overlapping PFM starting at the first position of each of the 3 or 5nt sliding window, we systematically kept the most informative one, maximizing the overall information content of the sequence. To preserve the relative distance between motifs during the comparison, we insert a "spacer box" every n consecutive nucleotides in regions free of predictions. In summary, we convert each input DNA sequence into a sequence of highly informative non-overlapping TFBS, which is used next in the comparative searches.

4.2 Comparative searches

For our searches, we have modified the classical local alignment Smith-Waterman algorithm (Smith and Waterman, 1981) to deal with sequences of TFBSs (associating a unique three-letter combination to each TFBS) and to provide the best scoring local alignment (i.e with the highest density of conserved TFBSs) for each of the comparisons performed between the reference sequence with each of the others (see pseudocode in supplementary information). The overall stringency of the searches, the reliability of the resulting predictions and the conservation of the TFBSs between the input sequences can produce different predictions. Instead of selecting a universal and fixed set of parameters for each of the searches, which would yield one unique prediction, we chose to run recursively each pairwise comparison (reference sequence against each of the other input sequences) with a different set of parameters generating a collection of preliminary predictions. A set of posterior selection filters (see below) is then applied on these preliminary predictions to come up with a final prediction of the promoter region.

Each of these pair-wise comparisons is carried out in two different scoring scenarios (10/-1 and 20/-1 match/mismatch scores, both with an open and extension gap penalties of -2 and with two overlapping thresholds to remove redundant sites (using a window of three or five nucleotides, see above); i.e. a total of four comparisons are performed on each pair of sequences and each set of matrices defined. These specific combinations of parameters were determined by monitoring and maximizing the relationship between sensitivity and specificity using a collection of 10 known promoters of the ABS database (Blanco, *et al.*, 2006a), (see supplementary information and www.bsc.es/cg/rela/downloads). These 10 regions were excluded from the performance evaluation. We also observed that the best predictions obtained during the benchmarking were those with sizes between 200 and 600 nt. Preliminary predictions covering shorter regions usually involved too few conserved TFBSs, while larger predictions tend to connect distant and, apparently, unrelated binding sites. For this reason, preliminary predictions outside this range of sizes are not considered during the generation of the final prediction.

4.3 Output generation

As part of the results, ReLA generates a graph showing the distribution of all accepted preliminary predictions on the reference sequence. From the analysis of the overlap among these preliminary predictions we generate the final prediction by selecting the region, between 200 and 1,000 nt long that contains the highest number of preliminary candidate predictions. Together with the final prediction on the reference sequence, additional consensus regions are also defined in each of the other sequences,

which correspond to those (if any) that match the final predicted promoter region. We also provide the list of all conserved TFBSs. From this list, a subset of the most conserved TFBSs (specifically, those within the top 10% of conservation) is selected and used to generate a multiple alignment in graphical format.

4.4 Web server

We have implemented ReLA as a web server that can be accessed at http://www.bsc.es/cg/rela. The underlying search engine is distributed also as a standalone program. We have designed the ReLA web site to meet the requirements of non-expert users. The web version provides a graphical representation of the putative promoter region predicted in all the input sequences (see Figure 5) and a plain text description of the results. Up to two suboptimal solutions can be provided through the web on each set of input sequences to potentially predict alternative promoters.

4.5. Selection of databases for evaluation

To validate the results obtained and to compare our method with other existing programs in similar searching conditions, three different working subsets or reported regulatory regions were generated from three different databases: Annotated regulatory Binding Sites database (ABS; (Blanco, *et al.*, 2006a)), Eukaryotic Promoter Database (EPD; (Schmid, *et al.*, 2006)) and Vista Enhancer Database Browser (VISTA; (Visel, *et al.*, 2007)). To follow common criteria and to be consistent with the annotation of ABS (as 500nt promoters), we transformed these TSS into regions by considering as promoter region 500 bp upstream from the EPD TSS. VISTA Enhancer Browser is a database of regions containing experimentally validated human and mouse enhancers tested in transgenic mice.

The working subsets were generated according to three different filters to facilitate the automation of the validation process and to ensure reliability of the evaluation protocol: (1) genes associated to selected regulatory regions must have at least 3 orthologous one-to-one genes according to ENSEMBL orthology data, (2) the promoter fragments selected should not overlap with coding regions, and (3) they have to be in our scanned region: as described in the results section, for ABS and EPD it is 5,000bp upstream of the first codon of the gene, and for Vista, 50,000bp upstream of the closest gene. (see below)

Applying these three filters we obtained 75 human and mouse promoters from ABS, and 740 from EPD. In both cases, 5,000 bp upstream from the first methionine were used to check and compare the accuracy of the method. From VISTA Enhancer Browser, we ended up with 44 enhancers laying in the 50,000bp upstream of a known gene.

4.6 Evaluation protocol

For the evaluation of the promoter prediction programs, we followed a modified version of the Distance-based validation evaluation protocol from (Abeel, *et al.*, 2009). Taking into account that we were evaluating promoter genes and we were considering distances, we calculated recall and precision values as

Precision = correct predictions / total predictions

Recall = discovered genes / total genes

For those programs that provide single positions as outputs (TSS predictors, ARTS, Eponine and Promoter Explorer), we considered the sequence ± 500 from TSS for the evaluation. In the case of TFM, we obtained conserved binding sites as result and considered the fragment between the first and the last one for evaluation. A correct prediction was considered if there was an overlap between the prediction and the 500 bp upstream of the defined TSS. For all programs we considered the unique or the best prediction, except for Promoter Explorer that does not rank the results. Since we were using already filtered promoter genes instead of big DNA fragments, we did not discarded any prediction further of 500 nt from the TSS as it is done elsewere (Abeel, *et al.*, 2009). For all programs of our evaluation, we used default settings defined by the corresponding developers. For EEL runs, we used the mouse and human sequences for each of the orthologous groups and applied the parameters described for this species pair elsewhere (Palin, *et al.*, 2006). All the data used for the validation procedure is available at www.bsc.es/cg/rela/downloads.

5 ACKNOWLEDGEMENTS

We thank Mar Albà, Steven Laurie and our entire group for constructive feedback during the development of this work and during the writing of the manuscript. We also thank Jan and Aina Sagristà for designing ReLA's logo. This work was supported by the Ministerio Español de Ciencia e Innovación [BIO2006-15036].

6 REFERENCES

Abeel, T., *et al.* (2008) Generic eukaryotic core promoter prediction using structural features of DNA, *Genome Res*, **18**, 310-323.

Abeel, T., *et al.* (2009) Toward a gold standard for promoter prediction evaluation, *Bioinformatics*, **25**, i313-320.

Altschul, S.F., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

Arnosti, D.N. and Kulkarni, M.M. (2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?, *Journal of cellular biochemistry*, **94**, 890-898.

Berezikov, E., *et al.* (2005) CONREAL web server: identification and visualization of conserved transcription factor binding sites, *Nucleic Acids Res*, **33**, W447-450.

Birney, E., *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, **447**, 799-816.

Blanchette, M. and Tompa, M. (2003) FootPrinter: A program designed for phylogenetic footprinting, *Nucleic Acids Res*, **31**, 3840-3842.

Blanco, E., *et al.* (2006a) ABS: a database of Annotated regulatory Binding Sites from orthologous promoters, *Nucleic Acids Res*, **34**, D63-67.

Blanco, E., *et al.* (2007) Multiple non-collinear TF-map alignments of promoter regions, *BMC Bioinformatics*, **8**, 138.

Blanco, E., *et al.* (2006b) Transcription factor map alignment of promoter regions, *PLoS Comput Biol*, **2**, e49.

Down, T.A. and Hubbard, T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA, *Genome Res*, **12**, 458-461.

Goni, J.R., *et al.* (2007) Determining promoter location based on DNA structure first-principles calculations, *Genome Biol*, **8**, R263.

Guigo, R., *et al.* (2006) EGASP: the human ENCODE Genome Annotation Assessment Project, *Genome Biol*, **7 Suppl 1**, S2 1-31.

Hsiao, K.M., *et al.* (1994) Multiple DNA elements are required for the growth regulation of the mouse E2F1 promoter, *Genes Dev*, **8**, 1526-1537.

Hubbard, T.J., *et al.* (2009) Ensembl 2009, *Nucleic Acids Res*, **37**, D690-697.

Johnson, D.G., *et al.* (1994) Autoregulatory control of E2F1 expression in response to positive and negative regulators of cell cycle progression, *Genes Dev*, **8**, 1514-1525.

Kel, A.E., *et al.* (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences, *Nucleic Acids Res*, **31**, 3576-3579.

Kent, W.J., *et al.* (2002) The human genome browser at UCSC, *Genome Res*, **12**, 996-1006.

Loots, G.G. and Ovcharenko, I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites, *Nucleic Acids Res*, **32**, W217-221.

Matys, V., *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res*, **34**, D108-110.

Palin, K., *et al.* (2006) Locating potential enhancer elements by comparative genomics using the EEL software, *Nature protocols*, **1**, 368-374.

Pavesi, G., *et al.* (2007) WeederH: an algorithm for finding conserved regulatory

motifs and regions in homologous sequences, *BMC Bioinformatics*, **8**, 46.
Puomila, K., *et al.* (2007) Two alternative promoters regulate the expression of lysinuric protein intolerance gene SLC7A7, *Mol Genet Metab*, **90**, 298-306.
Schmid, C.D., *et al.* (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms, *Nucleic Acids Res*, **34**, D82-85.
Sebestyen, E., *et al.* (2009) DoOPSearch: a web-based tool for finding and analysing common conserved motifs in the promoter regions of different chordate and plant genes, *BMC Bioinformatics*, **10 Suppl 6**, S6.
Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences, *J Mol Biol*, **147**, 195-197.
Sonnenburg, S., *et al.* (2006) ARTS: accurate recognition of transcription starts in human, *Bioinformatics*, **22**, e472-480.
Tokovenko, B., *et al.* (2009) COTRASIF: conservation-aided transcription-factor-binding site finder, *Nucleic Acids Res*, **37**, e49.
Tompa, M., *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites, *Nat Biotechnol*, **23**, 137-144.
Tonon, L., *et al.* (2010) TFM-Explorer: mining cis-regulatory regions in genomes, *Nucleic Acids Res*, **38 Suppl**, W286-292.
Van Loo, P. and Marynen, P. (2009) Computational methods for the detection of cis-regulatory modules, *Brief Bioinform*, **10**, 509-524.
Visel, A., *et al.* (2007) VISTA Enhancer Browser--a database of tissue-specific human enhancers, *Nucleic Acids Res*, **35**, D88-92.
Xie, X., *et al.* (2006) PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm, *Bioinformatics*, **22**, 2722-2728.

Publication 4

Unravelling the hidden DNA structural/physical code provides novel insights on promoter location.

7220–7230 Nucleic Acids Research, 2013, Vol. 41, No. 15 Published online 12 June 2013
doi:10.1093/nar/gkt511

Unravelling the hidden DNA structural/physical code provides novel insights on promoter location

Elisa Durán[1,2], Sarah Djebali[3], Santi González[2,4], Oscar Flores[1,2], Josep Maria Mercader[2,4], Roderic Guigó[3], David Torrents[2,4], Montserrat Soler-López[1,2] and Modesto Orozco[1,2,4,5,*]

[1]Institute for Research in Biomedicine (IRB Barcelona), Barcelona 08028, Spain, [2]Joint IRB-BSC Research Program on Computational Biology, Barcelona 08028, Spain, [3]Bioinformatics and Genomics Group, Center for Genomic Regulation and Universitat Pompeu Fabra, Barcelona 08003, Spain, [4]Barcelona Supercomputing Center, Barcelona 08034, Spain and [5]Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona 08028, Spain

Received October 23, 2012; Revised February 15, 2013; Accepted April 30, 2013

ABSTRACT

Although protein recognition of DNA motifs in promoter regions has been traditionally considered as a critical regulatory element in transcription, the location of promoters, and in particular transcription start sites (TSSs), still remains a challenge. Here we perform a comprehensive analysis of putative core promoter sequences relative to non-annotated predicted TSSs along the human genome, which were defined by distinct DNA physical properties implemented in our ProStar computational algorithm. A representative sampling of predicted regions was subjected to extensive experimental validation and analyses. Interestingly, the vast majority proved to be transcriptionally active despite the lack of specific sequence motifs, indicating that physical signaling is indeed able to detect promoter activity beyond conventional TSS prediction methods. Furthermore, highly active regions displayed typical chromatin features associated to promoters of housekeeping genes. Our results enable to redefine the promoter signatures and analyze the diversity, evolutionary conservation and dynamic regulation of human core promoters at large-scale. Moreover, the present study strongly supports the hypothesis of an ancient regulatory mechanism encoded by the intrinsic physical properties of the DNA that may contribute to the complexity of transcription regulation in the human genome.

INTRODUCTION

Gene expression in eukaryotes is a complex process regulated by a myriad of molecular mechanisms. The protein recognition of specific DNA sequence motifs located on promoter regions, upstream of transcription start sites (TSSs), has been traditionally considered as the most important regulatory element in transcription (1,2). Nevertheless, after one decade of the postgenomic era, the location of promoters and in particular TSSs still remains surprisingly challenging (3–6). Classical assumptions such as their location 5′ upstream of transcribed regions or their one-to-one correlation with coding genes might actually be oversimplistic. Indeed, sequence signals like transcription factor–binding sites (TFBSs) show little predictive power when applied at the entire genome level. Furthermore, massive annotation projects (7–9) have provided further evidence about the complexity of promoter location and its occurrence in rather unusual genomic regions. These difficulties illustrate that the mechanisms regulating gene expression are not exclusively based on specific interactions between nucleobases located upstream TSSs and regulatory proteins, as they would lead to detectable sequence signals otherwise. Conversely, it seems that the world of DNA regulation is much more intricate and probably involves a myriad of mechanisms, such as the modulation of chromatin structure or epigenetic signatures (10,11).

We and others (12–15) have suggested the existence of a physical code imprinted onto the DNA fiber, which could account for an ancient regulatory mechanism of basal gene expression. Indeed, core promoters and associated TSSs are DNA segments with an intrinsic ability to act as regulatory regions, as they are depleted in nucleosomes and need to bind to a large number of regulatory proteins, which certainly require special physical properties of the DNA fiber. According to this paradigm, we consider that promoters can be defined as regions of unusual physical deformability (13,15,16), which (even in the absence of traditional sequence motifs) might favor either a suitable nucleosome positioning pattern for protein recognition

*To whom correspondence should be addressed. Tel: +34 934037155; Fax: +34 934037157; Email: modesto.orozco@irbbarcelona.org

(17) or an effective binding of core promoter-binding proteins and RNA polymerase (12,18). Notwithstanding, genome-wide analysis of the DNA physical properties (13) revealed that 'promoter-like' physical signals appear in regions without evidence for real promoters, challenging the existence of a regulatory physical code in DNA, or alternatively, suggesting the presence of many hidden promoter regions in the human genome.

In this manuscript, we have revisited our presumptions about the existence of a physical code involved in gene activity regulation. To this end, we have evaluated de novo promoter predictions arising from the location of regions with unusual physical properties (13). A representative set of suggested (but not annotated) promoters have been analyzed by applying a combination of medium and high-throughput experimental techniques and analyses. Our study demonstrates that a strikingly large number of theoretical predictions, which were considered 'false positives' based on the 2007 knowledge, are indeed true promoters. Therefore, we have been able to determine many novel TSSs and core promoters, which were neither detectable by alternative methods nor presenting orthologous sequence signals with known promoters. Most importantly, the present study enables us to redefine promoter signatures and analyze the diversity, evolutionary conservation and dynamic regulation of human core promoters at large-scale. Overall, our findings provide a solid support to the hypothesis that a primitive physical code imprinted in the DNA fiber constitutes a first level of regulation of gene activity.

MATERIALS AND METHODS

ProStar promoter predictor

Our ProStar promoter prediction program is able to predict TSSs based on the presence of an unusual profile of physical properties (particularly the DNA helical stiffness) (13), simplifying previous algorithms that use a variety of empirical descriptors with complex translation to mechanistic models (12). As described elsewhere (13,15), stiffness parameters were derived from atomistic molecular dynamics simulations using model oligonucleotides, annotated at the dinucleotide level, and averaged linearly along 500 bp size windows. In short, we performed a large number of molecular dynamics (MD) simulations, computing then the covariance matrices in the helical space at the dinucleotide step $[d(X \cdot Y)/d(Z \cdot T)]$. Inversion of such matrix yields a 6×6 stiffness matrix for each dinucleotide step (13,15). To keep the model as simple as possible, we only considered diagonal elements of the matrix, i.e. the stiffness of DNA in front of pure 'twist', 'roll', 'tilt', 'rise', 'slide' and 'shift' deformations. The average physical property profiles were defined from the analysis of two genomic sequence sets (NCBI36/hg18 human genome release, March 2006), corresponding to known promoters (positive set) or randomly selected sequences (negative set) according to the reference GENCODE annotation (19). ProStar scores a given DNA sequence as 'promoter' or as 'background' depending on its similarity to the two reference profiles. This is computationally measured by the Mahalanobis distance—a simple statistical metrics widely implemented in clustering and classification analyses (20)—to both promoter and background reference profiles. Using ProStar default parameters, 500 bp long DNA sequences were analyzed at the genome-wide level to locate potential TSSs (13). In this work, putative human core promoters were identified as regions within a window of −1000/+200 bp relative to the ProStar-predicted TSS locations.

Selection of TSS prediction sets

To be coherent with the ProStar training, we applied our predictor using ENSEMBL (v47) (21) as a reference annotation to select TSSs located at least 1200 bp away from any other annotated TSS. As a result, we obtained a set of putative 'false positive', i.e. regions predicted as promoters by their unusual physical properties but which were not experimentally known. We then filtered out those regions that presented >70% of repetitive elements according to the RepeatMasker algorithm (http://www.repeatmasker.org), or that did not allow unique polymerase chain reaction (PCR) primer localization to the human genome assembly by *in silico* PCR BLAT search (http://genome.ucsc.edu). This process yielded 119 genomic regions (1200 bp long) located around 72 putative TSS (note that it was not always technically possible to study promoters located in both directions).

As a negative prediction set, we randomly selected 100 positions, where ProStar suggested no TSS in a 1200 bp window, and for which unique PCR primers could be located. To make the test unbiased, we did not perform any filtering based on the presence of 2006 known promoters in these ProStar negative predictions. Both ProStar-positive and ProStar-negative predicted promoters were subjected to experimental validation.

The positive set was further compared against the latest gene and transcript reference annotations GENCODE (v7) (19) and ENSEMBL (v56) (21) to determine the true positives.

Luciferase transcription activity assays

We designed hybridization primers suitable for high-GC content regions. The presence of a unique hybridization site was subsequently verified by a BLAT genome alignment (http://genome.ucsc.edu). Primers were ordered in 96-well plates to Sigma-Aldrich. PCR was performed in a 96-well format using AccuPrime GC-rich DNA polymerase (Invitrogen) for the amplification of selected regions. PCR products were analyzed in a 1% agarose gel. Successfully amplified regions were inserted into the promoterless pGL4.21 (luc2P/Puro) vector and ligated through Sfi I restriction sites (Rapid DNA ligation Kit, Roche) that enable directional cloning. *Escherichia coli* competent cells (DH5α, Invitrogen) were transformed with the ligation products. Two independent colonies were selected from each transformant and were verified by sequencing from both the 5′ and 3′ ends. The experimental approach for luciferase activity assays in a high-throughput approach is outlined in Supplementary Figure S1.

Cos-7, Hek293, U2OS, MIA PACA and MDA231 cells were cultured in Dulbeccós Modified Eaglés Media (DMEM) supplemented with 10% of fetal calf serum (FBS). All cultures were grown as a monolayer in a humidified incubator at 37°C in an atmosphere of 5% CO_2. One day before co-transfection, $2–6 \times 10^4$ cells per well were plated in 96-well plates with 100 µl of DMEM without antibiotics. Confluence of 90–95% was achieved by the second day. Transient DNA co-transfections were performed with 0.1 µg of the corresponding pGL4.21/construct plasmid and 0.02 µg of the pGL4.74 (*hRluc/TK*) vector (Promega) using TransFact reagent (Promega) according to the instructions of the manufacturer. DMEM supplemented with 10% FBS was added to the cells 1 h after co-transfection to allow correct growth and protein expression. Dual Luciferase Reporter Assay (Promega) was performed 36 h after co-transfection using a GloMax Multidetection Luminometer (Promega) with dual injector system allowing rapid reagent addition. Light emission was measured 2 s after addition of each of the substrates and integrated over a 10-s interval. The firefly luciferase activity results were normalized with the renilla luciferase activity from the pGL4.74 (hRluc/TK) plasmid to account for differences in transfection efficiency. The previously characterized *SPG4* gene promoter (22) was used to generate positive (S−621/−1) and negative (S−1290/−424) promoter region controls, respectively. Promoter activity was assessed in duplicates and was considered active if it exceeded 3-fold the score of negative control sequences from the normalized threshold value.

After luciferase assays, 80 regions from both the positive and negative promoter sets were further divided into four subsets for further analysis: subset 1 contains 20 high-confidence ProStar sequences with high luciferase activity (PS+L+); subset 2 contains another 20 high-confidence ProStar sequences with low luciferase activity (PS+L−); subset 3, 20 low-scored ProStar sequences with luciferase activity (PS−,L+); and subset 4, 20 low-scored ProStar sequences with no luciferase activity (PS−L−).

CAGE analysis

To measure transcription initiation in the different promoter subset regions, profiles of cap analysis gene expression (CAGE) 5′-ends were computed. For this purpose, ENCODE stranded CAGE data from polyadenylated cytosolic RNA of seven different cell lines (GM12878, H1-hESC, HUVEC, HeLa-S3, HepG2, K562 and NHEK) and generated in two bio-replicates were used (23–25). For each cell line, CAGE mappings of quality >20 from each of the two bio-replicates were merged, and their distinct 5′-ends extracted (redundancy was removed to avoid considering reverse transcriptase-PCR artifacts as true signal). Every region was subjected to two CAGE analyses, either considering the luciferase-tested 1200 bp region or a 2000 bp equivalent expanded region centered at the TSS. For every cell line, each time a CAGE tag 5′-end was located within and on the same strand as one of the promoter regions, the distance between the CAGE tag 5′-end and the promoter region 5′-end was computed, and the CAGE frequency corresponding to this distance (further normalized using percentage distance bins) was increased.

RNA-seq analysis

To measure transcription activity in the different promoter subset regions, profiles of RNA-seq 5′-ends were computed. For this purpose, ENCODE CSHL stranded paired-end RNA-seq data from polyadenylated cytosolic RNA of seven different cell lines (GM12878, H1-hESC, HUVEC, HeLa-S3, HepG2, K562 and NHEK) were used (25). For each cell line, all the mappings of the second bio-replicate were considered, and their distinct most 5′-ends extracted. Every subset region was expanded to a final length of 2000 bp centered at the TSS, similarly to the CAGE analyzed sequences. For every cell line, each time an RNA-seq mapping 5′-end was located within and on the same strand as one of the promoter regions, the distance between the RNA-seq 5′-end and the promoter region 5′-end was computed, and the RNA-seq frequency corresponding to this distance (further normalized using percent distance bins) was increased.

Chromatin structure and epigenetic signals

The chromatin structure was inferred from DNase I hypersensitivity sites as reported in ENCODE through the UCSC Table Browser data retrieval tool (26). From these data, we calculated the average of Dnase I hypersensitiviy clusters within 1200 bp regions of the different CAGE analyzed subsets, considering a positive cluster when overlapped with the reference promoter elements. We also explored potential epigenetic markers in the suggested promoter regions by looking at the occurrence of histone variants H3KMe1, H3K27Ac and H3K4Me3 in seven different cell lines (GM, H1, HSMM, HUVEC, K562, NHEK and NHLF). For each CAGE analyzed subset of 1200 bp regions, we calculated the number of regions that overcome a certain average alignment density (intensity signal) in any of the different cell types. Using a threshold of 10-fold, 92% of PS+ sequences contained stronger signals compared with the 37% of PS−. Increasing the threshold, up to 50, produced a reduction of the total number of regions, but increased the difference between PS+ and PS− in the same direction.

TFBS enrichment evaluation

We investigated if different subsets, including the PS+ predictions (17 909 in total), the experimentally tested PS+ predictions (119 sequences) and PS− predictions (100 sequences) or luciferase positive (49 sequences) and negative (23 sequences) regions, were enriched within the 1200 bp in any of the currently annotated 885 TFBSs. To this end, we systematically compared them with a full list of transcripts described in the BioMart database (http://www.biomart.org) (76 905 transcripts) as a background control. To determine the significant enrichment, we used a Fisher's exact test and represented the magnitude of enrichment as odds ratios, which is the ratio of enrichment for a given TFBS. The corrected significant *P*-value after applying a Bonferroni's correction for all tests was $0.05/885 = 5.65 \times 10^{-5}$. The analyses were performed using the R statistical environment (http://www.r-project.org).

Core region DNA element conservation and sequence-based signals

The conservation of the four different CAGE analyzed 1200 bp regions was evaluated by the comparison with available vertebrate genomes using the University of California, Santa Cruz (UCSC) Table Browser data retrieval tool (26). The level of conservation for each particular fragment was calculated according to Vertebrate Basewise Conservation by PhyloP as the average of conservation of all nucleotides comprising the region. TFBS conservation was determined from the comparison of boxes among human, mouse and rat according to the UCSC TFBS conservation track using matrices obtained from TRANSFAC database (27). In addition, we used Regulatory region Local Alignment (ReLA) algorithm (28), a footprinting-based program for the detection of conserved clusters of TFBSs, to determine whether the regions predicted by ProStar would also be detectable as sequence-based only promoter signals.

RESULTS

Selection of the TSS prediction sets based on DNA physical properties

If physical signals were indeed significant in regulatory regions, as we presume, we would expect a high proportion of ProStar predictions to be promoters, despite the lack of experimental annotation. As described elsewhere (13), promoter sequences provide a distinct profile for six descriptors of the DNA stiffness in front of 'twist', 'roll', 'tilt', 'rise', 'slide' and 'shift' deformations, particularly in regions spanning −250/+900 bp relative to TSSs (that is, covering core and proximal promoter distances).

Therefore, to validate our hypothesis, we first defined a TSS prediction set for experimental screening from the ProStar genome-wide calculations. To better validate the prediction power, we used the original ProStar outcome based on the 2006 release of the human genome (13), without retraining the software with more recent releases of genome data. We selected regions with unusual physical properties suggested to be promoters albeit they were not annotated in reference databases, i.e. ProStar 'false positives' (PS+, see 'Materials and Methods' section). Furthermore, even though the algorithm recognizes the directionality of transcription, predictions might also account for bidirectional regulatory elements. Thereby, we selected 72 high-scored putative TSSs that allowed unique PCR primer hybridization to the human genome assembly on the sense-strand (16 TSSs), antisense (9 TSSs) or in both senses (47×2 TSSs), yielding 119 different putative promoter regions in total (Figure 1a, Supplementary Table S1). We additionally defined a negative set consisting of 100 sequences corresponding to nonsignaled promoter regions by ProStar (PS−, 91 on direct-sense and 9 anti-sense due to PCR constraints) (Figure 1b, Supplementary Table S1).

Comparison of the physical deformability properties between both sets revealed the distinct underlying features that had allowed ProStar to recognize the positive TSS set as putative promoter regions, as described above (15). When we further compared our positive set against the latest transcript reference annotations GENCODE (v7) (19) and ENSEMBL (v56) (21), 24 predicted TSSs appeared to be functional (i.e. they are certainly true positives), giving an unexpected support to the quality of our physical de novo predictions (Supplementary Table S1). Yet, up to 48 ProStar predicted regions are not proximate (<1.2 kb) to any 2012-annotated TSS. Intriguingly, the attempt of validating ProStar predicted regions using methods based on interspecies sequence conservation, such as ReLA (28), yielded a low success rate (9%), providing further evidences that ProStar locates putative promoters in genomic regions where phylogenetic footprinting finds no signal.

Identification of functional promoters

We evaluated the ability of the selected putative regions to activate transcription in mammalian cells by using luciferase reporter gene expression assays (Supplementary Figure S1). By applying a threshold of at least 3-fold higher activity than the control vehicle, 85 putative promoters regions were scored as functional, with a validation rate of 71.4%, while only 34% of the analyzed regions in the negative set displayed activity (Figure 1; Supplementary Table S1). From those 85 positively active sequences, 8 correspond to sense strand, 5 to antisense and 72 to both directions (i.e. putative bidirectional promoters or alternative regulatory elements), accounting for 49 distinct TSSs. Interestingly, a significantly large number of the suggested promoters (37.8%) displayed high activity (10-fold above the vehicle). Furthermore, almost 98% of active promoters in one cell line also displayed activity in three additional cell lines, indicating that the identified regions would mainly generate transcripts involved in housekeeping activities, rather than in tissue-specific processes. Taken together, these findings suggest that physical properties would signal promoters of the loosely regulated 'housekeeping' genes, whereas highly specific sequence signals would be required for the activation of development or tissue-specific genes.

CAGE and RNA-Seq analyses in support of predicted TSSs

Luciferase measurements showed that the vast majority of ProStar TSS-derived regions function as promoters when coupled to a reporter gene and transfected to mammalian cells (Figure 1; Supplementary Table S1). Nevertheless, we should also consider that the resulting activity could be an artifact for some regions, as the activity measurements were based on plasmid-inserted regions rather than on their native structure like in bulk chromatin. Alternatively, the activity might result from the absence of methylation or other posttranslational modifications of a true cellular environment, which can modify the DNA physical properties and ultimately lead to a transcription repression *in vivo* (29–31).

Consequently, we complemented our first validation with a CAGE (7,32,33) to examine the transcription start activity of the experimentally tested 1200 bp regions in living cells (25). We selected 80 regions showing

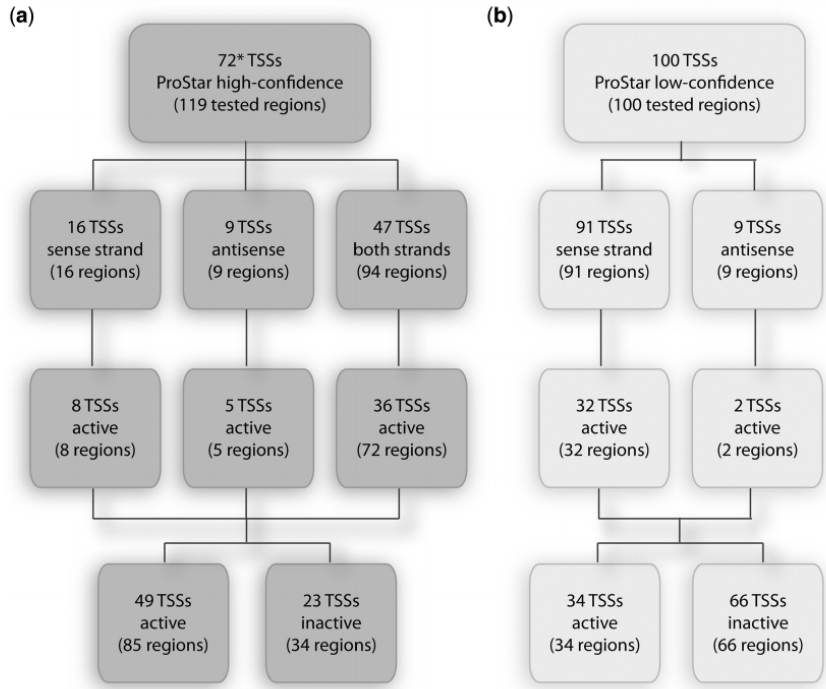


Figure 1. Identification of functional promoters. Summary scheme of TSS selection for both positive (**a**) and negative (**b**) ProStar sets, classified based on luciferase activity (3-fold) and directionality of tested regions: sense, antisense or both sense strands. (Asterisk) 24 out of 72 TSSs were annotated on recent transcriptome reference annotations (21) based on the 2009 genome release (GRChR37/hg19), i.e. true positives.

different levels of luciferase activity and classified them into four representative categories. Subset 1 contains 20 ProStar high-scored sequences with high luciferase activity (PS+L+); subset 2 contains another 20 ProStar high-scored sequences with low luciferase activity (PS+L−); subset 3, 20 low-scored ProStar sequences with luciferase activity (PS−L+); and subset 4, 20 low-scored ProStar sequences with no luciferase activity (PS−L−) (Supplementary Table S1).

The results summarized in Figure 2 show that regions from subsets 1 (PS+L+) and 2 (PS+L−) were dramatically enriched for CAGE tags that could be confidently mapped to single positions (Figure 2a and b), as compared with the ProStar negative subsets 3 (PS−L+) and 4 (PS−L−) (Figure 2c and d). Subset 1 displayed the highest proportion of sequences with CAGE tagged 5′-ends around 750 bp, indicating that those regions contained reliable TSS marks (Figure 2a, around 60th distance bin). CAGE tags were detected in most of the human cell type experiments, but a particular enrichment was found for polyadenylated (polyA+) transcripts, suggesting that active regions might correspond to promoter elements regulating protein-coding genes.

Interestingly, subset 3 (PS−L+) regions contain few cage tags, although they showed some activity in luciferase expression assays (Figure 2c). We could simply assume that this subset contains luciferase-false positives. However, it has been reported that the structure of promoters on different chromosomes varies and these variations might not be well covered by whole-genome promoter prediction algorithms (6). Thus, we cannot rule out the possibility that promoters located in anomalous positions, and hence harboring a divergent pattern of physical properties, could have been overlooked by ProStar (13,15). If these regulatory elements turn out to be under tight regulation in bulk chromatin (which would explain why no CAGE tags are detected), they could well show transcriptional activity in luciferase assays, which ignore activation or inhibition signals imprinted in the native chromatin structure.

Even more intriguingly, subset 2 regions (PS+L−) did show clear CAGE enrichment although they did not provide a luciferase response (Figure 2b). These discrepancies could simply result from luciferase-false negatives. However, the strength and the profile of CAGE signals (Figure 2b) indicated that other factors could also account for the low luciferase/high CAGE response. Comparison of the CAGE profiles indicated that subset 1 peaks are located at the expected TSSs (i.e. around 60th bin; Figure 2a), while subset 2 peaks are upstreamly

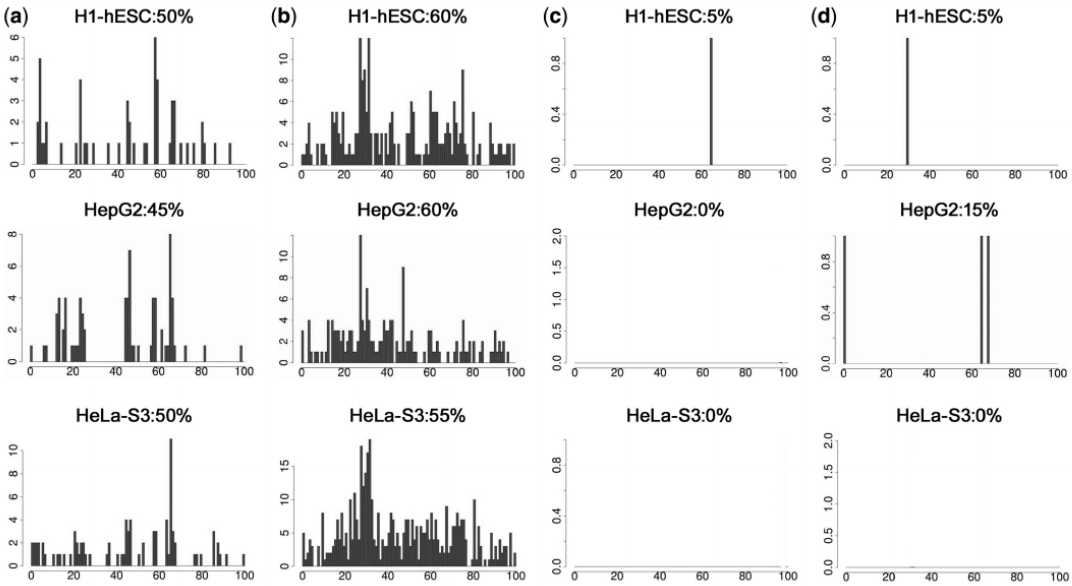


Figure 2. Orthogonal support of predicted TSSs: CAGE analysis. Distribution of distinct 5′-ends of CAGE tags from several representative CAGE experiments in H1-hESC, HepG2 and HeLa-S3 cell types based on cytosolic polyA+ transcripts. For every distinct most 5′-end of CAGE tag detected within and on the same strand as a particular promoter region, we increased the CAGE frequency of the percent distance bin corresponding to the distance between the CAGE tag 5′-end and the promoter region 5′-end. As the predicted promoter regions were 1200 bp long, each % distance bin includes 12 bp, and thereby the TSS is expected to be located on the 84th distance bin (i.e. at 1000 bp from the region 5′-end). (**a**) PS+L+ subset 1. For most of the cell types, the major peak appears around the 63th bin (i.e. 750 bp), closely matching with the prediction (**b**) PS+L− subset 2. We observe undefined peaks around the 30th–50th bins (350–600 bp). On the other hand, the number of CAGE tags is significantly higher than for subset 1 (**c**) PS−L+ for subset 3, (**d**) PS−L− for subset 4. ProStar negative PS− subsets clearly show an almost inexistent CAGE signal.

displaced from the original prediction (around 30th bin; Figure 2b). These findings suggest that, under certain conditions, physical properties are able to signal promoter regions although the prediction of the TSS location can be upstreamly displaced from the true site. In this scenario, CAGE experiments would still detect transcript 5′ in the −1000/+200 bp analyzed genomic window. On the other hand, this displacement would have led us to amplify truncated promoter constructs undetectable by the conservative luciferase test we initially applied in our experimental workflow (Supplementary Figure S1).

To validate this hypothesis, we carried out RNA-sequencing (RNA-seq) analysis to survey the transcription profiles of the selected regions and to identify putative exons near the suggested TSSs. We performed the analysis of 2000 bp regions centered on the predicted TSSs, using RNA-seq data of subcellular-fractionated RNAs from the ENCODE Consortium (Supplementary Figure S2, see 'Materials and Methods' section for details) (9,25). Interestingly, the profiles of subset 1 presented a sharp RNA-seq peak at 800 bp, which coincided with the CAGE major peak around 750 bp (Figure 3a, around 40th bin, orange frames). Furthermore, this TSS putative peak was corroborated with a downstream peak corresponding to an exon (50–80th bins, i.e. from 1000 to 1400 bp) in most of the cell lines. Conversely, subset 2 profiles showed two sharp RNA-seq peaks at 200 and 400 bp, respectively, which matched CAGE major peaks around 360 bp (Figure 3b, 10–20th bins, highlighted with orange frames). Moreover, a downstream broad peak likely corresponding to an exon (Figure 3b, 20–40th bins, i.e. from 400 to 800 bp, highlighted with purple frames) could further confirm the TSS displaced positions at ∼700–500 bp upstream relative to predictions.

We further interrogated this potential TSS displacement in the prediction by analyzing new genomic fragments but now centered on the observed CAGE peaks. To this end, we picked up regions from subset 2 and placed the TSS 500 bp upstream to the original ProStar TSS prediction, as indicated by the CAGE/RNA-seq profiles (Figure 4a). As expected, CAGE profiles exhibited a major peak around 800–900 bp, resembling subset 1 sequences (Figure 4b, around 45th bin). Similarly, RNA-seq profiles also presented a single peak at the expected position (Figure 4c, around 50th bin, 1000 bp). We then re-amplified four of these genomic regions by PCR, spanning 2000 bp but centered at the newly located TSS, as similarly done with previous subsets (Figure 4a; see Supplementary Figure S1 for method details). Interestingly, luciferase assays measured a 4-fold higher activity on average than the original sequences (Figure 4d), providing further evidence that subset 2 segments (PS+L−) do contain true TSSs.

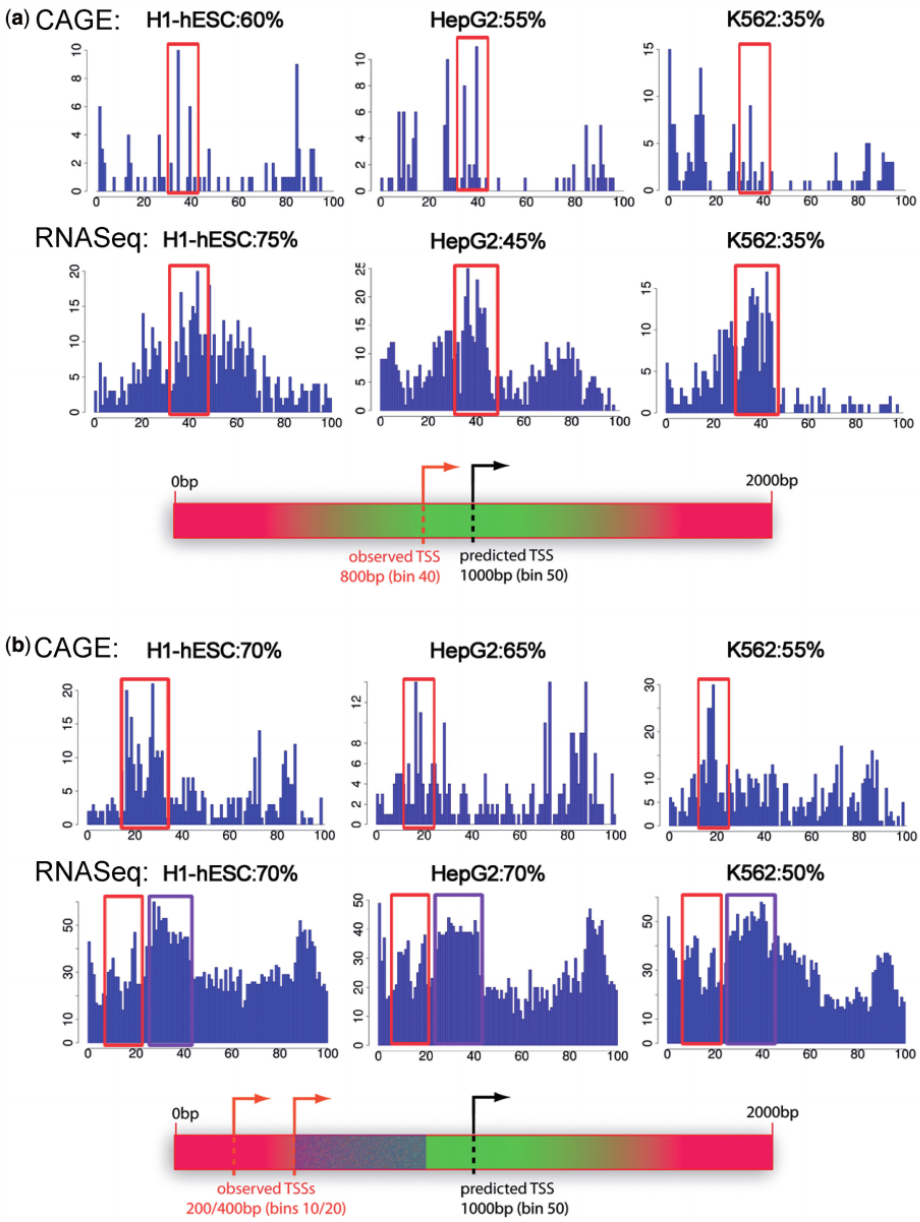


Figure 3. Orthogonal support of predicted TSSs: CAGE vs RNA-seq analyses. Distribution of 5′-ends of CAGE/RNA-seq tags from representative CAGE/RNA-seq experiments in H1-hESC, HepG2 and K562 cell types based on cytosolic polyA+ transcripts. The profiles were constructed similarly to the 1200 bp CAGE analysis. However, as the predicted promoter regions were now 2000 bp long, each % distance bin includes 20 bp and hence the predicted TSS should be located on the 50th distance bin (i.e. 1000 bp), as it is indicated in the promoter region schematic representations below the profiles (**a**) PS+L+ subset 1. The observed TSSs extrapolated from CAGE and RNA-seq profiles appear around the 40th bin (i.e. 800 bp, highlighted with orange frames), and closely match the predictions (**b**) PS+L− subset 2. We observe two sharp RNA-seq peaks around the 10th–20th bins (200–400 bp) that match with CAGE peaks around the 20th bin (highlighted with orange frames). Furthermore, a broad peak is observed right after the observed TSSs, indicating that it may correspond to a transcription active region (i.e. an exon, highlighted in purple) but not necessarily a transcription start region.

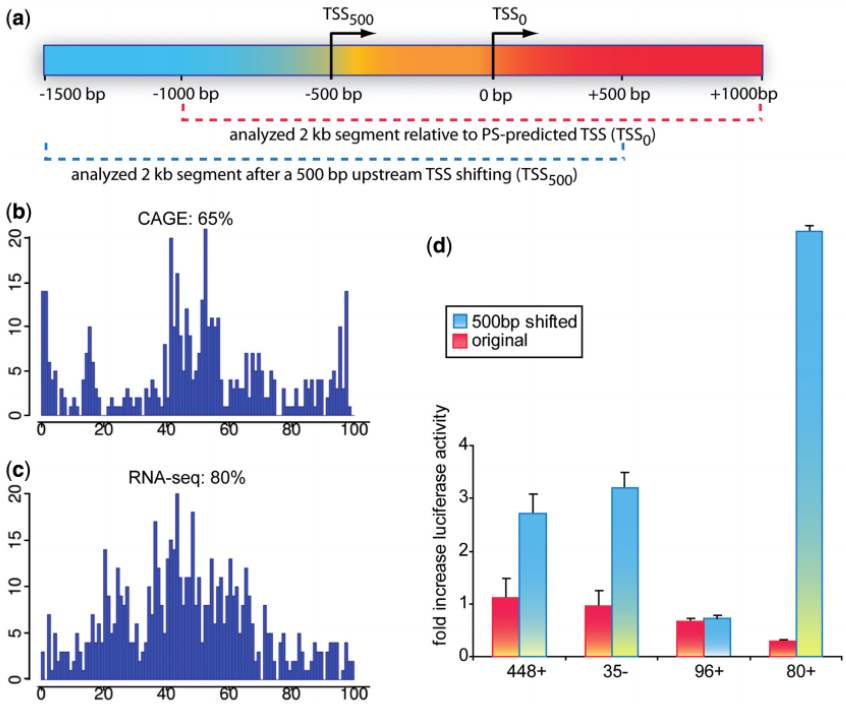


Figure 4. Evaluation of PS+L− sequences on centering the TSS 500 bp upstream from the prediction. (**a**) Subset 2–shifted regions were reconstructed by first re-locating the TSS 500 bp upstream from the relative prediction in the human genome, and subsequently selecting the flanking ±1000 bp upstream and downstream regions, respectively (**b**) Distribution of CAGE tags in H1-hESC cells for the 2000 bp regions centered in relocated TSSs (**c**) RNA-seq analysis profiles of the same regions. X-axes show % distance bins, each one including 20 bp. Y-axes display the number of detected tags. Here we observe a major peak from both analyses around the 50th bin (1000 bp), indicating that it may correspond to a transcription start region (**d**) We confirmed the transcription ability of those regions by additional luciferase assays in four representative PS+L− sequences (three in sense strand and one in anti-sense), showing a significant higher activity (green bars) as compared with the original predictions (red).

Core promoter activity landscape

We subsequently analyzed the four subsets of ProStar predictions to seek for correlations between structural or epigenetic motifs and the promoter activity status. To this end, we used data repositories publicly available from the ENCODE Consortium (9) (See 'Materials and Methods' section for details).

We first analyzed chromatin accessibility to DNase I degradation profiles, as DNase I hypersensitive sites (DHS) are expected to correlate with loosely packed regions in bulk chromatin and hence with gene transcriptional activity (34–36). Analysis of ENCODE data (Figure 5a) highlighted a similar DHS density for ProStar positive subsets (PS+L+ and PS+L−), which turned out to be much larger than the observed density for the negative subsets (PS−L+ and PS−L−). These observations indicate that ProStar-predicted regions are indeed open and thereby associated with transcriptionally active chromatin. Of note, those predictions cannot be simply explained on the basis of sequence-dependent rules such as the presence of CpG islands, as the CG content provides a disperse prediction signal and leads to large number of false positives (13). It should also be noted that ProStar is able to detect promoters located at a large distance to any annotated CpG island (37), as this is the case for 20% of the positive predictions analyzed by CAGE (Supplementary Table S2).

We also evaluated the occurrence of histone modifications correlated with epigenetic modulation of gene transcription, in particular H3K4Me1, H3K27Ac and H3K4Me3, which are specifically prevalent in regulatory regions (38). The results shown in Figure 5b revealed that these histone marks were actually more overrepresented in ProStar positive regions (PS+L+ and PS+L−) than in ProStar negative regions (PS−L+ and PS−L−), providing accumulating evidence about the attainable implication of ProStar regions in the regulation of gene activity.

Furthermore, as PS+ regions are located on regulatory elements, we quested for potential associations to specific functions by a TFBS enrichment evaluation, using the Transfac database (27). To this end, we examined diverse region subsets, including all PS high-scored predictions (PS+, 17909 sequences), the experimentally tested

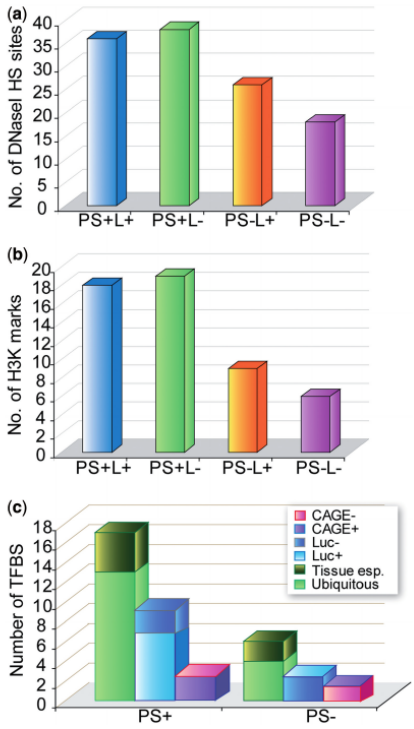


Figure 5. Putative promoter activity landscape. (**a**) Average DHS enrichment within 1200 bp regions of the different CAGE analyzed subsets in a large collection of cell types available in ENCODE (**b**) Average plots of detected Histone 3 variants correlating with transcriptional activity: H3K4Me1, often observed near regulatory elements; H3K27Ac, occurring near promoters; H3K4Me3, near active regulatory elements (**c**) TFBS enrichment evaluation of PS+ and PS− predictions according to gene expression (i.e. ubiquitous vs tissue-specific; left column in the respective PS+ and PS− groups). Further evaluation of TFBS enrichment in the experimentally tested PS+ and PS− sequence sets according to luciferase transcription activity (Luc+ or Luc−, middle bars) and CAGE mapping analysis (CAGE+ or CAGE−, right bars).

regions (containing 119 PS+ predictions and 100 PS− predictions, respectively) and the CAGE-analyzed sets (subsets 1 and 2 on the one hand, and subsets 3 and 4 on the other hand). The enrichment for a given TFBS was considered to be significant when $P < 5.65 \times 10^{-5}$ (Supplementary Table S3). Again, PS− regions showed little enrichment, whereas 17 human TFBSs were found to be overrepresented in at least one of the PS+ groups, the larger part being annotated as ubiquitous (Figure 5c, left column in PS+ and PS− groups, respectively) and mostly related to vital cellular functions (Supplementary Table S3). Interestingly, TFBSs overrepresented in the regions capable of driving luciferase transcription were also enriched in PS+ predictions (Figure 5c, middle bars, Luc+) and CAGE tagged sequences (Figure 5c, right bars, CAGE+). In addition, the identified TFBSs presented high binding affinity to GC-rich sequences, representing truly active TFBSs and thereby supporting our hypothesis that ProStar accurately predicts promoters of housekeeping genes.

Lastly, although the vast majority of the PS+ regions were not detectable using phylogenetic footprinting-based methods (as we previously discussed), we further investigated the conservation of ProStar regions across species, as biologically relevant sequences should display some level of sequence conservation. As expected, PS+ regions were enriched for conserved DNA elements (Figure 6a), particularly for TFBSs (Figure 6b), as compared with PS− regions.

DISCUSSION

A comprehensive analysis of ProStar predicted TSSs has enabled us to identify novel functional core promoters in the human genome exclusively detected by their differential physical deformability pattern and not simply by sequence-based signals such as the CG content alone. A large percentage of ProStar seemingly 'false positives', i.e. regions with unusual physical properties but not associated to any annotated promoter, are indeed transcriptionally active. In particular, highly active regions containing a differential physical pattern display typical chromatin features of housekeeping gene promoters involved in cell survival and maintenance, as proven by an overwhelming amount of direct (luciferase assays, CAGE or RNA-seq mapping) and indirect evidence (profile analysis such as DNaseI sensitivity, epigenetic markers, TFBS enrichment or DNA element conservation). Interestingly, physical signaling also appears to be able to detect promoter activity even in cases where the TSS is located 500 bp upstream of the prediction. Whether this displacement is indicative of a particular feature of genes with closely related alternative TSSs, as indicated by massive CAGE and RNA-seq mappings (Figures 3b and 4), will nevertheless require further investigation. Taken together, these observations reinforce the evidence that high-confidence ProStar predicted regions, sharing a defined pattern of physical features, truly behave like physiologically active TSSs.

We have also observed that most of the active core regions signaled by physical properties do not exhibit directionality in transcript initiation, indicating that physical properties might signal zones where the binding of regulatory proteins and the deformation of DNA are less intricate, as we had previously suggested (39–42). Yet, this signaling might not be sufficient to determine the correct sense of transcription. Intriguingly, more than half of all human promoters are bidirectional, and hence directionality of promoter activity may be regulated to some degree in a cell type–specific manner (43).

On the whole, our study provides insights into the role of DNA physical properties in ascertaining an ancestral coarse regulatory mechanism. Thereby, regions with high chance of undergoing spontaneous transcription would be recognized by protein effectors and favor nucleosome depletion aside from the purely sequence-based signals encoded as H-bond patterns in the DNA major and

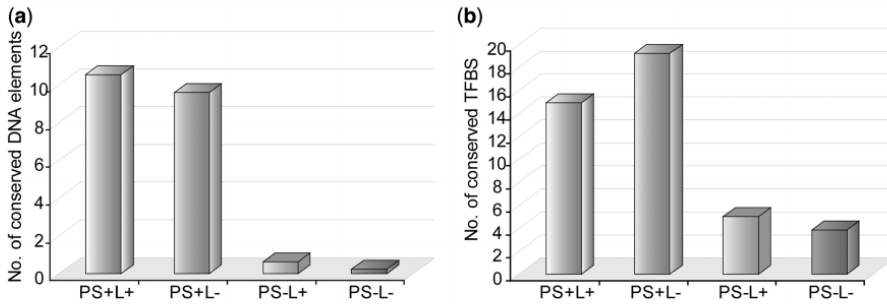


Figure 6. Putative promoter interspecies conservation. (**a**) Average plots of identified DNA elements conserved among human, mouse and rat using the 'Vetebrate PhyloP' algorithm within 1200 bp regions of the different CAGE analyzed subsets (**b**) Similarly, plot histograms showing the average number of identified TFBSs that are conserved across species.

minor grooves. In fact, recent genome-wide nucleosome mapping analyses from our group have revealed that housekeeping genes display unique nucleosome architectures, with large nucleosome refractory regions upstream the TSS (unpublished data). In general then, the interplay between DNA physical properties and regulatory regions could be rationalized in terms of nucleosome positioning (16), favoring the presence of sequences with unique deformation properties in promoter regions, although this might not be the only underlying mechanism, and this would probably vary from gene to gene.

Yet, the physical code type of mechanism could have been evolutionary deactivated in specific genes where fine regulation is required, but seems to be still active in many other cases, where such a stringent regulation is not essential. This convoluted regulatory signaling present in complex organisms could partially explain the failure of traditional promoter location methods to identify a significant number of TSSs, implying the presence of many hidden promoter regions in the human genome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3 and Supplementary Figures 1 and 2.

ACKNOWLEDGEMENTS

We thank Carles Fenollosa and Ramon Goñi for technical support with ProStar.

FUNDING

Spanish Ministry of Science and Innovation [BIO2012-32868 and Consolider E-Science Project]; Instituto de Salud Carlos III (Instituto Nacional de Bioinformática); European Research Council (ERC) Advanced Grant; Fundación Marcelino Botín. M.O. is an Institució Catalana de Recerca i Estudis Avançats (ICREA) Academia Researcher. Funding for open access charge: Fundación Marcelino Botín.

Conflict of interest statement. None declared.

REFERENCES

1. Hélène,C. (1981) Recognition of base sequences by regulatory proteins in procaryotes and eucaryotes. *Biosci. Rep.*, **1**, 477–483.
2. Baumann,M., Pontiller,J. and Ernst,W. (2010) Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview. *Mol. Biotechnol.*, **45**, 241–247.
3. Hannenhalli,S. and Levy,S. (2001) Promoter prediction in the human genome. *Bioinformatics*, **17**, S90–S96.
4. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nat. Genet.*, **29**, 412–417.
5. Bajic,V.B., Brent,M.R., Brown,R.H., Frankish,A., Harrow,J., Ohler,U., Solovyev,V.V. and Tan,S.L. (2006) Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol.*, **7**, S3.
6. Bajic,V.B., Tan,S.L., Suzuki,Y. and Sugano,S. (2004) Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.*, **22**, 1467–1473.
7. Cooper,S.J., Trinklein,N.D., Anton,E.D., Nguyen,L. and Myers,R.M. (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.*, **16**, 1–10.
8. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigó,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T. and Thurman,R.E. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
9. Raney,B.J., Cline,M.S., Rosenbloom,K.R., Dreszer,T.R., Learned,K., Barber,G.P., Meyer,L.R., Sloan,C.A., Malladi,V.S. and Roskin,K.M. (2011) ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.*, **39**, D871–D875.
10. Schueler,M.G. and Sullivan,B.A. (2006) Structural and functional dynamics of human centromeric chromatin. *Annu. Rev. Genomics Hum. Genet.*, **7**, 301–313.
11. Butcher,L.M. and Beck,S. (2008) Future impact of integrated high-throughput methylome analyses on human health and disease. *J. Genet. Genomics*, **35**, 391–401.
12. Pedersen,A.G., Baldi,P., Chauvin,Y. and Brunak,S. (1998) DNA structure in human RNA polymerase II promoters1. *J. Mol. Biol.*, **281**, 663–673.
13. Goñi,J.R., Pérez,A., Torrents,D. and Orozco,M. (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.*, **8**, R263.
14. Abeel,T., Saeys,Y., Bonnet,E., Rouzé,P. and Van de Peer,Y. (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, **18**, 310–323.
15. Goñi,J.R., Fenollosa,C., Pérez,A., Torrents,D. and Orozco,M. (2008) DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics*, **24**, 1731–1732.

16. Deniz,O., Flores,O., Battistini,F., Pérez,A., Soler-López,M. and Orozco,M. (2011) Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics*, **12**, 489.
17. Boeger,H., Griesenbeck,J., Strattan,J.S. and Kornberg,R.D. (2003) Nucleosomes unfold completely at a transcriptionally active promoter. *Mol. Cell*, **11**, 1587–1598.
18. Gross,P. and Oelgeschläger,T. (2006) Core promoter-selective RNA polymerase II transcription. *Biochem. Soc. Symp.*, **73**, 225–236.
19. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R. and Swarbreck,D. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7**, S4.
20. Marques de Sa,J.P. (2001) *Pattern Recognition: Concepts, Methods, and Applications*. Springer Verlag, Berlin.
21. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. and Down,T. (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
22. Mancuso,G. and Rugarli,E. (2008) A cryptic promoter in the first exon of the SPG4 gene directs the synthesis of the 60-kDa spastin isoform. *BMC Biol.*, **6**, 31.
23. Consortium,E.P. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology*, **9**, e1001046.
24. Djebali,S., Lagarde,J., Kapranov,P., Lacroix,V., Borel,C., Mudge,J.M., Howald,C., Foissac,S., Ucla,C. and Chrast,J. (2012) Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS One*, **7**, e28213.
25. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A.M., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
26. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
27. Wingender,E., Dietze,P., Karas,H. and Knüppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
28. González,S., Montserrat-Sentís,B., Sánchez,F., Puiggròs,M., Blanco,E., Ramirez,A. and Torrents,D. (2012) ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. *Bioinformatics*, **28**, 763–770.
29. Daniel,F.I., Cherubini,K., Yurgel,L.S., de Figueiredo,M.A.Z. and Salum,F.G. (2011) The role of epigenetic transcription repression and DNA methyltransferases in cancer. *Cancer*, **117**, 677–687.
30. Hawkins,P. and Morris,K.V. (2008) RNA and transcriptional modulation of gene expression. *Cell Cycle*, **7**, 602.
31. Fukuda,H., Sano,N., Muto,S. and Horikoshi,M. (2006) Simple histone acetylation plays a complex role in the regulation of gene expression. *Brief. Funct. Genomics Proteomics*, **5**, 190–208.
32. Kodzius,R., Kojima,M., Nishiyori,H., Nakamura,M., Fukuda,S., Tagami,M., Sasaki,D., Imamura,K., Kai,C. and Harbers,M. (2006) CAGE: cap analysis of gene expression. *Nat. Methods*, **3**, 211–222.
33. Trinklein,N.D., Karaöz,U., Wu,J., Halees,A., Aldred,S.F., Collins,P.J., Zheng,D., Zhang,Z.D., Gerstein,M.B. and Snyder,M. (2007) Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome. *Genome Res.*, **17**, 720–731.
34. Sabo,P.J., Humbert,R., Hawrylycz,M., Wallace,J.C., Dorschner,M.O., McArthur,M. and Stamatoyannopoulos,J.A. (2004) Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc. Natl Acad. Sci. USA*, **101**, 4537.
35. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
36. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
37. Meyer,L.R., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Kuhn,R.M., Wong,M., Sloan,C.A., Rosenbloom,K.R., Roe,G. and Rhead,B. (2013) The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
38. Rice,J.C. and Allis,C.D. (2001) Histone methylation versus histone acetylation: new insights into epigenetic regulation. *Cur. Opin. Cell. Biol.*, **13**, 263–273.
39. Goñi,J.R., Vaquerizas,J.M., Dopazo,J. and Orozco,M. (2006) Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. *BMC Genomics*, **7**, 63.
40. Noy,A., Pérez,A., Lankas,F., Javier Luque,F. and Orozco,M. (2004) Relative flexibility of DNA and RNA: a molecular dynamics study. *J. Mol. Biol.*, **343**, 627–638.
41. Pérez,A., Noy,A., Lankas,F., Luque,F.J. and Orozco,M. (2004) The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res.*, **32**, 6144–6151.
42. Orozco,M., Noy,A. and Pérez,A. (2008) Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Cur. Opin. Struct. Biol.*, **18**, 185–193.
43. Trinklein,N.D., Aldred,S.F., Hartman,S.J., Schroeder,D.I., Otillar,R.P. and Myers,R.M. (2004) An abundance of bidirectional promoters in the human genome. *Genome Res.*, **14**, 62–66.

Manuscript 1

Deciphering the genomic architecture of IGH-ZFP36L1 fusion in mature B-cell lymphomas with del(14)(q24q32) reveals cooperating molecular mechanisms

Deciphering the genomic architecture of IGH-ZFP36L1 fusion in mature B-cell lymphomas with del(14)(q24q32) reveals cooperating molecular mechanisms

I Nagel*§[1], I Salaverria*[2], S Gonzalez*[3], B Rodríguez[3], G Clot[2], D Martin-García[2], I Vater[1], M Sczepanowski[4], A Navarro[2], C Royo[2], Judit Pinteño[3], JI Martin-Subero[1,2], W Klapper[4], J Richter[1], M Kreuz[5], M Ritgen[6], E Callet-Bauchu[7], MJ Calasanz[8], F Sole[9], E Schroers[10], M Kneba[6], Martin J.S. Dyer[11], Julio Delgado[2], A López-Guilllermo[2], XS Puente[12], C López-Otín[12], E Campo[2], D Torrents#[3], S Beà#[2], R Siebert#[1]

1Institute of Human Genetics, Christian-Albrechts-University Kiel & University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany
2Hematopathology Unit, Hospital Clínic, Institutd'InvestigacionsBiomèdiques August Pi Sunyer (IDIBAPS), Barcelona, Spain
3Programa Conjunto de Biología Computacional, Barcelona Supercomputing Center (BSC), Institut de Recerca Biomèdica (IRB), Spanish National Bioinformatics Institute, Universitat de Barcelona, Barcelona, Spain
4Institute of Pathology, Hematopathology Section and Lymph Node Registry, University Hospital of Schleswig-Holstein, Campus Kiel, Kiel, Germany
5Institut für Medizinische Informatik, Statistik und Epidemiologie, Leipzig, Germany
6Second Department of Medicine, University Hospital of Schleswig-Holstein, Campus Kiel, Kiel, Germany
7Centre Hospitalier Lyon-Sud - Laboratoired'Hématologie, CH Lyon Sud, Lyon, France
8Department of Genetics, University of Navarra, Pamplona, Spain
9MDS Research Group, Institut de Recerca Contra la Leucèmia Josep Carreras, ICO-Hospital Germans Trias i Pujol, Universitat Autònoma de Barcelona, Badalona, Spain
10MVZ Dortmund, Dr. Eberhard und Partner, Cytogenetik, Dortmund, Germany
11Ernest and Helen Scott Haematological Research Institute and Department of Cancer Studies and Molecular Medicine, University of Leicester, UK
12Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología (IUOPA), Universidad de Oviedo, Oviedo, Spain

*contributed equally
#corresponding authors
§present address: Institute of Experimental and Clinical Pharmacology, Christian-Albrechts-University Kiel & University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany

Most mature B-cell lymphoid neoplasms are associated with specific immunoglobulin chromosomal translocations, whereas comparable rearrangements in chronic lymphocytic leukemia (CLL) detected by conventional cytogenetics are rare and mainly involve *BCL2*, *BCL3*, and *BCL11A* oncogenes albeit at low frequency[1]. Recently, two large studies of whole-genome sequencing (WGS) of CLL cases[2, 3] reported data on structural variants of 148 and 30 cases, respectively. The only recurrent rearrangements involved *BCL2* with inmunoglobulin (*IG*) genes as well as 13q14 rearrangements with different partners. Additionally, other non-recurrent rearrangements with *IG* genes and different chimeric genes were detected. Noteworthy, in both studies one of the most rearranged chromosomes was chromosome 14, mainly involving losses of different sizes.

B-cell malignancies, mainly CLL/small lymphocytic leukemia (SLL) carrying del(14)(q24q32) deletions have been described in the literature and the presence of 14q deletions in these cases was related with shorter treatment-free survival time, *NOTCH1*-mutations and trisomy 12[4-7]. Deletion del(14)(q24q32) occurs in around 2% of CLL[6, 8] being rare compared to the most common cytogenetic aberrations in this malignancy as 13q14-deletion (57%), trisomy 12 (14%), 11q-deletion (12%) and 17-pdeletion (7%)[2, 9]. Notably, the incidence of 13q14-deletions is only 15% in 14q-deleted CLL/SLL[7] and the reason why 14q24-q32 and 13q14 deletions seem to be mutually exclusive is still unknown.

Most of the molecular breakpoints of 14q deletion have been shown to cluster in a region around the *ZFP36L1* gene in the 14q24 chromosomal band and within the immunoglobulin heavy chain (*IGH*) locus in 14q32[7]. However, the exact breakpoints of the del(14)(q24q32) aberration, as well as the biological consequences of the deletion have not been yet described. The involvement of the *IGH*-locus, known as oncogene activator[10] and the clustering of breakpoints in 14q24 led to the hypothesis that the identified del(14)(q24q32) might activate an oncogene in 14q24 through juxtaposition to the *IGH* enhancer. Nevertheless, Pospisilova et al.[5] and Cosson et al.[7] failed to show an upregulation of *ZFP36L1* or the *RAD51B* gene, which is located centromeric to *ZFP36L1*. An alternative hypothesis was an inactivation of putative tumor suppressor gene/s in the deleted region del(14)(q24q32). In fact, biallelic inactivation of the *TRAF3* gene in chromosomal region 14q32 has been shown in 9/41 (22%) B-cell neoplasms with deletion del(14)(q24q32)[11].

In the present study, we have analyzed a total of 52 mature B-cell malignancies, mainly CLL, with 14q-deletion/*ZFP36L1-IGH* fusion by high-throughput genetic and transcriptomic sequencing, cloning or fine-mapping of genomic breakpoints and we have identified fusion-transcripts of this aberration. Moreover, we have comprehensively characterized the secondary aberrations, gene expression profile (GEP) and the clinical impact associated with 14q deletions.

Results

Identification of del(14)(q24q32)-positive B-cell lymphoma by SNP6.0 array, cytogenetics and fluorescence *in situ* hybridization (FISH) analysis. Cases included in the study are summarized in Supplementary Fig. 1. SNP6.0 array analysis depicted seven CLL carrying a deletion 14q24.1-q32.33 (cohort 1) out of 637 CLL/SLL cases. In five cases, centromeric breakpoint mapped within *ZFP36L1* gene whereas in two cases (cases 382 and 793) breakpoints were located 5' of the gene. In all seven cases telomeric breakpoint mapped to the *IGH* locus (Fig. 1a, Supplementary Fig. 2a). The global profile of copy number alterations (CNA) showed additional aberrations in all cases except one (case 1431) (Fig. 1b, Supplementary Fig. 2b, Supplementary Table 1). Screening of an additional subset of 98 lymphoma and leukemia cases harboring a cytogenetic 14q2 aberration or a loss of the proximal *IGH* signal by FISH revealed 45 cases (cohort 2) carrying a 14q-deletion with centromeric breakpoint within the *ZFP36L1* gene region and telomeric breakpoint within the *IGH* locus (Supplementary Fig. 3 and 4). CLL was the prevalent diagnosis in these cases (30/45; 67%) (SupplementaryTable 2).

Mapping of genomic *ZFP36L1-IGH* fusion breakpoints using whole-genome sequencing (WGS), Sanger sequencing, and custom array-comparative genomic hybridization (aCGH). Available WGS data from four cases of cohort 1[2] were reanalyzed for structural variants using SMUFIN[12] (Supplementary Table 3). Concordant with SNP6.0 data, the four cases presented the recurrent 37 Mb deletion (from 69 Mb to 106 Mb, Build GRCh37/hg19) in chromosome 14, which connects the *ZFP36L1* gene with the *IGH* gene (Fig. 1b-c, Fig. 2a). Fine-mapping of the chromosomal breakpoints in 14q24 was performed by

FISH in six cases from cohort 2 (Supplementary Fig. 5). The FISH pattern suggested recurrent involvement of the *ZFP36L1* gene region. Based on these findings and the knowledge about *IGH* involvement in 45 of the cases with deletion del(14)(q24q32), long distance (LD)-PCR was performed on 34 cases with available genomic DNA. By this strategy, deletion junctions could be amplified and sequenced in 17 cases, confirming the genomic fusion of *ZFP36L1* and *IGH* loci (Supplementary Table 4).

With the detailed analysis of both series, cohort 1 (4 cases, fusion positive by WGS) and cohort 2 (17 cases, fusion positive by long distance LD-PCR) we could determine the exact coordinates of the 14q deletion for 21 patients (coordinates between chr14:69256850 and chr14:69259241, Build GRCh37/hg19). All centromeric breakpoints were within the *ZFP36L1* gene in a 2.4 kb genomic window, 15 cases within the sole intron and six at the beginning of the second exon of the gene (Fig. 1c). The telomeric breakpoints affected different *IGH* segments (coordinates between chr14:106212409 and chr14:106329876, Build GRCh37/hg19). All *IGH* breakpoints clustered into the constant region, predominantly switch μ (16 cases) with the exception of case 3 showing a breakpoint affecting the J region. Two breaks arose in a switch γ1 and two breaks in a switch γ3 segment (Supplementary Table 4).

In 16 cases from cohort 2 in which LD-PCR approach was not successful, we mapped the breakpoints by high-resolution custom aCGH for chromosomal region 14q24 (Supplementary Table 5). Again, all breakpoints were located within the *ZFP36L1* gene, 14 within the intron and two within the second exon of *ZFP36L1* (Fig. 1c). Finally, FISH was performed using probes for the centromeric part of the

14q24 breakpoint region and the *IGH* locus in cases with no amplification by LD-PCR. Signal patterns indicating *ZFP36L1-IGH* could be demonstrated in all 28 analyzed cases in which amplification of junctional fragments by LD-PCR failed or with lack of available DNA (Supplementary Fig. 6). In two of the three cases from cohort 1 with no WGS available, *ZFP36L1-IGH* fusion could also be detected by FISH. In the remaining case (813) there was no suitable material left for FISH.

In total, we could demonstrate a genomic fusion of *IGH* and *ZFP36L1* in 37 cases using WGS, LD-PCR, custom aCGH and FISH analysis. The 37 Mb deletion leads to a loss of the first exon and the upstream cis-regulatory region of *ZFP36L1*, a gene encoding for the AU-rich element (ARE)-binding protein that triggers the degradation of several mRNAs.

Identification of chimeric *ZFP36L1-IGH* transcripts. We searched for the presence of chimeric *ZFP36L1-IGH* mRNAs. The RNAseq data of patients 802, 1169 and 1191 (cohort 1) revealed the existence of potentially coding fusion mRNAs in all three cases (Fig. 2a). For each of these patients, the second exon of *ZFP36L1* was involved in the chimeric transcripts, whereas the *IGH*-part varied in the 5' end, in agreement with the rearrangements observed at genomic level. The predicted longest open reading frame (ORF) contained most of the reading frame described for the *ZFP36L1* mRNA. In the 10 cases of cohort 2 with available RNA the search for the presence of *ZFP36L1-IGH* fusion transcripts was based on the knowledge about published *IGH*-oncogene fusion transcripts. In those cases, the transcripts frequently initiate from the non-translatable I exons of *IGH* germline transcripts, e.g. Iμ upstream of IGHM[13-15]. Indeed, using a reverse

transcription (RT)-PCR approach with primers targeting *Iμ* and *ZFP36L1* led to the amplification of fusion transcripts in 5 out of 10 cases. In two cases (cases 4 and 11), in which the genomic breakpoints were within the *ZFP36L1* intron, a fusion between the *Iμ* exon and *ZFP36L1* exon 2 was detected, indicating usage of the regular splice sites. Similarly to cohort 1, the predicted longest ORF contained most of the coding sequence of *ZFP36L1* mRNA. In three cases with genomic breakpoints within the *ZFP36L1* exon 2 (cases 13, 15 and 17), the *Iμ* exon fused with the 3'-terminal region of this exon using an alternative acceptor splice site at genomic position chr14:69256386; Build GRCh37/hg19 (Fig. 2a, Supplementary Table 6).

In order to assess the potential transcriptional effect of the *ZFP36L1-IGH* fusion on the chimeric mRNAs, we analyzed the RNA-seq data of samples 802, 1169, and 1191 from cohort 1 and compared the levels of expression that could be unambiguously assigned to either the rearranged allele or the wild type allele (Supplementary Results). In all three cases we detected a fraction of reads (between 22 and 39%) covering the fused mRNA region compared to the normal *ZFP36L1* mRNA (Fig. 2b). Furthermore, the reconstruction of the chimeric mRNAs in cases 802 and 1191 using read-clustering and paired-end information predicted the existence of different *ZFP36L1-IGH* isoforms derived either from the use of different transcription start sites, or from alternative forms of splicing. The read count could be specifically assigned to each of the isoforms detected in these two cases and demonstrated the presence of one form clearly predominant over the other mRNA species. In particular, case 802 showed two isoforms, one representing 90% of all *ZFP36L1-IGH*

expression. In case 1191 we detected three isoforms with relative abundances of 65, 22 and 13% (Fig. 2b).

Finally, in five out of eight cases with del(14)(q24q32) with detected *ZFP36L1-IGH* fusion transcripts (cases 802, 1191 and 1169 from cohort 1 and patients 4 and 11 from cohort 2) we identified chimeric mRNA forms with the potential to encode for a 316-385 amino acid (aa) truncated ZFP36L1 protein. These proteins lack 19-22aa compared to the wildtype ZFP36L1 protein and, thereby, harbor a truncated TIS11B domain. Interestingly, the physiological function of TIS11B domain is to recruit mRNA decay enzymes[16] (Fig. 2b). The two zinc finger domains of ZFP36L1 would remain intact. The isoforms detected involve different *IGH* gene fragments fused with part of the *ZFP36L1* mRNA. The *IGH* parts involved in these chimeric mRNAs would probably have a minor contribution to the coding potential, adding a maximum of 69 aa (isoform 2 of case 802). In the three cases with Iμ-fusion to the 3' terminal region of *ZFP36L1* a potential chimeric protein would consist of 18 aa from Iμ and 44 aa from ZFP36L1.

Aberrations accompanying del(14)(q24q32). In both cohorts with 14q-deletion affecting *IGH* and *ZFP36L1* additional genetic aberrations were identified using several methods (Supplementary Table 2 and Table 7). The six patients with either WGS or WES were screened for the presence of somatic mutations[2]. Overall, the mean number of somatic mutations in coding regions was 29 (range 14-43) (Supplementary Table 2). The only recurrently mutated gene in the 14q deleted region was *TRAF3* in two cases. *NOTCH1*, *CHD2*, *TYR*, *PHYHIPL*, *MKLN1*, and *GALNTL2* were also found to be mutated in two cases each affecting chromosomal regions outside the 14q-deleted

region (Fig. 1b). Notably, the remaining allele of *ZFP36L1* was not mutated in any case analyzed. In cohort 2, *NOTCH1* exon 34 (n = 27), *ZFP36L1* (n = 24) and *TRAF3* (n = 30) were analyzed by Sanger sequencing. *NOTCH1* was mutated in eight cases (30%) (7 out of 24 (29%) CLL) (Supplementary Tables 2 and 7). *ZFP36L1* was mutated in one CLL (case 7), showing a deletion of twelve bp within the second exon of *ZFP36L1* (c.779_781del12) (Supplementary Fig. 7). In three cases of cohort 2 (10%) we identified *TRAF3* mutations by direct sequencing. As alternative mechanism for *TRAF3* inactivation we detected homozygous deletion of *TRAF3* in 6 out of 42 (14%) cases of cohort 2 using FISH analysis [11]. Overall, 12/35 (34%) and 5/38 (13%) cases with *ZFP36L1-IGH* fusion have *NOTCH1* (activating mutation) and *TRAF3* (biallelic inactivation) alterations, respectively (Supplementary Fig. 7).

Because all seven cases of cohort 1 and most of the cases of cohort 2 have a CLL diagnosis we determined the incidence of CLL typical aberrations (13q14-deletion (13q-del), trisomy 12 (Tri12), 11q-deletion (11q-del) and *TP53* deletion (*TP53*-del) by SNP6.0, FISH and cytogenetics in cohort 1 and FISH and cytogenetics in cohort 2. Results and applied technologies are shown in Supplementary Table 2. Trisomy 12 was enriched in *ZFP36L1-IGH* CLLs compared to CLL: 5/7 (71.4%) in cohort 1 and 13/33 (39.4%) CLLs in cohort 2 compared to 63/444 (14.2%) (P<0.001) considering the ICGC CLL series. Conversely, 13q-del is depleted in *ZFP36L1-IGH* CLLs compared to 14q-wild type CLL: 2/7 (28.6%) in cohort 1 and 3/33 (9.1%) in cohort 2 compared to 220/444 (49.6%) (*P<0.001*) considering the ICGC CLL series.

Translin motif enrichment in *ZFP36L1-IGH* fusion cases. We explored the possible underlying mechanism for the generation of the interstitial 14q deletion. First, we searched for short microhomologies around the breakpoints, but could not detect them. Next, we searched for recurrent sequence motifs around the junction breakpoints and identified a single 13 bp motif significantly enriched within 200 bp windows around the breakpoints ($p = 0.000005$, see "Methods") in both genes (*ZFP36L1* and *IGH*) compared to the genomic background significantly enriched within 200 bp windows around. This 13 bp motif matches with the sequence GCCC[A/T][G/C][G/C] known to be recognized by the DNA-binding Translin protein (Fig. 2c), which has been previously pointed as a potential mediator in the fusion of *IGH* genes and the *BCL1 (CCND1)*, *BCL2*, *BCL6*, *IL3* and *MYC* genes [17, 18].

Deregulated gene expression in cases with *ZFP36L1-IGH* fusion. To identify candidate genes potentially deregulated through the *ZFP36L1-IGH* fusion, we compared the *ZFP36L1-IGH* fusion cases *vs* all 14q-wild type CLL, and found a total of 571 differentially expressed probe sets (405 up-regulated and 166 down-regulated) (Supplementary Table 8). Of note, 63% of the down-regulated probe sets belong to 51 genes located in the commonly deleted region of chromosome 14 (Supplementary Table 9; Fig. 2d). Only, eight genes located in 14q were up-regulated, among them *RAD51B* and *ZFP36L1,* the nearest genes to the fusion breakpoint in chromosomal region 14q24 (Fig. 2d). However, in the *ZFP36L1-IGH* cases of cohort 2 a significant up-regulation of *RAD51B* or *ZFP36L1* could not be observed using quantitative RT-PCR in eight *ZFP36L1-IGH* cases, compared to four CLLs without 14q-aberration (Supplementary Fig. 8). Differential gene expression analysis in 14q-deleted CLL cases

without *ZFP36L1-IGH* fusion (n = 6) is presented in Supplementary Results and Supplementary Fig. 9).

Next, we performed differential expression analysis from RNA-seq data from three *ZFP36L1-IGH* cases (cohort 1) compared to cases with no chromosome 14 deletion. In the three *ZFP36L1-IGH* cases we found 90/460 (19%) of the total downregulated genes correspond to genes of chromosome 14. Of note, 81 out of 90 genes (90%) were within the 14q deleted region (Supplementary Table 10). Overall, these results suggest that the deletion itself seems to be the main consequence of most of the downregulation observed. Forty-seven genes located in the minimal deleted region were found commonly downregulated by both microarray and RNA-seq analysis.

In a pathway enrichment analysis using the deregulated genes (excluding the genes affected by the deletion) we found a five-fold enrichment on mRNA processing gene ontology (GO) biological processes, related with ZFP36L1 function as transcriptional regulator at mRNA level[19-21] (Supplementary Table 11). Finally, we also explored deregulation of the ZFP36L1 targets according to Zekavati *et al*. publication[19]. Only 3 out of the 69 genes were significantly deregulated (*SSX2IP, LSM11* downregulated and *RAD1* upregulated).

Clinical characteristics of *ZFP36L1-IGH* cases. The seven CLL patients with the *ZFP36L1-IGH* fusion from cohort 1 had a median age of 62 years (range 44-85 years), were mainly female (71%), and present with Binet Stage A (71%) and B (29%), but none was in stage C. *IGHV* was unmutated in all cases except one. Only one of the patients had died, and all of them had been treated, five cases had a complete response whereas in two cases the response could not be assessed. CLL patients with *ZFP36L1-IGH* fusion had significantly

shorter time to first treatment (TTT) than patients without that fusion (*P*< 0.001; Hazard Ratio [HR] = 4.52; 95% Confidence Interval [CI] 2.00-10.24). Interestingly, CLL patients with 14q24 loss but no *ZFP36L1-IGH* fusion have also a significantly shorter TTT than patients with no alterations in chromosome 14, similar to cases carrying *ZFP36L1-IGH* fusion (*P* = 0.64; HR = 1.33; 95% CI 0.43-4.16) (Supplementary Fig. 10). In the 33 CLL from cohort 2, the median age at diagnosis was 64 years (range 47-82 years) with a slight male predominance (55%), and *IGHV* was unmutated in 76% CLL patients. Since these patients were derived from multiple institutions with different treatment regimens the clinical data cannot be compared accurately.

Discussion

Deletions in the long arm of chromosome 14 are recurrent events in B-cell malignancies, especially in CLL[5-7]. This study describes a global genetic characterization of mainly CLL cases with 14q deletions focusing on cases with breakpoints in the *ZFP36L1* region (14q24) and the *IGH* locus (14q32). Using several experimental and computational techniques we have molecularly characterized a deletion of 37 Mb within the long arm of chromosome 14 in 37 CLL patients and twelve other B-cell lymphomas. Through independent approaches, such as LD-PCR and the analysis of WGS using SMUFIN algorithm[12], the exact coordinates of the *ZFP36L1-IGH* fusion for 18 CLL as well as three other B-cell lymphomas were determined. The resulting 37 Mb deletion in these cases brings together different parts of *IGH* with the *ZFP36L1* gene, which has lost its first exon as well as any upstream cis-regulatory region.

Additionally, SNP6.0, custom array and FISH analysis fine-mapped the deletion breakpoints in 14q24 in 14 CLLs and five B-cell lymphomas with deletion *ZFP36L1-IGH*. Taken together, 29 of the breakpoints are located in the intron and nine in the second exon of the two exons-comprising *ZFP36L1* gene.

The recurrent occurrence of 14q-deletion breakpoints in the *ZFP36L1* gene region in 14q24 has been described by other groups without determining the exact breakpoint position by sequencing[5-7]. Even though 14q-deletions with breakpoints centromeric and telomeric of the *ZFP36L1* region are described in mature B-cell neoplasms, the clustering of the centromeric breakpoints in the *ZFP36L1* gene region in 57-62% of the 14q-deletion cases is remarkable[6, 7]. The clustering of the breakpoints in the *ZFP36L1* gene could be due to either the presence of sequence motifs in the affected chromosomal region or to biological mechanisms that lead to a gain of the cells fitness. Notably, by translocation-capture sequencing in mice it has been shown that the *ZFP36L1* gene belongs to the 83 genes with activation induced cytidine deaminase (AID) dependent hotspots for translocations[22]. Furthermore, ChIP-Seq using anti-AID antibodies revealed that AID associates with *ZFP36L1* in human B-lymphocytes[23]. The susceptibility of *ZFP36L1* for AID may also be due to the fact that in diffuse large B-cell lymphoma *ZFP36L1* is among the 44 targets of somatic hypermutation, a process initiated by AID[24]. Given that AID is absolutely required for class switch recombination (CSR) and the 20 out of the 21 sequenced del(14)(q24q32) breakpoints in our study are located within the switch regions of *IGH,* one might assume that the underlying mechanism of the del(14)(q24q32) could be an illegitimate CSR.

Interestingly, the precise fine-mapping of the breakpoints allowed us to detect consensus recognition motifs of translin around the breakpoints in the 17 cases tested. Translin is a single-stranded DNA and RNA binding protein suggested to be involved in chromosomal translocations, telomeres, mRNA transport and translation[18, 25-27]. Translin and Translin-associated factor X (TRAX) have a high homology and act together mediating several crucial biological processes, such as chromosomal translocations, regulation of genome stability[28-30] and telomere metabolism[25]. Initially, Translin was isolated as a protein that binds to a consensus motif in the breakpoints of 91 lymphoid malignancies, comprising 16 different tumor types, involving *IGH* (with *MYC, CCND1, BCL2, BCL6, IL3,* etc.), or *TCR* (with TAL1, TCL3, TTG2, TAN1 etc.), as well as different fusion genes (*BCR/ABL, HLF/E2F*)[18]. Moreover, translin motifs have also been found in solid tumors such as liposarcomas with the t(12;16)(q13;p11)[31], and alveolar rhabdomyosarcoma with the t(2;13)(q35;q14)[32].

Biologically the *ZFP36L1* gene encodes for the ZFP36L1 protein which is a member of the TIS11 family of early response genes. ZFP36L1 has been shown to mediate decay of AU-rich element mRNAs (ARE-mRNAs) encoding proteins involved in proliferation and apoptosis like BCL2, API2 (Apoptose inhibitor 2), BLIMP1, CDKN1A (p21) and NOTCH1[19, 33-35]. This occurs by binding of ZFP36L1 to AU-rich elements (ARE) in the 3' untranslated region of the target-mRNAs, mediating its accumulation in processing (P-)bodies, sites of mRNA decay, as well as their decay itself. A role of ZFP36L1 as tumor suppressor is postulated, given that it is required for Rituximab-mediated apoptosis of CLL cells[21] and expression of *ZFP36L1* increases after induction of anti-CD20 and B-cell receptor-

mediated apoptosis in B-cell lymphoma cell lines[36]. Moreover, Rapamycin, an inhibitor of mTOR, has a reduced ability to suppress the senescence-associated secretory phenotype (SASP) in cells lacking ZFP36L1[37]. A dominant negative effect through truncation of the ZFP36L1 protein as a result of the *ZFP36L1-IGH* fusion is conceivable as it is composed of different domains mediating either the decay (N-terminal domain), the ARE-mRNA-binding (zinc finger domains) or the transport of the ARE-mRNAs to P-bodies (C-terminal domain). The detection of *ZFP36L1-IGH* fusion transcripts with truncated *ZFP36L1* supports this hypothesis, whereas the fact that in three cases with breakpoint within the second exon of *ZFP36L1* the detected fusion transcripts might merely encode for 44 amino acids makes it questionable. Moreover, RNA-seq showed that the predominant expressed *ZFP36L1* transcript in CLL with *ZFP36L1-IGH* was the non-rearranged wildtype allele. However, additional facts argue for a role of *ZFP36L1* in CLL with *ZFP36L1-IGH*. It is remarkable that in general more than half of CLL patients show 13q14-deletions by FISH[9], but in the CLL with *ZFP36L1-IGH* described herein only in 12% of the cases 13q14 deletions have been detected. The genes in the 13q14 minimal deleted region are the miRNAs miR-15 and miR-16, which are known to be involved in mRNA-decay, particularly in AU-rich element mediated mRNA decay[38]. It has been shown that miR-16 requires the presence of ZFP36 (another TIS11 familiy member) and ZFP36L1 in order to bind to ARE-containing mRNA[39]. Deletion 13q14 and *ZFP36L1-IGH* might represent alternative mechanisms for the inhibition of ARE-mediated mRNA decay. In contrast to 13q14 deletions, which are significantly underrepresented in CLL with *ZFP36L1-IGH*, Cosson et al.[7] and our results have shown that *NOTCH1* mutations are enriched in CLL with del(14)(q24q32) (47% and 29%,

respectively) compared to CLL in general (12%)[2]. Interestingly, *NOTCH1* mRNA is a target of ZFP36L1[35]. The effect of the gain-of function mutation of *NOTCH1* might be enhanced by a putative stabilizing effect of a truncated ZFP36L1 protein in *ZFP36L1-IGH* cases. However, a significant deregulation of the ZFP36L1 targets according to Zekavati et al.[19] in *ZFP36L1-IGH* CLL compared to CLL without 14q-aberration could not be stated based on RNA-seq data, but the differentially expressed genes were enriched on genes related with RNA and mRNA processing.

Interestingly, most genes located in the 14q deleted region in *ZPF36L1-IGH* CLL cases were clearly downregulated, whereas *ZFP36L1* and *RAD51B,* the nearest non-truncated genes centromeric to the 14q24-breakpoint,were significantly overexpressed. A significant upregulation of *ZFP36L1* and *RAD51B* in *ZFP36L1-IGH* CLL compared to CLL without 14q-aberration has not been seen in Cohort 2 of the present study neither in the studies of Cosson et al. and Pospisilova et al.[5, 7]. This might be due to contamination of non-B-cells in the latter as they have not been sorted for B-cells. RAD51B protein is as a member of the RAD51 family involved in the repair of double strand breaks by homologous recombination[40]. In tumors like breast or pancreatic cancer these proteins have been shown to be down regulated due to deletion or truncation as well as upregulated by e.g. amplification of the gene[41, 42]. The close proximity of the strong *IGH* enhancers to the *RAD51B* locus through *ZFP36L1-IGH* fusion could result in overexpression of *RAD51B*. Given that the genes in the deleted region del(14)(q24q32) are downregulated, the biological consequence of the del14q/*ZFP36L1-IGH* might be an inactivation of a tumor suppressor. Indeed, mutations in the remaining allele of *TRAF3* in chromosomal region 14q32 have been identified in 9/41 (22%)

ZFP36L1-IGH cases from Cohort 2[11] and confirmed in two out of six cases from Cohort 1. In some of the cases the biallelic inactivation of *TRAF3* was present in small sub clones arguing for a secondary event in tumorigenesis.

Similar to previous studies, the *ZFP36L1-IGH* fusion CLL presented herein are associated with Tri12 and unmutated *IGHV*[5-7]. Clinically, CLL cases with *ZFP36L1-IGH* as well as CLL with 14q-deletion with different breakpoints showed shorter TTT. All patients needed treatment, and most achieved complete response.

In summary, our data have shown that the recurrent deletion del(14)(q24q32) with breakpoints within *ZFP36L1* and *IGH* occurring predominantly in CLL lead to *ZFP36L1-Iμ*-fusion transcripts, reduced expression of the genes in the deleted region and overexpression of *ZFP36L1* and *RAD51B*. Moreover, by cloning 17 *ZFP36L1-IGH* breakpoints, consensus recognition motifs of translin around the breakpoints have been identified giving a hint to an underlying mechanism behind the *ZFP36L1-IGH* fusion. The biological consequence of this recurrent deletion might be a truncated ZFP36L1-protein that reduces AU-rich element mediated decay, overexpression of RAD51B, and inactivation of tumor suppressors in the deleted region in 14q24-q32, like *TRAF3,* all potentially promoting cancer development.

Methods

Patients and samples. Cohort 1: From 637 cases of the CLL - International Cancer Genome Consortium (ICGC) analyzed by SNP6.0 arrays, fifteen cases (2.4%) displayed 14q deletions (Supplementary Fig. 1A). Seven of them, diagnosed as CLL or SLL carrying breakpoints at 14q24 and 14q32 involving *ZFP36L1* and *IGH* genes, were selected as cohort 1[2, 43]. Patient characteristics are described in Supplementary Table 2. Two additional patients with a similar pattern of chromosome 14q deletion (498 and 1193) were not included in the analysis due to the lack of WGS or cytogenetic material to validate the exact breakpoints. Whole-genome sequencing (WGS) and whole-exome sequencing (WES) data were available from six of the seven cases (WGS for cases 802, 1169, 1191 and 1431, and WES for cases 382, 802, 813, 1169) (Supplementary Table 2-3)[2]. The tumor samples were obtained before administration of any treatment. All patients gave informed consent for their participation in the study following the ICGC guidelines[44]. Sequencing, expression and genotyping array data have been deposited at the European Genome-Phenome Archive (EGA, https://www.ebi.ac.uk/ega/), which is hosted at the European Bioinformatics Institute (EBI), under accession number EGAS00000000092. All patients from cohort 1 gave informed consent according to International Cancer Genome Consortium (ICGC) guidelines and ethics policy committee[44].

Cohort 2: B-cell malignancies with cytogenetically identified translocations, deletions or additions of material of unknown origin in which chromosomal region 14q2 was affected (n = 73) were selected from the cytogenetic databases of the laboratories involved in the present study. The aberrations were detected by conventional G- or

R-banding chromosomal analysis performed according to routine methods in each of the institutions. Twenty-five B-cell neoplasms with available cytogenetic pellets harboring a specific FISH-pattern (loss of the proximal signal by using the LSI IGH Dual Color, Break Apart-probe (Abbott/Vysis)) but lacking a clonally aberrant karyotype were also included in the screening-cohort. Supplementary Fig. 3 illustrates the selection of cases for cohort 2. Characteristics of the 45 cases with FISH-proven break in the *ZFP36L1*- and *IGH*-gene regions that make up cohort 2 are listed in Supplementary Table 2. All cases analyzed and techniques applied are provided in Supplementary Fig. 1 and Table 2).

Copy number arrays. SNP-array experiments on cohort 1 were outsorced at CeGen (www.cegen.org). Nexus version 7.5 Discovery Edition software (Biodiscovery, El Segundo, CA) was used for global analysis and visualization of results. Array CGH on cohort 2 was performed using the Human Genome CGH Microarray Kit 44A and a 44K-Custom-Array designed with the eArray Software 6.2 (Arrays and software provided by Agilent) (Supplementary Methods).

Whole-exome and whole-genome sequencing. Whole-genome and -exome sequencing (Agilent SureSelect Human All Exon 50 MB) were performed as previously described[2, 45, 46]. Sequence data analysis was performed using the Sidrón mutation caller[46] and SMUFIN algorithm for structural variants[12]. The results were further verified by manual inspection of the corresponding BAM files obtained, as described elsewhere[46].

Cloning of the del(14)(q24;q32) breakpoints. For cloning of genomic breakpoints in cohort 2 different combinations of 18 forward primers in 14q24 and 13 reverse primers in 14q32 (Supplementary Table 12) were used for LD-PCR. The polymerase TaKaRa LA Taq (TAKARA BIO INC.) and a touchdown PCR program was applied according to manufacturers instruction. A nested PCR reaction using 1µl of 1:100-diluted amplicon was applied to enhance specificity.

Molecular cytogenetic and cytogenetic analyses. Conventional cytogenetics (CC) was performed on G-banded (cohort 1) or R-banded (cohort 2) chromosomes obtained after short term culture without stimulation or with stimulation with Phorbol 12-Myristate 13-Acetate. Results were described according to the International System for Human Cytogenetic Nomenclature[47]. FISH was performed on cytogenetic suspensions according to standard protocols[48]. The bacterial artificial chromosome (BAC) and Fosmid clones selected from the Human Genome Browser Gateway (http://genome.ucsc.edu/cgi-bin/hgGateway) and used for FISH are described in Supplementary Table 13. The commercial FISH probes applied in the study are listed in Supplementary Table 14. For each test, the signal constellations of a minimum of 100 nuclei were counted. Slides were evaluated by two observers.

***IGHV-, ZFP36L1-, TRAF3* and *NOTCH1*-mutation analysis.** *IGHV* and *NOTCH1* exon 34 mutational analyses (cohorts 1-2) and TRAF3 (cohort 2) were performed using direct Sanger sequencing as previously described[11, 45, 49, 50]. In cohort 2, the coding region of *ZFP36L1* including exon/intron-boundaries were PCR amplified

(Supplementary Table 15) according to standard protocols and subsequently sequenced.

Identifying translin motifs. Agnostic repetitive motif analysis was performed using standard MEME[51] parameters within 200 bp windows of sequence fragments spanning the 14q breakpoints junctions identified in 17 patients. The program was forced to provide only those motifs present in all tested samples at least once. The motif identified as recurrent in all samples was then mapped on all patients by performing matrix-guided alignments on the corresponding 200 bp windows using MatScan with a threshold of 0.75[52].

Identification of fusion transcripts using RT-PCR. Information on the genomic breakpoints was used to choose primer pairs for the detection of fusion transcripts (Supplementary Table 6). Depending on the expected product size the Gold Star Taq Polymerase (Eurogentec, Seraing, Belgium) or the Expand High Fidelity PCR System (Roche Diagnostics, Mannheim, Germany) were used under conditions recommended by the manufacturers. Depending on the PCR-results the conditions have been modified and a touch-down-PCR was done.

TOPO TA cloning. A subset of PCR products obtained from the breakpoint PCR and the RT-PCR to detect fusion transcripts in Cohort 2 have been cloned using the TOPO TA cloning kit with the pCR 2.1-TOPO Vector and chemically competent One Shot TOP10 E. coli cells (Life Technologies). Clones were PCR amplified and Sanger sequenced.

Gene expression profiling. Total RNA was extracted with the TRIzol reagent following the recommendations of the manufacturer (Invitrogen Life Technologies) and hybridized to Affymetrix GeneChip Human Genome U219 arrays, as previously described[53]. We studied the GEP of CLL cases (802,1169,1191) with *ZFP36L1-IGH* fusion compared to the remaining 14q-wildtype 458 CLL cases. To evaluate the impact of *ZFP36L1-IGH* fusion/del(14)(q24q32) deletion on gene expression, summarized expression values were computed using the robust multichip average (RMA) approach implemented in the Expression Console Software (Affymetrix Inc.). Limma was used to detect probe sets differentially expressed between two or more groups. P-values were adjusted for multiple comparisons using the Benajmini-Hochberg method. Probe sets with an adjusted P-value below 0.15 were considered significant.

RNA-seq analysis. RNA-seq libraries from 3 cases with ZFP36L1-IGH fusion (802,1169,1191) and 129 CLL cases with no 14q deletions were prepared from total RNA using the TruSeq™ RNA Sample Prep Kit v2 (Illumina Inc.,) as previously described[53]. RNA-seq data was processed with the ENCODE pipeline for long RNAs v2.0.0 (https://github.com/ENCODE-DCC/long-rna-seq-pipeline). The 76-bp paired-end reads were aligned to the reference genome (hs37d5) and long transcriptome (subset of GENCODE v19[54]) corresponding to long transcripts) with STAR [55]. The mapping was performed using a sex-specific reference sequence including the Y chromosome for males but not for females. Gene and transcript expression levels were quantified in FPKM (Fragments per Kb of exon per Million mapped reads)[56] from transcriptome mappings with RSEM package[57].

Differential gene expression analysis of RNA-seq data. To evaluate the impact of the *ZFP36L1-IGH* fusion/del(14)(q24q32) on global gene expression, RNA-seq data of *ZFP36L1-IGH* fusion CLL cases (n = 3) vs CLL cases without 14q-alterations (n = 129) were compared. RNA-seq differential expression analysis on normalized FPKM data was performed using the limma package[58, 59] using a P-value ≤ 0.05 for significance. Enrichment analyses were done over three categories: all significant differentially expressed genes, significantly overexpressed and significantly down-regulated genes.

Allele- and isoform- specific expression. Allele-specific expression (ASE) for *ZFP36L1-IGH* fusion and normal *ZFP36L1* alleles was computed. The ASE quantifies the contribution of each allele to the global expression of *ZFP36L1* gene. In order to count: (I) the number of split-mapped reads spanning the chimeric splice-junctions for the fusion mRNA isoforms (fusion junction reads) and (II) the number of split-mapped reads spanning the splice-junction specific and common for all the normal mRNA isoforms (normal junction reads), read-pairs to the normal and chimeric mRNA isoforms were calculated. ASE fusion was computed (ASEF) as the ratio between the fusion reads and fusion reads plus normal reads. Finally, normal ASE (ASEN) was calculated as 1-ASEF. The isoform- specific expression (ISE) for normal ZFP36L1 and chimeric *ZFP36L1-IGH* isoforms was calculated in normal B cells from 3 healthy individuals and the 3 CLL cases (cohort 2). To compute the ISE, the number of split-mapped reads spanning the splice-junction specific for each of the isoforms (junction spanning reads) were counted. The ISE for each isoform was calculated as the ratio between the number of junction spanning reads for the isoform and the sum of junction spanning reads for all the possible normal or chimeric isoforms (Supplementary Table 16).

Statistical analysis. Fisher's exact test was used to compare the frequencies of trisomy 12 and 13q deletion in CLL with *ZFP36L1-IGH* fusion (cohorts 1 and 2) with the general frequency of these aberrations in the ICGC CLL series. TTT of cases with *ZFP36L1-IGH* fusion, cases with del(14)(q24) and CLL cases without any 14q abnormality was evaluated. TTT curves from date of sampling were plotted by the Kaplan and Meier method and compared by the log-rank test.

FIGURE LEGENDS

Figure 1. Genetic features of CLL/SLL and B-cell lymphomas with 14q-deletion. (a) Copy number profile of chr14 in the seven cases with 14q-deletion identified by SNP array (cohort 1). (b) Circos plot representation of seven samples analyzed by WGS and SNP6.0 array. Blue inner lines represent intrachromosomic translocations and black inner lines represent interchromosomic translocations. The histogram depicts gains and losses of the seven samples colored in blue and red respectively. Finally the outer dots represent all exonic point mutations in the seven whole genomes. The size of the points correlate with mutation frequency in this series. (c) *ZFP36L1* representation showing all breakpoints obtained from different platforms: breakpoints from SNP6.0 array in orange, from whole-genome-sequencing in yellow, from LD-PCR sequencing in grey and from 14q-custom array in red (Case 46 is not a *ZFP36L1-IGH* case as it has its distal breakpoint centromeric to *IGH*). Blue squares depict exons and the green one depicts an intron.

Figure 2. Expression of *IGH-ZFP36L1* transcripts and isoforms and deregulated gene expression profile. (a) Recurrent deletion of chromosome 14 produces chimeric *IGH-ZFP36L1* mRNAs. Recurrent deletions between chromosomal regions 14q24 and 14q32 identified in CLL/B-NHL and the corresponding chimeric mRNA products are shown. For the sake of clarity, the involved 35 Mb region is shown in reverse. Dashed blue lines indicate, for each of the deletions characterized, the positions of the breakpoints in the immunoglobulin heavy chain (*IGH*) gene region in 14q32 (left) and the *ZFP36L1* gene in 14q24 (right).The resulting chimeric mRNA forms (center) are shown with the predicted protein underneath. Most abundant isoform for each fusion gene is represented (see Figure 2b). The truncated Tis11B domain, as well as downstream complete zinc finger regions, are shown as they have been described for the normal ZFP36L1 protein (top right panel). In patients 13, 15 and 17 the predicted protein has no annotated functional domains to represent. On the *IGH* locus, we indicate different immunoglobulin regions as annotated in the IMGT database, above the 300 Kb black line, and the genomic position of the exons (as grey boxes below), which are incorporated into the spliced chimeric mRNAs forms. These regions have been annotated according to the IMGT, GENCODE and UCSC databases (see Methods). Regarding the *ZFP36L1* region, we display (in yellow) the genomic structure of two isoforms described in RefSeq and as expressed in *ZFP36L1-IGH* cases according to our RNAseq (802, 1169 and 1191) and RT-PCR-based sequencing data. (b) Expression of different chimeric *IGH-ZFP36L1* mRNA isoforms in CLL patients. Representation of the identified isoforms deriving from the different deletions and reconstructed using RNAseq read data for three CLL samples. The chimeric mRNA isoforms and their

corresponding predicted proteins are shown. Allele and isoform specific expression levels were inferred from the relative count of unambiguous read abundance across splice junctions and are depicted as chart pies. The length for each of the resulting longest potential coding regions is shown in number of amino acids, indicating the fraction derived from the *IGH* and *ZFP36L1* regions. Protein domains are shown for each of the isoforms following the colour code used in figure 1a. When the Tis11B domain is truncated the percentage of the remaining fragment is specified. (c) Schematic view of the distribution of the 13 bp motif sequence around the breakpoint region in *ZFP36L1* in cohorts 1 and 2. This 13bp motif is overrepresented in the *ZFP36L1* and *IGH* breakpoint regions and matches with the sequence GCCC[A/T][G/C][G/C] known to be recognized by the DNA-binding Translin protein. (d) Heatmap of the expression of the genes (microarray data) of chromosome 14 in CLL patients grouped on the basis of *ZFP36L1-IGH* fusion (red) and cases with wild type chromosome 14 (14q- wt) (blue). In the group "14q wt" only 40 random CLL cases (out of 458) were represented. Cases with 14q deletion have downregulation of genes of the deleted region, whereas the three cases with *ZFP36L1-IGH* fusion have upregulation of a few genes, including *RAD51B* and *ZFP36L1*, juxtaposed to the *IGH* due to the deletion.

Figure 1

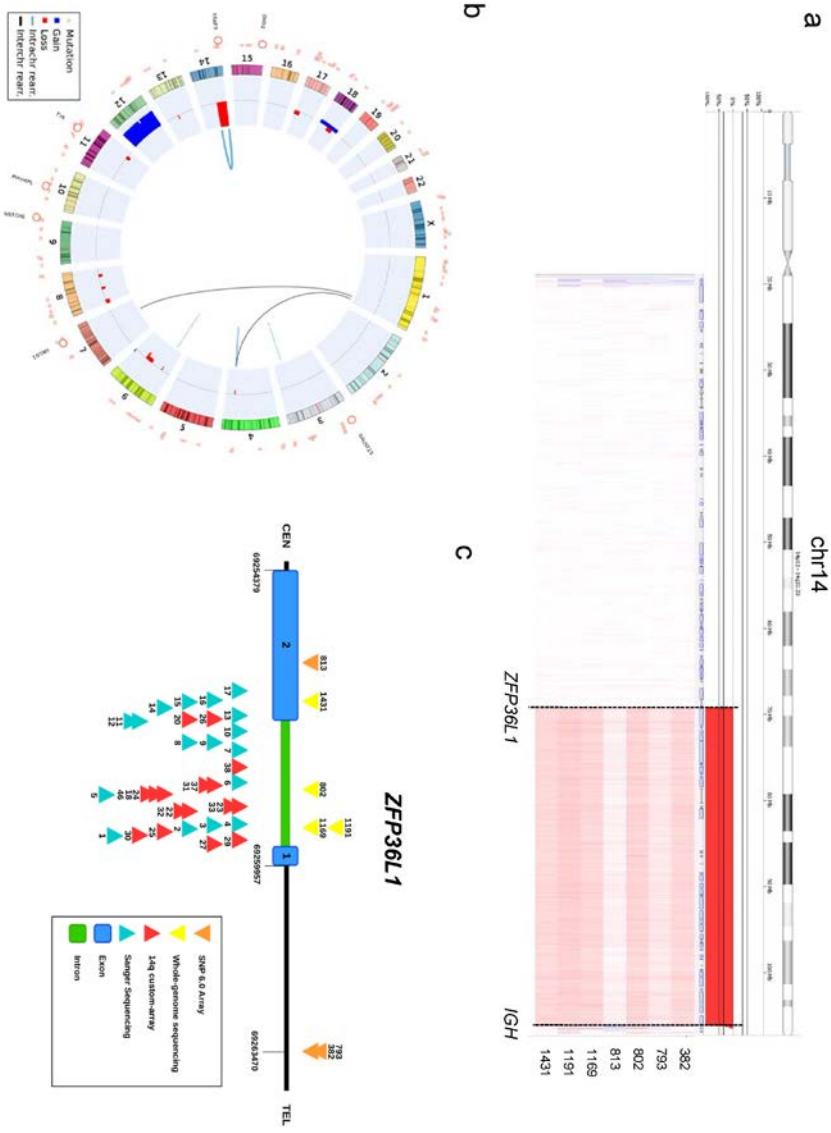
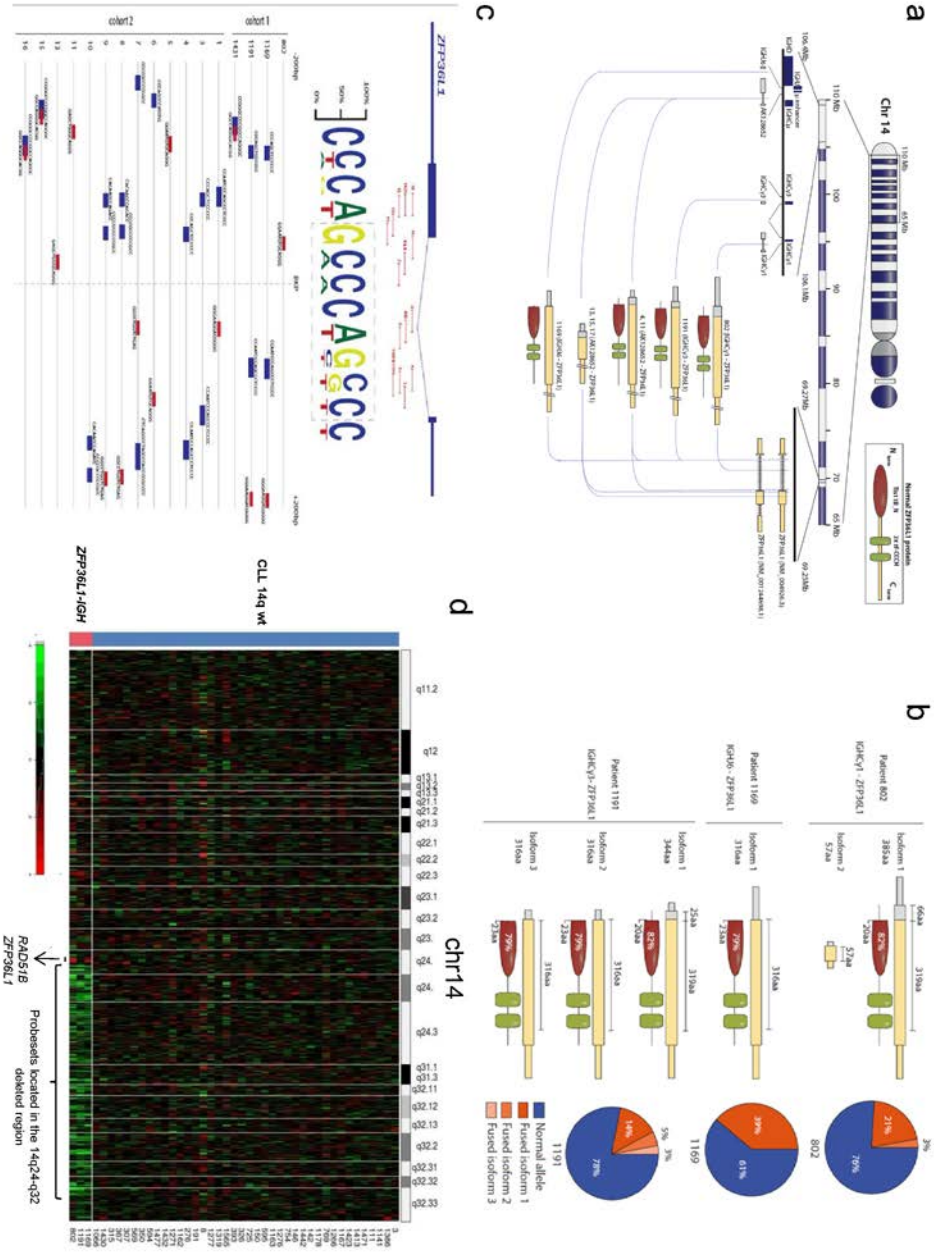
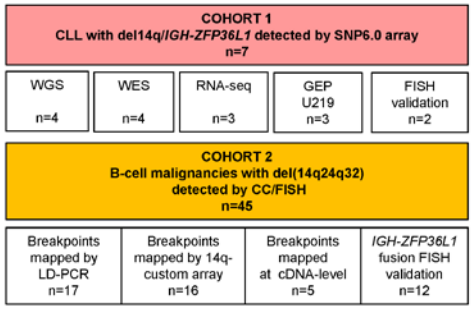

a
chr14
ZFP36L1
IGH
382
793
802
813
1169
1191
1431
b
Mutation
Gain
Loss
Intrachr rearr.
Interchr rearr.
c
ZFP36L1
CEN
TEL
69254379
69259957
69263470
SNP 6.0 Array
Whole-genome sequencing
14q custom-array
Sanger Sequencing
Exon
Intron

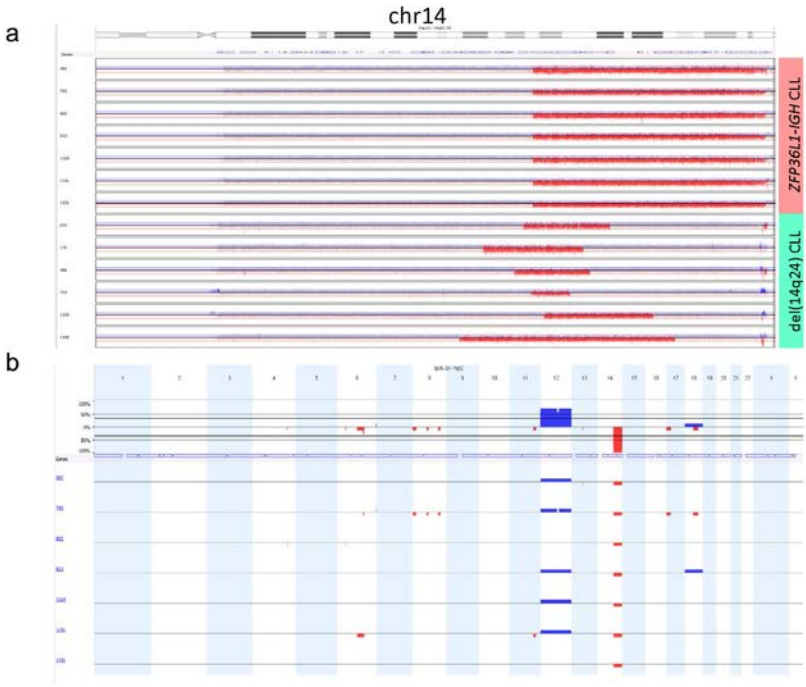
Figure 2


a
Chr 14
Normal ZFP36L1 protein
Zn-F-CCCH
b
Patient 802
IGHCγ1 - ZFP36L1
Isoform 1
385aa
Isoform 2
527aa
Patient 1169
IGHJH6 - ZFP36L1
Isoform 1
316aa
Patient 1191
IGHCγ3- ZFP36L1
Isoform 1
344aa
Isoform 2
316aa
Isoform 3
316aa
Normal allele
Fused isoform 1
Fused isoform 2
Fused isoform 3
c
ZFP36L1
cohort 1
cohort 2
d
CLL 14q wt
ZFP36L1-IGH
chr14
RAD51B
ZFP36L1
Probesets located in the 14q24-q32
deleted region

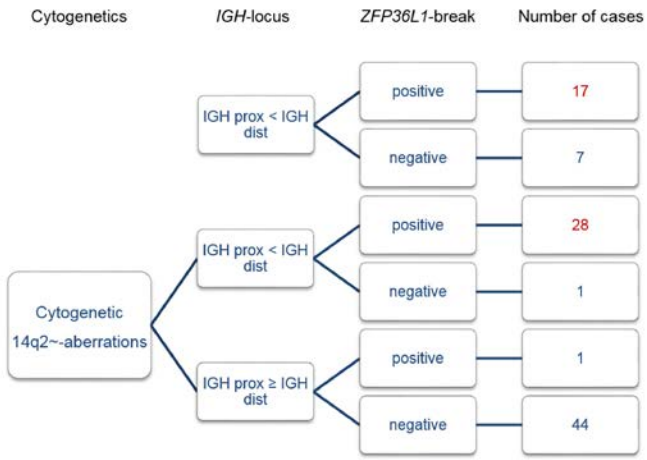
Suppl. Figure S1



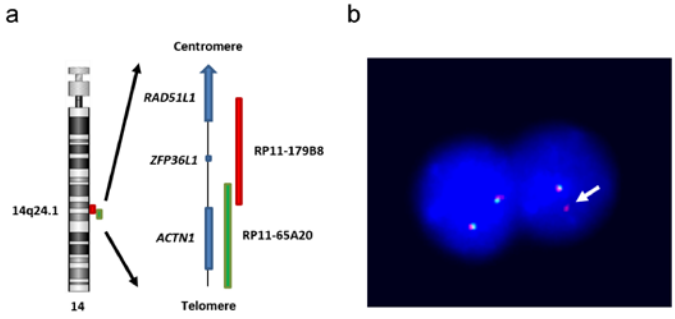
Suppl. Figure S2



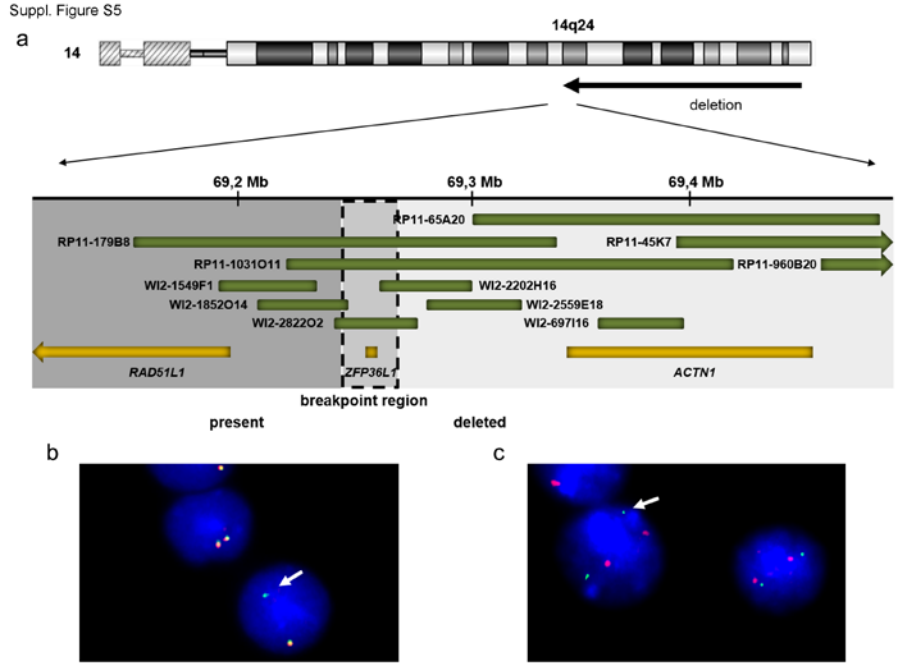
Suppl. Figure S3



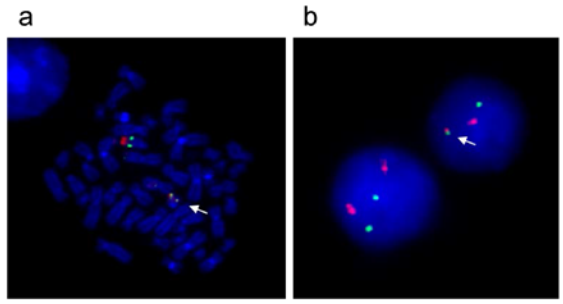
Suppl. Figure S4



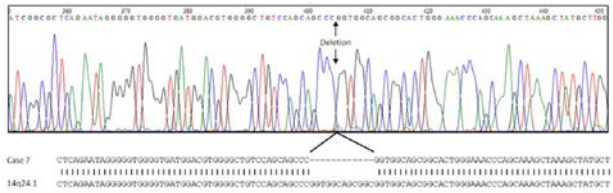
Suppl. Figure S5



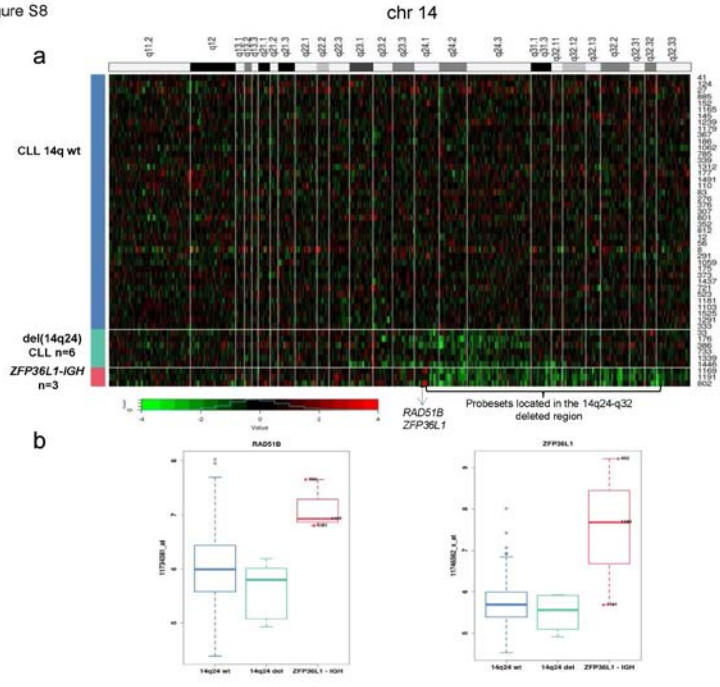
Suppl. Figure S6



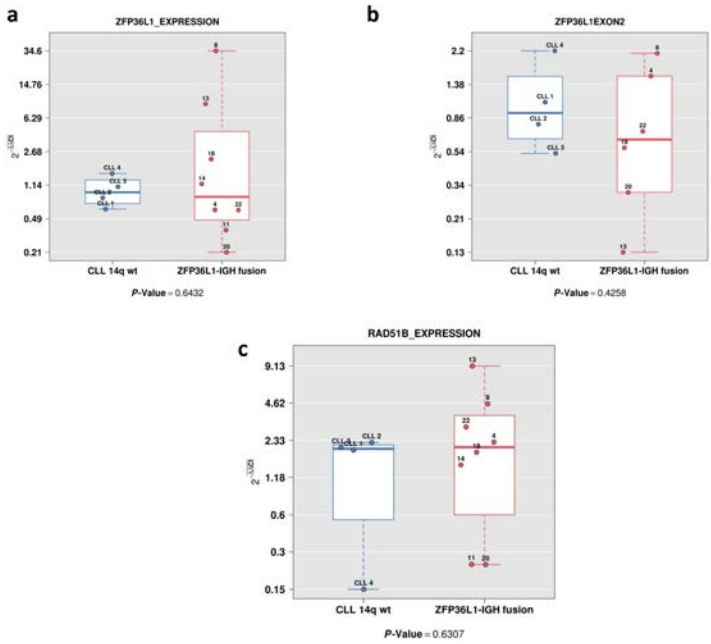
Suppl. Figure S7



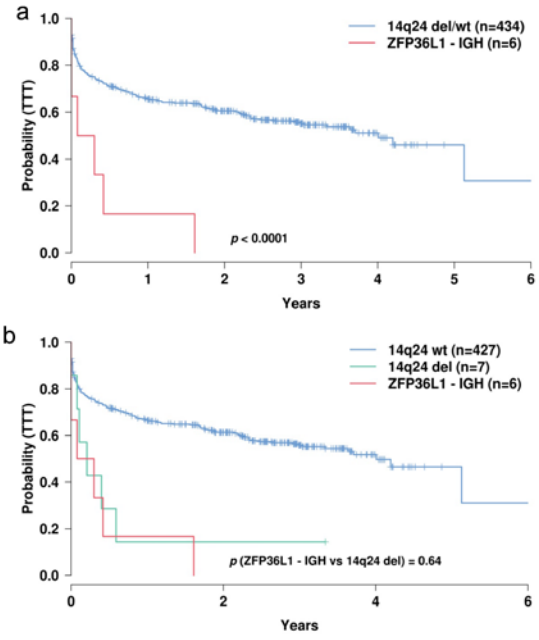
Suppl. Figure S8



Suppl. Figure S9



Suppl. Figure 10



References

1. Dyer,M.J. & Oscier,D.G. The configuration of the immunoglobulin genes in B cell chronic lymphocytic leukemia. *Leukemia* 16, 973-984 (2002).

2. Puente,X.S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526, 519-524 (2015).

3. Kasar,S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* 6, 8866 (2015).

4. Tilly,H. *et al.* Del(14)(q22) in diffuse B-cell lymphocytic lymphoma. *Am. J. Clin. Pathol.* 89, 109-113 (1988).

5. Pospisilova,H. *et al.* Interstitial del(14)(q) involving IGH: a novel recurrent aberration in B-NHL. *Leukemia* 21, 2079-2083 (2007).

6. Reindl,L. *et al.* Biological and clinical characterization of recurrent 14q deletions in CLL and other mature B-cell neoplasms. *Br. J. Haematol.* 151, 25-36 (2010).

7. Cosson,A. *et al.* 14q deletions are associated with trisomy 12, NOTCH1 mutations and unmutated IGHV genes in chronic lymphocytic leukemia and small lymphocytic lymphoma. *Genes Chromosomes. Cancer* 53, 657-666 (2014).

8. Gunnarsson,R. *et al.* Array-based genomic screening at diagnosis and during follow-up in chronic lymphocytic leukemia. *Haematologica* 96, 1161-1169 (2011).

9. Haferlach,C., Dicker,F., Schnittger,S., Kern,W., & Haferlach,T. Comprehensive genetic characterization of CLL: a study on 506 cases analysed with chromosome banding analysis, interphase FISH, IgV(H) status and immunophenotyping. *Leukemia* 21, 2442-2451 (2007).

10. Kuppers,R. Mechanisms of B-cell lymphoma pathogenesis. *Nat. Rev. Cancer* 5, 251-262 (2005).

11. Nagel,I. *et al.* Biallelic inactivation of TRAF3 in a subset of B-cell lymphomas with interstitial del(14)(q24.1q32.33). *Leukemia* 23, 2153-2155 (2009).

12. Moncunill,V. *et al.* Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* 32, 1106-1112 (2014).

13. Kobayashi,S. *et al.* Identification of IGHCdelta-BACH2 fusion transcripts resulting from cryptic chromosomal rearrangements of 14q32 with 6q15 in aggressive B-cell lymphoma/leukemia. *Genes Chromosomes. Cancer* 50, 207-216 (2011).

14. Ye,X.S. *et al.* The NIMA protein kinase is hyperphosphorylated and activated downstream of p34cdc2/cyclin B: coordination of two mitosis promoting kinases. *EMBO J.* 14, 986-994 (1995).

15. Ruminy,P. *et al.* Two patterns of chromosomal breakpoint locations on the immunoglobulin heavy-chain locus in B-cell lymphomas with t(3;14)(q27;q32): relevance to histology. *Oncogene* 25, 4947-4954 (2006).

16. Lykke-Andersen,J. & Wagner,E. Recruitment and activation of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1. *Genes Dev.* 19, 351-361 (2005).

17. Kasai,M. *et al.* The translin ring specifically recognizes DNA ends at recombination hot spots in the human genome. *J. Biol. Chem.* 272, 11402-11407 (1997).

18. Aoki,K. *et al.* A novel gene, Translin, encodes a recombination hotspot binding protein associated with chromosomal translocations. *Nat. Genet.* 10, 167-174 (1995).

19. Zekavati,A. *et al.* Post-transcriptional regulation of BCL2 mRNA by the RNA-binding protein ZFP36L1 in malignant B cells. *PLoS. One.* 9, e102625 (2014).

20. Nasir,A. *et al.* ZFP36L1 negatively regulates plasmacytoid differentiation of BCL1 cells by targeting BLIMP1 mRNA. *PLoS. One.* 7, e52187 (2012).

21. Baou,M., Jewell,A., & Murphy,J.J. TIS11 family proteins and their roles in posttranscriptional gene regulation. *J. Biomed. Biotechnol.* 2009, 634520 (2009).

22. Klein,I.A. *et al.* Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* 147, 95-106 (2011).

23. Yamane,A. *et al.* Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. *Nat. Immunol.* 12, 62-69 (2011).

24. Khodabakhshi,A.H. *et al.* Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget.* 3, 1308-1319 (2012).

25. Jacob,E., Pucshansky,L., Zeruya,E., Baran,N., & Manor,H. The human protein translin specifically binds single-stranded microsatellite repeats, d(GT)n, and G-strand telomeric repeats, d(TTAGGG)n: a study of the binding parameters. *J. Mol. Biol.* 344, 939-950 (2004).

26. Finkenstadt,P.M., Jeon,M., & Baraban,J.M. Trax is a component of the Translin-containing RNA binding complex. *J. Neurochem.* 83, 202-210 (2002).

27. Eliahoo,E., Marx,A., Manor,H., & Alian,A. A novel open-barrel structure of octameric translin reveals a potential RNA entryway. *J. Mol. Biol.* 427, 756-762 (2015).

28. Meng,G. *et al.* Genomic structure and chromosomal localization of the gene encoding TRAX, a Translin-associated factor X. *J. Hum. Genet.* 45, 305-308 (2000).

29. Bray,J.D., Chennathukuzhi,V.M., & Hecht,N.B. KIF2Abeta: A kinesin family member enriched in mouse male germ cells, interacts with translin associated factor-X (TRAX). *Mol. Reprod. Dev.* 69, 387-396 (2004).

30. Wang,J.Y., Chen,S.Y., Sun,C.N., Chien,T., & Chern,Y. A central role of TRAX in the ATM-mediated DNA repair. *Oncogene*(2015).

31. Hosaka,T. *et al.* A novel type of EWS-CHOP fusion gene in two cases of myxoid liposarcoma. *J. Mol. Diagn.* 4, 164-171 (2002).

32. Chalk,J.G., Barr,F.G., & Mitchell,C.D. Translin recognition site sequences flank chromosome translocation breakpoints in alveolar rhabdomyosarcoma cell lines. *Oncogene* 15, 1199-1205 (1997).

33. Ciais,D. *et al.* Destabilization of vascular endothelial growth factor mRNA by the zinc-finger protein TIS11b. *Oncogene* 23, 8673-8680 (2004).

34. Lee,S.K. *et al.* Butyrate response factor 1 enhances cisplatin sensitivity in human head and neck squamous cell carcinoma cell lines. *Int. J. Cancer* 117, 32-40 (2005).

35. Hodson,D.J. *et al.* Deletion of the RNA-binding proteins ZFP36L1 and ZFP36L2 leads to perturbed thymic development and T lymphoblastic leukemia. *Nat. Immunol.* 11, 717-724 (2010).

36. Ning,Z.Q., Norton,J.D., Li,J., & Murphy,J.J. Distinct mechanisms for rescue from apoptosis in Ramos human B cells by signaling through CD40 and interleukin-4 receptor: role for inhibition of an early response gene, Berg36. *Eur. J. Immunol.* 26, 2356-2363 (1996).

37. Herranz,N. *et al.* mTOR regulates MAPKAPK2 translation to control the senescence-associated secretory phenotype. *Nat. Cell Biol.* 17, 1205-1217 (2015).

38. Klein,U. *et al.* The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell* 17, 28-40 (2010).

39. Jing,Q. *et al.* Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell* 120, 623-634 (2005).

40. Suwaki,N., Klare,K., & Tarsounas,M. RAD51 paralogs: roles in DNA damage signalling, recombinational repair and tumorigenesis. *Semin. Cell Dev. Biol.* 22, 898-905 (2011).

41. Thacker,J. The RAD51 gene family, genetic instability and cancer. *Cancer Lett.* 219, 125-135 (2005).

42. Klein,H.L. The consequences of Rad51 overexpression for normal and tumor cells. *DNA Repair (Amst)* 7, 686-693 (2008).

43. Salaverria,I. *et al.* Detection of chromothripsis-like patterns with a custom array platform for chronic lymphocytic leukemia. *Genes Chromosomes. Cancer* 54, 668-680 (2015).

44. Hudson,T.J. *et al.* International network of cancer genome projects. *Nature* 464, 993-998 (2010).

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

46. Puente,X.S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* 475, 101-105 (2011).

47. Shaffer,L.G., McGowan-Jordan,J., & Schmid,M. ISCN (2013): An International System for Human Cytogenetic Nomenclature(Karger, Basel, 2013).

48. Nagel,I. *et al.* Deregulation of the telomerase reverse transcriptase (TERT) gene by chromosomal translocations in B-cell malignancies. *Blood* 116, 1317-1320 (2010).

49. Villamor,N. *et al.* NOTCH1 mutations identify a genetic subgroup of chronic lymphocytic leukemia patients with high risk of transformation and poor outcome. *Leukemia* 27, 1100-1106 (2013).

50. Delgado,J. *et al.* Genomic complexity and IGHV mutational status are key predictors of outcome of chronic lymphocytic leukemia patients with TP53 disruption. *Haematologica* 99, e231-e234 (2014).

51. Bailey,T.L. & Elkan,C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28-36 (1994).

52. Blanco,E., Guigo,R., & Messeguer,X. Multiple non-collinear TF-map alignments of promoter regions. *BMC. Bioinformatics.* 8, 138 (2007).

53. Ferreira,P.G. *et al.* Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res.* 24, 212-226 (2014).

54. Harrow,J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760-1774 (2012).

55. Dobin,A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 29, 15-21 (2013).

56. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L., & Wold,B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621-628 (2008).

57. Li,B. & Dewey,C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC. Bioinformatics.* 12, 323 (2011).

58. Ritchie,M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47 (2015).

59. Alonso,R. *et al.* Babelomics 5.0: functional interpretation for new generations of genomic data. *Nucleic Acids Res.* 43, W117-W121 (2015).

Results and discussion

This section contains a summary of the results of each of the studies represented by the publications included in this thesis and other unpublished results that are part of the work in progress. For the sake of clarity, the results are not described in chronological order, but instead, following the order of the work flow of the general protocols for genome analysis: (i) development of bioinformatic tools for the identification of somatic variation in tumor genomes; (ii) the application of these protocols to large datasets of cancer genomes; (iii) the development of new strategies for the annotation of gene regulatory regions; and (iv) the characterization of chromosomal rearrangements in cancer genomes, including the study of the underlying mechanisms, as well as their potential functional and clinical impacts.

Development of bioinformatic to identify somatic variation in cancer genomes

This particular study (Moncunill et al. 2014) was carried out in the context of our participation in the International Cancer genome Consortium, in particular, within the Chronic Lymphocytic Leukemia Spanish consortium. While the original participation and tasks of the BSC within this consortium were in relation to the management and primary analysis of all the generated CLL exomes and genomes, our group took this collaboration as an opportunity to develop solutions for the identification of somatic variants in cancer, which was a major bottleneck within this type of studies. At that time, the available analysis tools for identify somatic variants in tumors were restricted to the use of different programs developed by different groups and

focusing in the detection of specific type and range of somatic variation. Some programs were restricted to point mutations of small indels, others to specific size range of SVs, forcing their combination into complex pipeline for the complete analysis of tumor genomes. The overall specificity values for these programs, particularly those centered in SVs were quite low (<60%).

To overcome these limitations and to generate solutions accessible to all types of groups and computing environments, even those with no specific expertise in bioinformatics, we generated SMUFIN. Publication 1 describes the underlying search mechanism of SMUFIN, as well as the results obtained using in-silico and real tumor data, focusing on the reconstruction of complex chromosomal rearrangements. All this work has been done in close collaboration with the groups of Elias Campo (Hospital Clinics, IDIBAPS) and Jan Korbel (EMBL), which have been involved in the experimental validations of SMUFIN's findings.

In summary, compared with previous methods SMUFIN offers several novelties and improvements: (i) because SMUFIN is based on direct tumor and normal genome sequence comparison, the user can analyze his data without previous preparation, either of alignments or filters. Just this simple improvement avoids the use of several programs with different computational requirements, which constitute an important barrier for non-experts in the computational field. (ii) Furthermore, SMUFIN can detect different type of variants (point mutations, insertions, deletions and translocations) without size restriction and at base pair resolution, which allows a more precise interpretation of the results and a better inference of the potential functional impact of the variation. (iii) In terms of reliability,

after experimental validation, our program has demonstrated better specificity results than the other available methods, including those specialized in particular mutation type, particularly on large structural variation.

At the level of immediate use for research, this tool allows to easily perform somatic analysis of any genome sequence, covering all aspects of sequence modification, from point mutations to large reorganizations within the genome. We demonstrate on mantle cell lymphoma and paediatric medulloblastoma samples the potential of this application for the detailed characterization of structural variation often occurring in cancer genomes. In addition, considering all the advantages that SMUFIN provides, it also appears, as a realistic solution for the expected needs when the genomic analysis will become a regular practice within healthcare systems and will be extended to thousands and millions of individuals.

Application of SMUFIN for the analysis of large structural variation in large datasets

From the development of SMUFIN and through a strong collaboration generated with the ICGC-CLL consortium, in particular with the group of Elias Campo (Hospital Clínic, UB), I was directly involved in the aspects concerning the computational analysis of the structural variation associated to whole genome sequences of 148 CLL tumors. This study was part of a larger study with the aim of a wide characterization of the CLL genome that comprises, in addition to these 148 whole genome sequences, other 440 exome sequences,

expression data (RNAseq), genome arrays, and other information regarding the clinics of the pathology and the correlation with the molecular aspects identified (Publication 2).

In collaboration with Marta Munar, a student that I was guiding within the group, and with the group of Elias Campo, we have determined the landscape of chromosomal rearrangements in CLL through the use of SMUFIN and further manual and detailed analysis on the 148 whole CLL genomes. We can observe recurrent structural events (Figure 12).

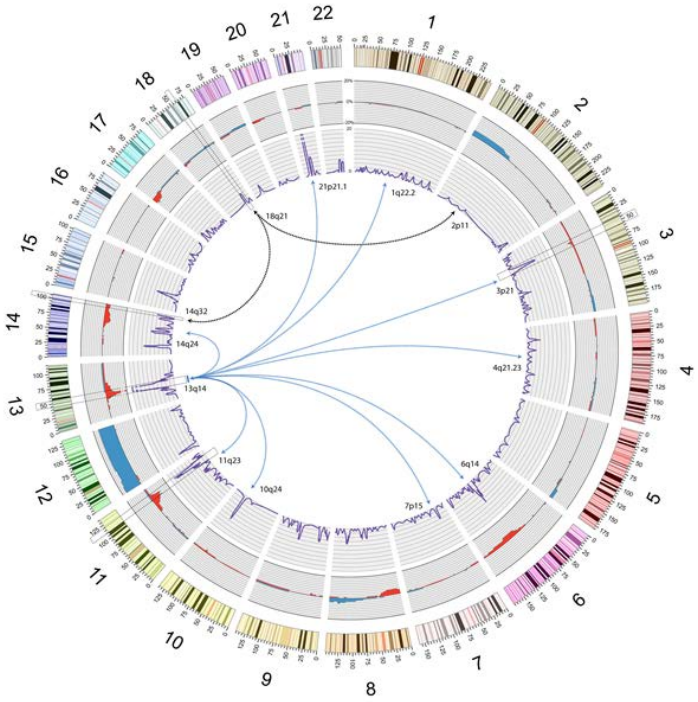


Figure 12. Circular diagram representation of the distribution of structural variants detected in 148 WGS CLL samples. Displayed in the outer layer we show recurrence in Copy number Alterations (CNAs) below of each of the represented chromosomes, followed by all the breakpoints derived from large (> 100 bp) intra- and inter-chromosomal rearrangements (dark blue) in the inner layer. For clarity, the scale of CNAs is set to 20%, as the

maximum, showing sequence gains and losses, as positive (blue) and negative (red) values, respectively. Rearrangements are displayed in absolute counts, indicating that the values in each of the regions do not reflect the recurrence among samples, as some regions with high values derive from one or two cases, normally with complex karyotypes. We highlighted with dashed squares those regions (3p21, 11q23, 13q14, 14q32 and 18q21) with rearrangements observed in more than 5% of cases with WGS. As to rearrangement events, of a total of 358 breakpoints were detected across all 148 samples, 41% of them correspond to interchromosomal translocations, while 59% occurred within chromosomes. Chromosomes 11 and 13 appear as the most rearranged, entailing 25% of all the breaks, followed by chromosomes 3 and 6 (with 8% each). Regarding interchromosomal rearrangements chromosomes 6, 8, 13 and 14 appear as the most translocated, being involved in 32% of all translocations observed. Recurrent breakpoints are indicated by arrows: black arrows for rearrangements affecting 18q21 and BCL2 (four cases with 14q32 and one case with 2p11) and blue arrows for rearrangements affecting 13q14 (nine cases with different chromosomes).

In six of the cases, we also detected recurrent patterns of chromosomal reorganization, similar to those described before and known as Chromoplexy and chromothripsis (Baca et al. 2013; Shen 2013; Moncunill et al. 2014; Rode et al. 2015). In these CLL cases, as shown in figure 13, we observed that some restricted regions in the genome (at least three) and different in each of the patients, translocate with each other in an all-with-all way (see tumors 141 and 853). This is consistent with an scenario where all three regions are close in the space, or physically interacting, further experiments are required to validate this hypothesis.

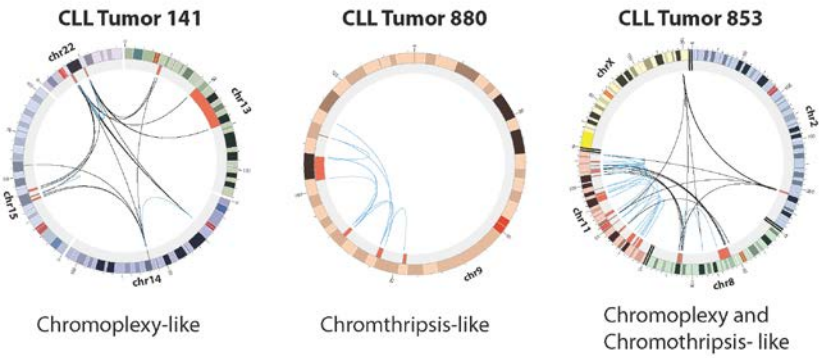


Figure 13. Circular representation of structural variants detected in three CLL tumours with complex rearrangements including chromoplexia (sample 141), chromothripsis (sample 880) and combined (sample 853). Chromosomes are represented in the outer layer, regions lost (red) and gained (blue) detected by SNP arrays are shown in the inner layer. Inter and intrachromosomal rearrangements are represented as black and blue lines, respectively.

Interestingly, tumor 853 clearly presents these two kinds of events simultaneously. By carefully organizing all the different breaks identified in this patient, together with intensive manual inspection of the sequence directly we could reconstruct the complex karyotype resulting from the chromothripsis and chromoplexy events (figure 14 A). Using chromosome painting techniques (figure 14 B), we could verify the existence of four derivative chromosomes, as we have predicted organizing the different breaks identified. As far as we can detect using SMUFIN and confirm by the chromosome painting, the translocations identified in these genomes, both intra and interchromosomal, appear to affect one allele only, leaving the other one intact.

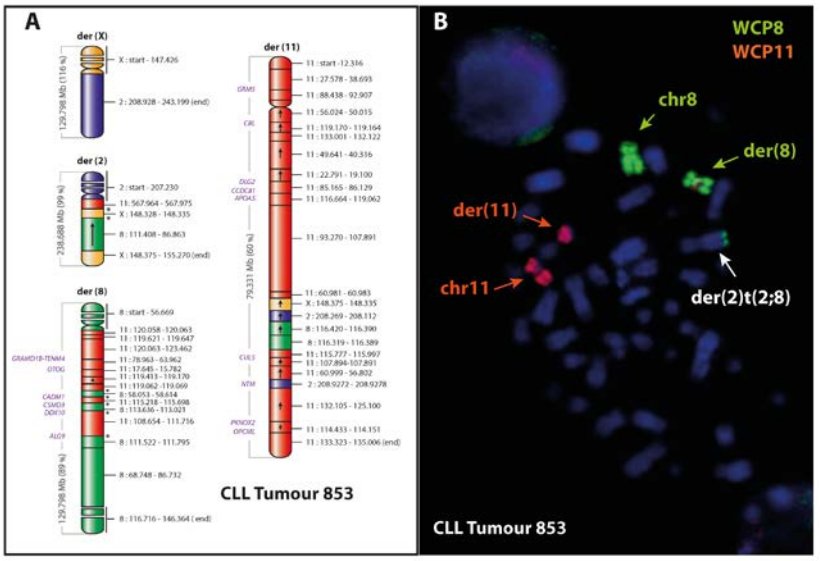


Figure 14. A) Reconstruction at base pair resolution of the resulting reorganized chromosomes in case 853 including der(X) in yellow, der(2) in dark blue, der(8) in green, and der(11) in red. In these reconstructions, only reorganized fragments larger than 100 bp are represented unless they involve interchromosomal translocations. Rearranged regions are not drawn to scale. Arrows denote inverted fragments relative to their normal and original orientation. Flanking portions of the derivative chromosomes without detected rearrangements are collapsed and shown as broken boxes. Estimated sizes (in Mb) for the resulting derivative chromosomes are shown on the left side, including the fraction (percentage) relative to the corresponding normal chromosome size. Asterisks indicate breakpoints that have been experimentally studied and verified. Genes disrupted by breakpoints are displayed on the left side of each of the proposed derivative chromosomes in purple. B) Whole-chromosome painting confirmed the sequencing reconstruction proposed in b. Simultaneous painting of chromosome 8 (green) and 11 (red) shows a normal chromosome 11 and a shorter chromosome der(11) as well as a normal chromosome 8 and der(8) that contains a fragment of chromosome 11 inserted below the centromeric region. In addition, a small fragment of chromosome 8 is detected in the telomeric region of derivative chromosome 2.

Development of new strategies for the annotation of gene regulatory regions

Following the annotation of all variants affecting a given genome, the evaluation of their local impact helps us to determine which functions of the cell are potentially affected. For this, it is necessary a good annotation of the functionality of all the regions in the genomes. The identification of regulatory regions in eukaryotic genomes has been always a challenge, particularly before all the generation of epigenetic data from the ENCODE project, for example. But, still the identification of the regulatory regions and the functional binding sites are not completely solved. I here describe the results we have obtained for the development of ReLA (REgulatory region Local Alignment tool; Publication 3), which also involved other members of the group: Barbara Montserrat and Montserrat Puiggròs. This study, the first in which I was involved in the group, was part of the annotation efforts done in the group and follows its general interest of correlating genome variation with functional impact and, ultimately, with disease.

In summary, the underlying search mechanisms of ReLA is the conservation of transcription factor binding sites (TFBS) among orthologous regions from different genomes, in contrast to previous bioinformatic methods that were based on sequence conservation to infer functionality. ReLA maps known TFBS in different orthologous regions and finds common patterns and sequences of motifs (not nucleotides). Similarly as BLAST does with amino acids, or nucleotides, ReLA uses the smith-waterman algorithm to find the best combination of conserved binding sites among all target regions

(Smith and Waterman 1981). We describe the possibilities of ReLA to identify proximal promoter regions, even improving the annotation of 5' gene regions and the potential transcription start site as well as enhancers. Finally, we also generated a server (http://www.bsc.es/cg/rela/), where the user can execute ReLA remotely and infer the regulatory potential of a set of provided orthologous regions.

Functional and mechanistic inference of somatic structural variation in cancer

Beyond providing general descriptions of tumor related events through the classification of recurrent rearrangements evens (see Publication 2), the identification of the exact position of somatic structural variation in cancer can also provide, in combination with the annotation of the genome, insights into their mechanism of formation, as well as into the potential functional consequences within the cell.

One of the studies that fall into this section aim to understand the potential mechanisms and consequences behind a recurrent and already known 14q deletion observed in CLL patients. In close collaboration with the groups of Silvia Bea (Hospital Clinic, IDIBAPS) and Reiner Siebert (Institute of Human Genetics at Christian-Albrechts-Universität), we have characterized in detail the genomic architecture of this 14q24.1-q32.33 deletion and determined the formation of a gene fusion event between an *IGH* locus and the *ZFP36L1* gene (Manuscript 1). These patients develop a more

aggressive form of the tumor. The detailed evaluation of the break points identified with SMUFIN over the whole genome sequence of 50 CLL samples identified three clear cases that suffered a deletion connecting different points of the 14q24.1 *IGH* region, with the first intron of the *ZFP36L1* gene. The analysis of transcription, done by Bernanrdo Rodriguez in our group, showed actual expression of different forms of chimeric transcripts containing a short 5' segment of the *IGH* region and a large portion of the *ZFP36L1* coding sequence. All these transcripts have been seen to potentially code for a fusion protein with a disrupted TS11B domain within the ZFP36L1 protein that is involved in the interaction with mRNAs and response to growth factors (Bustin et al. 1994) .

In addition, and in order to uncover the potential molecular mechanisms underlying this, and maybe other rearrangements in cancer, we also searched for recurrent sequence patterns around the break points of this 14q24.1-q32.33 deletion. The systematic inspection of 200bp around the breaks has resulted in the detection of a recurrent motive. Although the position of the motive is not fixed relative to the position of the break, the conservation of the sequence and its enrichment within these regions compared to random models is significant (see Manuscript 1). This motive agrees with the sequence recognized by the Translin protein, which is involved in other known *IGH* translocations (Aoki et al. 1997).

This collaborative study is an example of the power of integrating different data and expertise to uncover the biology behind rearrangement events in cancer.

Recurrent mutated regulatory regions in CLL

As a follow up of the general characterization of the structural variation of the CLL genome, we also obtained preliminary results on the potential impact of somatic variants in gene regulation that still require further and more detail study. From the detailed study of all the rearrangements identified in CLL tumors, we have clustered non-coding variants from all the different patients and identified some regions with a clear recurrence among samples that could indicate a functional impact at the level of regulation.

Among all the regions identified, I here highlight one. It corresponds to a recurrent mutated region upstream of the proto-oncogene *BCL6* (figure 15). Mutations in this gene, even in their promoter region, have been demonstrated to be a driver event in CLL (Pasqualucci et al. 2003). That is the reason why the presence of mutations in several CLL patients 150kb upstream of BCL6 rapidly suggests a possible interaction between this region and the oncogene.

Analyzing the region in more detail we can observe a peak on H3K4Me1 histone mark, usually associated with regulatory elements. This specific region has been described as a candidate enhancer after chromatin conformation capture assays demonstrated their interaction with the promoter of BCL6 (Ramachandrareddy et al. 2010).

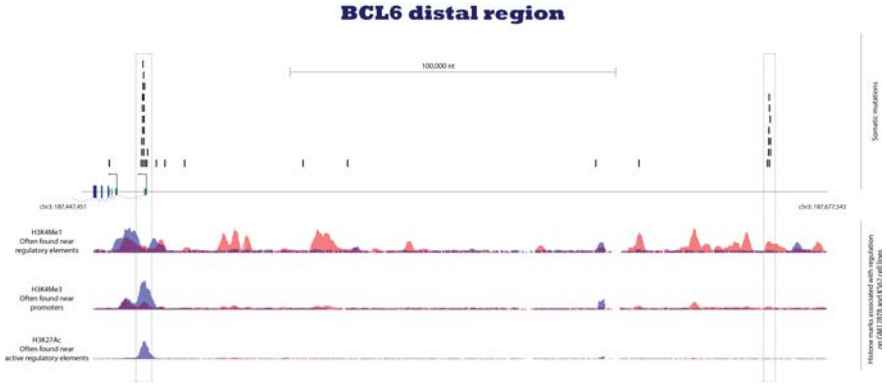


Figure 15. Overall representation of BCL6 upstream area. First row corresponds to the different mutations annotated across 148 CLL patients, both regions that cluster most of the variants are framed. Information of histone marks associated with regulatory potential is also shown for histones H3K4Me1, H3K4Me3 and H3K27Ac.

These results are not published yet and together with other promising candidate regions are currently being studied as new driver regulatory regions involved in CLL.

General Conclusions

This thesis has had the opportunity to experience the transformation from the first large NGS studies to the establishment of these techniques in many of the current genomic studies. Partly, the developed tools and the obtained results reflect that evolution.

Nowadays seems obvious that in a close future the whole genome sequencing will not be limited to research studies and it will take a relevant role in health care systems. First initiatives have been recently launched, such as the 100.000 genomes from Genomics England, as a first step to integrate this analysis in the hospitals for a personalized medicine solution.

To reach this objective a combination between technology accessibility, analysis capabilities and knowledge about the different diseases will be needed. This thesis covers somehow part of this process that begins from the whole genome sequence of an individual and goes throw the identification of their different variants and the potential functional impact in the disease.

Although all the material and methods used to develop this thesis are reflected in the different published papers, I would like to finish my thesis with some considerations about their role and impact in my research.

Often, the design of efficient software can improve the speed up of the analysis, even more than the addition of more computing power. A clear example was ReLA, originally designed using graphs approximations. By using, instead, a Smith&Waterman dynamic

programming algorithm (Smith and Waterman 1981) the efficiency for the detection changed drastically, not only improving the predictions, but also reducing the execution time from few days to some minutes.

In a similar way and discussed in the previous sections, to select the correct DNA alignment algorithm is key in order to obtain successful results. While Smith&Waterman provides you with the most optimal result, BLAST and other BLAST-like algorithms (Altschul et al. 1990; Kent 2002) are suitable for high identity searches on large databases such as through the human genome. An extreme adaptation of the alignment algorithm must be used to perform millions of alignments from sequencing data; these methods were introduced in the "Analysis of NGS data" section.

The different available programming languages can also offer plasticity for adapting to the final goals of the method. For the most text oriented tasks (parsing) I have chosen to use Perl, as language program. It allows the comparison of text files, such as genes or pathways lists, as well as for other text and numerical data in a simple and quick way. However, Perl shows some limitations when dealing with large amount of data and with the management of objects. RELA, for example, is developed in Perl because it has to deal with relatively small pieces of the genome. C++ is the opposite program language in terms of accessibility and optimization. Large datasets and costly computing analysis are more successfully lead with C++ than with Perl. C++ allows a more accurate memory management and a deeper manipulation and understanding of the complete computational process. SMUFIN uses this last programming language to deal with the millions of reads obtained from NGS platforms.

Lastly, despite the access to previous knowledge and studies reinforces the research, the accessible data is large and complex, due to the available formats and the weaknesses associated to each of the datasets, usually generated using highthroughput approaches. The proper use of this data requires the understanding of the underlying strategies and their limitations. All these data can be accessed through genome browser that collect and display data over the genome. These datasets are by far the most used material in all the work developed during this thesis. Public collector databases such as ENSEMBL, UCSC and NCBI (Hubbard et al. 2002; Kent et al. 2002; Cooper et al. 2010) offer an intuitive and value system to organize and filter all this information. Most of the figures presented on this thesis and their companion publications derive from the analysis and interpretation of the data obtained from these public resources.

Conclusions

I. Through the development of SMUFIN, we conclude that the direct comparison of sequence reads from whole genomes allows a more accurate identification of somatic variants in the analysis of cancer genomes.

II. SMUFIN allows the identification and characterisation of somatic chromosomal rearrangements in tumours, including the complete reconstruction of complex karyotypes at base pair resolution level.

III. Complex chromosomal reorganisation events, such as chromothripsis and chromoplexy, are also found in blood tumours, such as Mantle Cell Lymphomas and Chronic Lymphocytic Leukaemia.

IV. Chronic Lymphocytic Leukaemia shows several recurrent structural variation, which, in part, correlate with a more aggressive progression of the tumour.

V. A recurrent deletion identified in chromosome 14 produces a potentially coding chimeric mRNA resulting from the expression of the fusion between IGH parts and the ZFP36L1 gene.

VI. Translin is a candidate effector triggering the recurrent deletion observed in chromosome 14 of CLL patients.

VII. The analysis of conservation of transcription factor binding sites improves the prediction of regulatory regions in eukaryotic genomes compared to classical approaches based on direct sequence conservation.

Bibliography

Abbott A. 2011. Europe to map the human epigenome. *Nature* **477**(7366): 518.

Abeel T, Van de Peer Y, Saeys Y. 2009. Toward a gold standard for promoter prediction evaluation. *Bioinformatics* **25**(12): i313-320.

Affer M, Chesi M, Chen WD, Keats JJ, Demchenko YN, Tamizhmani K, Garbitt VM, Riggs DL, Brents LA, Roschke AV et al. 2014. Promiscuous MYC locus rearrangements hijack enhancers but mostly super-enhancers to dysregulate MYC expression in multiple myeloma. *Leukemia* **28**(8): 1725-1735.

Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, Myeroff L, Lutterbaugh J, Jarrar A, Kalady MF et al. 2012. Epigenomic enhancer profiling defines a signature of colon cancer. *Science* **336**(6082): 736-739.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500**(7463): 415-421.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* **215**(3): 403-410.

Aoki K, Inazawa J, Takahashi T, Nakahara K, Kasai M. 1997. Genomic structure and chromosomal localization of the gene encoding translin, a recombination hotspot binding protein. *Genomics* **43**(2): 237-241.

Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S et al. 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic acids research* **33**(Database issue): D459-465.

Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M et al. 2013. Punctuated evolution of prostate cancer genomes. *Cell* **153**(3): 666-677.

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**(3): 268-276.

Bustin SA, Nie XF, Barnard RC, Kumar V, Pascall JC, Brown KD, Leigh IM, Williams NS, McKay IA. 1994. Cloning and characterization of

ERF-1, a human member of the Tis11 family of early-response genes. *DNA and cell biology* **13**(5): 449-459.

Callinan PA, Batzer MA. 2006. Retrotransposable elements and human disease. *Genome dynamics* **1**: 104-115.

Carninci P Kasukawa T Katayama S Gough J Frith MC Maeda N Oyama R Ravasi T Lenhard B Wells C et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309**(5740): 1559-1563.

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. 2009. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research* **69**(16): 6660-6667.

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31**(3): 213-219.

Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. 2013. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics* **45**(10): 1127-1133.

Consortium EP. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**(5696): 636-640.

-. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Cooper PS, Lipshultz D, Matten WT, McGinnis SD, Pechous S, Romiti ML, Tao T, Valjavec-Gratian M, Sayers EW. 2010. Education resources of the National Center for Biotechnology Information. *Briefings in bioinformatics* **11**(6): 563-569.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* **6**(9): 677-681.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295**(5558): 1306-1311.

Dubchak I, Munoz M, Poliakov A, Salomonis N, Minovitsky S, Bodmer R, Zambon AC. 2013. Whole-Genome rVISTA: a tool to determine enrichment of transcription factor binding sites in gene promoters from transcriptomic data. *Bioinformatics* **29**(16): 2059-2061.

Eisenstein M. 2015. Big data: The power of petabytes. *Nature* **527**(7576): S2-4.

Erez A, DeBerardinis RJ. 2015. Metabolic dysregulation in monogenic disorders and cancer - finding method in madness. *Nature reviews Cancer* **15**(7): 440-448.

Escaramis G, Docampo E, Rabionet R. 2015. A decade of structural variants: description, history and methods to detect structural variation. *Briefings in functional genomics* **14**(5): 305-314.

Friedberg EC. 2003. DNA damage and repair. *Nature* **421**(6921): 436-440.

Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M. 2014. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome biology* **15**(10): 480.

Fullwood MJ, Wei CL, Liu ET, Ruan Y. 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome research* **19**(4): 521-532.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R et al. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**(7414): 91-100.

Gonzalez-Perez A, Lopez-Bigas N. 2012. Functional impact bias reveals cancer drivers. *Nucleic acids research* **40**(21): e169.

Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E et al. 2006. EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome biology* **7 Suppl 1**: S2 1-31.

Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J. 2006. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**(1): 47-59.

Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. *Cell* **100**(1): 57-70.

-. 2011. Hallmarks of cancer: the next generation. *Cell* **144**(5): 646-674.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D et al. 2006. GENCODE: producing a reference annotation for ENCODE. *Genome biology* **7 Suppl 1**: S4 1-9.

Herz HM, Hu D, Shilatifard A. 2014. Enhancer malfunction in cancer. *Molecular cell* **53**(6): 859-866.

Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. 2013. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**(6122): 957-959.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T et al. 2002. The Ensembl genome database project. *Nucleic acids research* **30**(1): 38-41.

Huttenhofer A, Schattner P, Polacek N. 2005. Non-coding RNAs: hope or hype? *Trends in genetics : TIG* **21**(5): 289-297.

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

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830): 1497-1502.

Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon R, Zheng J et al. 2010. Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**(7308): 869-873.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA et al. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* **502**(7471): 333-339.

Kazazian HH, Jr., Moran JV. 1998. The impact of L1 retrotransposons on the human genome. *Nature genetics* **19**(1): 19-24.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome research* **12**(4): 656-664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome research* **12**(6): 996-1006.

Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL et al. 2012. Exome sequencing and the genetic basis of complex traits. *Nature genetics* **44**(6): 623-630.

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome research* **8**(5): 464-478.

Klein U, Lia M, Crespo M, Siegel R, Shen Q, Mo T, Ambesi-Impiombato A, Califano A, Migliazza A, Bhagat G et al. 2010. The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer cell* **17**(1): 28-40.

Korbel JO, Campbell PJ. 2013. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**(6): 1226-1236.

Kulis M, Heath S, Bibikova M, Queiros AC, Navarro A, Clot G, Martinez-Trillos A, Castellano G, Brun-Heath I, Pinyol M et al. 2012. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nature genetics* **44**(11): 1236-1242.

Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R et al. 2015. The European Genome-phenome Archive of human data consented for biomedical research. *Nature genetics* **47**(7): 692-695.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Liu Y, Hermanson M, Grander D, Merup M, Wu X, Heyman M, Rasool O, Juliusson G, Gahrton G, Detlofsson R et al. 1995. 13q deletions in lymphoid malignancies. *Blood* **86**(5): 1911-1915.

Marco-Sola S, Sammeth M, Guigo R, Ribeca P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature methods* **9**(12): 1185-1188.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG* **24**(3): 133-141.

Marx V. 2015. The DNA of a nation. *Nature* **524**(7566): 503-505.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**(9): 1297-1303.

Mitelman F, Johansson B, Mertens F. 2007. The impact of translocations and gene fusions on cancer causation. *Nature reviews Cancer* **7**(4): 233-245.

Moncunill V, Gonzalez S, Bea S, Andrieux LO, Salaverria I, Royo C, Martinez L, Puiggros M, Segura-Wang M, Stutz AM et al. 2014. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nature biotechnology* **32**(11): 1106-1112.

Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* **45**(1): 81-94.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* **31**(13): 3812-3814.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA et al. 2010. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* **42**(1): 30-35.

Palazzo AF, Lee ES. 2015. Non-coding RNA: what is functional and what is junk? *Frontiers in genetics* **6**: 2.

Palin K, Taipale J, Ukkonen E. 2006. Locating potential enhancer elements by comparative genomics using the EEL software. *Nature protocols* **1**(1): 368-374.

Pasqualucci L, Migliazza A, Basso K, Houldsworth J, Chaganti RS, Dalla-Favera R. 2003. Mutations of the BCL6 proto-oncogene disrupt its negative autoregulation in diffuse large B-cell lymphoma. *Blood* **101**(8): 2914-2923.

Puckelwartz MJ, Pesce LL, Nelakuditi V, Dellefave-Castillo L, Golbus JR, Day SM, Cappola TP, Dorn GW, 2nd, Foster IT, McNally EM. 2014. Supercomputing for the parallelization of whole genome analysis. *Bioinformatics* **30**(11): 1508-1513.

Puente XS, Bea S, Valdes-Mas R, Villamor N, Gutierrez-Abril J, Martin-Subero JI, Munar M, Rubio-Perez C, Jares P, Aymerich M et al. 2015. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**(7574): 519-524.

Ramachandrareddy H, Bouska A, Shen Y, Ji M, Rizzino A, Chan WC, McKeithan TW. 2010. BCL6 promoter interacts with far upstream sequences with greatly enhanced activating histone modifications in germinal center B cells. *Proceedings of the National Academy of Sciences of the United States of America* **107**(26): 11930-11935.

Rausch T, Jones DT, Zapatka M, Stutz AM, Zichner T, Weischenfeldt J, Jager N, Remke M, Shih D, Northcott PA et al. 2012a. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**(1-2): 59-71.

Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. 2012b. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**(18): i333-i339.

Ren S, Peng Z, Mao JH, Yu Y, Yin C, Gao X, Cui Z, Zhang J, Yi K, Xu W et al. 2012. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell research* **22**(5): 806-821.

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**(17): e118.

Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, Consortium WGS, Wilkie AO, McVean G, Lunter G. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics* **46**(8): 912-918.

Rode A, Maass KK, Willmund KV, Lichter P, Ernst A. 2015. Chromothripsis in cancer cells: An update. *International journal of cancer Journal international du cancer*.

Ronaghi M, Uhlen M, Nyren P. 1998. A sequencing method based on real-time pyrophosphate. *Science* **281**(5375): 363, 365.

Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG et al. 2013. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic acids research* **41**(Database issue): D56-63.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**(12): 5463-5467.

Seumois G, Chavez L, Gerasimova A, Lienhard M, Omran N, Kalinke L, Vedanayagam M, Ganesan AP, Chawla A, Djukanovic R et al. 2014. Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. *Nature immunology* **15**(8): 777-788.

Shen MM. 2013. Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer cell* **23**(5): 567-569.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *Journal of molecular biology* **147**(1): 195-197.

Smonskey MT, Block AW, Deeb G, Chanan-Khan AA, Bernstein ZP, Miller KC, Wallace PK, Starostik P. 2012. Monoallelic and biallelic deletions of 13q14.3 in chronic lymphocytic leukemia: FISH vs miRNA RT-qPCR detection. *American journal of clinical pathology* **137**(4): 641-646.

Stintzing S, Stremitzer S, Sebio A, Lenz HJ. 2015. Predictive and prognostic markers in the treatment of metastatic colorectal cancer (mCRC): personalized medicine at work. *Hematology/oncology clinics of North America* **29**(1): 43-60.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. *Nature* **458**(7239): 719-724.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**(7571): 75-81.

Sun H, De Bie T, Storms V, Fu Q, Dhollander T, Lemmens K, Verstuyf A, De Moor B, Marchal K. 2009. ModuleDigger: an itemset mining framework for the detection of cis-regulatory modules. *BMC bioinformatics* **10 Suppl 1**: S30.

Teles Alves I, Hartjes T, McClellan E, Hiltemann S, Bottcher R, Dits N, Temanni MR, Janssen B, van Workum W, van der Spek P et al. 2015. Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer. *Oncogene* **34**(5): 568-577.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K et al. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome research* **18**(7): 1051-1063.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends in genetics : TIG* **30**(9): 418-426.

Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L et al. 2011. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature methods* **8**(8): 652-654.

Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research* **22**(9): 1798-1812.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics* **10**(1): 57-63.

Weatherall DJ. 2001. Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. *Nature reviews Genetics* **2**(4): 245-255.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O et al. 2007. A unified classification system for eukaryotic transposable elements. *Nature reviews Genetics* **8**(12): 973-982.

Xiao-Jie L, Hui-Ying X, Qi X, Jiang X, Shi-Jie M. 2015. LINE-1 in cancer: multifaceted functions and potential clinical implications. *Genetics in medicine : official journal of the American College of Medical Genetics*.

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

Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B et al. 2011. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics

data. *Database : the journal of biological databases and curation* **2011**: bar026.

Ziats MN, Rennert OM. 2013. Aberrant expression of long noncoding RNAs in autistic brain. *Journal of molecular neuroscience : MN* **49**(3): 589-593.