UNIVERSITÀ DEGLI STUDI DI TORINO

Department of Molecular Biotechnology and Health Sciences
PhD Program in Biomedical Sciences and Oncology (XXXI Cycle)

PhD Thesis

Computational methods to investigate the genetic determinants of intermediate molecular phenotypes

Supervisor:
Prof. Paolo Provero

Coordinator:
Prof. Emilio Hirsch

Candidate:
Elisa Mariella

Academic Years 2015-2019

Abstract

Elisa Mariella

Computational methods to investigate the genetic determinants of intermediate molecular phenotypes

In this thesis, after a broad introduction to the scientific background of my research, I will present two computational methods for the detection of molecular quantitative trait loci (QTLs). In both projects we exploited the GEUVADIS dataset that includes whole-genome sequencing (WGS) and RNA sequencing (RNA-seq) data from 373 European (EUR) individuals. Notably, these studies have been motivated by the idea that the investigation of the genetic determinants of intermediate molecular phenotypes, such as gene expression and transcript structure, is a possible way to uncover the mechanisms that underlie genotype-phenotype associations that are identified by genome wide association studies (GWAS).

In the first study, we propose a new strategy to analyse the effects of regulatory variants on gene expression, leveraging on the idea that all genetic variants within a regulatory region can contribute to a global perturbation of transcription factor (TF) binding and thus determine an alteration to the expression of target genes. Unlike the standard expression QTL (eQTL) mapping analysis, our approach takes advantage of the current knowledge of the regulatory code and it naturally incorporates the effect of multiple variants within regulatory regions. In particular, we show that it is able to reveal eQTLs that are not identified when studying the correlation between gene expression and individual variants. In addition, it can help formulating hypotheses on the mechanism behind eQTLs by indicating the TFs whose binding perturbation mostly contribute to the gene expression variation.

In the second study, we move from gene expression to transcript structure, focusing on alternative polyadenylation. We indeed present a new computational strategy to discover genetic variants that specifically affect the relative expression of alternative 3′ untranslated region (UTR) isoforms and we analyse possible mechanisms of action of these variants. Notably, our results point to an important role for genetically determined alternative polyadenylation in affecting predisposition to complex diseases, thus suggesting new ways to extract functional information from GWAS data.

Contents

Chapter 1

Cracking our genetic code

> Nearly two centuries ago, in this room, on this floor, Thomas Jefferson and a trusted aide spread out a magnificent map, a map Jefferson had long prayed he would get to see in his lifetime. The aide was Meriwether Lewis and the map was the product of his courageous expedition across the American frontier all the way to the Pacific. It was a map that defined the contours and forever expanded the frontiers of our continent and our imagination. Today the world is joining us here in the East Room to behold the map of even greater significance. We are here to celebrate the completion of the first survey of the entire human genome. Without a doubt, this is the most important, most wondrous map ever produced by human kind.
>
> Former USA President William F. "Bill" Clinton, 26 June 2000

In the first chapter of this thesis, I would like to provide the reader with a broad overview of the scientific background of my research. During the last years, I have mainly dealt with the development of computational methods to investigate the genetic determinants of intermediate molecular phenotypes, namely gene expression and alternative polyadenylation, that can provide a mechanistic link between human genetic variants and complex human diseases. Nowadays, thanks to the extraordinary technological advancement that has been fuelled by the Human Genome Project, around $1,000 is enough to sequence the whole genome of an individual. Consequently, we have at our disposal an unheard-of quantity of genetic data that give us an unrivalled opportunity to understand the genetic causes of human diseases. I believe that a deep understanding of the results of the current post-genomic era needs the awareness of the events that have brought us from a gene-centric vision, that has characterized the initial phases of molecular biology, to the disclosure of many secrets of our genetic code in just approximately three decades. In particular, we must remember that systematic analyses of human genetic variants would not be possible without a reference sequence of our genome and a good knowledge of the activity of its building blocks. Therefore, I will start telling the story of the sequencing and exploration of the human genome and then I will gradually move to my research topics, talking about the core mechanisms of gene regulation and then the efforts that have been done to identify and characterize human genetic variants.

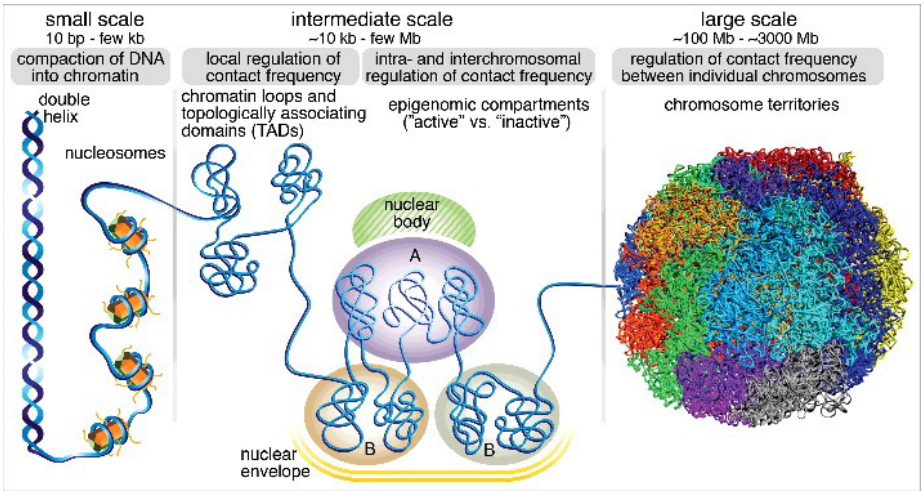


Figure 1.1 – The organization of the human genome can be described at three different scales. At a small scale, the DNA assumes its characteristic double helix structure and wraps around histone proteins generating the nucleosomes. The complex resulting from the interaction of chromosomal proteins, like histones, with the DNA is called chromatin and it can have different levels of compaction. At an intermediate scale, the chromatin is organized into topologically associating domains (TADs) that are characterized by an elevated frequency of local interactions. In addition, other intra- and interchromosomal interactions mediate the formation of epigenomic compartments. The A compartment includes active genomic regions that are positioned near the nuclear body, instead inactive genomic regions preferentially localize in the B compartment near the nuclear envelope. Finally, the existence of chromosome territories can be appreciated at a large scale: although chromosomes in the nucleus are frequently depicted like spaghetti in a bowl, in reality each chromosome always occupies its own discrete region. Figure adapted from [1].

1.1 The Human Genome

The information required to build any living being are stored in the DNA, or deoxyribonucleic acid, that is formed by two complementary and antiparallel strands that wrap around to form its iconic double helix structure (Figure 1.1). Each DNA strand can be described as a long chain of nucleotides held together by the formation of covalent bonds and every nucleotide is composed by a characteristic nitrogenous base (adenine [A], thymine [T], guanine [G] and cytosine [C]), a 5-carbon sugar (deoxyribose) and a phosphate group. On the contrary, the two strands are connected by the formation of hydrogen bonds between complementary base pairs (i.e. A pairs with T and C with G). The set of all the genetic material present inside the nucleus of an eukaryotic cell is called genome. In the case of our species, the genome is packed into twenty-three pairs of homologous chromosomes and its length exceeds 3 billion nucleotides.

The discovery of nucleic acids dates back to 1869, when the chemist Friedrich Miescher extracted a previously uncharacterized substance, that he originally called "nuclein", from the nucleus of leucocytes [2]. Some years later, its chemical composition was determined by Albrecht Kossel who found that it is a relatively simple molecule containing only four different types of bases. This result had

long fuelled the idea that it could not be the carrier of genetic information, function that was usually attributed to proteins due to their more complex structure. The definitive proof that DNA is the hereditary molecule eventually came in 1944, thanks to the brilliant experiments conducted by Oswald Avery, Colin MacLeod and Maclyn McCarty. Another milestone in the history of DNA was laid in 1953 when James Watson and Francis Crick, working with crystallographic data produced by Rosalind Franklin and Maurice Wilkins, uncovered its characteristic double helix structure. At that time, probably no one imagined that after only fifty years we would know the complete nucleotide sequence of the human genome.

Sequencing the human genome has surely been one the most ambitious and exciting scientific projects of the twentieth century. The idea was born indicatively in May 1985 when Robert Sinsheimer hosted a meeting at the University of California, Santa Cruz, to discuss the feasibility of sequencing the human genome. However, the project started only a few years later, after an exciting period of discussion and technological advancement [3]. The essay that the Italian scientist Renato Dulbecco published on Science in 1986 is particularly effective in giving us back the atmosphere of those years [4]. At that time, significant results were obtained in cancer research using model systems of limited complexity. In particular, experiments with oncogenic viruses had allowed the identification of the first oncogenes providing an overview of the initial cancer events. On the other hand, the process of cancer progression, that is intimately linked with the onset of heterogeneity and the accumulation of genetic abnormalities, was much less understood. Dulbecco decisively promoted the sequencing of the human genome, describing it as an essential step to push forward our comprehension of cancer mechanisms. Nevertheless, he anticipated that the knowledge of our genome would probably have a much larger impact on our comprehension of human physiology and pathology, well beyond cancer research. Furthermore, he pointed out that it would have to be a national effort because "its significance would be comparable to that of the effort that led to the conquest of space, and it should be carried out with the same spirit". He went even further writing that "even more appealing would be make it an international undertaking, because the sequence of the human DNA is the reality of our species, and everything that happens in the world depends on those sequences".

The publicly funded Human Genome Project (HGP) officially started on October 1990 with the aim of completing the sequencing of the human genome within 15 years and it was initially directed by James Watson. Although the U.S. Department of Energy (DOE) and National Institutes of Health (NIH) have been the two major founding agencies, numerous groups worldwide were quickly involved leading to the birth of the International Human Genome Sequencing Consortium that included researchers from the United States, the United Kingdom, France, Germany, Japan and China (for further information, see `https://www.genome.gov/human-genome-project`). Therefore, in addition to the undeniable contribution to scientific progress, the HGP has also had the merit of forever changing the way of doing research in biomedicine, inaugurating a long season of large international projects. Furthermore, it has laid a solid foundation for genomic data sharing. In 1996, during a meeting held in Bermuda, representatives of the major sequencing centres established the rules by which to ensure the rapid and public diffusion of the DNA sequence

data [5]. The most important point was the commitment to release any data within twenty-four hours, providing for their uploading to a public database. Getting to this decision was not easy because it was in contrast with both existing laws in some of the involved countries and the common practice in the scientific community to make experimental data publicly available only after publication. Although the primary intent was to allow an efficient collaboration between the different groups involved in the project, the "Bermuda Principles" have been considered as a significant precedent in favour of open science and have become a reference point for the subsequent regulations regarding the sharing of biological data.

Starting from 1998, a parallel project was conducted by a private company, Celera Genomics, headed by the visionary Craig Venter. In disagreement with some choices made by the public consortium, Craig Venter wanted to make the sequencing of the human genome faster and less expensive, intending to complete its whole sequence within just three years. The entry into the scene of a private competitor undoubtedly gave impetus to the HGP which eventually succeeded in completing a first draft of the human genome in advance of the initial plans. The Britain's Wellcome Trust played a crucial role in this phase, significantly increasing the funding of the project and assuming the task of completing a third of the sequencing.

On June 2000 the production of rough drafts of the human genome was jointly announced by the USA President Bill Clinton and the Britain's Prime Minister Tony Blair, with the participation of both Francis Collins, the former HIH director who succeeded James Watson as head of the Human Genome Project, and Craig Venter. This announcement took place several months earlier than the release of peer-reviewed papers describing the results obtained by the two projects. Indeed, an extensive analysis of the first HGP results appeared on Nature on 15 February 2001 [6] and during the same week Science published an article reporting the results obtained by Celera Genomics [7]. The second paper was accompanied by a fiery debate as a result of the refusal by the private company to place its data in a publicly accessible dataset, in clear contrast with both the ideal of sharing that animated the public effort and the historical Science policy [8]. After months of negotiations, Celera Genomics and Science signed an agreement according to which Celera could maintain the data on its own website but undertook to guarantee free access to academic and non-profit researchers, albeit with some limitations. The contrast ended definitively in 2005, when Celera Genomics agreed to upload its data to a public database [9]. This decision stems from the realization that profiting from the commercialization of genomic data was made very difficult by the large amount of publicly available data. From that moment, Celera Genomics officially abandoned genomics to devote itself to the development of new drugs, a transition that had already begun a few years earlier and that coincided with the abandonment of Craig Venter in 2002 [10].

From a technical point of view, the sequencing of the human genome has been enabled by the DNA sequencing method that was developed by Frederik Sanger in 1977 [11]. This technology was based on the incorporation of radiolabelled chain-terminating dideoxynucleotides (ddNTPs) during in vitro DNA replication; the ddNTPs are chemical analogues of the deoxynucleotides (dNTPs) that normally constitute DNA molecules, but they lack the $3'$-OH group and therefore cannot bind the phosphate group of another nucleotide. It follows that the

addition of ddNTPs at low concentration during a DNA chain extension reaction determines the generation of DNA fragments of any possible length, corresponding to the incorporation of ddNTPs in different position of the sequence. Four parallel reactions were required, one for each of the possible bases; then, the resulting fragments were separated by electrophoresis and finally the sequences were read through the analysis of autoradiography images. Although in the original implementation the Sanger sequencing was essentially a manual procedure, in the following years a series of improvements allowed the construction of automated DNA sequencing machines. The most important innovations were the replacement of radiolabelled ddNTPs with fluorescently-labeled ddNTPs, that allow all reactions to take place in the same vessel, and the introduction of capillary based electrophoresis. Furthermore, several other technical advancements have been critical for the genomics revolution [11]. For example, the polymerase chain reaction (PCR), that can be described as "molecular photocopying system" and was invented by Kary Mulles in 1985, and recombinant DNA technologies resulted to be essential, because they allowed to reach the high DNA concentration that was required for sequencing.

An important limitation of DNA sequencing machines has long been the possibility to read only relatively short sequences. For that reason, the sequencing of entire genomes required the adoption of shotgun strategies [12]. Basically, the extracted DNA molecules are randomly fragmented in order to obtain pieces of enough small size, then each fragment is sequenced and finally overlapping sequences are assembled by computer programs. In the case of the hierarchical shotgun strategy, the genome is initially decomposed into a tiling path of overlapping bacterial artificial chromosomes (BACs); then, the sequence of each BAC is determined by shotgun sequencing and finally the sequences of adjacent and overlapping clones are merged. Despite the not negligible amount of preliminary work that it requires, the hierarchical shotgun strategy was chosen by the HGP because it ensures a quite secure path to obtain an accurate genome sequence, since the assembly, that is the most critical phase, is done locally and it is anchored to the genome. On the contrary, Celera Genomics decided to apply to the human genome the alternative whole-genome shotgun strategy that they previously used to produce a draft sequence for the *Drosophila* genome [13]. The whole-genome shotgun strategy entails performing shotgun sequencing directly on the entire genome and then relying on very sophisticated computer programs to assemble the obtained sequences without leveraging any physical map. Although this procedure is intrinsically faster than hierarchical shotgun, the assembly process is greatly complicated and the risk of long-range miassembly increases. It is noteworthy that for the human genome Celera Genomics did not publish any assembly obtained using only its own data, but performed only two joint assemblies that incorporated the data that were generated by the public consortium. For this reason, three HGP leaders questioned whether the private group led by Craig Venter really demonstrated the possibility of applying the whole-genome shotgun method to the complex genome of mammals [12].

At the beginning of the new millennium, the genome sequence was known only for a handful of other species [14]. *Haemophilus influenzae* was the first free-living organism to have its genome completely sequenced in 1995 and the genome sequence was subsequently obtained for about thirty other species of bacteria. The genome of *Saccharomyces cerevisiae* was published in 1996 and it was the first eukaryotic genome sequence to be released. Instead, the first

genome of a multicellular organism to be sequenced was that of the worm *Caenorhabditis elegans* in 1998, followed by the publication of the *Drosophila melanogaster* genome in 2000. Finally, the genome of *Arabidopsis thaliana*, that is commonly used as a model in plant biology, was completed in 2000. Sequencing the human genome was much more challenging for two main reasons: it is much larger than all the previously sequenced genomes and it contains a high number of repetitive sequences. It follows that the draft sequence of the human genome that was published in 2001 was highly imperfect: it covered only about 90% of the euchromatic genome, was interrupted by about 25000 gaps and contained many errors in the nucleotide sequence [14]. Therefore, there was an urgent need to provide a high-quality reference sequence that could really become a solid foundation for biomedicine and in 2004 the HGP consortium published an improved sequence that covered >99% of euchromatic human genome with a very high accuracy [15].

The Human Genome Project, that was declared concluded in 2003, provided the blueprint of our species and for this reason it has often been celebrated as a point of arrival. However, today we know that it was rather a starting point. Indeed, it soon became clear the mere availability of the sequence of the human genome was not enough: to make a great leap forward in our understanding of human physiology and pathology we needed to understand how to interpret it. The road ahead has turned out to be no less complex and fascinating. In this regard, it is important to point out that at the beginning of the twenty-first century our knowledge of the content of the human genome was dramatically limited [14]. First, the total number of genes, the majority of which were assumed to be protein-coding, was unknown and usually it was largely overestimated. The HGP paper suggested a total of 30,000-40,000 protein-coding genes, while the current estimates point to a much smaller value (about 21,000 protein-coding genes). Second, it was thought that the regions of the genome that code for proteins were much more abundant than regulatory regions. However, the decades that followed the publication of the first draft of human genome have totally overturned this idea, revealing that most of the active sequences in the human genome do not encode for proteins. The current collection of non-coding elements includes several untranscribed regions involved in the gene expression modulation (i.e. promoters and enhancers) and increasing numbers of non-coding transcripts with regulatory functions. Finally, the idea that a substantial portion of the genome was simply garbage was still widely accepted.

The c-value (i.e. constant or characteristic value) is the amount of DNA that is present within the haploid genome of a species and it is reasonable to expect that the c-value correlates with the degree of complexity of an organism. Although comparing prokaryotes (archaea and bacteria) with eukaryotes this assumption seems to be correct, among eukaryotes the situation immediately appears much more complicated [16]. Indeed, the eukaryotic organisms can vary widely in the quantity of DNA in their genomes and this is true even for closely related species, suggesting that the genome dimension can change much more rapidly than genes and regulatory sequences during evolutionary timescales (Figure 1.2). In 1971 C.A. Thomas Jr referred to this situation as the "c-value paradox" and it led to the idea that genomes can contain a substantial excess of DNA with respect to coding genes and their regulatory regions. In addition, given the high mutation rate that was observed in hu-

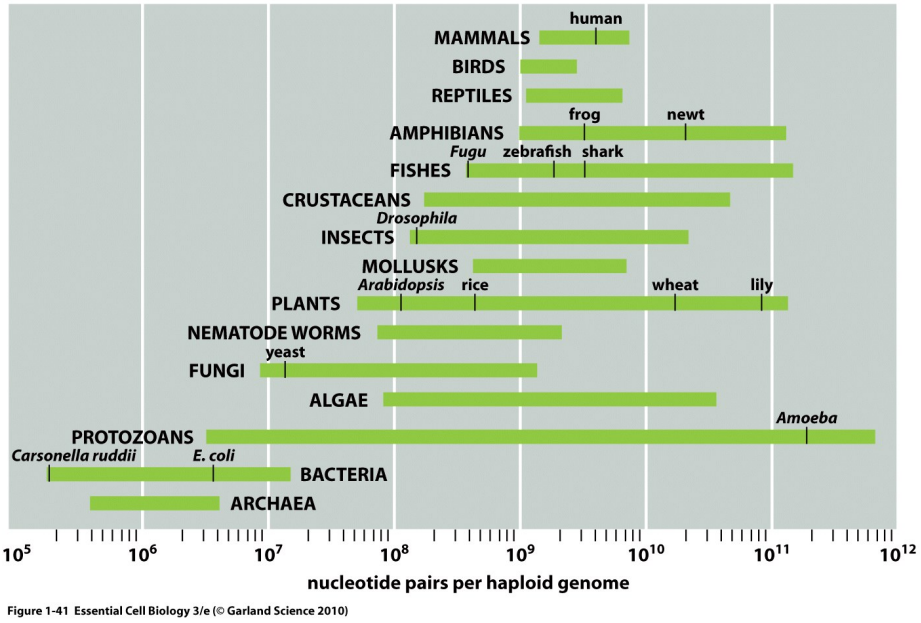


Figure 1.2 – The genome size (i.e. the number of nucleotide pairs per haploid genome) greatly varies across organisms, even between closely related species. Notably, although, as expected, a significant increase is observed moving from prokaryotes (archea and bacteria) to eukaryotes, among eukaryotes the genome size does not correlate with the organism complexity, a situation that is known as "the c-value paradox".

man, it was believed that, if the human genome were entirely functional, the mutational load (i.e. the number of deleterious mutations that are accumulated per generation) would be too much elevated. In 1972 Susumu Ohno formalized a persuasive explanation for both the c-value paradox and the mutational load issue: all genomes contain a certain amount of DNA that does not provide any adaptive advantage to the organism, or whose contribute is very modest, and he attributed to it the provocative name of "junk DNA". According to this hypothesis, only a small fraction of eukaryotic genomes is *functional*, while the large majority is undergoing neutral evolution. At that time, a convincing model was still missing to explain how the amount of junk DNA can change rapidly during evolution and later works identified the activity of transposons, “selfish” DNA elements that are able to replicate and increase their number at the expense of the host organism, as a likely mechanism. Transposon derived elements occupy a substantial part of the genome of many eukaryotes (for example, it has been estimated that around 45% of the human genome derives from transposable elements), however most of them were active in an evolutionary past and subsequently became partial or defective elements because of neutral mutational drift. Evolution has often drawn on this material to create something new, for example the expansion of transposable elements has been related to the appearance of new regulatory elements recognized by DNA-binding proteins [17].

Scientists who have worked towards the characterization of the non-coding genome have mainly relied on two possible approaches [14]. First, comparing

the human genome with that of other species we can identify the sequences that have been conserved during evolution and that therefore are supposed to be functional. Comparative genomics analyses were made possible by the release of several vertebrate genomes, starting from the publication of a high-quality draft sequence for the mouse genome in December 2002 [18]. In addition, important information can be obtained through the investigation of the epigenome. The DNA is not naked in the nucleus, but it interacts with different types of chromosomal proteins. In particular, it wraps around complexes made by eight histone proteins (histone octamer), thus assuming the characteristic "beads on a string" structure whose fundamental units are called nucleosomes (Figure 1.1). More in general, the complex resulting from the union of DNA and chromosomal proteins is called chromatin: it can reach different levels of compactness depending on the cell cycle phase and the regulatory needs of the cell. Basically, it is possible to distinguish two types of chromatin: the euchromatin, that contains active genomics regions, and the more compact heterochromatin, where there are the silenced regions. Both DNA and histones are subjected to several reversible modifications that do not involve changes in the DNA sequence (i.e. DNA methylation and methylation or acetylation of lysine residues within histone proteins) and these “epigenetic” modifications are essential to define the identity and the activity of each domain in the genome.

Few months after the conclusion of the Human Genome Project, the National Human Genome Research Institute (NHGRI) established the international ENCyclopedia of DNA Elements (ENCODE) consortium with the purpose of carrying out a new ambitious project that would aim to identify all the functional elements in the human genome [19]. Since different definitions are possible, it is important to stress out that, according to the ENCODE consortium, operationally a functional element is "a discrete genome segment that encodes a defined product (for example, protein or non-coding RNA) or displays a reproducible biochemical signature (for example, protein binding, or a specific chromatin structure)" [20].

During the pilot phase, that began in September 2003, the existing technologies were rigorously compared to select those that would allow to obtain an economically effective and high-throughput exploration of the whole genome. The attention was focused on a carefully selected group of regions covering a total of approximately 30 Mb, corresponding to around 1% of the genome, and a limited number of cell lines [21]. Each group of the consortium was required to analyse the entire set of ENCODE targets, thus ensuring that the results obtained by different groups or with different approaches were comparable. The concurrent technology development phase was instead devoted to the development of new laboratory and computational methods for the identification of functional elements, addressing the shortcomings that emerged during the pilot phase. Finally, in the production phase all the acquired knowledge was exploited to attack the remaining 99% of the genome.

The whole-genome analysis officially started in 2007 and involved more than 400 researchers coming from 32 laboratories around the world. Scientists of the ENCODE consortium applied 24 different technologies to characterize 147 different cell lines (corresponding to 1648 datasets) and finally, in September 2012, 30 papers simultaneously appeared on Nature, Genome Research and Genome Biology [20]. These initial results had the undeniable value of providing an outstanding first overview of multiple aspects of the genome, including transcribed

regions, open chromatin regions, DNA methylation, histone modifications and transcription factor binding. However, the most important outcome of the ENCODE project was the public sharing of this dizzying amount of biological data in order to feed future discoveries.

Among many important observations, one is particularly remarkable: a reproducible biochemical function was associated to around 80% of the genome, including the broad expanse of non-coding sequences. The publicity of the ENCODE project has often distorted this result, attributing to it the credit for having dealt a mortal blow to the old myth of the "junk DNA". The same ENCODE leaders frequently promoted this interpretation in press releases and interviews, attempting to popularize the result and the whole project in a more seductive way [22]. However, the fact that a region of the genome is active and characterized by a reproducible biochemical function and therefore is supposed to be active is not sufficient to assume that it is *functional* (i.e. being under selective pressure). There are indeed several possible explanations for reproducible biochemical activities that not necessarily experience selective pressure, including pseudogenes, mobile elements and biological noise. Straightforwardly, it is not an issue that the ENCODE experiments have directly addressed [16,22]. This speculation finds a further confirmation in the discrepancy between biochemical and evolutionary estimates of the abundance of functional elements in the human genome. Comparative genomics studies have indeed found that a strong evolutionary constraint applies only to the 5-15% of the mammalian genome and it is reasonable to argue that some of these "biochemically active but selectively neutral" elements are probably non-functional [23]. The question of how much of the human genome is functional still remains open.

On the other hand, the ENCODE results probably underestimate the total amount of biochemically active genome regions, because their collection is intrinsically limited by the biochemical assays that have been performed and the cell types that have been analysed. More in general, this type of analysis only takes cell snapshots, failing to catch the dynamic nature of gene regulation. In the last years, several other efforts have been done to further improve the characterization of the human genome generating an increasing number of large biological datasets. For example, the NIH Roadmap Epigenomics Mapping Consortium, continuing on the road that had been undertaken by ENCODE, in 2015 released 111 human reference epigenomes from primary cell lines and tissues that are representative of all major lineages in the human body [24].

In summary, the decades between the end of the twentieth and the beginning of the twenty-first century have been greatly characterized by huge efforts to obtain a gold-standard sequence of the human genome and the subsequent commitment to uncover its secrets. It is noteworthy that throughout this period the human genome has usually been considered as a one-dimensional entity, whereas numerous recent studies have contributed to reveal that it adopts a complex three-dimensional (3D) conformation within the nucleus (Figure 1.1) [25]. According to the current knowledge, the 3D structure of the human genome can be described in a hierarchical way. Starting from the highest level, during anaphase chromosomes are confined into specific regions of the nucleus, called "chromosome territories". Despite this clear separation, there are two interchromosomal contact hubs: regions of active (euchromatic) chromatin preferentially cluster around nuclear speckles, instead regions of inactive (heterochromatic) chromatin are usually located near the nucleolus. From a technical point of

view, chromosome territories can be visualized thanks to powerful microscopy techniques. More recently, the development of chromosome conformation capture technologies, that systematically quantify the interaction between genomic regions, has allowed to investigate the internal organization of chromosomes. First, within each chromosome active and inactive regions segregate into two epigenomic compartments that differ also in the localization within the nucleus. The A compartment predominately includes euchromatic regions that occupy the most inner part of the nucleus, instead the heterochromatic segments localizes into the B compartment near the nucleolus and the nuclear lamina. At an even lower level, the chromatin seems to be fragmented into topologically associating domains (TADs), self-interacting genomic regions that are separated by low-contact regions called "TAD boundaries". Finally, chromatin loops can be detected within TADs and they correspond to the dynamic interactions occuring between regulatory regions. To date we have only begun to shed light on the molecular mechanisms that mediate the formation of the different compartments or domains and on their functional meaning. Hence, in conclusion, achieving a good comprehension of the principles governing the 3D folding of the human genome is probably one the most important challenge that genomics will have to face in the coming years.

1.2 Overview of gene regulation mechanisms

One of the most fascinating aspect of biology is the generation of multiple cell types starting from the same genome sequence. The "epigenetic landscape", that was proposed by Conrad Waddington in 1957, is a well known metaphor of cell differentiation [26]. Differentiating cells are compared with marbles rolling down an hill: following the landscape shape, each cell pursues a slightly different trajectory and at the end it arrives to a distinct point at the hill base, corresponding to a particular differentiated cell state. Notably, rolling down the hill the cell genome experiences an important epigenetic remodelling that definitely reduces its expression potential towards increasingly specialized functions. This process is allowed by complex and intricate mechanisms that modulate the gene expression, establishing when and where each gene must be expressed. As I discussed in the previous section, in eukaryotes the genome size does not correlate well with the organism complexity and a similar observation applies to coding sequences (number of protein-coding genes), that is essentially unchanged across Metazoa [27]. On the contrary, the proportion of non-coding sequences increases exponentially with the rise of organism complexity, suggesting that living beings have become increasingly more complex through the adoption of progressively more sophisticated strategies to control the gene expression [27]. The mechanisms of gene expression regulation are the topic of this section and, after a general overview, I will focus on two aspects that have been crucial for my doctoral work: the computational investigation of transcription factor binding and alternative polyadenylation.

According to the central dogma of molecular biology, that was proposed by Francis Crick in 1958, the flux of biological information goes from DNA to RNA and finally to proteins [28]. This implicates that the gene expression involves two key stages: during transcription, the DNA sequence of the gene is copied into an RNA molecule (the messenger RNA or mRNA) that is subsequently

Histone modification	Functional meaning
H3K4me1	enrichment at enhancers
H3K4me3	enrichment at promoters
H3K36me3	enrichment within gene bodies that are actively transcribed
H3K27ac	enrichment at active enhancers
H3K9ac	enrichment at active promoters
H3K27me3	associated with temporarily repressed regions
H3K9me3	associated with permanently silenced heterochromatic regions

Table 1.1 – Description of some commonly studied epigenetic marks.

used for the protein synthesis (translation). Although a great number of more recent evidence has induced a profound revision of this simple model, it can still be considered as a reference for the purpose of this introduction.

Despite the existence of different levels of gene expression regulation, the most important part of this control happens at the transcriptional level (Figure 1.3). In the genome each gene is preceded by a regulatory region (promoter) that signals the starting point for transcription and to which the RNA polymerase (i.e. the enzyme responsible for mRNA synthesis) and the general transcription factors bind. In addition, each gene is associated with several proximal and distal regulatory regions whose combinatorial activity allows a precise expression control. They contain many regulatory motifs that are recognized by transcription factors (TFs), that is DNA binding proteins that can increase (activators) or decrease (repressors) the likelihood that the transcription of a particular gene occurs. Distal regulatory regions that are recognized by activator proteins are also called enhancers and they can contact the promoter region thanks to chromatin looping. Furthermore, each genomic region that is involved in transcription is characterized by a distinct pattern of epigenetic modifications that also reflect its active or inactive state; notably, histone acetylation is associated with an open chromatin structure that allows TF binding, thus promoting the transcription activation (Table 1.1).

In eukaryotes, transcription and translation takes place in two different cellular compartments, respectively the nucleus and the cytoplasm. Therefore, each mRNA molecule must be transported from the nucleus to the cytoplasm in order that the synthesis of the encoded protein may occur. This transfer requires that newly synthesized mRNA molecules (primary mRNA, or pre-mRNA) undergoes some modifications that, at least in part, occurs simultaneously with transcription (mRNA maturation). In order of time, the RNA capping is the first one: a 7-methylguanosine cap, whose primary function is protecting the nascent transcript from degradation, is added to the $5'$ end of the pre-mRNA while it is still being synthesized. In addition, the sequence of eukaryotic genes is characterized by the alternation of coding sequences (exons) and non-coding regions (introns). During co-transcriptional RNA splicing the introns are removed and the adjacent exons are joined to generate the mature mRNA. Finally, the last modification is the cleavage and polyadenylation of the transcript: when the RNA polymerase exceeds the transcription end point, the nascent molecule is cut and a poly(A) tail is added to the its $3'$ end. In the cytoplasm, both

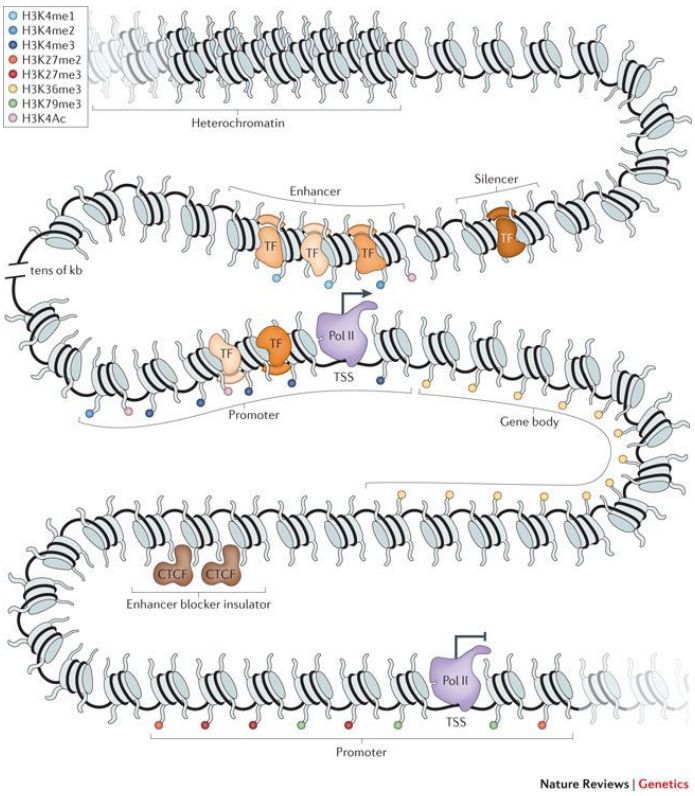


Figure 1.3 – Several genetic and epigenetic mechanisms contribute to gene expression regulation. First, cell type-specific active and inactive genomic regions are characterised by the presence of distinct patterns of histone modifications. In addition, a precise control is achieved thanks to the combinatorial activity of several proximal and distal regulatory regions that contains many sequence motifs that are recognized by transcription factors (TFs). Furthermore, insulator elements, that are usually recognized by CTCF, prevent both spurious interactions between promoters and enhancers (enhancer-blocking activity) and the spread of inactive chromatin (barrier activity). Figure adapted from [29].

the 5′-cap and the 3′-tail are recognized by ribosomes: it is an important step in controlling the mRNA quality before starting its translation. Furthermore, both splicing and polyadenylation process can lead to the generation of multiple alternative isoforms for the same gene in a cell type-specific manner, thus greatly contributing, together with the usage of alternative transcription start sites, to transcriptome diversity.

1.2.1 Computational investigation of TF binding

Given their primary role in the orchestration of cell type-specific gene expression profiles, knowing the position of TF binding sites (TFBS) in the genome is vitally important for a good comprehension of biological systems. From an experimental point of view, this knowledge is commonly achieved through chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments. After the

cross-linking of bound proteins to the DNA and the chromatin shearing, bead-attached antibodies are used to immunoprecipitate the DNA-protein complexes that includes the target protein. The recovered DNA is subsequently sequenced and the derived sequences are aligned to the genome, thus obtaining a map of all the binding sites of the target protein. Notably, the same technology may be exploited to characterize the epigenome, using antibodies that specifically recognize different epigenetic marks. On the other hand, several bioinformatic tools are available to predict TFBS. This task is not straightforward because the sequences that underlie TFBS are usually short and highly degenerate. This results in the common usage of positional weight matrices (PWMs) to provide a quantitative description of the TF preferences in a probabilistic framework [30, 31].

Basically, any DNA binding motif may be represented by a matrix in which each row corresponds to one of the four nucleotides and whose number of columns is equal to the motif length. Starting from a collection of sequences that are recognized by the investigated TF, a simple position frequency matrix (PFM) can be obtained reporting the number of occurrences of each nucleotide in each position. Then, these counts may be divided by the total number of analysed sequences, obtaining a position probability matrix (PPM). Finally, the PWM is generated transforming the PPM elements into log-likelihood values using a background model:

$$w_{r,i} = \log_2 \frac{P(r,i)}{P(b,r)} \tag{1.1}$$

where $w_{r,i}$ is the PWM value for the nucleotide r in the position i, $P(r,i)$ is the probability that the nucleotide r is present in the position i and $P(b,r)$ is the frequency of the nucleotide r in the background. The simplest background model assumes that all nucleotides appear with the same probability in the reference sequences, but more sophisticated models may be obtained depending on the situation; for example, non-coding sequences of the genome may be considered in case of TF binding motifs. PWMs are often graphically represented as sequence logos that consist of a stack of letters for each position, where the size of each letter is proportional to its frequency and the total height at each position indicates its information content, that measures how far it is from an uniform distribution (Figure 1.4).

In addition, PWMs may be easily used to associate a score to any sequence that has the same length of the motif. This score reflects the degree of similarity between the analysed sequence and the PWM, or, from a more physical point of view, the binding energy of the sequence. The highest score among those obtained for the two DNA strands is usually taken into account, consistently with the idea that TFs usually recognize both strands and the absence of evidence supporting the idea that the binding orientation is relevant for their regulatory activity:

$$S(w,r) = \log \max\left(\prod_{j=1}^{l} \frac{P(w_j, r_j)}{P(b, r_j)}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_j)}{P(b, r'_j)}\right) \tag{1.2}$$

where $S(w,r)$ is the score of the sequence r for the PWM w, l is the length of sequence, r_j is the nucleotide at the position j of the sequence r on the

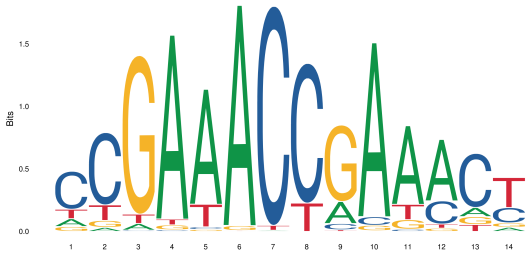


Figure 1.4 – Sequence logo describing the regulatory motif that is recognized by the IRF5 (Interferon Regulatory Factor 5) transcription factor. At each position the size of each letter is proportional to the frequency of the corresponding nucleotide and the total height indicates the information content in bits. Retrieved from JASPAR 2018 [32].

plus strand, r'_j is the nucleotide in the same position but on the other strand, $P(w_j, r_j)$ is the probability to observe the given nucleotide at the position j of the PWM w and $P(b, r_j)$ is the background probability to observe the same nucleotide. Then, in order to establish if the analysed sequence matches the examined PWM, its score is usually compared with an arbitrary threshold: if $S_{max}(w)$ is the maximum possible score for the PWM w, only sequences scoring better than $C \times S_{max}(w)$, with $0 < C < 1$, are considered as putative binding sites. In the case of sequences that are longer than the motif, all the possible sub-sequences may be evaluated sliding the PWM over the sequence with 1-bp increments.

However, this approach is unsatisfactory for at least two reasons. First, it requires the introduction of an arbitrary parameter (i.e. the cutoff that is used to select the binding sites). In addition, the TF-DNA interaction has been described as a thermodynamic process in which also the transient binding to low-affinity sequences plays an important role [33], thus making the distinction between bound and unbound DNA sequences artificial. A better solution is to consider the total binding affinity (TBA) of the whole DNA sequence, catching the contribution of both high- and low-affinity sites [34–36]. The TBA a_{rw} of a PWM w for a sequence r is given by:

$$a_{rw} = \log \sum_{i=1}^{L-l} \max\left(\prod_{j=1}^{l} \frac{P(w_j, r_{i+j-1})}{P(b, r_{i+j-1})}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_{i+j-1})}{P(b, r'_{i+j-1})} \right) \tag{1.3}$$

where L is the sequence length and l is the PWM length.

In order to compare the performance of TBA-based approaches with that of strategies that rely on the identification of discrete binding sites, a cutoff-dependent occupancy can be computed similarly to the TBA, but limiting the sum to sites scoring better than $C \times S_{max}$:

$$t_{rwc} = \log \sum_{i=1}^{L-l} \max\left(\prod_{j=1}^{l} \frac{P(w_j, r_{i+j-1})}{P(b, r_{i+j-1})}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_{i+j-1})}{P(b, r'_{i+j-1})} \right) \times \phi(C, w, r, i) \tag{1.4}$$

where

$$\phi(C, w, r, i) = \begin{cases} 1 & \text{if} \quad S(w, r, i) > C \times S_{max}(w) \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

The TBA concept was originally introduced by works that were focused on transcription regulation in yeast and derives from a statistical mechanical modeling of TF-DNA interactions [34, 35]. Subsequent applications include the study of the evolution of regulatory regions in mammals, where the TBA has the advantage of naturally taking into account the widespread phenomenon of binding site turnover (i.e. during evolution the position of a functional binding site can change without determining a difference in the affinity of the whole sequence for the TF) [36]. In addition, a more recent paper showed that TBA performs better than occupancy in the prediction of both *in vivo* binding of human TFs to regulatory regions and gene expression levels in different human cell types [37]. In the second chapter of this thesis I will show that the TBA can also be successfully exploited to study the genetic-driven variation of gene expression levels across individuals.

Despite their widespread usage, PWMs have some drawbacks [31]. For example, each motif position is considered independent from the other ones, thus ignoring additional constraints that may arise from interactions between nucleotides. Therefore, alternative representations of DNA binding motifs have been proposed, such as hidden Markov models (HMM) in which each node (state) corresponds to a different motif position; notably, HMM can also take into account the presence of a variable space within the motif, thus allowing the representation of more complex binding scenarios like TF dimers. However, all the computational methods that rely on existing models to predict TFBS are currently limited by a serious lack of information: several hundreds of TFs are indeed still uncharacterised. This trouble is overcome by more advanced machine learning methods that do not depend on motif collections. They are commonly trained on datasets including both known regulatory sequences and random sequences and leverage on k-mer vocabularies (i.e. all possible DNA sequences of length k) to infer the sequence features that are relevant for TF-DNA interaction. This makes them more flexible than other models, allowing in particular to catch also distinguishing properties of the sequences that surround the binding sites. In addition, machine learning methods can integrate multilayered information to further enhance the modelling of TF-DNA interaction through a better description of the local chromatin landscape. On the other hand, the resulting models are more specific than others in terms of cell type, differentiation stage, or species, because, unlike TF binding preferences, the regulatory activity is variable and context dependent. Moreover, their interpretation is usually difficult and this inherent "black box" nature may dampen their broad application.

1.2.2 Alternative polyadenylation

As I mentioned above, before being transported from the nucleus to the cytoplasm all the pre-mRNAs must undergo some crucial modifications, including the endonucleolytic cleavage and the addition of a poly(A) tail at the $3'$ end. In addition, it is particularly noteworthy that most human genes have multiple poly(A) sites that may be used in a condition-specific manner, therefore

alternative polyadenylation (APA) extensively contributes to the transcriptome diversification through the generation of alternative mature transcripts. In the first part of this section, I will provide an overview of our current knowledge of this molecular process, mainly referring to a recently published review [38], and then I will move to the computational methods that can be applied to the detection of APA events.

The polyadenylation process is mainly operated by a core protein complex that is composed by four subcomplexes: cleavage and polyadenylation factor (CPSF), cleavage stimulation factor (CSTF), cleavage factor I (CFI) and CFII. In addition, the poly(A) polymerase (PAP) is responsible for the synthesis of the poly(A) tail whose length is controlled by poly(A) binding protein 1 (PABPN1). At the sequence level, the polyadenylation signal (PAS), that is located ∼21 nucleotides upstream with respect to the poly(A) site, is the most important element for the definition of 3′ end processing. The hexamers AAUAAA and AUUAAA are the canonical PAS, but several close variants, that are recognized with a lower efficiency by the polyadenylation machinery, have been identified in mammals. However, also other sequence motifs are bound by the core factors and contribute to the regulation of poly(A) site usage: for example, a G/U-rich region is usually present immediately downstream of the poly(A) site and the UGUA motif is commonly found upstream of the PAS. Notably, all these regulatory motifs are enriched at a specific distance with respect to the cleavage site and this is particularly evident for distal poly(A) sites, suggesting that they usually are stronger than the proximal ones. This speculation is also supported by the observation that canonical PAS motifs are preferentially observed at poly(A) sites that are distal to the coding sequence compared to sites that are near the 3′ UTR beginning. This pattern is probably the result of selection, because, if the proximal sites were optimal, the "first come first served" rule would be slavishly followed and thus the processing would never occur at distal sites.

Different types of 3′ UTR isoforms can be generated by alternative cleavage and polyadenylation (Figure 1.5) [38]. In the simplest and most common situation, tandem poly(A) sites are located within the same terminal exon and their differential usage is associated with the generation of mature transcripts that have the same coding sequence and 3′ UTRs of different length. In addition, other types of APA events are associated with the coding sequence variation. When the processing occurs at intronic poly(A) sites that are located immediately downstream of an annotated exon, it results in the appearance of composite terminal exons. For example, an event of this type allows the switch from membrane-bound to secreted IgM following the B cell activation during immune adaptive response [39]. Moreover, an alternative splicing event that uncovers an alternative poly(A) site within an exon that is normally skipped leads to the generation of a shorter isoform with a cassette terminal exon.

The variable length of 3′ UTRs can have relevant functional consequences because they are major docking platforms for factors involved in all the different layers of post-transcriptional regulation, including mRNA stability, localization and translation rate. The 3′ UTR shortening can lead to the loss of miRNA binding sites, thus potentially making short isoforms more stable than the corresponding long isoforms. For example, this mechanism has been implicated in oncogene activation in cancer cells [40]. However, other results question the idea that long isoforms are generally characterized by a higher decay rate than

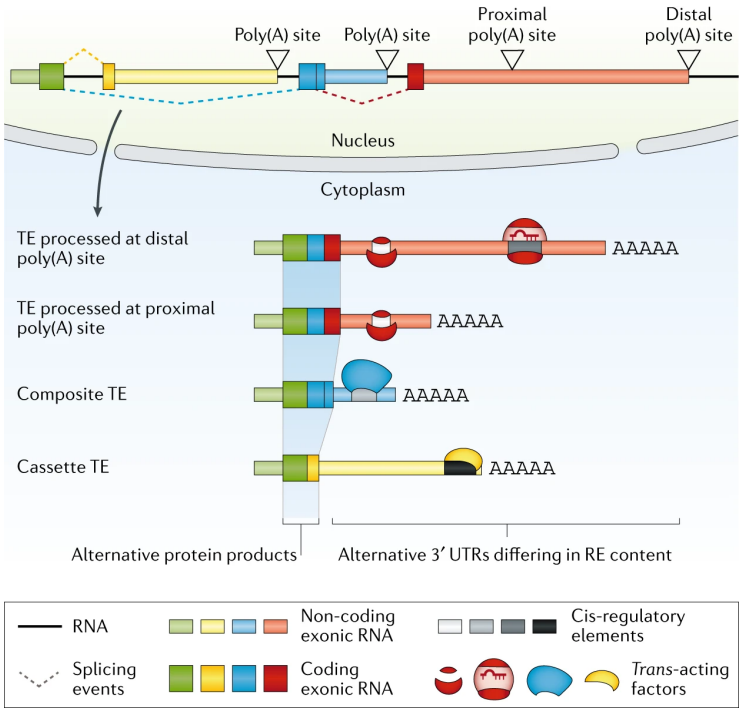


Figure 1.5 – APA leads to the generation of different types of alternative 3′ UTR isoforms. The simplest and most common modality regards tandem poly(A) sites that are located within the same terminal exon (TE), thus determining the production of isoforms that have the same coding sequence and different 3′ UTRs. Notably, short and long 3′ UTRs differ in the content of cis-regulatory elements (RE) that are recognized by trans-acting factors that can have pervasive effects on both the transcript and the encoded protein. Nevertheless, APA events can give rise to alternative protein products, through the alteration of the coding sequence. Figure adapted from [38].

the short ones. In particular, 3′ UTR shortening can actually potentiate repression mediated by miRNA binding sites that are located in the centre of long isoforms, because the miRNA-mediated repression is supposed to be stronger when miRNA binding sites are located near 3′ UTR boundaries [41]. Moreover, despite the known role of miRNAs in translation repression, during T cell activation, a process that is greatly associated with the preferential expression of short isoforms, changes in expression ratio between short and long isoforms are not reflected into consistent variation of protein levels [42]. On other hand, regulatory elements in 3′ UTRs can also influence the localization of the encoded protein (3′ UTR-dependent protein localization) and therefore APA can redirect proteins into different subcellular compartments [43]. For example, within the 3′ UTR of the transmembrane protein CD47 there are uracil-rich elements that are recognized by heterogeneous ribonucleoprotein C (hnRNPC) in the nucleus. The hnRNPC binding prevents the processing at a distal poly(A) site, thus leading to the generation of the short isoform. The long isoform is produced in the absence of hnRNPC and it is characterized by the presence of adenine and uracil-rich elements (AU-rich elements, or AREs) that are recognized by embryonic lethal abnormal vision-like protein 1 (ELAVL1). This interaction is

sufficient to induce the assembly of a protein complex that mediates the CD47 translocation from the endoplasmic reticulum to the plasma membrane.

Many analyses regarding APA have focused on the comparison of the expression of alternative 3′ isoforms in different cell types or conditions. These studies have allowed a significant advancement of our knowledge of this molecular process, highlighting in particular that the APA regulation is strongly tissue-specific and uncovering intriguing context-specific trends towards the preferential expression of short or long isoforms. For example, genes that are expressed in testis, ovary and stem cells preferentially express isoforms with short 3′ UTRs. Neurons seem to have an even more peculiar patter: recent evidence suggest that isoforms with long 3′ UTRs are enriched within the soma and instead isoforms with short 3′ UTRs preferentially localize within neurites [44]. Furthermore, cell proliferation is associated with the preferential expression of short isoforms, whereas the higher expression of long isoforms can be appreciated during cell differentiation [45] and the induction of pluripotency in somatic cells consistently determine a switch towards the expression of transcripts with short 3′ UTRs [46]. In addition, 3′ UTR shortening is observed following both immune cell and neuronal activation [47, 48].

Several mechanisms have been proposed to explain the regulation of poly(A) site choice. The first opportunity for modulation is offered by the expression levels of 3′ end processing factors. For example, their global upregulation, driven by the E2F transcription factor, may be responsible for the generation of short 3′ UTR isoforms in proliferating cells [49] and the increased CSTF2 concentration after B cell activation leads to the preferential usage of an alternative proximal poly(A) site within the IgM transcript, hence allowing the protein secretion [39]. Furthermore, several factors that have been originally implicated into other steps of RNA processing can take part into APA regulation and many of the involved RNA binding proteins (RBPs) have effects that depend on the location of their binding sites with respect to cleavage sites, thus making our comprehension of the molecular details even more tricky. From this perspective, the interplay between splicing and polyadenylation is particularly noteworthy. The U1 snRNP is usually more abundant than other splicing factors and this is probably due to its double life, indeed it also protects long transcripts with long introns for premature cleavage and polyadenylation (telescripting) [50]. This mechanism may contribute to definition of promoter directionality (i.e. in each locus transcripts are predominantly generated from just one of the two strands), because antisense transcripts that lack U1 binding sites are rapidly terminated and degraded. Moreover, U1 levels are supposed to underlie the pattern of 3′ end processing that is observed following neuronal activation [48]. In that cellular context, as a consequence of the considerable increase of transcriptional activity, the U1 concentration is no longer enough to protect nascent transcripts from premature cleavage and polyadenylation, thus determining the emergence of short isoforms.

Given its interaction with virtually all aspects of post-transcriptional regulation, it is not surprising that APA perturbation plays a primary role in human diseases [38]. The dysregulation can derive from both the loss or gain of individual poly(A) sites and global changes in poly(A) site usage. Mutations within PAS motifs that lead to an abnormal 3′ end processing have been identified in several pathological conditions, including neonatal diabetes, α-thalassaemia and β-thalasseamia, systemic lupus erithematosus (SLE) and other hematologi-

cal disorders. In addition, a single-nucleotide polymorphism (SNP) that breaks a canonical PAS within *TP53* has been associated with an higher susceptibility for different cancer types [51]. Another SNP increases the risk of SLE as a consequence of the enhanced expression of *IRF5* that derives from the appearance of a new canonical PAS in a proximal position of the transcript and the subsequent production of a shorter and more stable isoform [52]. Furthermore, global changes in poly(A) sites usage have been reported in almost all cancer types where the cleavage at proximal position is generally preferred, consistently with the known association between proliferation and $3'$ UTR shortening [40]. Finally, intronic polyadenylation is common in cancer and can lead to the generation of truncated proteins that lack tumor suppressor activity or gain oncogenic properties [53].

From a computational point of view, a central task, preliminary to virtually all APA investigations, is the quantification of alternative $3'$ UTR isoforms. Various strategies have been implemented to this end, from custom pipelines for microarray data analysis [54], to the development of next-generation sequencing technologies specifically targeted to the $3'$ end of transcripts, such as the serial analysis of gene expression (SAGE) [45] and sequencing of APA sites (SAPAs) [55], allowing also the identification of previously unannotated APA sites. More recently, tools able to capture APA events from standard RNA sequencing (RNA-seq) data have been developed. In general, these approaches can be divided into two categories: those that exploit previous annotation of poly(A) sites [56,57], such the ones provided by PolyA_DB2 [58] and APASdb [59], and those that instead try to infer their location from the data [60]. Although the latter approach potentially allows analyzing also previously unannotated sites, the former leads to higher sensitivity [56,57]. Undoubtedly, approaches based on standard RNA-seq are not as powerful and accurate as technologies that specifically sequence the $3'$ ends. However, they allow studying this phenomenon in an incomparably larger number of samples and conditions.

Few years ago, I contributed to the development of a bioinformatic software (Roar) that is able to catch APA events from standard RNA-seq data (Figure 1.6) [37]. It requires a collection of alternative poly(A) sites and RNA-seq data for the two conditions that must be compared. In the original work, melted gene structures were obtained collapsing all the transcripts assigned to the same gene and then, using an external annotation of poly(A) sites, for each gene we define two segments of interest: the PRE segment, extending from the beginning of the last exon to the proximal poly(A) site, and the POST segment, from the proximal poly(A) site to the end of the gene. The PRE fragment is assumed to be contained into both the long and the short isoform, while the POST segment should be contained exclusively into the long isoform. The RNA-seq reads that fall respectively over the PRE and the POST segment are then counted, so as to obtain for each gene, in each of the two considered conditions, the expression ratio (m/M) between the short and the long isoform:

$$m/M_{a,i} = \frac{l_{POST_a} \times \#r_{PRE_{a,i}}}{l_{PRE_a} \times \#r_{POST_{a,i}}} - 1 \tag{1.6}$$

where $m/M_{a,i}$ is the m/M of the gene a in the condition i, l_{PRE_a} and l_{POST_a} are respectively the length of the PRE and POST segment of the gene a, $\#r_{PRE_{a,i}}$ and $\#r_{POST_{a,i}}$ are respectively the number of reads mapped on the PRE and

the POST segment of the gene a in the condition i. Therefore, when $m/M > 1$ it suggests that in the examined condition the short isoform is more expressed, instead the production of the long isoform is supposed to be preferred in the opposite case ($m/M < 1$). In addition, the ratio of the two m/M values is obtained to compare the two different conditions: this parameter is called "ratio of a ratio" (roar) and represents the tendency of the first condition to express relatively higher levels of the short isoform (when roar > 1) or higher levels of the long one (when roar < 1). Finally, in order to provide a statistical evaluation of this difference, a Fisher test is performed for each gene, comparing the imbalance between the PRE and POST read counts in the two conditions.

I would like to emphasize that this description of Roar's algorithm refers to the simplest situation in which only one proximal poly(A) site is considered for each gene, comparing its usage with the transcript end. Although several genes have a higher number of annotated poly(A) sites, even genes with many reported polyadenylation sites predominantly use only two of them. Therefore, this strategy is usually enough to get much of the signals and thus uncover if alternative polyadenylation is relevant in the context of interest. Nevertheless, the possibility to take into account all the annotated poly(A) sites of a gene may be useful for more in depth analyses and indeed Roar also implements a computationally efficient strategy to evaluate the expression of all the possible alternative 3'UTR isoforms of a gene (Figure 1.6).

Although it was originally designed to detect APA events comparing two alternative conditions, this approach can be also exploited to study the variation of the relative expression of alternative $3'$ UTR isoforms across individuals, treating the m/M values as a quantitative measure of alternative polyadenylation. In the third chapter of this thesis, I will present an efficient computational strategy to pinpoint genetic variants that specifically affect this quantitative molecular phenotype, focusing on alternative isoforms that are generated through the differential usage of tandem poly(A) sites.

1.3 Characterization of human genetic variation

Virtually all phenotypic differences that we can observe between two individuals derive, at least partially, from more or less extensive germline variations in the sequence of their genome. Genetic variants can impact both coding and non-coding sequences, and they are commonly classified into four main categories [61]. Single nucleotide variants (SNVs) are point mutations that strike single positions of the genome and can consist in any nucleotide substitution. SNVs that are detectable at an appreciable frequency within a population (>1%) are also called single nucleotide polymorphisms (SNPs). Instead, an indel is the insertion or deletion of a short stretch of DNA (<50 bp) in the genome sequence. Globally, SNVs and indels represent the higher proportion of genetic variants that may be detected in any human genome. In addition, structural variants (SVs) involve much wider regions of the genome and include copy number variations (CNVs), chromosomal rearrangements, and mobile element insertions (MEIs). Moreover, tandem repeat variations, such as short tandem repeats (STRs), can be extremely abundant, but so far they have received little attention.

Shortly after the end of the race to obtain the first draft sequence of the

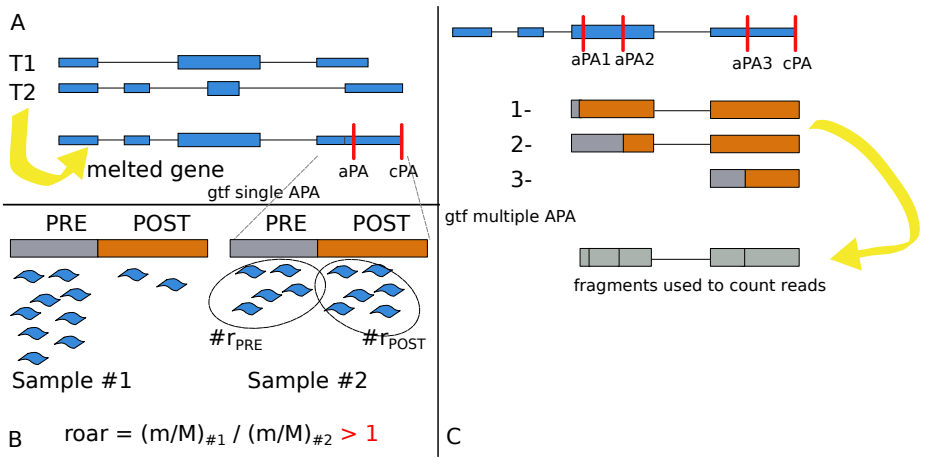


Figure 1.6 – This figure provides an overview of the algorithm that is implemented by Roar to detect APA events through the comparison of two different conditions for which standard RNA-seq data are available. (A) Melted gene structures are obtained collapsing all the transcripts assigned to the same gene. (B) Then, for each gene two segments of interest are defined (the PRE and POST segments), exploiting an external annotation of poly(A) sites. The RNA-seq reads falling over these regions are subsequently counted to compute a m/M value for each gene in each condition. At the end, another parameter (roar) is obtained to catch the alternative $3'$ UTR isoforms whose relative expression is different among the two examined conditions. (C) In the simplified version of the algorithm, only one proximal poly(A) site is taken into account for each gene, comparing its usage with the transcript end. However, Roar also implements a computationally efficient strategy to examine all the annotated poly(A) sites of a gene. Figure adapted from [37].

human genome, serious efforts began to be devoted to the characterization of the genetic diversity both within and between human populations. Although the collection of human genetic variants started much earlier, with the creation of rudimentary genetic maps that in 1980s were used to uncover several genetic variants that cause Mendelian diseases, this process has made substantial progresses in the first decades of the new millennium. Initially, the development of genotyping arrays, that allow to interrogate millions of genomic positions simultaneously and at a relatively low cost, and the discovery of the haplotype structure of the human genome have been crucial [14]. In eukaryotes, during gametogenesis, an exchange of genetic material between homologous chromosomes (crossing-over) combines with the independent assortment of chromosomes to further increase the genetic variability within the population. Notably, the frequency of recombination between two genomic loci is inversely proportional to their distance. Therefore, in the case of nearby genomic loci, their alleles are transmitted together much more frequently than expected if they were independent and the association was random: this concept is known as linkage disequilibrium (LD). For that reason, nearby genetic variants are tightly linked together into DNA sequences that are almost always transmitted identical from one generation to another (haplotypes), suggesting that a relatively low number of genetic variants may be sufficient to catch ∼90% genetic variance that is present within any population. This principle has been the foundation of the

International Haplotype Map (HapMap) Project that was launched on October 2002 with the aim to provide a public resource to speed the discovery of the genetic basis of complex human diseases. In the first phases of the project (Phase I and Phase II) they analysed samples from 269 individuals who came from four geographically diverse population: Yoruba in Ibadan (Nigeria), Japanese in Tokyo (Japan), Han Chinese in Beijing (China) and the CEPH (Utah residents with ancestry from Northern and Western Europe). Globally, they genotyped over 3.1 million SNPs, with a frequency of about one SNP per kilobase [62]; in addition, a subsequent release (HapMap 3) included 1.6 million SNPs that were genotyped for 1184 samples from 11 populations [63].

A further leap forward has been allowed by the advent of high-throughput next-generation sequencing (NGS) technologies that significantly reduced the time required for DNA sequencing and resulted in a progressive and awesome drop in costs (Figure 1.7) [64]. NGS technologies have introduced some key features, including multiplexing (i.e. DNA templates that derive from multiple libraries, that were obtained from different samples, can be subjected to a common sequencing reaction), different strategies for *in vitro* amplification and the adoption of sequence-by-synthesis (SBS) approaches, in which a biochemical reaction is associated with imaging of the surface on which DNA templates are immobilized. The first NGS instrument was marketed in 2005 and the first years were characterized by the competition between multiple companies. Nonetheless, starting from 2012 the Illumina's platforms have progressively became dominant. Consequently, although different strategies have been developed for SBS, the polymerase-mediated incorporation of fluorescently labelled deoxynucleotides (dNTPs) is the only one currently available. It relies on the usage of reversible chain-terminators and a suitable enginereed polymerase that should allow only one dNTP to be incorporated during each cycle. However, the efficiency of this process is far from optimal, thus significantly limiting the length of the sequenced molecules that are much shorter than those obtained by Sanger sequencing.

Thanks to the availability of NGS technologies it has been possible to undertake ambitious re-sequencing projects in human populations, that have given rise to incredible large collection of human genetic variants. The 1000 Genomes Project has been the first initiative to sequence the whole-genome of thousands of individuals [65, 66]. The project, that was active from 2008 to 2015, was launched aiming to identify most genetic variants with a frequency of at least 1%. Whereas the first datasets that have been released were mainly focused on people with an European ancestry, in the conclusive phase they analysed 2,504 genomes from individuals that came from 26 world-wide populations, thus covering five continental regions: East Asia (EAS), South Asia (SAS), Europe (EUR), Africa (AFR) and Americas (AMR). This resource provides an unique opportunity to deepen our knowledge of recent human evolution. Furthermore, the inclusion of data from individuals with different ancestries is really valuable in medicine, even in the big cosmopolitan cities where clinicians frequently come into contact with people from many ethnic backgrounds. Notably, the last dataset includes also recently admixed populations that derive from the union of previously unrelated populations. The African American populations are maybe the most known example: their genetic heritage results from the combination of components that derive from African, European and Native American populations. These data may be also useful to investigate the genetic variation within

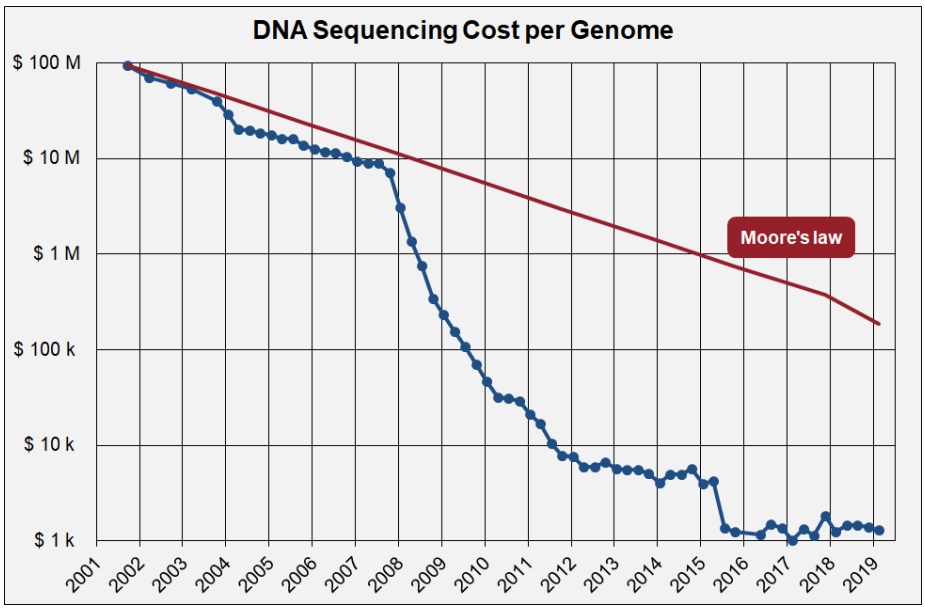


Figure 1.7 – For many years the NHGRI has monitored the costs associated with DNA sequencing at the centers that it supports and this graph shows the cost per human-sized genome from February 2001 to February 2019. In addition, it reports hypothetical data reflecting the famous Moore's law, according to which the "computer power" should double about every two years, while halving the costs. Taking into account that computer hardware industries that keep up with Moore's law are considered to have excellent performance, the gap between real and hypothetical data that starts from 2008 is really impressive: it represents the time at which Sanger sequencing gave way to NGS technologies. Data from "Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)". Available at: `www.genome.gov/sequencingcostsdata`. Accessed 08/08/2019.

populations for whom to date only few genomes are available, like the Native Americans. In addition, while in the earlier phases only bi-allelic events were examined, in the third and final phase the 1000 Genomes Project investigated a broader spectrum of human genetic variants, identifying more than 88 million variants that include multi-allelic SNPs, indels and structural variants. Their results suggest that a typical genome contains 4.1-5 million sites in which it differs from the reference sequence and, although they mainly consist of SNPs and indels, the number of structural variants is not negligible (∼2100-2500 structural variants are present in each genome). Furthermore, they observed that the total number of genetic variants is variable between different populations: the highest genetic variability is observed among African individuals, consistently with the "Out of Africa" theory (i.e. anatomically modern humans originated in Africa and they subsequently migrated towards other world regions about 100,000 years ago). On the other hand, individuals from recently admixed population show a great variability in the number of genetic variants, that, coherently with the previous observation, is roughly proportional to the degree of recent African origin in their genomes. Finally, the analysis of genetic variant sharing among populations can provide useful insights for the reconstruction of their recent

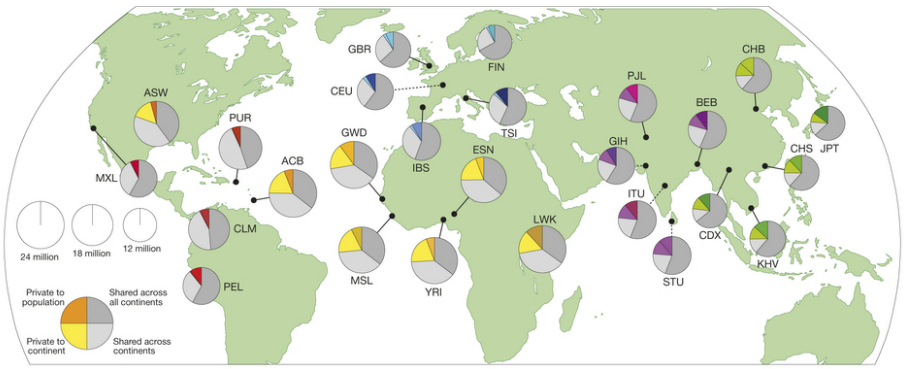


Figure 1.8 – The 1000 Genomes Project has produced the most comprehensive view of global human variation, through the sequencing of whole genomes in 2,504 individuals that came from 26 world-wide populations. In the figure the area of each pie is proportional to the number of genetic variants that were identified in the corresponding population and each of them is subdivided into four slices referring to different categories of genetic variants (private to population, private to continent, shared across continents and shared across all continents). Dashed lines denotes populations that were sampled outside their ancestral continental region (for example, the CEU are Utah residents with Northern and Western European ancestry). Figure adapted from [65].

history: common genetic variants are usually shared across continents, while rare variants are normally detectable only within closely related populations and 86% of the identified variants are present only in a single population.

The 1000 Genomes Project has produced one of the most comprehensive view of global human genetic variation. At the same time, other sequencing efforts have been devoted to the characterization of the genetic structure of specific populations. For example, Chiang *et al.* [67] analyzed whole-genome sequencing (WGS) data from 3,514 modern individuals that came from different regions of Sardinia, a large Italian island in the Mediterranean Sea. Different types of data, including genetic, cultural, linguistic and archaeological findings, indicate that, as a consequence of its prolonged isolation, Sardinia has experienced a major differentiation from the mainland. In particular, ancient DNA studies have suggested that the ancestral composition of its population is much different from that of other European populations[a]. Although Chiang *et al.* confirmed the known prevalence of early Neolithic ancestry across all Sardinia populations, they also revealed an intriguing within-island variation. The major genetic differentiation was observed between individuals that came from the Ogliastra province, that is included within the mountainous Gennargentu region and that experienced the greatest isolation, and those that live in rest of the island. While an enrichment of Neolithic farmer and pre-Neolithic hunter-gatherer

[a]The genetic structure of modern European populations is usually modelled referring to three ancestral populations: the pre-Neolithic hunter-gatherers, the Neolithic farmers that arrived in Europe from Near East and Anatolia, and finally the steppe pastoralists that expanded into Europe during the Bronze Age. Notably, the highest genetic similarity to ancient Neolithic farmers was observed for Sardinia population, suggesting that the island was mainly colonized by early Neolithic farmers, with minor contributions from pre-Neolithic groups, and it subsequently remained largely isolated.

ancestries was revealed in the Gennargentu region, higher levels of Bronze Age steppe pastoralist ancestry were observed elsewhere. In addition, comparing Sardinia with mainland populations, they detected a remarkable genetic similarity with Basque people, especially for individuals from the more historically isolated regions. On the other hand, some mainland components may be relevant for the ancestry of individuals living outside Ogliastra: their data suggest past admixture events with both Eurasian and Sub-Saharan African populations. Globally, a greater knowledge of Sardinia genetic history may be useful to better understand the high prevalence of some hematological and immunological diseases, like multiple sclerosis, type 1 diabetes and β-thalassaemia, in the island; in addition, its unique history suggests that Sardinia may keep some genetic variants that have been lost in the mainland European populations.

Although few recent works have evaluated the impact of rare and structural variants (e.g., [68,69]), current technologies have mainly limited population studies of genome-wide human genetic variation to common SNVs and indels [61]. Thanks to the progressive lowering of costs, this limitation will be probably overcome in the coming years and WGS will definitely become the standard technology for genetic investigations. However, it is important to remember that also whole exome sequencing (WES) has been extensively exploited for both population genetics studies and the diagnosis of genetic diseases [70]. WES is a targeted sequencing technique in which the DNA sequencing step is preceded by the capture of all the exons that are present within a genome (i.e., the exome). The genome coverage is significantly lower in WES data than WGS data ($\sim$1-2% of the genome); therefore, WES can combine a much higher sequencing depth (100x is a common depth for WES, while 30x is frequent for WGS) with a greatly reduced per-genome cost. Notably, the elevated depth of WES data provides an unrivalled opportunity to catch also the rare and low-frequency variants that are usually missed by current whole-genome investigation.

The Exome Aggregation Consortium (ExAC) has done an admirable effort to aggregate and harmonize existing collections of WES data [71]. The ExAC data set, that was publicly released in 2016, resulted from the collection and joint analysis of high-quality exome sequencing data from 60,706 individuals of different geographic ancestry, exceeding by nearly an order of magnitude any previously available exome-wide variant database (e.g., the NHLBI Exome Sequencing Project (ESP) database, that has been available since 2012, includes $\sim$7000 exomes [72]). This effort has been subsequently extended to genome sequencing data, giving rise to the Genome Aggregation Database (gnomAD) [73]. The v2 release of the gnomAD browser aggregates 125,748 exomes and 15,708 genomes from unrelated individuals, but it has been recently expanded with the addition of an impressive number of new genomes (the v3 release includes 71,702 genomes).

As I will extensively discuss in the next sections, in the last decades genetic information has been broadly related with multiple human phenotypes, including several pathological conditions. In this regard, today we are witnessing a new revolution that coincides with the spread of biobank projects. While previous population genetics projects, like HapMap and 1000 Genomes, only collected genetic information for few hundreds or thousands individuals, current initiatives are getting both genetic and biomedical data for an incredibly higher number of individuals, potentially for entire populations of relatively small dimension. For example, the UK Biobank has obtained genome-wide genetic data, pheno-

typic information and biological samples, that may be used to perform different kind of genetic, proteomic and metabolomic essays, for ∼500,000 volunteers across the United Kingdom [74]. At the recruitment participants were subjected to a wide variety of physical measures and they answered to questions about socio-demographics and lifestyle factors. The study also includes the possibility of follow-up analyses because participants provided consent to linkage to electronic health records; moreover, repeated assessments have been done in subsets of the cohort. The initial recruitment lasted from 2006 to 2010, but over time other initiatives have been undertaken, thus further increasing the number of evaluated parameters: for example, objective measures of physical activity have been obtained for 100,000 participants and a multi-modal imaging assessment, that at the end will reach the same number of individuals, is still ongoing. The UK Biobank has generously released the whole dataset and the resulting research findings as an open access resource that may be exploited by academic and private researchers from around the world in the public health interest. Moreover, many other nations, including Estonia, Japan, Canada and Finland, have launched biobank projects: the huge amount of data that will be generated by these initiatives provides a unique chance to change our way of approaching health care, making precision medicine a viable goal in the coming years. While for a long time the one-size-fits-all solution has been the dominant approach in medicine, future healthcare will have to be customized for specific subgroups of patients that will be defined taking into account genetic characteristics, lifestyle and environment.

1.4 The genetic landscape of human diseases

Understanding the relationship between genetic variation and human diseases has always been one of the major interest in biomedicine. In the last decades of the previous century, several genes that cause Mendelian diseases (i.e. monogenic illnesses that are caused by the alteration of a single gene and whose pattern of inheritance follows Mendel's laws) were uncovered thanks to linkage analyses followed by positional cloning [75]. Linkage analyses require the availability of entire families in which the disease of interest is present and consists of looking for DNA markers that tend to be inherited together with the disease; the frequency with which this occurs is considered an estimate of the distance between the marker and the gene responsible for the disease. This allows to obtain an approximate information about the position of the searched gene in the genome, but it is not sufficient to locate it with precision. Chromosome walking is a method of positional cloning and consists in moving along a chromosome by identifying overlapping DNA fragments until the destination is reached, for example the beginning of a gene that may be implicated in a Mendelian disease. Given the short length of the DNA fragments that were used, this procedure could be implemented in a reasonable time only after the identification of two flanking markers that were supposed to be at a low distance from the searched gene. In this regard, the chromosome jumping technique has been a substantial improvement because it allows to move faster avoiding to be stopped by unclonable DNA sequences. While "walking" or "jumping" along chromosomes, the gene start site may be identified for example through the comparison of the reconstructed sequences with those of other organisms, under the assumption

that coding sequences should be conserved. However, additional work was usually necessary to clone the entire gene and the definitive confirmation of having found the right one resulted from the identification of mutations in patients.

In 1989 this strategy allowed to identify alterations of the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene as the cause of cystic fibrosis, an autosomal recessive disease that affects approximately 70,000 individuals worldwide [75–77]. The protein that is encoded by *CFTR* is an epithelial ion channel that regulates the chloride transit through the plasma membrane. *CFTR* mutations result into abnormal salt and water regulation, with consequent problems especially in lungs, pancreas and sweat glands. Pulmonary infections are the most serious challenge for the clinical management of cystic fibrosis and death usually follows the arise of antibiotics resistance. Few years after the discovery of *CFTR*, a major success was obtained also for the Huntington's disease (HD), a rare and adult-onset neurodegenerative disease that is characterized by an autosomal dominant pattern of inheritance [78]. Although HD was the first disease to be mapped to a human chromosome in 1983, the HD gene was eventually identified ten years later: it is also known as huntingtin (*HTT*) and it is characterized by the presence of some repetitions of the CAG nucleotide triplet. The number of repetitions has an elevated prognostic value: while in healthy individuals there are 6-35 repeats, the presence of 40 or more repetitions is associated with the certain development of the disease and intermediate values confer a higher risk. Furthermore, the number of repetitions is negatively correlated with the age of onset and pathogenic repeats largely expand during gametogenesis, especially in the case of male transmission, thus giving rise to a progressive anticipation of the disease onset from one generation to the next. In the protein, whose cellular function is still not known with certainty, CAG repeats are translated into a long sequence of glutamine residues. Notably, these polyglutamine (polyQ) tracks can self-aggregate forming toxic compounds that are structurally similar to the amyloid deposits that are revealed in other neurodegenerative diseases, like Alzheimer's and Parkinson's disease.

Subsequent studies have uncovered that different mutations within *CFTR* are present in patients and this genetic knowledge has began to be exploited to provide better treatments. While traditional therapies have mainly focused on the secondary consequences of the disease, in the most recent years targeted therapies, that aim to restore the normal protein functionality, are available [77]. Ivacaftor has been the first drug of this kind to be approved and it works increasing the opening probability of the channel at the cell surface. Nevertheless, it is efficacious only in 5% patients in which, despite a defective regulation, the protein at least reaches the plasma membrane. On the contrary, it is ineffective in 85% patients that have the most common mutation because it is associated with the production of a misfolded protein that is normally degraded in the cytoplasm before reaching the cell surface; in addition, when it occasionally escapes from the proteasome degradation and reaches the membrane, its regulation is anomalous. Therefore, the combination of lumacaftor, that enhances the intracellular processing and delivery of the protein, and ivacaftor may be an optimal therapeutic option for these patients. However, although the existence of successful examples, an improved knowledge of the genetic basis of human diseases hardly translates into the development of new drugs and, for example, despite the efforts that have been done to clarify its molecular mechanisms, un-

fortunately we are still far away from the development of effective therapies to halt or slow the HD progression [78].

Genetic variants that are associated with human diseases can be placed on a two-dimensional space taking into account their allele frequency and the strength of their effect (Figure 1.9) [79]. As discussed above, family-based linkage studies allow the identification of rare alleles that cause Mendelian diseases (high effect size). On the opposite side, there are genome-wide association studies (GWAS) that look for genotype-phenotype associations testing thousands or millions genetic variants in large human cohorts [80]. GWAS are commonly performed to investigate the genetic determinants of common human diseases (e.g., asthma, diabetes, hypertension, schizophrenia and cancer) and several complex traits, such as body mass index (BMI) and height. Unlike Mendelian diseases, common diseases are influenced by several genetic or environmental factors and for that reason are also called "complex diseases" or "polygenic diseases". GWAS often rely on the comparison of two groups of individuals: cases, that are affected by a particular disease, and matched controls; basically, if a single genetic variant is much more frequent in cases than controls, we can say that there is an association between that variant and the investigated disease. Notably, according to the "common disease, common variant" hypothesis, the inheritance of common diseases is mainly attributable to genetic variants that have an appreciable frequency within human populations (MAF, i.e. minor allele frequency, >1-5%) and that individually have weak genetic effects. Nevertheless, few cases are known of high-effect common variants influencing common diseases: for example, the analysis of the genetic causes of age-related macular degeneration (AMD) has pointed to a relatively low number of genetic variants with large effects. The relative lack of high-frequency and high-effect variants may derive from natural selection forces that, during human evolution, have acted towards the elimination of variants that substantially reduce the reproductive fitness. Furthermore, also low-frequency variants could contribute to the inheritance of common diseases with substantial effect sizes: so far their discovery has been hindered by insufficiently large sample sizes, but future studies could shed light on their contribution. On the contrary, rare variants of small effect are very hard to identify with both GWAS and the classical linkage analyses, thus requiring ad-hoc approaches that, for example, can evaluate the "mutational load" of different genes.

Since the first GWAS was performed in 2005 for AMD, tens of thousands associations have been uncovered, thus revolutionizing our capacity to understand the genetic basis of human diseases and confirming that polygenic effects are relevant for any common disease. Notably, this has been frequently interpreted as the accumulation of weak effects on key genes and regulatory pathways; this hypothesis has been mainly endorsed by the observation that disease-associated variants are primarily enriched within genomic regions that are active in the important cell types, or near genes that are expressed in the relevant cellular context. However, in a recent perspective, Boyle *et al.* [81] reported several discording considerations, starting from the observation that putative causal genetic variants, among which they also included variants that did not reach genome-wide significance but whose directional effect may be replicated in independent cohorts, are extensively widespread in the genome (for example, according to their estimate, every 100-kb window contributes to height variation). They suggested an alternative model for complex traits that

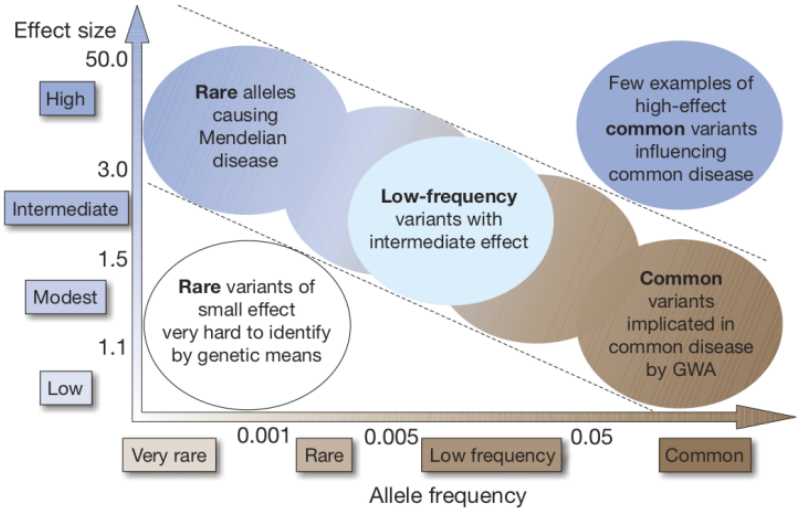


Figure 1.9 – The genetic landscape of human diseases is composed by genetic variants that span along a wide spectrum of allele frequencies and effect sizes. Therefore, its exploration has required the implementation of different approaches, including family-based linkage studies and GWAS. The diagonal dotted lines demarcate the variant categories to which most interest has been dedicated. Figure adapted from [79].

points to the involvement of *all* genes that are expressed in relevant tissues in determining disease risk. According to their "omnigenic model", each complex trait is directly affected by a low number of "core genes", whose biological role can be easily understood and whose perturbation is supposed having strong effects. In addition, all the other expressed genes (peripheral genes) are likely to have smaller but not negligible effects, as a consequence of the high connectivity of gene regulatory networks. Above all, since peripheral genes are usually much more abundant than core genes, the cumulative effect of genetic variants that weakly hit peripheral genes may exceed the contribution of variants that directly affect core genes.

Furthermore, the authors of a recent review suggested that "GWAS findings published to date represent only the tip of the iceberg" [80]: until now GWAS have been mainly performed within European populations, taking into account easy-to-measure phenotypes and assuming an autosomal additive model. First, given that GWAS discoveries are primarily limited by the sample size, taking into account larger groups of individuals will probably result in the pinpointing of additional loci. In this regard, it is important to note that, according to empirical evidence, for each trait the discovery of associations accelerates above a distinct sample size threshold and for many phenotypes the identification of risk loci has not reached a plateau yet. Furthermore, future GWAS cohorts will not only have to include an even greater number of individuals, but also to consider different ethnic groups and populations. Moreover, new discoveries may derive from the analysis of more sophisticated phenotypes. For example, gene-environment interactions seem to play a major role in determining several complex diseases and different gene subsets may be involved in response to

specific environmental exposures. In addition, further information may be obtained through longitudinal studies that examine the variation of quantitative phenotypes over time. Finally, to bring out the currently submerged part of the iceberg, alternative inheritance models should be taken into account. Notably, the genetic architecture of many complex diseases includes epistasis (i.e. statistical interaction between different genome loci in affecting a phenotypic trait) as a pivotal component. Although epistatic effects have so far proved elusive in humans, future studies could be able to uncover them thanks to greater statistical and computational power.

GWAS have received several criticisms and the enthusiasm for new discoveries has been accompanied by a certain scepticism [80]. One of the major issue regards the difficulty to translate GWAS results into an improvement in biological knowledge. First, as a consequence of linkage disequilibrium, multiple genetic variants in the same locus commonly give a significant association, thus complicating the discernment of causal genetic variants from bystander associations. Second, the pinpointed genetic variants may be rarely interpreted mechanistically without further investigations, in particular because GWAS findings are strongly enriched within non-coding genome regions. Even the identification of the causal gene is usually not trivial because it is seldom the one closest to the GWAS result. A striking example of this complexity involves the *FTO* gene [31, 80]. Its name derives from the appearance of fused toes in mouse after the gene deletion, however it has been renamed as "fat mass and obesity-associated gene" and has been deeply investigated for a role in energy homeostasis after the discovery of genetic variants associated with obesity within its first and second introns. More recent studies have shed light on the misunderstanding, showing that the pinpointed genetic variants are located inside a regulatory region that controls the expression of the TF-coding genes *IRX3* and *IRX5*, while it does not impact the *FTO* expression. More precisely, one variants destroys a TFBS recognized by ARID5B, thus determining the increased expression of both *IRX3* and *IRX5*. This molecular perturbation enhances the formation of white fat cells and maybe be responsible for the excessive fat accumulation that was observed in both children and adults.

1.5 Genomics of molecular traits

Quantitative trait locus (QTL) mapping analysis is an approach to relate genetic variants with the inter-individual variation of any quantitative trait. Notably, this strategy has been exploited to investigate the genetic determinants of a wide variety of molecular traits that may be measured trough genomics and proteomics assays, including gene expression and transcript structure [82], histone modifications [83], DNA methylation [84], chromatin accessibility [85] and protein abundance [86]. Since the effect of genetic variants on human diseases probably goes through the perturbation of quantitative molecular phenotypes, the functional interpretation of GWAS findings can be greatly facilitated by the integration of molecular QTLs, genomic annotations and other molecular data. For example, an integrative genomics approach has recently allowed to identify the regulatory circuits that underlie the effect of genetic variants on BMI and other obesity-related metabolic traits [87]. In another study, the quantification of the effect of genetic variants on all the major stages of gene regula-

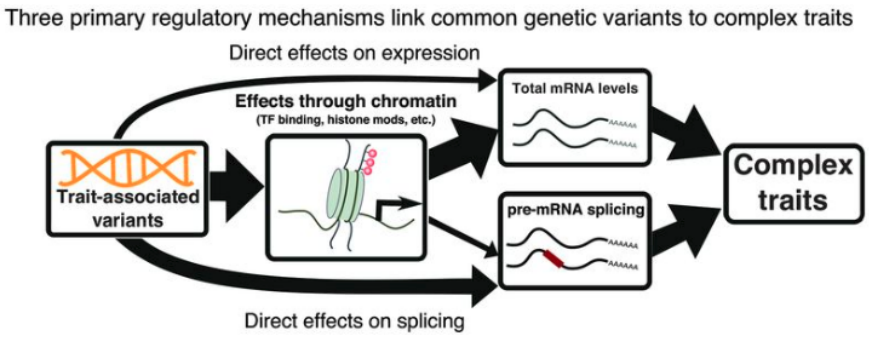


Figure 1.10 – Three main pathways mediate the impact of common genetic variants on complex traits: direct effects on gene expression, direct effects on splicing and effects that pass through chromatin modifications. Figure adapted from [88].

tion, from chromatin to proteins, allowed to devise an interesting model of the regulatory mechanisms through which common variants affect complex traits (Figure 1.10) [88]. According to the authors, genetic variants usually have independent effects on gene expression and pre-mRNA splicing and those affecting splicing have effects of similar or even larger magnitude on complex traits than those affecting gene expression. Furthermore, the effect of genetic variants on gene expression is usually mediated by chromatin modifications, although direct effects are also possible. On the contrary, in the case of splicing direct effects are supposed to be predominant, but genetic variants can affect splicing also by altering chromatin-level traits. More in general, this study, along with many others [89], suggests that genetic variants affecting gene expression or transcript structure are equally important in determining the genetic predisposition to complex human diseases. Therefore, in the last part of this introduction I will elucidate some important concepts about the investigation of the genetic determinants of both these molecular phenotypes.

1.5.1 Gene expression

Genomic loci that include genetic variants affecting gene expression levels are called expression QTLs (eQTLs). More precisely, *cis*-eQTLs are associated with the variation of the expression levels of nearby genes, whereas *trans*-eQTLs affect genes that are further away in the genome. While the alteration of local regulatory elements is the most likely mechanism of action for *cis*-eQTLs, *trans* associations may derive from the qualitative or quantitative perturbation of regulatory proteins. In the last years, the majority of studies has focused on *cis*-eQTL mapping, because *trans*-eQTLs are supposed having much smaller effects that are difficult to catch at the current sample sizes. The identification of *cis*-eQTLs requires the availability of both genotypic and gene expression data for a high number of individuals and it is usually performed through linear regression (Figure 1.11). First, for each target gene, a *cis*-window is commonly defined as the region spanning 1 Mb from both its transcription start site (TSS) and transcription end site (TES). Then, the fitting of several independent linear models is exploited to evaluate the association between each *cis*-variant and the gene expression level:

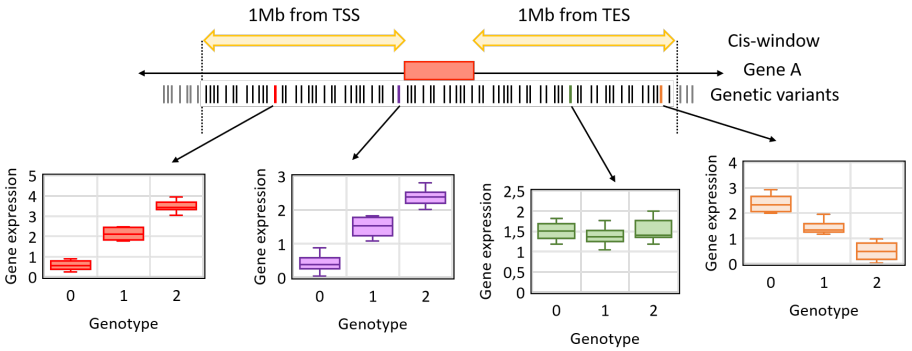


Figure 1.11 – *cis*-eQTL mapping analysis seeks for associations between genetic variants and gene expression levels. First, for each target gene, a *cis*-window is commonly defined as the region spanning 1 Mb from both TSS and TES. Then, the fitting of several independent linear models is exploited to evaluate the association between each *cis*-variant and the gene expression level. In these models genotypes are commonly represented using digits that reflects the copy number of the alternative allele in each individual; therefore, 0 means homozygous for the reference allele, 2 means homozygous for the alternative allele and finally 1 is used for heterozygotes. The same statistical approach allows to identify genetic variants that affect transcript structure instead of gene expression, when a quantitative measure of alternative transcript isoforms is available.

$$y \approx \beta_0 + \beta_1 \times x + cov \tag{1.7}$$

where y is a gene expression trait, β_0 is the model intercept, x is the independent variable, corresponding to the genotype of each individual for a single genetic variant, and β_1 is its regression coefficient. Several covariates, that in equation 1.7 are generically indicated as *cov*, are commonly included in these models, in order to take into account confounding sources of inter-individual gene expression variation. Notably, a correction for the population structure is usually done to avoid both false negatives and spurious associations that may derive from the belonging of individuals to different human populations. This correction usually consists in including the first principal components obtained from genotypic data in the models, since several examples have shown that a principal component analysis can generate a summary of genotypic information that greatly catches geographical differences among individuals [90].

In the context of this thesis, it is particularly noteworthy that the same statistical approach allows identifying genetic variants that affect transcript structure instead of gene expression, when a quantitative measure of alternative isoforms is available. Furthermore, eQTL mapping analysis has been largely complemented by the study of allele-specific expression (ASE), that is the quantification of expression differences between the two haplotypes of an individual, that may distinguished thanks to the presence of heterozygous coding sites; an allelic imbalance in measured expression levels can derive from multiple genetic and epigenetic effects, including the presence of *cis*-regulatory variants, nonsense-mediated decay triggered by variants causing a premature stop codon and imprinting [91].

The GEUVADIS (Genetic EUropean VAriation in DISease) consortium laid a milestone in this field, because, for the first time, they integrated RNA-seq with WGS data to perform a genome-wide mapping of genetic variants affecting both gene expression (eQTL) and the relative abundance of alternative transcripts from the same gene (transcript ratio QTL, i.e. trQTL). In particular, they performed RNA-seq experiments on lymphoblastoid cell lines (LCLs) that were derived from hundreds of individuals whose genome had been sequenced by the 1000 Genomes Project [82]. The GEUVADIS dataset includes WGS and RNA-seq data for 462 individuals from five populations: the CEPH (CEU), Finns (FIN), British (BGR), Toscani (TSI) and Yoruba (YRI). Until few years ago, the generation of RNA-seq data for such a large number of individuals was beyond the reach of a single laboratory and, from a technical point of view, one of the important contributions of this work was to show that the RNA-seq technology was ready for distributed data production. Transcriptome sequencing experiments were indeed performed in seven centres and the variation among laboratories, that is due to technical artifacts, was lower than the variation among individuals, that instead should reflect their different genetic makeup. Furthermore, they showed that eQTLs and trQTLs are equally common but largely independent, suggesting that gene expression and transcript structure are controlled by different regulatory elements.

As stated above, very large collections of WES data are currently available (e.g., the ExAC catalogue of human genetic variants [71]). Although in principle these data may be exploited for the identification of eQTLs and other molecular QTLs, we must be aware of some critical limitations. In the first place, the identification of molecular QTLs requires the integration of genetic information with molecular data, such as gene expression values, that usually are not available for the large cohorts of exome-sequenced individuals. Furthermore, several studies have reported that non-coding genetic variants have important effects of both gene expression and transcript structure (e.g., [92]), as shown also in this thesis. In this regard, it is noteworthy that over half of the sequences that are generated by WES fall outside the target exon sequences [93]. The unexpected sequences may result from the random fragmentation of genomic DNA that precedes the isolation of target sequences with exon-specific probes. Notably, they can be exploited for the identification of genetic variants that fall within coding-exon connected non-coding sequences, such as core promoters [94]. Nevertheless, WES data might cover only a small fraction of the proximal regulatory regions and do not allow the investigation of the distal ones.

For several years, molecular QTL studies have been restricted to few cell types that are easily accessible, thus limiting the possibility to take advantage from them to clarify disease mechanisms at the molecular level. The Genotype-Tissue Expression (GTEx) program, that was launched in 2010, has given a prominent contribution in overcoming this limit, collecting post-mortem samples from many individuals and a wide variety of "normal", non-diseased, tissues [95]. The GTEx project is still ongoing and they plan to collect samples from approximately 1,000 donors upon the end of the program. For example, the GTEx v6p release included gene expression data for 449 human donors, 42 distinct tissues (31 solid-organ tissues, 10 brain subregions and whole blood) and two cell lines derived from blood and skin samples. Genotype information were available for the same individuals, thus enabling the multi-tissue identification of both local (*cis*-eQTLs) and distal (*trans*-eQTLs) genetic effects on gene ex-

pression. The pairwise comparison of the results obtained in the different tissues revealed a substantial concordance with the degree of tissue similarity; notably, the same conclusion was drawn considering *cis*- or *trans*-eQTLs. The most evident broad pattern consisted in a high correlation of effect sizes among brain tissues and among non-brain tissues, with a much lower correlation between the two groups; at a finer scale, within non-brain tissues strong correlations were observed among closely related tissues, for example skeletal muscle and heart tissues. A bimodal pattern of tissue sharing was observed for *cis*-eQTLs: they were usually shared across most of the analysed tissues or specific to a small subset of tissues. In addition, those that were discovered in tissues with larger sample sizes were more likely to be tissue-specific and it could be due to the difficulty in replicating small effect-size eQTLs in tissues with fewer samples. On the contrary, a more predominant tissue-specificity was observed for trans-eQTLs and they consistently showed a greater enrichment within enhancers, that are supposed to be more involved in tissue specific regulation than promoters. However, this property could also result from their under-powering in the detection of trans-eQTLs that may derive not only from low sample sizes, but also from the analysis of bulk tissue samples. It is indeed possible that some regulatory effects can be revealed only within specific cell types, rather than inside an entire heterogeneous tissue. Future studies that will take advantage of both computational (deconvolution methods) and experimental (single-cell RNA-seq techniques) approaches to study gene expression patterns at the single-cell level will probably reach a sufficient resolution to catch also more precise effects. In the meantime, other studies are also contributing to bring out the context-specificity of eQTL effects. For example, the comparison of the associations that were detected before and after stimulation of $CD14^+$ human monocytes with interferon-γ (INFγ) or lipopolysaccharide (LPS) revealed that the functionality of many regulatory variants largely depends on the type and duration of stimulus [96]. The authors of a recent paper have pushed this concept even further, analysing the variation of eQTL effects over time [97]. In particular, they collected RNA-seq at 16 time points during the differentiation of human induced pluripotent stem cells (iPSCs) into cardiomyocytes, starting from 19 Yoruba HapMap cell lines for which genome sequence data are available. Performing a joint analysis that leverages the whole multi-point data to quantify the effect of interactions between genotype and differentiation time over gene expression, they found 550 genes with significant dynamic eQTLs that were classified as early (the effect size decreases over time), late (the effect size increases over time), or switch (the effect size exhibits different directions over time). These effects were consistent with temporal changes of chromatin activity, as revealed by the enrichment of early and late eQTLs within cell type-specific enhancers. Furthermore, they revealed 693 genes with non-linear dynamic eQTLs that affect only intermediate stages of differentiation; for example, 28 genes were associated with variants that have the strongest effect in the middle of the differentiation time course (middle dynamic eQTLs). Many dynamic eQTLs, in particular those with non-linear associations, reflect transient effects that cannot be caught analysing stem cells or adult tissues; nevertheless, they may have phenotypic consequences, as suggested by the examples that are illustrated in the article.

When an eQTL mapping analysis detects associations between genetic variants and gene expression levels it does not provide any information about the

mechanism driving the gene expression variation. Nevertheless, given the prominent role of TF circuits in the orchestration of gene expression profiles, the perturbation of TF binding is the most straightforward hypothesis and it has been confirmed by many studies that have shown that eQTLs are significantly enriched within experimentally determined or computationally predicted TFBS (e.g., [98,99]). Several computational methodologies have been proposed to evaluate the impact of regulatory variants on TF-DNA interaction and, although they may differ in details, they usually share a common workflow [31]. Basically, for each investigated variant, the sequence of both the reference and the alternative allele is extracted, then both sequences are scored according to a model describing the TF binding preferences (S_{REF} and S_{ALT}) to obtain a differential score ($\Delta S = S_{REF} - S_{ALT}$) that reflect the difference in DNA binding affinity; the greater this difference, the greater the likelihood that the variant severely impacts *in vivo* binding. Moreover, valuable information can be obtained through the direct identification of genetic variants that affect DNA binding events (binding QTLs, or bQTLs). Different studies have revealed an extensive inter-individual variation of TF binding through the repetition of standard ChIP-Seq experiments in genotyped individuals, but unfortunately they have been strongly limited to small sample sizes, as consequence of both prohibitive costs and substantial experimental variation across samples (batch effects). Tehranchi *et al.* proposed the pooled ChIP-Seq technique as an efficient and cost-effective alternative [100]. Samples coming from different individuals can be pooled to perform a single ChIP-Seq experiment for each TF: alleles that enhance the TF binding will be enriched in the post-ChIP pool with respect to the pre-ChIP pool, while the frequency will be remain the same for alleles that do not have an impact on TF binding. This strategy has also the advantage to fully leverage the information that are present in heterozygotes; on the contrary, standard ChIP-seq experiments can only evaluate the average signal between the two alleles and therefore they are not able to catch any association in the extreme situation in which there are only heterozygous individuals. In addition, the same approach may be exploited to pinpoint genetic variants affecting epigenetic markers and they indeed considered one histone modification (H3K4me3) together with five TFs that are critical for immune cell differentiation and function (JunD, NF-κB, Pou2f1, PU.1 and Stat1), performing the experiments on LCLs that were obtained for 71 YRI individuals. In particular, they observed that only a very low proportion of bQTLs (<0.9%) overlap with predicted DNA motifs for the corresponding TF. This result suggested that also other features of the local sequence may influence TF binding. Consistently with this hypothesis, their data show that the GC content, in the region surrounding predicted binding sites, may be relevant in determining the binding affinity of some TFs. An alternative explanation is that many bQTLs reflect cooperative binding events, in which the variation of the binding of recruited TFs would be seen without the alteration of their corresponding DNA motif. Pairwise comparisons uncovered a significant overlap between bQTLs that were revealed for different TFs, suggesting that single genetic variants may perturb the binding of multiple TFs. Although it could also point to the presence of independent causal variants in the same loci, it is not likely to occur systematically and in addition it was denied by the detection of a broad directionality agreement (i.e. if bQTLs for two different TFs overlap, the same allele usually have higher binding for both TFs). The genetic-driven variation of the binding of multiple TFs may reflect

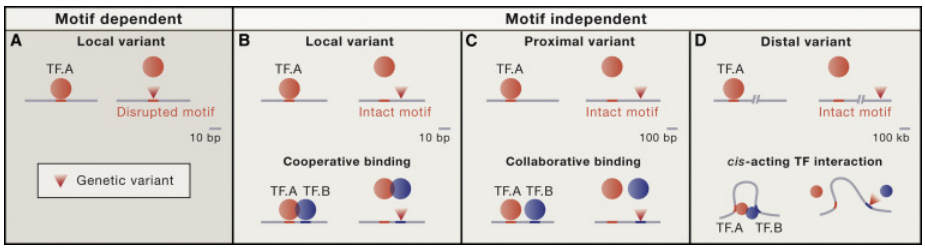


Figure 1.12 – Different mechanisms can explain the genetic-driven variation of TF-DNA interactions. Notably, in an seemingly counter-intuitive way, motif independent events, that may involve local, proximal and distal variants, are much more frequent than motif dependent event. See the main text for additional details about the different cases. Figure adapted from [31].

the alteration of a motif that is recognized by a "pioneer factor", that binds first to the DNA and then directly or indirectly promotes the binding of other TFs. For example, they observed several events of this kind for NF-κB and JUND that are known to act cooperatively. Furthermore, they found that overlapping bQTLs frequently reflect the perturbation of motifs that are recognized by CTCF, a protein that plays a crucial role in the organization of 3D chromatin structure. This result suggested that CTCF may be a major pioneer factor, able to recruit all the other TFs, but it did not provide information about the mechanism. Although it could result from direct physical interactions between CTCF and the other TFs, according to the authors an indirect effect is more likely: CTCF may model nucleosome occupancy and chromosome looping in such a way to make chromatin more accessible to other factors.

Many other studies have contributed to consolidate the idea that only a minority of variable TF-DNA binding events derives from the destruction of motifs that are recognized by the investigated TF, stressing on the contrary the importance of motif independent events (Figure 1.12) [31]. In presence of an intact motif for the studied TF, the perturbation of its binding ability may be due to variants that hit motifs that are recognized by other TFs. Local variants can impact cooperative binding events that involves physical interaction between TFs and the presence of overlapping, or closely located, binding sites. On the other hand, proximal variants may influence collaborative binding events, that does not require direct protein-protein interactions, but reflects an interdependence among TFs that, for example, must compete with nucleosomes to access DNA. Furthermore, the perturbation of TF-DNA binding is not restricted to the action of nearby genetic variants: the stability of interactions with DNA and between TFs may be indeed reduced in the presence of distal variants that alter 3D chromatin structure. Finally, two additional observations accentuate the complexity of TF-DNA interactions and contribute to further complicate our interpretations of genetic-driven perturbations. First, only a minority of motif-disrupting variants effectively causes the alteration of the binding of the corresponding TF, maybe as a consequence of an extensive buffering that derives from the presence of TFBS clusters in the genome [31]. Second, many regulatory QTLs do not overlap with eQTLs, consistently with the finding that changes in TF-DNA binding frequently have no measurable effects on gene expression [31, 101].

As I said above, standard eQTL mapping analyses correlate single genetic variants with gene expression levels: this approach is limited with respect to the complexity of gene expression regulation and it does not leverage our incomplete but substantial knowledge of the regulatory code. Conversely, we can speculate that all the genetic variants within a regulatory region contribute to determine a global perturbation of TF binding that eventually results in the gene expression variation, consistently with the knowledge that many genes are under the control of multiple independent eQTLs in their *cis*-regulatory region [99, 102]. In particular, the binding profile of any regulatory region can be obtained by computing TBA values for multiple TFs using public collections of PWMs; in addition, when both WGS and gene expression data are available for many individuals, the correlation between TBA profiles and gene expression levels can be evaluated to uncover new eQTLs. The TBA model will be better described in the second chapter, along with the results that were obtained by its application to the GEUVADIS dataset.

1.5.2 Transcript structure

Although many efforts have been focused on genetic variants that affect gene expression levels, recent studies have shown that genetic variants broadly impact the whole cascade of RNA processing that include both pre-mRNA processing (splicing and polyadenylation) and mRNA dynamics (translation, stability and mRNA localization). In addition, numerous examples support the idea that RNA processing variants can be important drivers of phenotypic variability in humans [89].

From a computational point of view, multiple approaches have been proposed to discover human genetic variants that are associated with changes in transcript structure. In this context, a primary distinction can be made between methods that focus on distinct molecular events (i.e. alternative splicing or polyadenylation) and those that instead generically investigate the expression of alternative isoforms. The results that were reported by the GEUVADIS project fall into the second category. As described above, they leveraged the integration of RNA-seq and WGS data to identify also genetic variants affecting the relative abundance of alternative transcripts from the same gene, using as quantitative phenotype the ratio between the expression of every transcript isoform and the total gene expression (transcript ratio). The AStalavista software [103] was subsequently exploited for the classification of detected events and they observed that $3'$ and $5'$ end modifications are much more common than exon skipping and intron retention events. Moreover, they found that, besides the expected enrichment at splice sites, trQTLs were also over-represented within $3'$ UTRs and promoters. Globally, these results indicate that alternative polyadenylation may be largely affected by genetic variants and suggest that also DNA regulatory elements may be involved in the regulation of post-transcriptional events, although they are usually supposed to be mainly regulated at the RNA level. The analysis of the same dataset with a more sophisticated strategy, that treats the relative expression of all the alternative isoforms of each gene as a multivariate phenotype and implements a distance-based approach to compare within-genotypes and between-genotypes variance, allowed to catch additional associations and confirmed the prevalence of $3'$ and $5'$ end modifications [104].

Taking advantage of the GEUVADIS dataset, other studies have investi-

gated the effect of genetic variants on specific mechanisms of RNA processing. For example, the Altrans software was exploited for the discovery of splicing QTLs (sQTLs), through the identification of genetic variants that are associated with the variable expression levels of exon junctions [105]. In addition, the computational analysis of DNA sequences around annotated splice sites and poly(A) signals allowed the identification of many genetic polymorphisms that potentially influence splicing or polyadenylation as a consequence of the alteration of the affinity of core RNA elements for regulatory factors [106]. However, they did not perform any quantitative evaluation of the relationship between the pinpointed genetic variants and the expression of alternative APA isoforms; moreover, considering only the PAS motif, they omitted other known regulatory mechanisms. On the other hand, previous studies found strong associations between human genetic variants and the expression of alternative 3′ UTR isoforms, but they were strongly limited in the number of variants and samples [107–109]. For example, a NGS technology that targets 3′ ends (DeepSAGE) was exploited to catch genetic variants affecting both gene expression and the usage of alternative poly(A) sites [109]. In that particular context, the identification of APA associated variants relied on searching genes that contained deepSAGE tags that were regulated in opposite directions by the same SNP. For instance, this pattern was observed for the *LPCAT2* gene where rs12934747 creates a new canonical polyadenylation signal (AAUAAA) at the beginning of the 3′ UTR, thus resulting in the reduced expression of the transcript that contains the full length 3′ UTR. However, the statistical power of this study was seriously compromised by the availability of data for less than one hundred individuals. In summary, a systematic investigation, specifically targeted to APA rather than generically to transcript structure, based on a large number of samples and variants, and unbiased in the choice of variants to examine, was lacking. In the third chapter, I will present a new computational strategy that aims to overcome the highlighted limitations for the specific purpose of correlating variants with the relative expression of alternative 3′ UTR isoforms in a large human population. Notably, it relies on standard RNA-seq data for the quantification of APA isoforms and it is statistically analogous to methods commonly implemented in eQTL mapping analysis.

In conclusion, it is worthwhile to note that QTL mapping analysis is not the only viable way to investigate the genetic determinants of transcript structure variation. Deep learning methods have been successfully exploited to generate predictive models for alternative splicing and polyadenylation. Besides providing new insights about regulatory mechanisms, these models allow the functional interpretation of genetic variants, without relying on the availability of population data. For example, Xiong *et al.* [110] implemented a machine-learning approach to generate tissue-specific computational models of splicing regulation using reference genome sequences and RNA-seq data from healthy human tissues. These models can be used to score how strongly genetic variants affect RNA splicing and, scoring more than 650,000 intronic and exonic variants, they observed that disease-annotated SNVs have higher scores than common variants, including SNPs that have been associated with diseases in GWAS. Analogously, a deep neural network (APARENT) [111] was trained on isoform expression data from millions of synthetic APA reporters to learn a comprehensive model of the *cis*-regulatory APA code and it enabled the identification of human genetic variants that act through APA dysregulation.

Chapter 2

A functional strategy to characterize expression Quantitative Trait Loci

In this chapter I will describe a new strategy to detect genes whose expression level is influenced by regulatory genetic variants. Using a large dataset of coupled genome and expression data, we correlated the TBA profile of individual regulatory regions with the expression level of target genes. Unlike the standard eQTL mapping analysis, our TBA model takes advantage of the current knowledge of the regulatory code and it naturally incorporates the effect of multiple variants within regulatory regions. Notably, we found that it allows the pinpointing of eQTLs that are not identified when studying the correlation between gene expression and individual variants. In addition, it can help formulating hypotheses on the mechanism behind eQTLs by indicating the TFs whose binding perturbation mostly contribute to the gene expression variation.

2.1 Results

2.1.1 The TBA model discovers new eQTLs

We computed the TBA profiles of both local and distal regulatory regions associated with ∼22,000 genes in 344 individuals of European descent whose genomes were sequenced by the 1000 Genomes consortium [65] and whose gene expression profiles in lymphoblastoid cell lines were measured by RNA-sequencing in the GEUVADIS project [82]. Each TBA profile describes the affinity of the regulatory region for the 640 human PWMs of the HOCOMOCO collection [113]. A TBA profile is, therefore, a summarization of a regulatory region in terms of its overall propensity to bind transcription factors, taking simultaneously into account the effect on binding affinity of variants located along the whole sequence.

To each gene, we associated a proximal regulatory region, defined as the region spanning from 1500 bps upstream to 500 bps downstream of the TSS (suit-

The content of this chapter was published as Ref. [112]

ably merged in the case of multiple TSSs, see Methods); and, when available, distal regulatory regions associated with the gene by the PreSTIGE tool [114], which is based on the correlation between cell-type specificity of epigenetic modifications and gene expression. The median length of the regulatory regions considered was 3,296 for proximal and 1,752 for distal. Note that each gene can be associated by PreSTIGE to more than one distal region, and conversely some distal regions are associated with multiple genes.

We then used principal component regression to detect correlations between variation in TBA and gene expression in individuals. Specifically, for each gene, we fitted a model for each associated regulatory region where the independent variables are the top Principal Components of TBA values for each TF that explain (at least) 95% of the original variance and the dependent variable is gene expression as measured by RNA-sequencing (Figure 2.1 and Supplementary Figure A.1). We reasoned that this approach could discover eQTLs not easily revealed by the analysis of correlations between expression and individual SNPs for two main reasons: first, we are able to take into account the combined effect of several variants lying in the same regulatory region; and second, we exploit our knowledge of the regulatory code by weighting variants based on their predicted effect on TF binding.

The number of genes for which a statistically significant correlation was found between TBA and gene expression is slightly higher than the number of genes for which at least one eQTL was found in [82] by univariate eQTL (uni-eQTL) analysis (Table 2.1). Importantly, while the overlap between the significant genes in the two methods is sizable (2,238), there are 1,543 genes showing significant correlation with TBA but no individually significant variants. These regions are, therefore, eQTLs that can be revealed only by taking into account the combined effect of multiple variants on the affinity for TFs of the cis-regulatory region. Conversely, for 980 genes one or more eQTLs were found in Ref. [82] but not with the TBA model: in 759 cases, all the significant SNPs lie outside the regulatory regions analysed by the TBA model. However, for the remaining 221 genes there is at least one individual SNP correlated with gene expression whose effect is not captured by the TBA model. This could be explained by the fact that even if the HOCOMOCO collection of TFs is very large it may be not exhaustive. However, using a much smaller collection of PWMs (the JASPAR Core Vertebrate Collection [115]), containing only 205 PWMs, resulted in a modest reduction in the number of significant models (3,208 vs 3,421 – this analysis was performed on proximal regulatory regions only). To understand why the loss in predictive power is so limited when the number of JASPAR PWMs is less than half compared to HOCOMOCO, we looked at the distribution of the number of principal components needed to explain 95% of the variance among subjects: the mean number of PCs needed decreases from 4.24 to 3.55 moving from HOCOMOCO to JASPAR (Supplementary Figure A.2). Therefore, while in theory the TBA spaces defined by the two PWM databases have very different dimensionality, the TBA profiles of actual regulatory sequences lie in spaces of much smaller dimensionality, and these are not very different between the two databases. Alternatively, significant eQTLs that are not captured by the TBA model could affect expression by mechanisms other than a change in affinity for a TF (e.g., a variant could affect gene expression through altering DNA methylation [116]).

An alternative way to test whether differences in TBA correlate with differ-

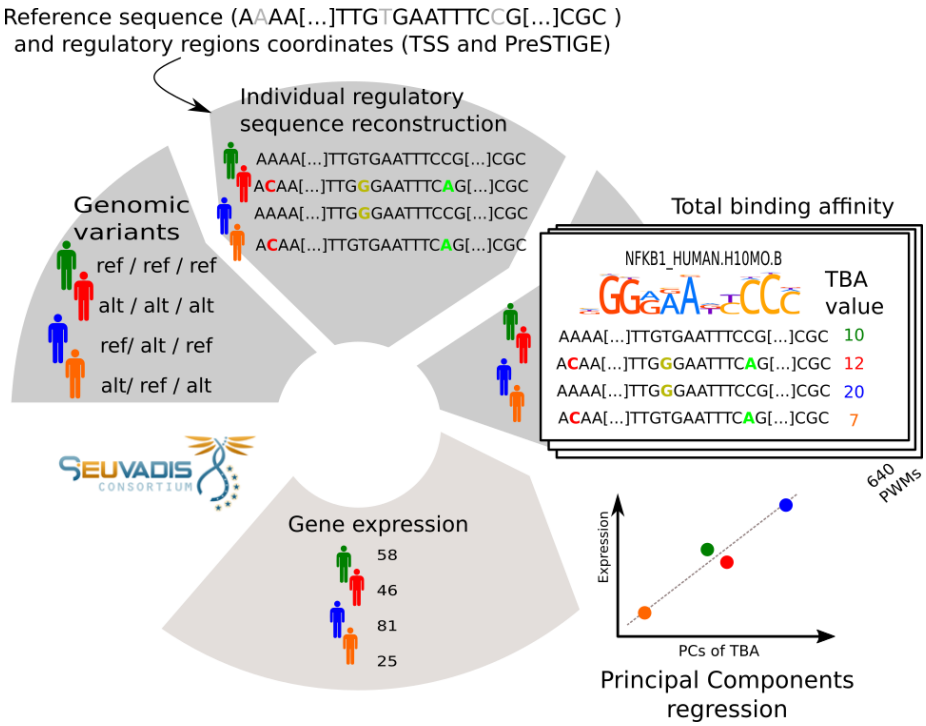


Figure 2.1 – Schematic representation of the proposed method. Using a dataset, like the Geuvadis one, where both WGS and expression data are available for the same set of individuals, for each TBA model, we obtain the individual sequences of a single regulatory region starting from the reference sequence and the individual genomic variants calls. Then, we compute TBA values for 640 HOCOMOCO PWMs on the whole sequences, without looking for individual binding sites. Finally, we seek correlations between individual variation in the TBA profile of the regulatory region and in the expression level of the target gene—for simplicity, we represented univariate regression with the TBA of a single TF as the independent variable while in reality we fit a multivariate model using principal component regression.

	uni-eQTL model	TBA model	
		Local RR	Distal RR
Median number of variants	10,243	19	10
Median sequence length	2,000,000	3,296	1,752
Number of significant genes	3,259	3,781	

Table 2.1 – Comparison of TBA and uni-eQTL models. For the uni-eQTL models, the sequence length was always the same and we reported the median number of variants that were considered for each gene. On the other hand, many TBA models may have been fitted for each gene, independently for each associated local or distal regulatory region (RR).

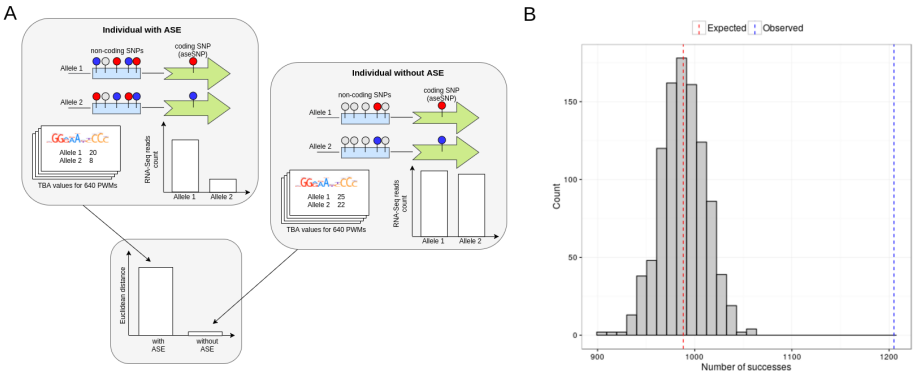


Figure 2.2 – Individuals with allele-specific gene expression (ASE) tend to have greater TBA differences between the maternal and paternal regulatory region. (A) Schematic representation of the analysis. (B) The number of genes for which the median distance between maternal and paternal TBA profiles is higher in individuals with ASE compared to individuals without ASE (blue line) is greater than expected by chance (red line). The histogram represents the distribution of such number in 1000 simulations in which the ASE/non-ASE status of individuals was independently randomized for each gene.

ences in gene expression relies on allele-specific expression (ASE). We expect to find larger differences in the TBA profiles of the maternal and paternal regulatory regions in subjects displaying ASE than in other subjects. To test this hypothesis, we considered the ASE data reported in Ref. [82]: for 2,740 genes for which robust ASE data were available (selected as described in the Methods) we computed, for all heterozygous subjects, the Euclidean distance between the TBA profiles of the paternal and maternal regulatory regions. We then separately computed the median distance for subjects with and without evidence of ASE. For 2,204 genes, such medians were different in the two groups, and for 1,205 the median distance in the ASE group was greater than the non-ASE group (binomial $P=5.1\times10-6$). The significance of the result is also confirmed by a permutation-based approach: we repeated the analysis on 1,000 permuted datasets obtained by randomizing the individuals with ASE, obtaining an empirical $P<0.001$. Therefore, differences in TBA profiles between maternal and paternal alleles in heterozygous individuals predict differential expression of the two alleles. The analysis pipeline and the result are summarized in Figure 2.2.

2.1.2 Knowledge of the regulatory code contributes to the TBA model

As discussed, two factors could in principle explain why the TBA model is able to identify some associations that are not uni-eQTLs: first, the TBA model can take into account the combined effect of more than one variant in a regulatory region; and second, it uses knowledge of PWMs to give more weight to variants likely to change the affinity of the sequence for TFs. However, the results shown above do not reveal whether both factors actually contribute to the model results.

To clarify this point, we built a third type of model, called the multivariate

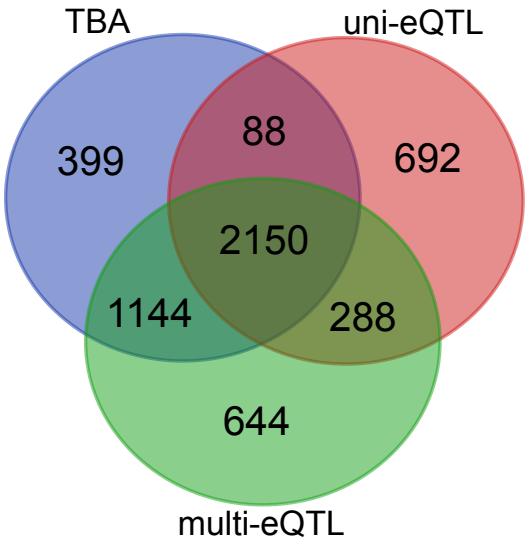


Figure 2.3 – Overlap between the genes with a significant uni-eQTL, multi-eQTL and TBA model.

eQTL (multi-eQTL) model, which, like the uni-eQTL model, uses the genotype as its independent variables, but, like the TBA model, considers together all the variants in a principal component regression. Having 399 genes significant only for the TBA model and not for the other two shows that the performance of the TBA model is not simply due to the simultaneous inclusion of several variants in a regulatory region, but also to the weight given to them based on their effect on affinity for TFs (Figure 2.3).

2.1.3 The TBA model detects TF-target interactions that depend on genetic variation

Once the TBA model has revealed an association between regulatory variants and gene expression, it can be further exploited to identify the TFs whose affinity for a specific regulatory region correlates with gene expression and, thus, hypothesize a mechanistic basis for the variation in gene expression. To identify the specific TFs whose TBA correlates with expression, for each gene with at least one significant TBA model, further linear models were fitted independently for each PWM on each linked regulatory region whose TBA model is significant. In these models, the regressors include the TBA of a specific PWM and the same covariates used in the PC regression, but for simplicity we refer to them as "univariate TBA models". By selecting models in which the PWM coefficient is significantly different from 0, we thus obtained, for each gene on each linked regulatory region, a list of putative TFs whose binding variation affects the gene expression in a sequence-dependent way, which we refer to as the "eQTL-predicted regulators" of the gene.

Thus, we obtained, for each significant regulatory region, the list of TFs predicted by the TBA model to affect the expression of target genes. We expect the binding of these TFs to depend on genetic variants found in the regulatory

regions. To validate this claim, we used a recent systematic evaluation of binding QTLs (bQTLs) in lymphoblastoid cells carried out for 5 TFs (JUND, NF-κB, PU.1, POU2F1 and STAT1) [100]. Specifically, we fitted, for each of the five TFs, a logistic model in which the independent variable is the presence of a bQTL in a regulatory regions, and the regressors are the length of the region and the significance of the appropriate PWM in the univariate TBA model. A significant coefficient for the latter regressor would imply that the TBA model is able to predict bQTLs, thus supporting its mechanistic interpretation.

For all five TFs, the coefficient of the TBA term is positive, as expected, and for two of them (NF-κB and STAT1) it is statistically significant (resp. P=1.9×10^{-4} and 3.4×10^{-3}). These results suggest that the TBA model is potentially able to provide information about the mechanism by which genetic variants influence expression. Variation in the motif sequence is not the only driver of variation in TF binding [31], and this could explain why our model is unable to predict bQTLs in a significant way for all the 5 TFs. Another possible reason is the difference between populations used as sources for the lymphoblastoid cell lines (Yoruba for bQTLs and EUR for the GEUVADIS samples used in our models). Finally, some of the TFs studied in [100], such as PU.1 and JUND, are pioneer transcription factors whose binding is required for the subsequent binding of signal-dependent transcription factors which in turn drive expression (e.g. [117]). As discussed below, our linear regression in its present form is not ideally suitable to deal with such interactions.

2.1.4 The TBA model allows the mechanistic interpretation of GWAS hits

Our approach can therefore be useful to generate mechanistic hypotheses to explain the association between regulatory variants and disease. While this would be ideally done on transcriptomics datasets including diseased individuals and healthy controls, some interesting examples could be obtained exploiting the vast number of GWAS results already available. Using the GRASP catalog [118] of GWAS hits and linkage disequilibrium data for CEU individuals we asked how many GWAS hits correspond to a variant found in our significant regulatory regions: out of 48,211 variants 7,231 were the same or in LD with variants included in our significant TBA models. This 15% overlap is in line with recent results from eQTL analyses [82] and is significant (p-value 0.0003, randomization test).

The authors of [100] showed that bQTLs are more enriched in GWAS hits than eQTLs, but also pointed out that this raises a problem, since it is not clear how a variant can affect a phenotype by altering the binding of a TF without being an eQTL. We reasoned that our new approach for the detection of eQTLs could provide a partial solution to this problem, by revealing eQTLs that are not detectable by standard analysis. To verify this, we considered the bQTLs that are in LD with a GWAS hit (GWAS-bQTLs) and asked whether they were enriched in regulatory regions corresponding to significant TBA models. We fitted a logistic model in which the presence of a GWAS-bQTL in a regulatory region was predicted by three regressors: the length of the region, the presence of a putatively causal eQTL found by ordinary eQTL analysis in [82], and the significance of the region in the TBA model. All three predictors were significantly and positively correlated with GWAS-bQTLs (Figure 2.4), implying in

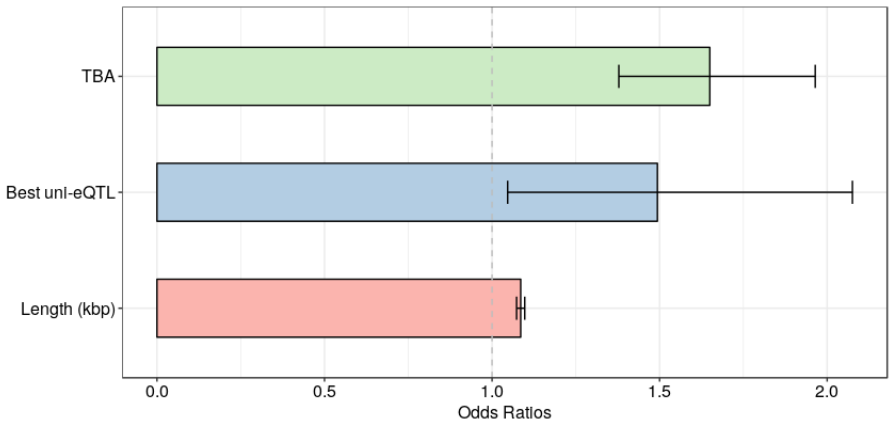


Figure 2.4 – The probability that a regulatory region contains a GWAS-bQTL is increased, independently, by the significance of the corresponding TBA model, by the presence of a putatively causal eQTL, and by the length of the region. The figure shows the odds ratios for each predictor in a multi-variate logistic regression and their 95% confidence intervals.

particular that the TBA model is a predictor of GWAS-bQTL independent of the other two ($P = 3.0 \times 10^{-8}$ for the significance of the TBA model in the multivariate model). Specifically, we found 144 regulatory regions containing bQTLs that are in LD with GWAS hits, but whose association with gene expression variation was unknown. Some individual examples are discussed below.

The TBA model shows a correlation between the expression of C-type lectin-like 1 (*CLECL1*) and the TBA profile of its local regulatory region, in which SNPs associated with type 1 diabetes [119] are found that are also bQTLs for NF-κB and JUND. Both these TFs are identified as relevant to *CLECL1* expression by our model, which also suggest a role for CREB1, a TF already known to have a relevant role in autoimmune diseases and glucose metabolism [120]. In addition, the TBA model evidenced that regulatory variants within 12 distal regulatory regions are significantly associated with the variation of CLECL1 expression; one of these enhancers contains another bQTL for JUND: it was associated with type 1 diabetes, but its association with gene expression variation was previously unknown.

The ability of the TBA model to make light on distal regulatory mechanisms that can contribute to the development of pathologies is also demonstrated by another example. The TBA model reveals a distal regulatory element whose TBA profile is strongly correlated with the expression of two genes (*BLK* and *FAM167A*): it includes a SNP that was detected as bQTL for several TFs (JUND, NF-κB, STAT1 and PU.1) and that was associated with both systemic lupus erythematosus (SLE) [121] and Kawasaki disease [122], but its effect of gene expression was not previously reported. *BLK* (B lymphoid tyrosine kinase) is a protein coding gene that encodes for a nonreceptor tyrosine-kinase that plays a role in B-cell receptor signaling and B-cell development; its putative involvement in the etiology of both SLE and Kawasaki disease is confirmed also by other studies [123, 124]. *FAM167A* is a much less characterized protein coding gene. The importance of three out of four TFs (JUND, NF-κB and STAT1)

for the expression of both genes was correctly predicted by the univariate TBA models on these distal regulatory elements.

Finally, *PTPRO* is a protein phosphatase that has been linked via three intronic SNPs to learning and memory in a cohort of patients with mental disorders (schizophrenia spectrum disorder and bipolar disorder) and normal controls [125] and that is highly expressed in mouse neural system during development [126]. Our results show correlation between *PTPRO* expression and the TBA profile of its local regulatory region where no individual SNPs were detected by uni-eQTL models; in addition, the univariate model includes POU2F1 among the significant TFs, in agreement with the bQTL evidence, thus suggesting a mechanistic link between SNPs, *PTPRO* expression and possibly neurocognitive functions.

2.2 Discussion

Variation in gene expression is a common effect of genetic variation, and GWAS studies suggest that often higher-level phenotypes are mediated by variants affecting gene regulation. The aim of eQTL studies is to determine how genetic variation influences gene expression, by seeking statistically significant correlations between genotypes and mRNA levels.

GWAS and eQTL studies are commonly carried out performing single-variant tests. Although an unbiased analysis of genetic variants is worthwhile and these methods turned out to be extremely powerful, they also have some important limitations. First, they do not leverage the current knowledge of the regulatory code. Second, they do not provide any suggestion for the mechanistic interpretation of the results. In addition, they do not explicitly model the allelic heterogeneity that concerns both gene expression [99, 102] and complex traits [127]. Shifting from SNP-based associations to SNP aggregation methods, that evaluate biologically informed aggregates of SNPs, can allow to overcome these issues. Gene-based approaches are particularly attractive, because, since genes are the primary functional unit of the genome, the result interpretation is more straightforward. In addition, gene-based strategies allow to substantially reduce the multiple testing burden, thus intrinsically enjoying greater statistical power, and they may also guarantee a higher grade of replication among different populations. In the last years, several statistical methods have been proposed to aggregate the results of SNP-based GWAS into gene-based measures of associations (e.g., GATES [128], VEGAS [129] and COMBAT [130]). Furthermore, in Gamazon *et al.* [131] the cumulative effect of regulatory SNPs is aggregated into a predicted expression value that is then used to identify genes whose dysregulation is involved in common diseases.

Here, we focused on eQTL mapping and we proposed a different SNP aggregation method that explicitly takes into account the predicted effect of variants on the binding of transcription factors, by evaluating the changes in TBA profiles induced by the variants found within regulatory regions. The TBA-based method can be applied to both local and distal regulatory elements and it is able to reveal hundreds of eQTLs that were not identified by single-SNP eQTL analysis, providing in this way new testable hypotheses about the mechanism leading from genetic variation to complex phenotypes.

Previous attempts to exploit TF sequence specificity in functionally inter-

preting GWAS hits [132–134] focused on ranking single SNPs using their predicted effect on binding. Our approach instead is able to consider cooperative effects of several SNPs together, and models TF binding without resorting to a cutoff on the PWM score, thus identifying associations not found by other methods. Using TBA to measure the effect of SNPs allows us to consider their role both on high scoring subsequences and on other DNA portions less similar to the perfect match but relevant in determining functional binding [37]; it would be interesting to expand this approach to more recent models of TF binding preferences that also account for interdependencies between nucleotides inside matches [31].

The main limit of our approach lies in our limited knowledge of the regulatory code. Indeed, the transcription factors whose binding preference we can describe as a PWM are probably a relatively small fraction of those coded by the human genome. Moreover, there is certainly genetic variation that influences gene expression by mechanisms other than changing the affinity for a transcription factor. Therefore, the best strategy is, currently, to use aggregation methods in combination with single-SNP eQTL analysis to achieve maximal power in detecting correlations between genetic variants and gene expression. In addition, until now we have taken into account SNPs and indels, but an interesting expansion of the approach would be to include copy number variations (CNVs), since they are known to have a role in several human diseases [79] and datasets for big cohorts will likely become available soon. Furthermore, our method cannot in its present form take into account interaction effects between TFs, such as cases in which pioneer TFs are required to bind a region before signal-dependent TFs can bind it and drive expression. Statistically, this could be modelled by interaction effects between TBA profiles of different TFs. Currently, this would be difficult to implement systematically because of the exponential growth of the number of regressor variables, but case studies in controlled contexts in which the number of relevant TFs is limited would be potentially interesting.

Another promising avenue of investigation is the application of TBA-based SNP summarization to complex diseases, in which regulatory variation plays a major role. Inspired by the gene-based approach that was proposed by Gamazon *et al.* [131], we are currently investigating statistical models in which TBA profiles are used to predict the genetic component of gene expression that can be subsequently correlated with any disease status (for further details see Chapter 4).

2.3 Material and Methods

2.3.1 Individual sequences of regulatory regions

TSS annotation was downloaded from Gencode v12 [135] consistently with the annotation used by the GEUVADIS Consortium [82]. For each individual, we considered the region spanning 1,500 bp upstream and 500 bp downstream from the Transcription Start Site (TSS). In the following, we will refer to these 2,000 bp regions as "promoters". One TSS is defined for each transcript in Gencode v12 and we associated a promoter to each transcript. When different promoters linked to the same gene shared overlapping regions, the promoters were merged with `bedtools merge` [136], in order to consider each genomic region only once.

Furthermore, in order to take into account distal regulatory regions, all the gene/enhancer associations obtained by the PreSTIGE algorithm [114] in lymphoblastoid cells were downloaded (GRCh37). The Gencode v12 annotation was used to associate Ensembl Id (ENSG) to gene symbols. In addition, for each gene, enhancers that overlap at least one of its alternative promoters were excluded. In this way, we were able to associate at least one enhancer to 4,291 genes with expression values.

Genomic sequence variation data for 373 European samples, obtained by the 1000 Genome Project [65], were downloaded from the GEUVADIS Data Browser [82] together with gene expression data. 29 samples were discarded because their genotypes were not phased by [82]; thus, a total of 344 European samples were analysed. In addition for efficiency reasons, only genetic variants that are variable among this subset of genotyped individuals were retained for the subsequent analysis. The reference genome (GRCh37, as in Gencode v12) and these whole genome sequencing data were exploited by the `vcf_rider` library [137] to efficiently reconstruct the diploid genomic sequence of regulatory regions in each individual and subsequently compute Total Binding Affinity (TBA) values on them (see below for details).

When the presence of indels led to regulatory regions (promoters or enhancers) differing more than 10% in length from the reference sequence the regulatory region was discarded for all individuals to prevent differences in TBA from being strongly driven by differences in the length of regulatory regions. Genes without any variant in all the associated regulatory regions or without an available expression value were discarded: a total of 22,129 genes were retained for further analysis.

2.3.2 Total binding affinity

Following Foat *et al.* [34], we computed the total binding affinity (TBA) of each regulatory region for 640 human PWMs derived from the HOCOMOCO-v10 database. In particular, Positional Count Matrices (PCMs) were downloaded in the TRANSFAC format (April 11, 2017). We then added a pseudocount of 1 only to zero counts and converted PCMs into the format accepted by the vcf_rider library, which internally performs the conversion to PWMs. The TBA a_{rw} of a sequence r for a PWM w is given by:

$$a_{rw} = \log \sum_{i=1}^{L-l} \max\left(\prod_{j=1}^{l} \frac{P(w_j, r_{i+j-1})}{P(b, r_{i+j-1})}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_{i+j-1})}{P(b, r'_{i+j-1})} \right) \tag{2.1}$$

where l is the length of the PWM w, L is the length of the sequence r, r_i is the nucleotide at the position i of the sequence r on the plus strand, r'_i is the nucleotide in the same position but on the other strand, $P(w_j, r_i)$ is the probability to observe the given nucleotide r_i at the position j of the PWM w and $P(b, r_i)$ is the background probability to observe the same nucleotide r_i. Background nucleotides frequencies were calculated on intergenic portions of the human genome (UCSC version hg19) as [37].

To avoid excessive complexity, we decided to collapse all the promoters linked to the same gene in a single local regulatory region; therefore, the TBA values computed on the alternative promoters of the same gene and referring to the

same PWM were summed. In addition, TBA values for the two alleles were always summed. In this way, for each individual we obtained one TBA value for each PWM on all the local and distal regulatory regions.

2.3.3 PC regression models to predict gene expression data

To study correlation between the TBA values of regulatory regions and the expression of their target genes across individuals, we used principal component regression as implemented in the CRAN package `pls` [138]. We chose this approach to avoid overdetermination due to the large number of independent variables (640 TBA values for different PWMs) compared to dependent variables (344 individuals). For each gene, a linear model was fitted independently for each linked regulatory region, selecting in each case the number of principal components (PC) of the log2 transformed TBA values that explained 95% of the total variance in TBA and using those PCs as independent variables and the gene expression e as the dependent variable:

$$e_i = \beta_0 + \sum_{j=1}^{N} \beta_j \times pcTBA_j^i + \sum_{l=1}^{3} \alpha_l \times cov_l + \epsilon_i \tag{2.2}$$

where e_i is the expression for the i_{th} individual in analysis, $pcTBA_j^i$ the PC for the i_{th} individual and for the j_{th} PC, β are the fitted coefficients, β_0 the intercept and ϵ the errors. N is the number of PCs needed to explain 95% of the variance of the log2 transformed TBA. *cov* are the first three principal components of genotypes added as covariates to correct for possible population stratification or sequencing biases, α are their fitted coefficients. We chose as covariates the first three Eigenstrat [139] principal component calculated on genotype calls, as in [82].

This is a strictly unsupervised procedure, as the dependent variable e is not used in selecting the PCs that end up in the model. It allows us to dramatically reduce the number of independent variables, since the average number of chosen PCs is 3.75 (min: 1, max: 24).

Other ways of overcoming the overfitting problem, such Lasso [140] and Ridge regression [141], are often used. We chose PC regression because it allows a rigorous evaluation of the statistical significance of the model, as detailed below. This is of course a crucial requirement for us when we compare our model to other ways of detecting eQTLs. To associate a measure of statistical significance to the regression models, while correcting for possible biases of the genotype values we fitted two nested linear models and compared them via an F test. The inner model uses as independent variables only the covariates defined above and the outer one is the one described in 2.2 that adds the TBA principal components. Thus, the F test identifies the genes for which the TBA is able to effectively predict expression of different individuals after correcting for differences in allele frequencies due to population stratification or sequencing biases.

The nominal P-values of these F-tests are not completely reliable due to possible violations in the underlying assumptions: therefore, for every model we computed the F-test P-values of models obtained with 1000 permutations of the expression values to obtain an empirical P-value. With this procedure, we

obtained a P-value for at least one model for 22,125 genes: for only 4 genes the model could not be fitted on any associated regulatory region due to insufficient TBA variance among individuals.

2.3.4 Analysis of allele-specific expression

Allele-specific expression data were obtained from the Supplementary Material of Ref. [82]. First of all, aseSNPs were attributed to the genes in which they lie using the gene annotation provided by Gencode V12. Then, only aseSNPs with at least 30 overlapping reads were selected to avoid potential bias in the ASE detection due to the low number of overlapping reads and only those with $P < 0.005$ were considered significant. Furthermore, only EUR and phased individuals were taken into account. A gene was considered subjected to ASE in a single individual when at least one of the associated aseSNP was significant. In this way for each gene, we obtained the lists of ASE and non-ASE subjects. Only genes with at least 10 individuals in each class were retained for the subsequent analysis (2,740 genes): for each of them, the Euclidean distance between the TBA profile of the two alleles was computed in all the individuals and then the median distance was obtained independently for the ASE and non-ASE groups.

2.3.5 Multivariate eQTL Models

To test the effect of considering many variants together in eQTL analyses, we fitted multivariate eQTL models using the same genetic information as in the TBA models. We computed the principal component regression models with the same procedure used for TBA but used as input for the PC analysis the genotype state of all the variants used to obtain the individual sequences of the regulatory regions (as described above).

Also in this case for each gene a linear model was done independently for each linked regulatory region:

$$e_i = \beta_0 + \sum_{j=1}^{N} \beta_j \times pcMultiSNP_j^i + \sum_{l=1}^{3} \alpha_l \times cov_l + \epsilon_i \tag{2.3}$$

where $pcMultiSNP_j^i$ is the j_{th} PC for the i_{th} individual and N is the number of principal components needed to explain 95% of the genotype calls variance.

The number of variants in the regulatory regions associated with genes ranges from 1 to 539, while N ranges from 1 to 68 with average 5.5. The significance of these models was evaluated with the same permutation-based procedure used for the TBA model.

2.3.6 Relevant transcription factors

To identify the transcription factors most strongly predictive of the gene expression we used linear models in which the regressor is the TBA for a single PWM, while correcting for the same covariates used in the PC regression.

$$e_i = \beta_0 + \beta \times \log a_i + \sum_{l=1}^{3} \alpha_l \times cov_l + \epsilon_i \tag{2.4}$$

For each gene these linear models were done independently for each significant regulatory region. The t-test p-values were corrected using the Benjamini-Hochberg procedure to account for multiple testing and PWMs with a significant ($P < 0.05$) coefficient were associated to the corresponding gene. An average of 310 PWMs were significantly correlated to the expression level of the target genes on each regulatory region.

2.3.7 Prediction of bQTLs

To establish if the results of univariate TBA models are consistent with bQTL data, for each TF studied in [100] we fitted a logistic model that predicts if a significant regulatory region contain bQTLs for that TF using as predictors the result of the univariate TBA model and the length of the regulatory region.

$$t = \beta_0 + \beta_1 \times univTBA_{i,j} + \beta_2 \times length_j \tag{2.5}$$

$$Pr(bQTL_{i,j}) = \frac{1}{1 + \exp^{-t}} \tag{2.6}$$

$Pr(bQTL_{i,j})$ is the fitted probability that the regulatory region j contains bQTLs for TF_i, $univTBA_{i,j}$ is a binary variable corresponding to the outcome of the univariate TBA model done for TF_i on the regulatory region j and $length_j$ is the length of the regulatory region j. The regulatory region's length was included to take into account that the likelihood that a regulatory region contains bQTLs increases as its length raises. The t-test p-values were corrected using the Benjamini-Hochberg procedure to account for multiple testing and those TFs with significant ($P < 0.05$) coefficients were considered predictive of the presence of bQTLs in significant regulatory regions. The following TF-PWM pairs were used:

- JUND/JUND_HUMAN.H10MO.A
- NF-κB/NFKB1_HUMAN.H10MO.B
- PU.1/SPI1_HUMAN.H10MO.A
- POU2F1/PO2F1_HUMAN.H10MO.B
- STAT1/STAT1_HUMAN.H10MO.A

2.3.8 bQTLs that are GWAS hits

To account for LD when considering matches between GWAS hits, bQTLs and uni-eQTLs we downloaded linkage data for the CEU population from HapMap ([63], downloaded 05/31/2016). We define two SNPs to be in LD when they have an $R^2 > 0.8$. GWAS results were downloaded from the GRASP catalog [118], downloaded 07/15/2016) – the catalog was filtered removing results for gene expression, methylation or differential splicing and keeping only SNPs with a nominal P-value $< 5 \times 10^{-8}$, following [100]. The significance of the overlap between GWAS hits and significant regulatory regions for the TBA model was determined by computing an empirical P-value based on 5000 randomized sets of regulatory regions.

To establish if the TBA model is predictive of the presence of bQTLs associated with GWAS hits within regulatory regions and if it gives additional information than those given by the uni-eQTL model, the following logistic model was fitted:

$$t = \beta_0 + \beta_1 \times TBA_j + \beta_2 \times uni - eQTL_j + \beta_3 \times length_j \tag{2.7}$$

$$Pr(bQTL - GWAS_j) = \frac{1}{1 + \exp^{-t}} \tag{2.8}$$

where TBA_j and $uni - eQTL_j$ are binary variables indicating if the regulatory region j is significant according to the TBA model or the uni-eQTL model (limited to putatively causal uni-eQTLs, namely those with the best P in the analysis of [82]), $length_j$ is the length of the regulatory region j and $Pr(bQTL - GWAS_j)$ is the fitted probability that the regulatory region j includes bQTLs associated with GWAS hits.

Chapter 3

The length of the expressed 3′ UTR is an intermediate molecular phenotype linking genetic variants to complex diseases

While the previous chapter was dedicated to the investigation of the genetic determinants of gene expression variation, here I will talk about transcript structure, focusing on alternative polyadenylation. I will indeed present a new computational strategy to discover genetic variants that specifically affect the relative expression of alternative 3′ untranslated region (UTR) isoforms, providing also an extensive analysis of the possible mechanisms of action of the pinpointed variants. Notably, our results point to an important role for genetically determined alternative polyadenylation in affecting predisposition to complex diseases, thus suggesting new ways to extract functional information from GWAS data.

3.1 Results

3.1.1 Genetic variants affect the relative expression of alternative 3′ UTR isoforms of thousands of genes

In order to investigate the effect of human genetic variants on the expression of alternative 3′ UTR isoforms, we developed a computational approach similar to the one commonly used for eQTL analysis (Figure 3.1). It was applied to a large dataset in which WGS data paired with RNA-Seq data are available for 373 European (EUR) individuals (GEUVADIS dataset [82]). A collection of known alternative poly(A) sites [58] was used, together with a compendium of human transcripts, to obtain an annotation of alternative 3′ UTR isoforms that

The content of this chapter was published as Ref. [142]

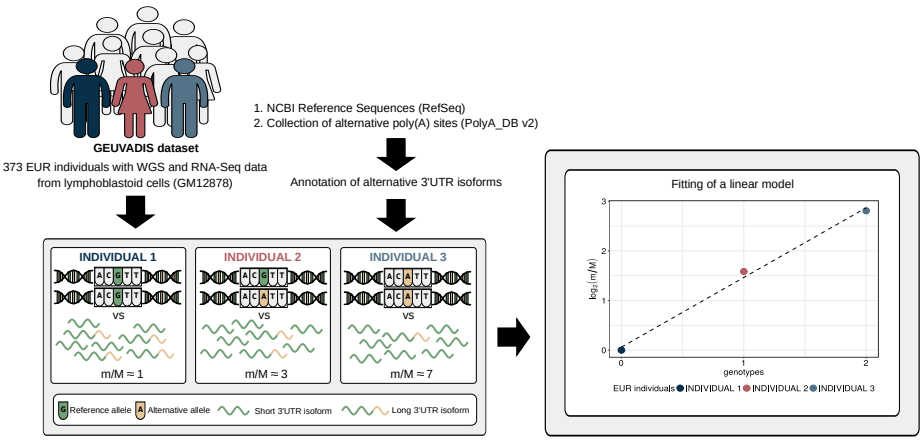


Figure 3.1 – Schematic representation of the method. Genotypic data paired with RNA-Seq data from a large cohort of individuals are required to perform apaQTL mapping analysis. RNA-Seq data are exploited, together with an annotation of alternative 3′ UTR isoforms, to compute for each gene the m/M value that is proportional to the ratio between the expression of its short and long 3′ UTR isoforms. Then, the association between the m/M values of a gene and each nearby genetic variant is evaluated by linear regression. Genotypes are defined in the standard way: 0 means homozygous for the reference allele, 1 means heterozygous and 2 indicates the presence of two copies of the alternative allele.

	Total	Significant
Models	30,136,480	192,715
Genes	6,256	2,530
Variants	5,309,860	160,223

Table 3.1 – Results of apaQTL mapping analysis

was then combined with RNA-Seq data in order to compute, for each gene, the expression ratio between short and long isoform (m/M value) in each individual.

Linear regression was then used to identify associations between the m/M values of each gene and the genetic variants within a cis-window including the gene itself and all sequence located within 1Mbp from the transcription start site (TSS) or the transcription end site (TES). This led to the fitting of ∼30 million linear models, involving ∼6,300 genes and ∼5.3 million variants. About 190,000 models, involving 2,530 genes and ∼160,000 variants, revealed a significant association (Figure 3.2 and Table 3.1).

Our set of significant genes shows only moderate overlap with genes for which eQTLs or transcript ratio QTLs (trQTLs) were reported in Ref. [82] from the same data (Figure 3.2). Alternative polyadenylation can result in changes in gene expression levels as a consequence of the isoform-dependent availability of regulatory elements affecting the stability of transcripts, such as microRNA binding sites [143]. In this case, apaQTLs should also be eQTLs. However, APA may also have effects that do not imply changes in expression levels, including the modulation of mRNA translation rates [144, 145] and localization [146],

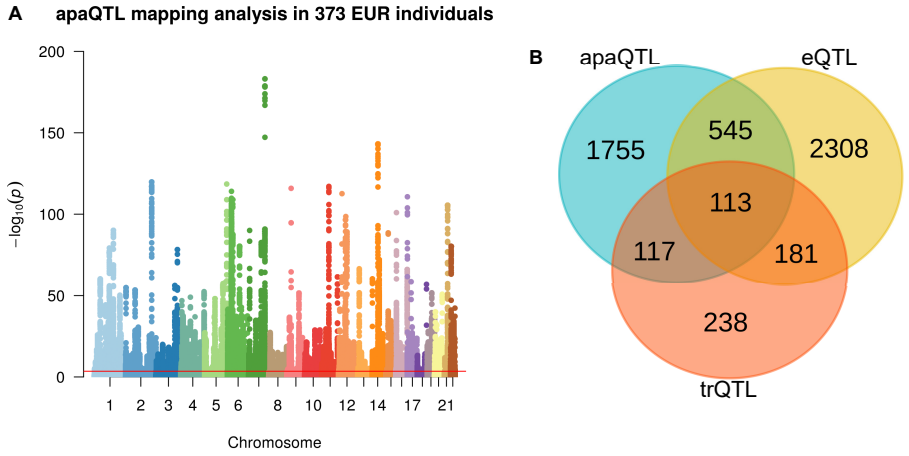


Figure 3.2 – (A) Manhattan plot illustrating the results of the apaQTL mapping analysis. For each fitted model, the -$\log_{10}$ nominal P-value is shown according to the position of the tested genetic variant. The red line indicates the threshold for genome-wide statistical significance, after multiple-testing correction (nominal P-value $< 3.1 \times 10^{-4}$, corresponding to corrected empirical P-value < 0.05). (B) Venn diagram showing the overlap between genes with significant alternative polyadenylation QTL (apaQTL), expression QTL (eQTL) and transcript ratio QTL (trQTL).

and protein cytoplasmic localization [43]. Similarly, a complete overlap with trQTLs is not expected, because they were identified by taking into account all the annotated alternative transcripts of a gene including alternative splicing and transcription initiation. The identification of apaQTLs for several genes for which trQTLs were not identified suggests that focusing on a specific class of transcript structure allows higher sensitivity.

These results show that a large number of genetic determinants of alternative polyadenylation can be inferred from the analysis of standard RNA-Seq data paired with the genotypic characterization on the same individuals.

3.1.2 apaQTLs are preferentially located within active genomic regions

Just like eQTLs, we expect apaQTLs be located within genomic regions that are active in the relevant cell type (lymphoblastoid cells for our data). In order to verify this hypothesis, we superimposed the apaQTLs to the ChromHMM annotation of the human genome for the GM12878 cell line [147], and used logistic regression, as detailed in the Methods, to determine the enrichment or depletion of apaQTLs for each chromatin state, expressed as an odds ratio (OR). As expected, significant ORs > 1 were obtained for active genomic regions, such as transcribed regions, promoters and enhancers, suggesting that genetic variants have a higher probability of being apaQTLs when they are located in active regions. Conversely, apaQTLs were depleted in repressed and inactive chromatin states. Similar results were obtained using broad chromatin states (Figure 3.3), defined following [147], or all 15 chromatin states reported by

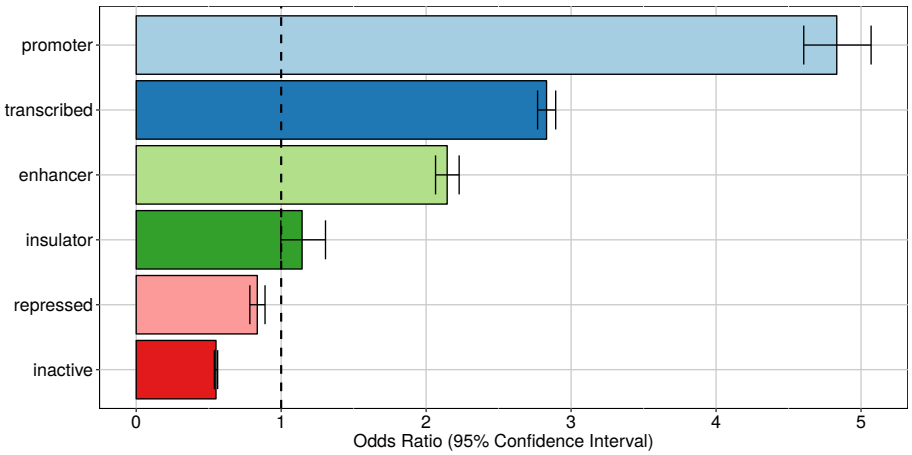


Figure 3.3 – Enrichment of apaQTLs within active genomic regions in the GM12878 cell line. For each broad state, that was defined starting from the ChromHMM annotation, the OR obtained by logistic regression and its 95% CI are shown.

ChromHMM (Supplementary Figure B.1).

As a control, the same enrichment analysis was performed with the chromatin annotation obtained in a different cell type, namely normal human epithelial keratinocytes (NHEK). All NHEK active chromatin states showed a reduced enrichment in apaQTLs compared with GM1278, and regions repressed in NHEK cells actually showed significant enrichment of lymphoblastoid apaQTLs (Supplementary Figure B.2 and Supplementary Figure B.3). Taken together, these results show that genetic variants affecting alternative polyadenylation tend to be located in cell-type specific active chromatin regions.

The detection of a significant apaQTL enrichment within promoters and enhancers suggests that also these genomic regions may be involved in the APA regulation, in agreement with the similar enrichment found, generically for trQTLs, in [82]. However, these results could also be explained, in principle, by linkage disequilibrium between promoters or enhancers and 3′ UTR regions. To evaluate the prevalence of this phenomenon, we observed that among 2113 (3192) significant genetic variants surviving LD pruning (see Methods) inside promoters (enhancers) only 288 (376) are in LD ($R^2 > 0.8$) with significant genetic variants within 3′ UTRs. Furthermore, the reported enrichments remained highly significant after the exclusion of these variants, supporting the idea that promoters and enhancers have an independent role in the genetic component of APA regulation.

In the following, we will divide apaQTLs in two classes: intragenic apaQTLs are those located inside one of the genes whose isoform ratio we are able to analyse, while all other apaQTLs will be referred to as extragenic (note that these might be located inside a gene for which we are unable to perform the analysis, for one of the reasons explained in the Methods).

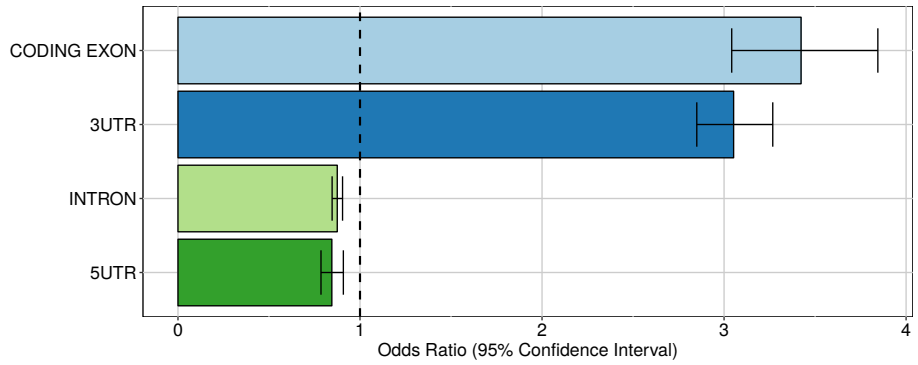


Figure 3.4 – Enrichment of intragenic apaQTLs within coding and non-coding transcript regions. For each gene region, the OR obtained by logistic regression its 95% CI are shown.

3.1.3 Intragenic apaQTLs are enriched in coding exons and 3′ UTRs

Having established that genetic variants have a widespread influence of the expression of alternative 3′ UTR isoforms, we turned to their putative mechanisms of action. First of all, we considered the distribution of intragenic apaQTLs among regions contributing to the mRNA vs. introns. As shown in Figure 3.4, intragenic apaQTLs are enriched in coding exons and 3′ UTRs, and depleted in introns and 5′ UTRs. The depletion of introns suggests that most intragenic apaQTLs exert their regulatory role at the transcript level, e.g. by modulating the binding of trans-acting factors to the mRNA.

Among mRNA regions, the enrichment of 3′ UTRs is expected, since these regions contain several elements involved in the regulation of both alternative polyadenylation and mRNA stability. The enrichment of coding exons could be ascribed to regulatory elements residing in these portions of the mRNAs, or to residual effects of linkage disequilibrium (LD) with variants located in the 3′ UTR, notwithstanding the LD pruning procedure implemented in the enrichment analysis (see Methods). Note that while several poly(A) sites are located upstream of the last exon [148], within both intronic sequences and internal exons, such sites were not taken into account in our analysis. Finally, the depletion of 5′ UTRs might be due to the distance of these elements from the polyadenylation loci, and to the fact that these regions are mostly involved in other regulatory mechanisms, such as translational regulation [149]. In the following, we examine in more detail three possible mechanisms by which intragenic apaQTLs could exert their action.

Creation and destruction of PAS motifs

The first possibility is direct interference with the APA regulation, favoring the production of one of the two isoforms in individuals with a particular genotype. A comprehensive atlas of high-confidence PAS has been recently reported [150]. In addition to the canonical PAS motifs (AAUAAA and AUUAAA) it contains 10 previously known signals and 6 new motifs. Exploiting this resource, we

were able to identify SNPs that cause the creation or the destruction of putative functional PAS motifs and, as expected, we found that they were enriched among apaQTLs (OR = 1.72, 95% confidence interval (CI) = 1.08 - 2.75, P-value = 0.0216). In total, 42 PAS-altering variants were found to be apaQTLs of the gene in which they reside. While expected, this result can be considered to validate our strategy.

A few examples are worth discussing in detail. SNP rs10954213 was shown by several studies [52,107,151] to determine the preferential production of the short isoform of the *IRF5* transcription factor through the conversion of an alternative PAS motif (AAUGAA) into the canonical one (AAUAAA) in a proximal position within the 3′ UTR. Consistently, we found that this variant is associated with higher prevalence of the short isoform (Figure 3.5). Moreover, the same variant was associated to higher risk of systemic lupus erythematosus (SLE), and higher *IRF5* expression, that could be due to the loss of AU-rich elements (ARE) in the short transcript isoform [107]. Globally, these findings are in agreement with the known involvement of *IRF5* in several pathways that are critical for the onset of SLE (Type I IFN production, M1 macrophage polarization, autoantibody production, and induction of apoptosis [152]).

A similar trend was detected in the case of the rs9332 variant, located within the 3′ UTR of the *MTRR* gene, encoding an enzyme essential for methionine synthesis (Figure 3.6). This variant was reported to be associated with a higher risk of spina bifida, along with other variants within the same gene [154]. We found that the variant is associated with the increased relative expression of the short isoform of the *MTRR* transcript, as a consequence of the creation of a proximal canonical PAS. We can thus speculate that, similarly to what was shown for *IRF5*, this post-transcriptional event could lead to a variation in the activity of the enzyme activity and ultimately to increased disease susceptibility.

The same mechanism might provide putative mechanistic explanations for associations found by GWAS studies. For example we found the variant rs5855 to be an apaQTL for the *PAM* gene (Supplementary Figure B.4), essential in the biosynthesis of peptide hormones and neurotransmitters [155–157]. No eQTLs or trQTLs for this gene were revealed by the analysis of the same data reported in [82]. This variant replaces an alternative PAS motif (AGUAAA) with the canonical AAUAAA, thus presumably increasing its strength. This PAS motif is located 26 bps upstream of an APA site corresponding to a 3′ UTR of ∼450 bps, instead of the ∼2,000 bps of the canonical isoform, lacking several predicted microRNA binding sites. Indeed, our analysis revealed a shortening of the 3′ UTR in individuals with the alternate allele, i.e. the canonical PAS motif. Notably, the variant is in strong LD ($R^2 = 0.90$) with the intronic variant rs10463554, itself an apaQTL for *PAM*, which has been associated to Parkinson's disease in a recent meta-analysis of GWAS studies [158].

Conversely, the destruction of a canonical, proximal PAS motif leads to shortening of the 3′ UTR of *BLOC1S2* (Supplementary Figure B.4). The variant rs41290536 replaces the canonical PAS motif AAUAAA with the non-canonical one AAUGAA 17 bps upstream of a poly(A) site corresponding to a UTR length of ∼750 bps compared to the ∼2,200 of the longest isoform. The variant is in complete LD ($R^2 = 1$) with two variants that have been associated to predisposition to squamous cell lung carcinoma (rs28372851 and rs12765052) [159].

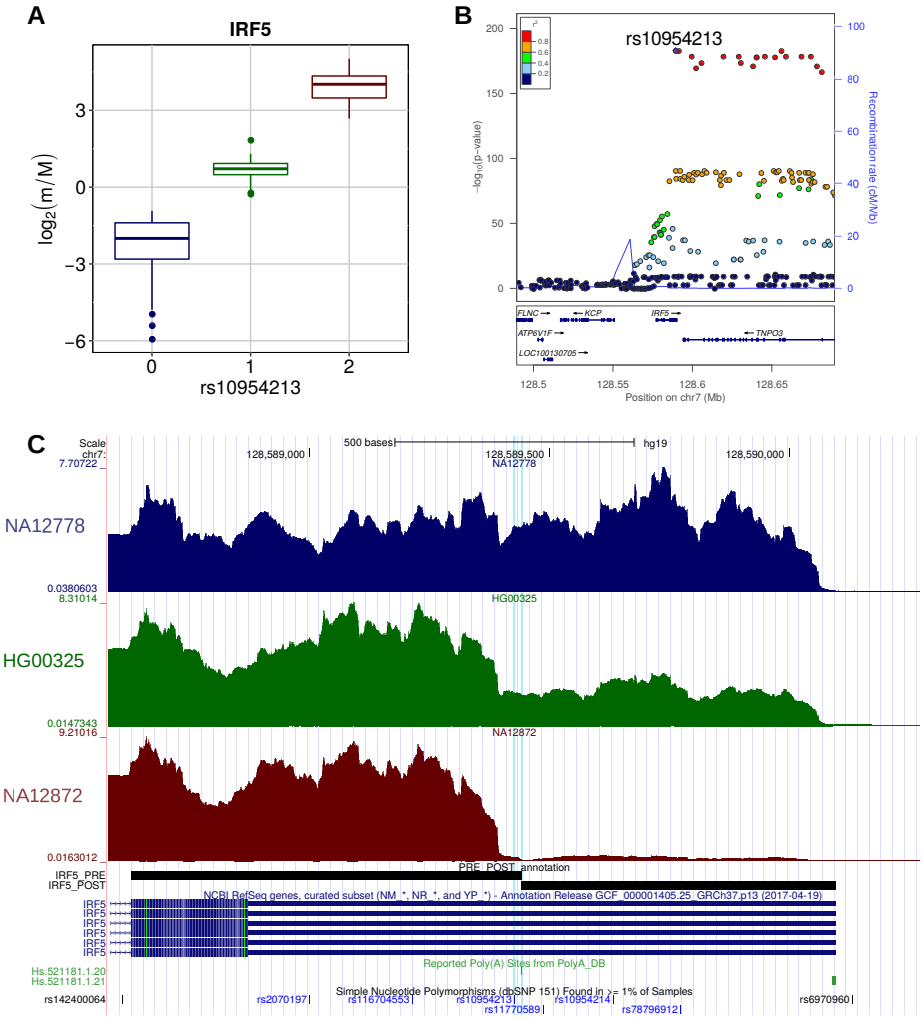


Figure 3.5 – (A) Boxplot showing the variation of the log2-transformed m/M values obtained for *IRF5* as a function of the genotype of the individuals for rs10954213. (B) LocusZoom plot [153] illustrating the results obtained for *IRF5* in the genomic region around rs10954213 (100kb both upstream and downstream its genomic location). In the top panel each tested genetic variant was reported as a function of both its genomic coordinate and its association level with *IRF5* (log10-transformed nominal P-value); the points color reflects the LD level (R^2) between rs10954213 and each of the other genetic variants in the locus. The bottom panel shows the genes and their orientation in the locus. (C) Figure adapted from the UCSC Genome Browser screenshot. RNA-Seq tracks, reporting coverage per million mapped reads, are shown for three representative individuals: NA12778 (homozygous for the reference allele), HG00325 (heterozygous) and NA12872 (homozygous for the alternative allele). *IRF5* RefSeq, *IRF5* PRE/POST segments, poly(A) sites and common SNPs are shown. The rs10954213 variant and the affected poly(A) site (Hs.521181.1.20) are highlighted.

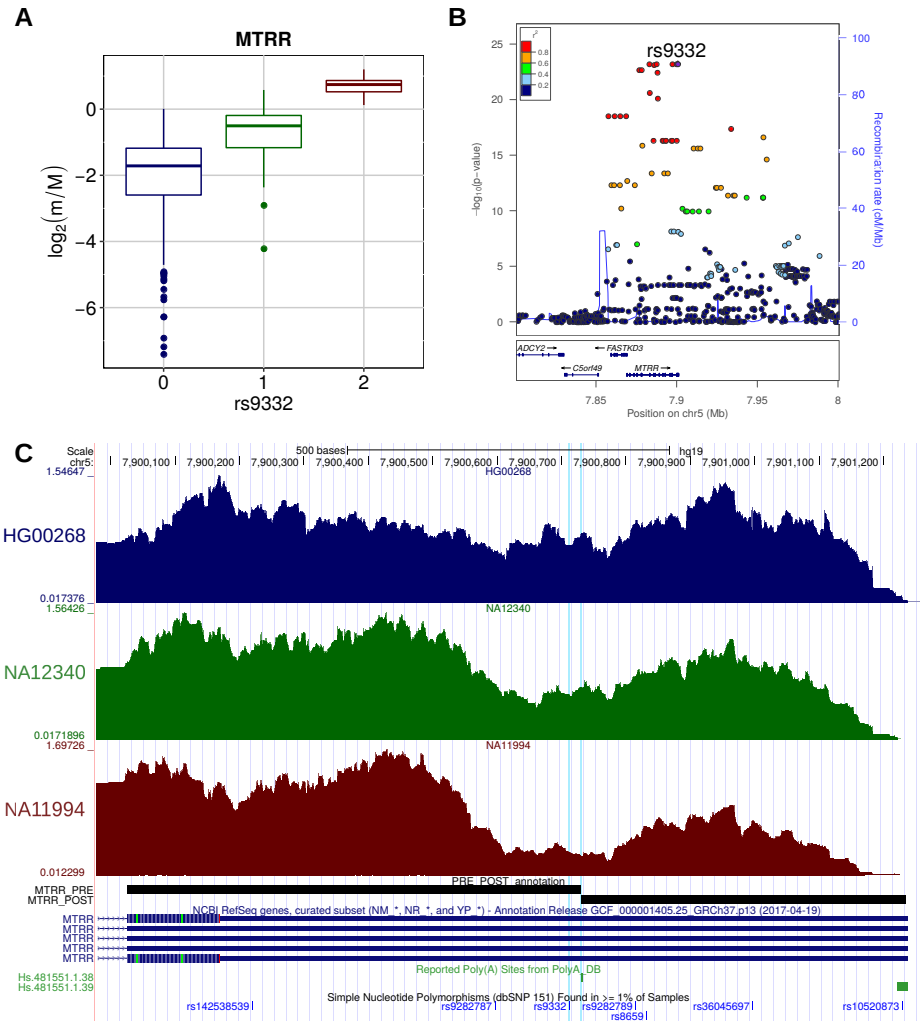


Figure 3.6 – (A) Boxplot showing the variation of the log2-transformed m/M values obtained for *MTRR* as a function of the genotype of the individuals for rs9332. (B) LocusZoom plot illustrating the results obtained for *MTRR* in the genomic region around rs9332 (100kb both upstream and downstream its genomic location). (C) Figure adapted from the UCSC Genome Browser screenshot. RNA-Seq tracks, reporting coverage per million mapped reads, are shown for three representative individuals: HG00268 (homozygous for the reference allele), NA12340 (heterozygous) and NA11994 (homozygous for the alternative allele). *MTRR* RefSeq, *MTRR* PRE/POST segments, poly(A) sites and common SNPs are shown. The rs9332 variant and the affected poly(A) site (Hs.481551.1.38) are highlighted.

Alteration of microRNA binding

In an alternative scenario, genetic variants can influence the relative expression of alternative 3′ UTR isoforms by acting on the stability of transcripts, for example through the creation or destruction of microRNA binding sites. For each gene with alternative 3′ UTR isoforms, we divided the 3′ UTR into two segments: the "PRE" segment, common to both isoforms, and the "POST" segment, contained only in the longer isoform. Variants altering microRNA binding sites located in the POST segment can result in the variation of the relative isoform expression since they affect only the expression of the long isoform.

For example, we found that the rs8984 variant is associated with an increased prevalence of the long transcript isoform of the *CHURC1* gene, an effect that could be due to the destruction of a binding site recognized by microRNAs of the miR-582-5p family within the POST segment of the gene (Supplementary Figure B.5). More generally, we found that apaQTLs are enriched, albeit slightly, among the genetic variants that create or break putative functional microRNA binding sites (OR = 1.15, 95% CI = 1.02 - 1.30, P-value = 0.022). However, we could not find significant agreement between the predicted and actual direction of the change in isoform ratios for these cases. Together with the marginal significance of the enrichment, this result suggests that the alteration of microRNA binding sites is not among the most relevant mechanisms in the genetic determination of 3′ UTR isoform ratios.

Alteration of RNA-protein binding

RNA-binding proteins (RBPs) play important roles in the regulation of the whole cascade of RNA processing, including co- and post-transcriptional events. Although many of them have not been fully characterized yet, a collection of 193 positional weight matrices (PWMs) describing a large number of RNA motifs recognized by human RBPs has been obtained through in-vitro experiments [160]. Here we exploited this resource to identify SNPs that alter putative functional RBP binding sites. Consistently with the involvement of RBPs in the regulation of alternative polyadenylation, mRNA stability and microRNA action, we found a highly significant enrichment of RBP-altering SNPs among intragenic apaQTLs (OR = 1.48, 95% CI = 1.31 - 1.66, P-value = 8.54×10^{-11}).

Specifically, we obtained a positive and significant OR for 20 individual RBP binding motifs (Supplementary Table B.1). Although in most cases the enrichment is modest, some of the enriched motifs correspond to RNA-binding domains found in RBPs with a previously reported role in polyadenylation regulation (members of the muscleblind protein family [57, 161], *KHDRBS1* [162] and *HNRNPC* [150]). Other enriched RNA-binding motifs are associated with splicing factors (*RBM5*, *SRSF2*, *SRSF9* and *RBMX*) and other RBPs that may be involved in RNA processing (such as members of the MEX3 protein family and HNRNPL). On the contrary, only one significant motif is associated with a RBP that may be involved in RNA degradation (*CNOT4* [163]). The involvement of several splicing factors is consistent with evidence supporting a mechanistic interplay between polyadenylation and splicing, that goes beyond the regulation of the usage of intronic poly(A) sites [164–168].

3.1.4 Extragenic apaQTLs act in-cis through the perturbation of regulatory elements

Understanding the function of extragenic apaQTLs is less straightforward because, although there are few examples of DNA regulatory elements contributing to APA regulation [169], it is commonly believed that APA is mainly controlled by cis-elements located within transcripts, both upstream and downstream of the poly(A) sites [143].

To further explore this aspect we took advantage of a different annotation of active genome regions, which includes the association between regulatory regions and target genes, namely the cis-regulatory domains (CRDs) identified in lymphoblastoid cell lines in Ref. [170]. Extragenic apaQTLs were indeed found to be enriched in CRDs (OR = 1.73, 95% CI = 1.69 - 1.78, P-value $< 10^{-16}$). The 3D structure of the genome is a key aspect of gene regulation [171], as it determines physical contacts between distal regulatory regions and proximal promoters. In particular, CRDs have been described as active sub-domains within topologically associated domains (TADs), containing several non-coding regulatory elements, both proximal and distal. The perturbation of those regulatory elements by genetic variants can lead to the alteration of gene expression and perhaps interfere with other processes such as alternative polyadenylation, as suggested by our results. Importantly, CRDs have been assigned to the nearby genes they regulate. We could thus observe that extragenic apaQTLs tend to fall within CRDs that have been associated with their target genes much more frequently than expected by chance. Indeed, this correspondence was verified for 27,527 extragenic apaQTLs, while the same degree of concordance was never obtained in 100 permutations in which each extragenic apaQTL was randomly associated to a gene in its cis-regulatory window (median number of correspondences 12,571). These results suggest an important role of genetic variants located in active, non-transcribed cis-regulatory regions in regulating alternative polyadenylation of the target genes.

3.1.5 A role for apaQTLs in complex diseases

Since common genetic variation is involved in complex diseases, often by affecting gene regulation, a natural question is whether apaQTLs can be used to provide a mechanistic explanation for some of the genetically driven variability of complex traits, thus adding 3′ UTR length to the list of useful intermediate phenotypes. Besides the specific examples discussed above, we found an overall striking enrichment among apaQTLs of genetic variants reported in the NHGRI-EBI GWAS Catalog [172] (OR = 3.17, 95% CI = 3.01 - 3.33, P-value $< 10^{-16}$).

We also investigated the enrichment of each trait category defined by the Experimental Factor Ontology (EFO) and then for each individual trait. In line with the fact that the apaQTL mapping was performed in lymphoblastoid cells, the strongest enrichment was observed for immune system disorders (OR = 5.41, 95% CI = 4.52 - 6.45, P-value = 2.50×10^{-77}) (Figure 3.7 and Supplementary Table B.2). However, a strong enrichment was also detected for almost all the other tested categories, including neurological disorders (OR = 4.32, 95% CI = 3.86 - 4.83, P-value = 2.47×10^{-142}) and cancer (OR = 3.96, 95% CI = 3.36 - 4.64, P-value = 4.15×10^{-63}).

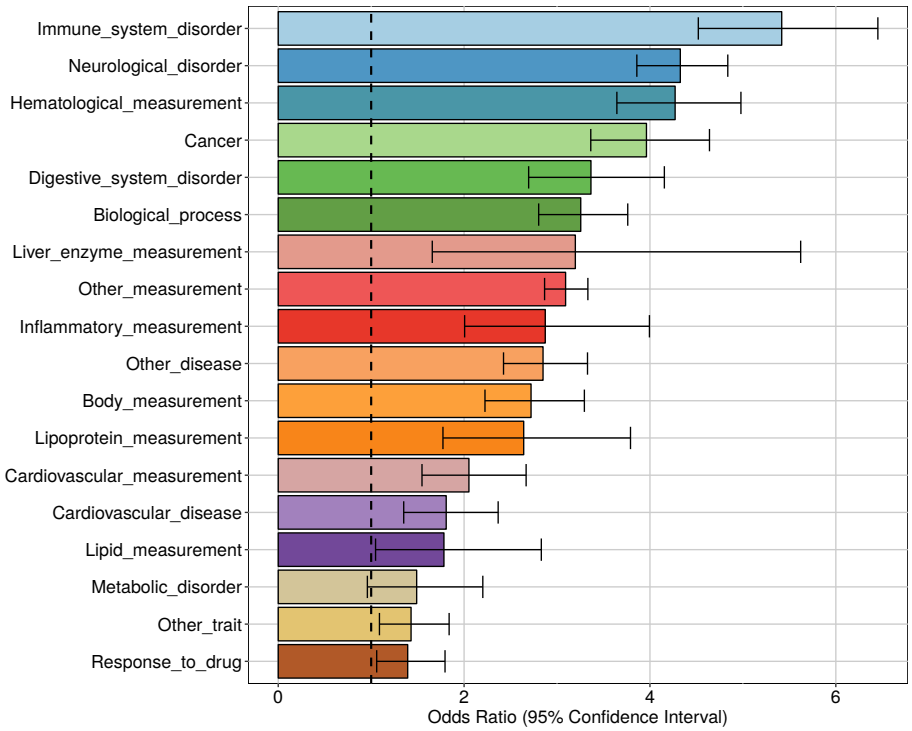


Figure 3.7 – Enrichment of GWAS hits among apaQTLs, for different categories of complex traits. For each category, the OR obtained by logistic regression and its 95% CI are shown.

A significant enrichment was detected for 95 individual complex traits, including several diseases. Among these, the largest ORs were observed for autism spectrum disorder (OR = 42.6, 95% CI = 32.9 - 55.5, P-value = 2.36×10^{-174}), squamous cell lung carcinoma (OR = 26.1, 95% CI = 15.7 - 43.3, P-value = 1.29×10^{-36}), lung carcinoma (OR = 17.9, 95% CI = 12.7 - 25.2, P-value = 9.63×10^{-62}), schizophrenia (OR = 10.6, 95% CI = 9.01 - 12.4, P-value = 1.25×10^{-182}), and HIV-1 infection (OR = 6.51, 95% CI = 3.75 - 10.8, P-value = 2.28×10^{-12}).

We observed that apaQTLs that are also GWAS hits often map to genes in the human leukocyte antigen (HLA) locus, suggesting that in at least some cases the enrichment could be mostly driven by this genomic region. Somewhat unexpectedly, this was particularly evident for neurological disorders. In order to clarify this point, we evaluated all enrichments after excluding the variants in the HLA locus. Although in some cases the OR decreased after removing HLA variants, for most GWAS categories the enrichment was still significant (Supplementary Figure B.6 and Supplementary Table B.3). For example, we found 155 apaQTLs associated with autism spectrum disorder, 116 of which affecting HLA genes. After the exclusion of HLA variants, the enrichment was still highly significant (OR = 10.66, 95% CI = 6.92 - 15.95, P-value = 7.05×10^{-29}). On the contrary, the enrichment of variants associated to pulmonary adenocarcinoma is driven by the HLA locus, and becomes non-significant after

excluding HLA variants (OR = 1.35, 95% CI = 0.22 - 4.39, P-value = 0.68).

3.1.6 The effect of genetic variants on APA can be confirmed in patients

As briefly discussed above, the rs10954213 variant is associated with a higher risk of SLE. Evidence about the related molecular mechanism arose from the analysis of cell lines derived from healthy individuals [52,151], and the effect of the variant on *IRF5* expression in blood cells was confirmed in SLE patients [173, 174]. However, direct evidence on the effect of this variant on APA regulation in SLE patients is still missing.

In order to assess whether rs10954213 affects *IRF5* APA regulation in SLE patients, we analyzed RNA-Seq data derived from whole blood cells in 99 patients [175]. After the exclusion of 52 individuals whose genotype cannot be determined with certainty from RNA-Seq reads, we detected a strong difference in *IRF5* m/M values among the three rs10954213 genotypes, with the alternative allele associated with higher m/M values, i.e. shorter 3′ UTR (Kruskal-Wallis test P-value = 2.49×10^{-8}; Figure 3.8). Therefore the variant has, at least qualitatively, the same effect in the whole blood of SLE patients as in lymphoblastoid cell lines of normal individuals.

3.2 Discussion

We used a new efficient strategy to study how human genetic variants influence the expression of alternative 3′ UTR isoforms. This issue has been previously investigated with different approaches [82, 104, 108, 109, 176]. The method we propose combines wide applicability, being based on standard RNA-Seq data, with the high sensitivity allowed by limiting the analysis to a single type of transcript structure variant, namely 3′ UTR length. Such higher sensitivity led us to discover thousands of variants associated with 3′ UTR length that were not identified in a general analysis of transcript structure from the same data in [82]. Moreover, the significant overlap between our apaQTLs and the eQTLs identified in [82] confirms the known relevant role of 3′ UTRs in gene expression regulation. However, the regulation of 3′ UTR length is known to affect regulatory processes that do not directly alter mRNA abundance, such as regulation of translation efficiency, mRNA localization and membrane protein localization [43,177]. Indeed most of the apaQTLs we found were not identified as eQTLs in [82].

The various mechanisms underlying the association between genetic variants and the relative abundance of 3′ UTR isoforms can be classified in two main classes based on whether they affect the production or degradation rates of the isoforms. The production related-mechanisms include the alteration of APA sites, of cis-regulatory elements located in promoters and enhancers, and of binding sites of RBPs involved in nuclear RNA processing; the degradation-related mechanisms include the alteration of the binding sites of microRNAs and cytoplasmatic RBPs affecting mRNA stability. Taken together, our results suggest that the genetic effects on 3′ UTR isoforms act prevalently at the level of production, as shown by the strong enrichment of apaQTLs in non-transcribed regulatory regions and among the variants creating or disrupting APA sites, and

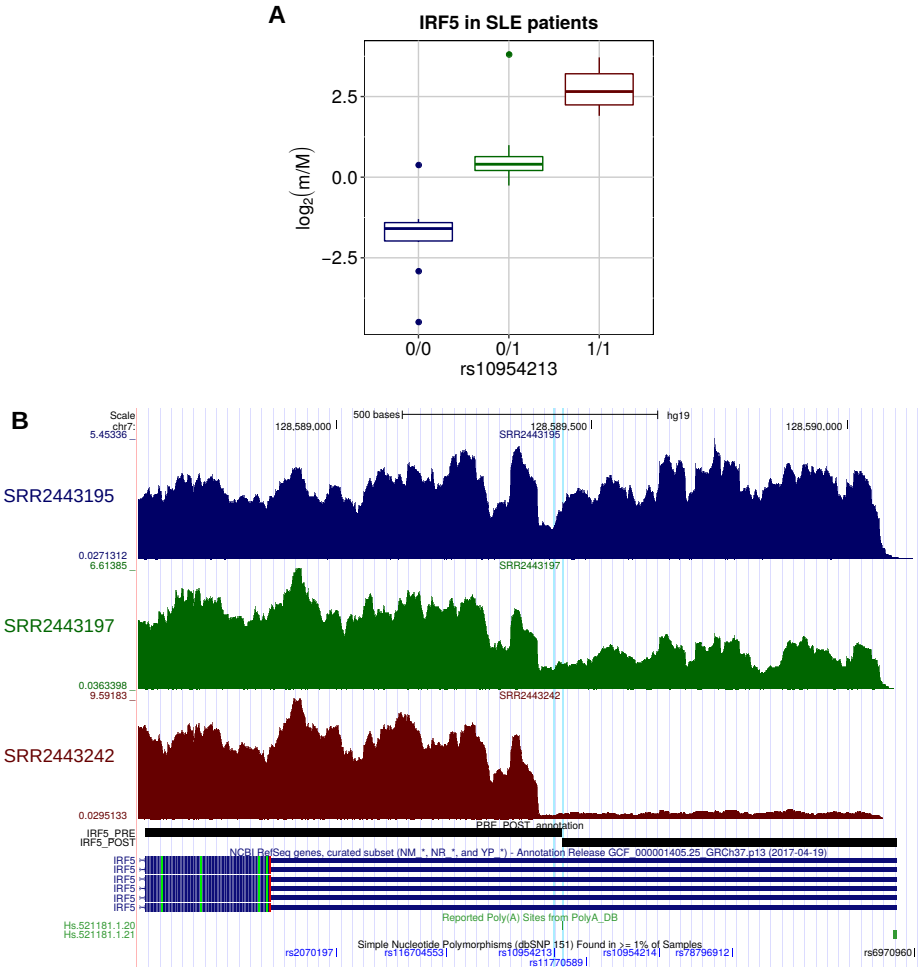


Figure 3.8 – (A) The effect of rs10954213 on the relative expression of the *IRF5* alternative isoforms was investigated also in a small cohort of SLE patients. The boxplot shows the variation of the log2-transformed m/M values obtained for *IRF5* as a function of the genotype of the individuals. (B) Figure adapted from the UCSC Genome Browser screenshot. RNA-Seq tracks, reporting coverage per thousand mapped reads, are shown for three representative individuals: SRR2443195 (homozygous for the reference allele), SRR2443197 (heterozygous) and SRR2443242 (homozygous for the alternative allele). *IRF5* RefSeq, *IRF5* PRE/POST segments, poly(A) sites and common SNPs are shown. The rs10954213 variant and the affected poly(A) site (Hs.521181.1.20) are highlighted.

by the relatively weak enrichment of variants creating or disrupting microRNA binding sites. Also the results on altered RBP binding sites confirm this picture, since most motifs altered by apaQTLs are associated to nuclear RBPs involved in nuclear RNA processing.

In particular, we identified several apaQTLs creating or destroying putative functional PAS motifs. However, it should be noted that our ability to detect these events is intrinsically limited by the motif repertoire that we used [150], which might miss some of the rarest alternative PAS motifs. For example, we found that the rs6151429 variant is associated with the increased expression of the long isoform of the transcript codified by the Arylsulfatase A (*ARSA*) gene (Supplementary Figure B.4), in agreement with previous evidence [178]. However, we did not include this variant among those disrupting a PAS motif since the disrupted motif (AAUAAC) is not included in the catalog that we used. In addition, we considered only PAS-altering single nucleotide substitutions, while also other types of genetic variants can modify the PAS landscape of a gene. For example, a small deletion (rs374039502) causes the appearance of a new PAS motif within the *TNFSF13B* gene, and has been associated with an higher risk of both multiple sclerosis and SLE in the Sardinia population [179].

We observed a strong enrichment of apaQTLs in regulatory regions such as promoters and enhancers, as previously found for variants generically affecting transcript structure in [82]. These results point to an important role of DNA-binding cis-acting factors in the regulation of 3′ UTR length, and to the existence of a widespread coupling between transcription and polyadenylation [177,180]. The mechanisms behind this coupling are thought to include the interaction between rates of Pol II elongation and alternative polyadenylation and the recruitment, by the transcription machinery, of trans-acting factors affecting PAS choice [143]. Moreover, it has been shown that RBPs involved in APA regulation can interact with promoters [169].

Regarding the effect of genetic variants on mRNA stability, we focused on the perturbation of microRNA binding, taking into account both the creation and the destruction of microRNA binding sites within transcripts. The relevance of mRNA stability seemed to be confirmed by a modest enrichment of microRNA-altering SNPs among intragenic apaQTL, however the direction of their effect on microRNA binding is not statistically consistent with the expected direction of the change in 3′ UTR isoform ratio. The same type of ambiguity has been previously reported with regard to the relationship between the effect of SNPs on microRNA binding and gene expression levels [181] and makes us doubt whether these microRNA-altering apaQTLs are truly causal for the associated gene. These results suggest that the alteration of microRNA binding may not be a predominant mechanism explaining the variation of the expression of alternative 3′ UTR isoforms across individuals. Limitations in the accuracy of predicted microRNA binding sites might also contribute to this result.

Another possible mechanism of action of intragenic apaQTLs is the perturbation of the regulatory action of RBPs, as indicated by the modest but highly significant enrichment of SNPs altering RNA-binding motifs. However, the lack of strong enrichments when considering each motif individually suggests that specific RBP motifs may have a small regulatory impact on APA that may also depend on the context, as recently suggested [57]. As in the case of microRNAs, also our limited knowledge of the binding preferences of RBPs might limit our power to detect their effects. More sophisticated models should take

into account the highly modular structure of RBPs that often include multiple RNA binding domains (RBDs), the emerging importance of both the binding context and the RNA structure and even more sophisticated modes of RNA binding [182, 183].

Furthermore, it is reasonable to assume that also non-canonical modes of APA regulation can be affected by genetic variants and therefore drive the detection of variable isoform expression ratios. For example, it has been recently suggested that an epitranscriptomic event, the m^6A mRNA methylation, can be associated with alternative polyadenylation [184]. In addition, recently published results suggest that genetic variants could affect APA regulation also in an indirect way, without affecting the regulatory machinery. Past studies have reported that a narrow range of 10-30nt between the PAS and the poly(A) site is required for efficient processing, however [185] suggested that also greater distances can sometimes be used thanks to RNA folding events that bring the PAS and the poly(A) site closer to each other. Therefore, we can speculate that if a genetic variant affects RNA folding in such a way as to modify the distance between the PAS and the poly(A) site, it could also influence APA regulation.

While the mechanisms discussed above act at the level of the primary or mature transcript, our results revealed a perhaps unexpectedly large number of extragenic apaQTLs, mostly located in regulatory regions. These apaQTLs point to an important role of DNA-binding elements such as transcription factors in regulating alternative polyadenylation through long-distance interactions with cleavage and polyadenylation factors. The investigation of these mechanisms is thus a promising avenue of future research.

Alternative polyadenylation can affect several biological processes, influencing mRNA stability, translation efficiency and mRNA localization [143]. Therefore, it is not surprising that its perturbation has been associated with multiple pathological conditions [89, 186]. In the present study, we detected a strong enrichment of GWAS hits among apaQTLs, supporting the idea that $3'$ UTR length is a useful addition to the list of intermediate molecular phenotypes that can be used for a mechanistic interpretation of GWAS hits. In particular, we identified genetic variants previously associated to neurological disorders, such as autism, schizophrenia and multiple sclerosis, which may act by affecting the regulation of polyadenylation. The importance of post-transcriptional events in the onset of neurological diseases has been recently confirmed by two studies demonstrating that genetic variants affecting alternative splicing (sQTL) give a substantial contribution to the pathogenesis of schizophrenia [187] and Alzheimer's disease [188]. We also observed that the relevant apaQTLs often map to HLA genes, but that the enrichment is not explained by the HLA locus alone. On the other hand, examples of APA events involving HLA genes have been reported [189, 190] and genes encoding antigen-presenting molecules account for the highest fraction of genetic risk for many neurological diseases [191].

We are aware of some limitations of this study. First, the simple model that we used for the definition of alternative $3'$ UTRs isoforms limits the type of events that can be detected, because we can see only events involving poly(A) sites located within the transcript segments taken into account for the computation of the m/M values (the PRE and the POST segments). Nonetheless, the adoption of this simple model significantly reduces the computational burden and might be sufficient to indicate general trends that can be subsequently further investigated with more sophisticated models. Indeed, it has been previ-

ously shown, in a slightly different context (i.e. the comparison of APA events detected in different cellular conditions or tissues), that the results obtained with our model are comparable with those obtained exploiting a more complex model that takes into account all the possible APA isoforms of a gene, especially because also genes with multiple poly(A) sites mainly use only two or a few of them [56]. Second, our strategy depends on a pre-existing annotation of poly(A) sites. Methods that infer the location of poly(A) sites from RNA-Seq data are available, but they can have lower sensitivity in the detection of APA events [56, 57]. In addition, although the method is generally able to successfully discriminate APA events from alternative splicing events, it may give rise to spurious associations when intron retention is present within the $3'$ UTRs and therefore such special cases should be inspected with particular attention. Furthermore, we examined only a single cell type (lymphoblastoid cells) to demonstrate the feasibility of apaQTL mapping analysis. A broader investigation, exploiting data such as those provided by Genotype-Tissue Expression (GTEx) consortium [95], would be particularly valuable. Indeed, APA regulation seems to be significantly tissue-specific and global trends of poly(A) sites selection in specific human tissues have been described: for example transcripts in the nervous system and brain are characterized by preferential usage of distal PAS, whereas in the placenta, ovaries and blood the usage of proximal PAS is preferred [177]. Finally, we analysed only alternative APA isoforms that derive from coding transcripts, because polyadenylation has been classically described in the context of mRNA processing. Nevertheless, recent evidence suggests that also other gene classes might be specifically investigated in future studies. Notably, in a comprehensive mapping of PASs in mammalian genomes, about 60% human PASs were assigned to mRNAs, while about 10% human PASs were attributed to long non-coding RNA (lncRNA) genes [192]. Indeed, although lncRNAs have several unique features, many of them are transcribed by RNA polymerase II (Pol II) and therefore are 5'-capped, spliced and polyadenylated [193].

In conclusion, we have identified thousands of common genetic variants associated with alternative polyadenylation in a population of healthy human subjects. Furthermore, our results suggest that alternative polyadenylation is a promising intermediate molecular phenotype for the mechanistic interpretation of genetic variants associated to phenotypic traits and diseases. Therefore, we are now working to integrate apaQTLs with GWAS summary statistics, thus aiming to directly identify genes whose contribution to diseases is mediated by APA dysregulation (for further details see Chapter 4).

3.3 Material and Methods

3.3.1 Data sources

Human genome and transcriptome

The coordinates of the NCBI Reference Sequences (RefSeqs) in the human genome (hg19) were downloaded from the UCSC Genome Browser (09/04/2015) [194, 195]. The corresponding transcript-gene map was downloaded from NCBI (version 69) and the Bioconductor R package `org.Hs.eg.db v3.4.0` [196] was used to associate each Entrez Gene Id to its gene symbol.

In addition, the reference sequence of the hg19 version of the human genome was downloaded from the ENSEMBL database and a collection of poly(A) sites was obtained from PolyA_DB2 (10/02/2014) [58].

ChromHMM annotations [147] were downloaded from the UCSC Genome Browser for the GM12878 and the NHEK cell lines (`http://genome-euro.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHmm`).

Finally, the coordinates of Cis Regulatory Domains (CRDs) and their association with genes were downloaded for lymphoblastoid cells from `ftp://jungle.unige.ch/SGX/` [170].

WGS and RNA-Seq data

We exploited the RNA-Seq data obtained by the GEUVADIS consortium in lymphoblastoid cell lines of 462 individuals belonging to different populations, but we considered only 373 individuals with European ancestry (EUR). BAM files were downloaded from the E-GEUV-1 dataset [82] in the EBI ArrayExpress archive (`https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/`). We also downloaded genotypic data for the same individuals and the results of the eQTL/trQTL mapping analyses. The downloaded VCF files include genotypes for 465 individuals: among the 462 of them for which also RNA-Seq data are available, the large majority had been previously subjected to Whole Genome Sequencing (WGS) by the 1000 Genomes Project (Phase 1) [65], but the GEUVADIS consortium additionally obtained genomic data for 41 of them through genotyping with Single Nucleotide Polymorphism (SNP) array followed by genotype imputation [82]. Furthermore, whole blood RNA-Seq data for 99 individuals affected by SLE were downloaded from the NCBI SRA database (SRP062966) [175, 197].

Regulatory motifs and related expression data

Different collections of regulatory motifs were downloaded. A list of 18 PAS motifs was obtained from [150], microRNA seeds were downloaded from TargetScan 7.2 [198] and Positional Weight Matrices (PWMs) describing the binding specificities of RNA-binding proteins were downloaded from the CISBP-RNA dataset [160], including both the experimentally determined motifs and those that were inferred from related proteins. In addition, the list of microRNAs and RBPs expressed in lymphoblastoid cells were obtained from the expression data available in the E-GEUV-2 and E-GEUV-1 datasets on the EBI ArrayExpress archive (`https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-2/`).

GWAS Catalog

A collection of genomic loci associated with human complex traits was obtained by downloading the NHGRI-EBI GWAS Catalog, v1.0.2 [172]. This resource is continuously updated: the version we used was downloaded on October 10^{th}, 2018 and it was mapped to GRCh38.p12 and dbSNP Build 151. From the same website, we also downloaded a file showing the mapping of all the reported traits to the Experimental Factor Ontology (EFO) terms [199], including the parent category of each trait (the version of the downloaded file was r2018-09-30). In addition, the dbSNP Build 151 [200] collection of human genetic variants was downloaded for hg19.

3.3.2 Annotation of alternative 3′ UTR isoforms

We considered the human mRNAs included in RefSeq and associated them with the corresponding Entrez Gene Id. Moreover, we collapsed together the structures of all the transcripts assigned to a gene, using the union of all the exons of the various transcripts associated to a gene and defining the 3′ or 5′ UTR using respectively the most distal coding end and the most proximal coding start.

The coordinates of the human poly(A) sites were converted from hg17 to hg19 using `liftOver` [201] and then combined with the gene structures defined above to define the alternative 3′ UTR isoforms. For the definition of alternative 3′ UTR isoforms we adopted a simple model taking into account only two alternative poly(A) sites for each gene, because previous evidence suggests that also genes with multiple poly(A) sites mainly use only two of them [56]. In particular, for each gene we selected the most proximal poly(A) site among those falling within exons, preferring those located within the 3′ UTR, and the end of the gene as the distal poly(A) site. In this way we were able to define two segments of interest for each gene: the PRE segment, extending from the beginning of the last exon to the proximal poly(A) site, and the POST segment, from the proximal poly(A) site to the end of the gene. The PRE fragment is assumed to be contained into both the long and the short isoform, while the POST segment should be contained exclusively into the long isoform.

The relative prevalence of the short and long isoforms are evaluated, as described below, based on the number of RNA-Seq reads falling into the PRE and POST regions. While the whole region from the transcription start site to the proximal poly(A) site could be taken, in principle, as the PRE region, we chose to limit it to the last exon to minimize the confounding effect of alternative splicing.

3.3.3 Computation of m/M values

Using the Bioconductor R package `Roar` [56], for each gene with alternative 3′ UTR isoforms we obtained an m/M value in each individual. The m/M value estimates the ratio between the expression of the short and the long isoform of a gene in a particular condition and the $m/M_{\mathrm{a,i}}$ of gene a in the i_{th} individual is defined as

$$m/M_{a,i} = \frac{l_{POST_a} \times \#r_{PRE_{a,i}}}{l_{PRE_a} \times \#r_{POST_{a,i}}} - 1 \tag{3.1}$$

where l_{PRE_a} and l_{POST_a} are respectively the length of the PRE and POST segment of the gene a, $\#r_{PRE_{a,i}}$ and $\#r_{POST_{a,i}}$ are respectively the number of reads mapped on the PRE and the POST segment of the gene a in the i_{th} individual.

The m/M values were computed for 14,542 genes for which we were able to define alternative 3′ UTR isoforms. Infinite and negative values of m/M (that happen when the POST region does not produce any reads, and when the POST region produces more reads than the PRE region after length normalization, respectively) were considered as missing values. Then only those on autosomal chromosomes (chr1-22) and with less than 100 missing m/M values were selected for the following investigation, leaving us with 6,256 genes.

3.3.4 Genotypic data pre-processing

Starting from the downloaded VCF files, we extracted genotypic data for 373 EUR individuals for whom also RNA-Seq data are available using `VCFtools` [202]. In addition, only common genetic variants with Minor Allele Frequency (MAF) higher than 5% were considered in all the following analyses. The MAF values were computed taking into account that the reference allele reported in the VCF file may not always be the most frequent one in the EUR population considered by itself and we conservatively attributed the most frequent homozygous genotype to individuals for which the genotype was missing, thus being sure to exclude all the less frequent variants from the analysis. We are aware that these MAF values may be an underestimate of the real ones and therefore in all the enrichment analyses (see below for details) we instead used MAF values obtained ignoring individuals with missing data.

3.3.5 Principal Component Analysis of genotypic data

It is known that special patterns of linkage disequilibrium (LD) can cause artifacts when a Principal Component Analysis (PCA) is used to investigate population structure [203]. We filtered out all the genetic variants falling within 24 long-range LD (LRLD) regions whose coordinates were derived from [203]. In addition, following [90], we performed an LD-pruning of the genetic variants using the `--indep-pairwise` function from `PLINK v1.9` [204] to recursively exclude genetic variants with pairwise genotypic R^2 >80% within sliding windows of 50 SNPs (with a 5-SNPs increment between windows). Also in this case `VCFtools` [202] was used to apply all these filters to the VCF files and finally `EIGENSTRAT v6.1.4` [139] was used to run the PCA on the remaining genotypic data at the genome-wide level.

3.3.6 apaQTL mapping

From a statistical point of view, we adopted the same strategy used in standard eQTL mapping analyses [82] to identify genetic variants that influence the expression level of the alternative 3′ UTR isoforms of a gene. For each of the 6,256 examined genes, we defined a cis-window as the region spanning the gene body and 1 Mbp from both its TSS and its TES. Then, for each gene a linear model was fitted, independently for each genetic variant within its cis-window, using the genotype for the genetic variant as the independent variable and the log2-transformed m/M value of the gene as the dependent variable:

$$\log_2(m/M_{a,i}) = \beta_0 + \beta_1 \times g_{j,i} + \beta_2 \times I_i + \sum_{n=1}^{3} \alpha_n \times gPC_{n,i} + \epsilon_a \qquad (3.2)$$

where $log_2(m/M_{a,i})$ is the log2-transformed m/M value computed for the a gene in the i_{th} individual, $g_{j,i}$ is the genotype of the i_{th} individual for the j_{th} genetic variant, I_i is the imputation status (0|1) of the i_{th} individual, $gPC_{n,i}$ is the value of the n_{th} Principal Component (PC) obtained from genotypic data for the i_{th} individual, β_0 is the intercept, β_1, β_2 and α_n are the fitted regression coefficients and ϵ_a is the error term for the gene a.

The fitting of the linear models was done using `MatrixEQTL` [205]. Genotypes were represented using the standard 0/1/2 codification, referring to the number of alternative alleles present in each individual, and matrices with genotypic information were obtained from VCF files exploiting the Perl API (Vcf.pm) included in the VCFtools suite [202]. Following [82], in all our models we included both the imputation status of the individuals and the first three PCs obtained from genotypic data as covariates, in order to correct for possible biases due to population stratification (Supplementary Figure B.7) or genotype imputation.

The observed distribution of nominal P-values was compared with the expected one in Quantile-Quantile plots (Q-Q plots), revealing the expected inflation due to the LD issue (Supplementary Figure B.8). A permutation-based procedure was implemented [206]: all the models were fitted again after the random shuffling of the m/M values of each gene across samples; then for each gene-variant pair we counted how many times we obtained a random P-value less than its nominal P-value and divided this value by the total number of random tests done. Finally, to control for multiple testing, the empirical P-values were corrected with the Benjamini-Hochberg procedure [207] and models with a corrected empirical P-value less than 0.05 were considered statistically significant. Manhattan plots were drawn using the CRAN R package `qqman` [208].

3.3.7 Comparison with other molecular QTLs

In order to compare the genes for which we detected one or more apaQTLs with those for which eQTL/trQTL were reported [82], we translated the Ensembl Gene IDs (ENSG) to NCBI Entrez Gene IDs using Ensembl v67 [209] retrieved using the Bioconductor R package `biomaRt v2.30` [210, 211]. 229 ENSGs could not be translated with this procedure and were therefore excluded from this analysis.

3.3.8 Enrichment analyses

In order to functionally characterize the apaQTLs, we analyzed the enrichment of several features among such variants, including their genomic location, their ability to alter known regulatory motifs, and their association with complex diseases. All enrichments were evaluated through multivariate logistic regression to allow correcting for covariates. In this section we provide an overview of the method, but refer to the following subsections for details about each analysis.

For each feature we first established which genetic variants were potentially associated with the feature (for example only variants in the 3′ UTR can alter microRNA binding sites). Therefore, each enrichment analysis started with the selection of the "candidate variants" that were subsequently subjected to an LD-based pruning, in order to obtain a subset of independent candidate variants (the same strategy was implemented for example in [212] to evaluate the enrichment of GWAS hits among eQTLs). LD-based pruning was always performed using `PLINK` with the same parameters used in the case of the PCA of genotypic data (see above), but applied in each case to the candidate variants only. To each candidate variant surviving pruning we attributed a binary variable indicating whether it has the feature under investigation. Finally, these variants are classified as apaQTLs (i.e. corrected empirical P-value < 0.05 for at least one gene) and null variants (i.e. nominal P-value > 0.1 in all the fitted

models). We excluded the "grey area" variants with nominal P-value < 0.1 but empirical corrected P-value > 0.05 as they are likely to contain many false negatives. Finally we fitted a multivariate logistic model in which the dependent variable is the apaQTL/null status of the variant, and the independent variables are the feature of interest and covariates. The latter always include the MAF of the variant, since variants with higher MAF are more likely to be found as significant apaQTLs, and possibly other covariates depending on the feature under examination (see below).

The logistic model can thus be written as:

$$t_j = \beta_0 + \beta_1 \times Feature_j + covariates + \epsilon_j \tag{3.3}$$

$$Pr(apaQTL)_j = \frac{1}{1 + \exp^{-t_j}} \tag{3.4}$$

where $Feature_j$ is a binary variable indicating whether the genetic variant j has the feature of interest, β_0 is the intercept, β_1 is the regression coefficient for the feature, ϵ_j is the error term and $Pr(apaQTL)_j$ is the fitted probability that the genetic variant j is an apaQTL. As expected, in our models the regression coefficient of the MAF was always positive. The regression coefficient of the $Feature$ term and its associated P-value were used to establish if having the feature under investigation influences the probability of being an apaQTL, and to compute the corresponding odds ratio (OR).

Chromatin states

This analysis was performed independently for two cell types (the GM12878 and NHEK cell lines). In both cases, the candidate variants were virtually all the genetic variants for which the apaQTL models were fitted, but we excluded those not associated with any chromatin state and all the structural variants, because their length can prevent them from being univocally associated with a chromatin state.

Each of the 15 chromatin states and 6 broad chromatin classes (promoter, enhancer, insulator, transcribed, repressed and inactive) defined in [147], separately for the two cell lines, was treated as a binary feature to be used as a regressor in Eq. 3.3, with value 1 assigned to the variants falling within a DNA region associated to the given chromatin state. Only the MAF was included in the covariates.

Gene regions

The candidate variants were all the intragenic variants for which the apaQTL models were fitted. We defined as intragenic all variants falling between the start and the end of the gene, plus 1,000 bps after the end (to take into account possible misannotations of the $3'$ UTR).

Independent enrichment analyses were performed for the following sequence classes: coding exons, introns, $5'$ UTR and $3'$ UTR. For each class the binary feature used as a regressor was assigned the value 1 for variants falling within the class and 0 otherwise. Only the MAF was included in the covariates.

Cis Regulatory Domains

The candidate variants were all the extragenic variants (i.e. all variants that are not intragenic according to the definition given above) for which an apaQTL model was fitted. The binary feature was given value 1 for variants falling within a CRD and 0 otherwise. Besides the MAF, the distance from the nearest gene was included as a covariate, since variants closer to a gene are more likely to be apaQTLs.

To verify that the apaQTLs tend to be included in the CRDs specifically associated to the gene on which they act, we translated the CRD-gene associations provided in [170] into Entrez Gene IDs, and we counted how many genetic variants fall within a CRD associated to at least one gene for which the variant is an apaQTL. This number was then compared with that obtained in the same way after randomly assigning a target gene to each extragenic variant within the cis-window used for apaQTL analysis (100 independent randomizations were used).

Alteration of putative functional motifs

Similar strategies were implemented to investigate the alteration of different types of putative functional motifs by intragenic variants. This analysis was restricted to Single Nucleotide Polymorphism (SNPs), excluding therefore both indels and structural variants. For all SNPs we reconstructed the sequence of both the reference (REF) and the alternative (ALT) allele in the 20 bp region around each candidate genetic variant to determine whether the ALT allele creates or destroys a functional motif with respect to the REF allele. The functional motifs analyzed included PAS motifs, microRNA binding sites, and RBP binding sites.

To each candidate variant surviving LD pruning we associated, using `PLINK`, a list of tagging variants with genotypic $R^2 > 80\%$, and a binary feature value of 1 if the candidate variant itself or any of its tagging variant altered a functional motif. The enrichment of apaQTLs among motif-affecting variants was then evaluated with the logistic model described by Eq. 3.3. In the following, we describe the details of the logistic model for each class of functional motifs.

PAS motifs. The PAS motif is always located upstream of its target poly(A) site. It has been suggested that a narrow range of 10-30 nt is required for efficient processing, but recent work suggests that also larger distances can be functional thanks to RNA folding processes bringing the poly(A) site closer to the PAS [185]. Assuming that a PAS-altering SNP would affect the usage of its nearest poly(A) site, we associated to each intragenic SNP the nearest downstream poly(A) site, selected those for which such poly(A) site was located within the PRE/POST segments, and retained as candidate variants only those whose distance from the corresponding poly(A) site was between 10 and 100 nt. PAS-altering variants were defined as those for which a particular PAS motif was found in either the REF or the ALT sequence, but not in both (note that the interconversion between PAS motifs is considered as well, assuming that they can have different strength).

microRNA binding sites. microRNA binding sites located downstream of a poly(A) site, and hence in the POST segment, can affect the relative abundance of the long and short isoforms by allowing the selective degradation of

the former by microRNAs. Therefore, we chose as candidate variants all the SNPs within the POST segment of the genes analyzed. Putative microRNA binding sites were classified, as in [198], in three classes: 8mer, 7mer-m8, and 7mer-A1 (matches classified as 6-mer were not considered). A variant was defined to alter a microRNA binding site if a putative binding site was present in either the REF or the ALT sequence, but not in both, or if the site class was different between the REF and the ALT sequences. Moreover, altering variants were classified as creating (destroying) a binding site if only the ALT (REF) sequence contained a binding site or if the ALT (REF) sequence contained a stronger binding site than the REF (ALT), according to the hierarchy 8mer > 7mer-m8 > 7mer-A1 match. Only microRNA families conserved across mammals or broadly conserved across vertebrates and expressed in lymphoblastoid cells were considered. Following [82], each microRNA was considered expressed if its expression value was greater than 0 in at least 50% of the samples, and each microRNA family was considered expressed if at least one of its microRNAs was expressed.

RBP motifs. The candidate variants were all the intragenic SNPs. REF and ALT sequences around each candidate variant were scanned by `FIMO` [213], using as background the nucleotide frequencies on the sequence of all the analyzed genes. A motif was considered altered if its score was greater than 80% the score of the perfect match in only one of two alleles. As in the case of microRNAs, only motifs corresponding to RBPs expressed in lymphoblastoid cell lines were considered. Enrichment was evaluated both for SNPs altering any RBP motif, and for each expressed RBP separately.

GWAS hits

We considered only the GWAS catalog records referring to a single genetic variant on autosomal chromosomes for which all the fields CHR_ID, CHR_POS, SNPS, MERGED, SNP_ID_CURRENT, and MAPPED_TRAIT_URI were available, as well as the RSID. The coordinates of the selected genetic variants in hg19 were derived from dbSNP Build 151. We thus obtained 56,672 genetic variants associated with at least one complex trait. Furthermore, starting from the EFO URI(s) reported for each association, we obtained the corresponding EFO Parent URI(s) from the EFO annotation file.

All variants examined as potential apaQTLs were considered as our candidate variants. A binary feature value of 1 was attributed to each candidate variant surviving LD pruning and associated to a trait, or with a tagging variant associated to a trait, as in the case of motif-altering variants. Enrichment was evaluated for all trait-associated variants together, for each single trait, and for trait categories defined based on the EFO ontology. Only traits and trait categories associated with at least 100 GWAS hits were analysed. The same analysis was also performed after excluding all variants within the HLA locus, as defined by The Genome Reference Consortium (`https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC?asm=GRCh37`).

3.3.9 The rs10954213 variant in SLE patients

In the analysis of SLE patient RNA-seq data, we were interested in the *IRF5* gene only. Therefore, RNA-Seq reads were aligned to a reduced genome com-

prising the gene sequence and an additional 50bp at its $3'$ end using Bowtie v2.2.3 [214] and TopHat v2.0.12 [215]. As genotypic data were not available for these individuals, we inferred the rs10954213 variant status from the relative proportion of A and G in the RNA-Seq reads. Initially, individuals were considered homozygous for the reference (G) or for the alternative (A) allele when the same nucleotide was present in all the reads, and a single read with a different nucleotide was considered sufficient to call an heterozygous individual. Then, genotype quality was assessed using VCFx version 1.2b [216,217] with default parameters to filter out low-confidence genotypes. In this way we obtained 11 homozygotes for the reference allele, 22 heterozygotes and 14 homozygotes for the alternative allele (Supplementary Figure B.9), while 52 individuals with missing genotype information were excluded from the subsequent analysis. Notably, the genotypes are in Hardy-Weinberg equilibrium (chi-squared P-value = 0.705). A Kruskal-Wallis test was then used to evaluate the differences in m/M values between genotypes.

Chapter 4

Conclusions

4.1 Moving from GWAS associations to function

GWAS have identified tens of thousands of associations between genetic variants and complex traits, thus totally revolutionizing the study of human disease genetics. While the number of associations will probably further increase thanks to the availability of even larger cohorts and samples from different ethnic populations, the mechanistic interpretation of the results is still disappointingly challenging. Several strategies have been proposed to integrate GWAS findings and reference genomic data with the purpose of gaining insight about disease biology. For example, the stratified LD score regression is a statistical method to estimate the proportion of genome-wide SNP heritability that is attributable to various functional categories, taking advantage of GWAS summary statistics and explicitly modelling the LD structure. In particular, this technique can be exploited to uncover which cell types are more relevant for a specific complex trait, as illustrated by the results that Finucane *et al.* [218] obtained partitioning the heritability of 17 complex diseases and quantitative traits taking into account cell type-specific annotations. First, they were able to confirm several well-known findings, like the importance of pancreatic islets for type 2 diabetes and fasting glucose levels; on the other hand, they also detected some intriguing enrichments, such as the relevance of brain cell types for several non-psychiatric phenotypes, such as BMI and age at menarche.

Since the effect of genetic variants on common diseases is probably mediated by the perturbation of intermediate molecular phenotypes, the integration of GWAS results with molecular QTLs is an additional strategy to shed light on the mechanisms that underlie genotype-phenotype associations. In particular, it should be noted that, compared to the usage of reference annotations, molecular QTLs can provide complementary information, among which the identification of relevant genes is the most important. In the previous chapters, I presented in details and extensively discussed two computational strategies to investigate the genetic determinants of gene expression (Chapter 2) and alternative polyadenylation (Chapter 3). In the first study, I showed that TBA-based regression allows the identification of eQTLs not revealed by traditional methods and that this additional power derives from both its distinguishing features: the evaluation of the combined effect of genetic variants within regulatory regions

and the exploitation of the available knowledge about the regulatory code to weigh each genetic variant according to its effect on TF binding. In addition, the TBA model can help in the eQTL interpretation in terms of altered transcription factor binding. In the second project, I found that human genetic variants have a widespread effect also on the relative expression of alternative 3′ UTR isoforms, mainly as a consequence of the perturbation of the regulation of alternative polyadenylation. Notably, in both studies the pinpointed genomic regions showed significant enrichment for genetic variants that were identified by GWAS, consistently with the underlying assumption. This result is particularly remarkable in the case of apaQTLs, because, although previous studies have definitely proved the importance of alternative polyadenylation in human diseases, until now it was not possible to exploit this knowledge to systematically extract functional information from GWAS data. However, I would like to point out that an experimental validation, that would be crucial to securely discriminate between real functional associations and false positives, is currently missing for the results of both projects; on the other hand, the lack of a direct and systematic experimental validation is a limit shared by this work with most existing studies in this field (e.g., [82, 95]).

The overlap between GWAS and QTL results is frequently considered as an evidence of common causality. However, the implementation of colocalization strategies, that correctly account for linkage disequilibrium, should be preferred to formulate reliable mechanistic hypotheses. Indeed, given a genomic region that includes multiple genetic variants that are tested for association with two distinct traits, for example a disease and a molecular phenotype, five possible configurations can be hypothesized: (1) no association with either trait; (2) association with the first trait, but not with the second one; (3) association with the second trait, but not with the first one; (4) association with both traits, but the causal SNPs are independent; (5) association with both traits and the casual SNP is shared. The commonly used COLOC package implements a Bayesian statistical procedure to associate a posterior probability to each of these configurations [219, 220]. Nevertheless, some issues make the evaluation of common causality even more challenging. First, COLOC assumes the presence of only one causal variant in any genomic locus for both GWAS and QTL studies, thus ignoring the widespread allelic heterogeneity (i.e. the presence in the same locus of multiple causal variants for the same phenotype) of both complex traits [127] and gene expression [99, 102]. Furthermore, every QTL effect can be detected in different tissues, but only one or few associations may be relevant for the investigated disease; therefore, a joint analysis of the results obtained in different tissues could be convenient. Both these features are implemented by eCAVIER that, starting from GWAS summary statistics, can identify target genes and relevant tissues [221].

Recent works have proposed transcriptome wide association studies (TWAS) as another way to integrate GWAS and QTL data. Although TWAS are based on the idea that many genetic variants influence complex traits by modulating gene expression and therefore they originally have exploited eQTL data, they can also be implemented using other molecular QTLs (see below). Measured gene expression values can be conceptually decomposed into different parts: the first component results from the effects of regulatory variants (genetically regulated expression, or GReX), the second one reflects reverse causality (i.e. gene expression changes that are induced by the disease under investigation) and the

last one is determined by other factors, including the environment. In the initial version, TWAS first exploit eQTL data obtained in any reference dataset (e.g., the GTEx dataset) to predict the genetic component of gene expression in cases and controls for which only genotypic data are available (GWAS cohorts); then, GReX values are correlated with the disease status, thus allowing the identification of susceptibility genes [131]. Focusing on the genetic component of gene expression has some key advantages compared to the usage of measured gene expression levels [222]. First, expression data are usually not available for large human cohorts. Second, it allows to overcome potential confounding that may derive from reverse causality. In addition, the elimination of environmental noise may result in an increased statistical power. Furthermore, as other gene-based strategies, TWAS benefit from a lower multiple-testing burden than GWAS, because they evaluate thousands of genes instead of millions of variants, and the interpretation of results is more straightforward. However, compared with GWAS, the power of TWAS may be reduced when the effect of genetic variants on diseases is not mediated by the perturbation of gene expression levels, or when expression data cannot be obtained for the relevant tissue as a consequence of lacking eQTL data [222].

We are now working to develop a TBA-based framework to perform TWAS[a]. Notably, switching from genetic variants to TF affinity, the TBA model should be less dependent on the variants that were used for training. Therefore, compared with other existing tools (e.g., PrediXcan [131] and TIGAR [224]), it could be superior in catching also the effects of low-frequency and rare variants, including variants that were not observed in the training cohort. In this regard, it must be noted that current reference datasets usually include hundreds of individuals with genotype and gene expression data, while commonly there are thousands of GWAS samples. In addition, the TBA-based model could better generalize to different populations. The latter feature would be particularly exciting because existing reference datasets mainly include individuals of European ancestry and the predictive performance of methods that directly evaluate genetic variants can greatly change within and among populations of different continents [225]. We are also wondering if it could be used to predict the transcriptome starting from ancient human genomes, thus providing an unprecedented opportunity to study gene expression evolution in the human lineage. This speculation relies on the assumption that TFs and their binding preferences have been conserved, consistently with the knowledge that TFs usually evolve much slower than their binding sites and the identification of human TF orthologs well beyond primates [226]. On the other hand, we aware that the generalization potential of the TBA model may be limited in some conditions: for example, if in the target population the regulation of a gene depended on a TF whose binding sites are completely absent in the regulatory region that was observed in the reference population, the power of the TBA model would be greatly reduced. Furthermore, a possible weakness of this method is the (appar-

[a] A preliminary R/Bioconductor package (AffiXcan) [223] is already available to perform TBA-based imputation of GReX values. The TBA model can be trained in a reference dataset where both real gene expression data and TBA values are known, and then it can be used to predict GReX values for other samples for which only TBA values must be provided. Some important technical improvements will be implemented in the next months. In particular, to facilitate its use, we have planned to integrate both the computation of TBA values and the evaluation of the association between GReX values and traits of interest. Furthermore, TBA models trained on all GTEx tissues will be released.

ently) mandatory usage of WGS data, while the majority of currently available GWAS data were obtained using SNP arrays. In this regard, it is important to point out that SNP array data are not enough for the implementation of TBA models, because the computation of accurate TBA values requires the reconstruction of regulatory regions as similar as possible to real ones. This issue will be probably resolved in the coming years, with the further spread of WGS approaches. In addition, we reasoned that, thanks to availability of large panels of genome-wide human variation, like the ones provided by the 1000 Genomes Project and the Haplotype Reference Consortium [227], it could be successfully overcome through genotype imputation[b]. Therefore, we decided to investigate if genotype imputation is actually a suitable solution exploiting the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[c] that contains genotypic data obtained both with WGS and SNP arrays, together with gene expression values measured in whole blood cells using microarrays, for about one hundred healthy individuals. First, despite the cell type and the technology used to measure gene expression levels were different, significant genes that were found by the TBA model using ADNI WGS data significantly overlap with those that were identified in the GEUVADIS dataset, supporting the reproducibility of the results. In addition, in the ADNI dataset the TBA values that were computed after genotype imputation were highly correlated with those obtained starting from WGS data and the results of the TBA model were virtually the same in the two analyses, suggesting that the TBA values resulting from genotype imputation are good enough for this kind of analysis. Furthermore, now we would like to perform a TBA-based imputation of GReX values for both cases and controls for which WGS data are available in the ADNI database, thus aiming to uncover new genes that are associated with Alzheimer's disease onset and progression. In particular, we believe that complementary information may be obtained correlating multi-tissue GReX values with multiple phenotypes, including disease status, quantitative biomarkers and brain imaging measurements.

As described above, the initial TWAS implementation requires GWAS data at an individual level (individual-level TWAS). However, the results of large-scale GWAS are frequently publicly available only as summary statistics. Therefore, several methods have been proposed to indirectly estimate expression-trait association statistics by integrating SNP-expression and SNP-trait correlation data, while accounting for LD among SNPs (summary-based TWAS) [222, 224, 230, 231]. Notably, this kind of analysis has also been performed exploiting sQTL collections, leading to the identification of new susceptibility genes for schizophrenia [232] and Alzheimer's disease [188]. In a similar way, apaQTLs could be integrated with GWAS summary statistics to discover cases

[b]Genotype imputation is the process of predicting genotypes that are not directly assayed in a group of individuals, taking advantage of reference haplotype sets. Basically, in each sample, phased haplotypes may be modelled as a mosaic of those present in the reference panel and then this representation can be exploited to predict missing genotypes. For further details, see for example [228].

[c]Data used in the preparation of this thesis were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`http://adni.loni.usc.edu/`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For a recent summary of publications using ADNI data, see [229].

in which the association between genes and diseases is driven by the alteration of the expression of alternative 3′ UTR isoforms. From a technical point of view, the computational strategy that I described in the previous chapter requires at least one important modification. While fitting models that evaluate the association between single variants and quantitative phenotypes allows the identification of molecular QTLs, TWAS require prediction models that should be generated by fitting additive models that jointly evaluate the effect of *cis*-variants. Different statistical approaches, including both variable selection (e.g., LASSO [140] and elastic net [233]) and shrinkage-based methods (e.g., BLUP [234] and BSLMM [235]), are commonly exploited for this purpose. Since we would like to perform APA-wide association studies (apaWAS), we are currently applying these strategies to the generation of prediction models for the expression of alternative 3′ UTR isoforms, using multi-tissue RNA-seq data obtained from the GTEx dataset.

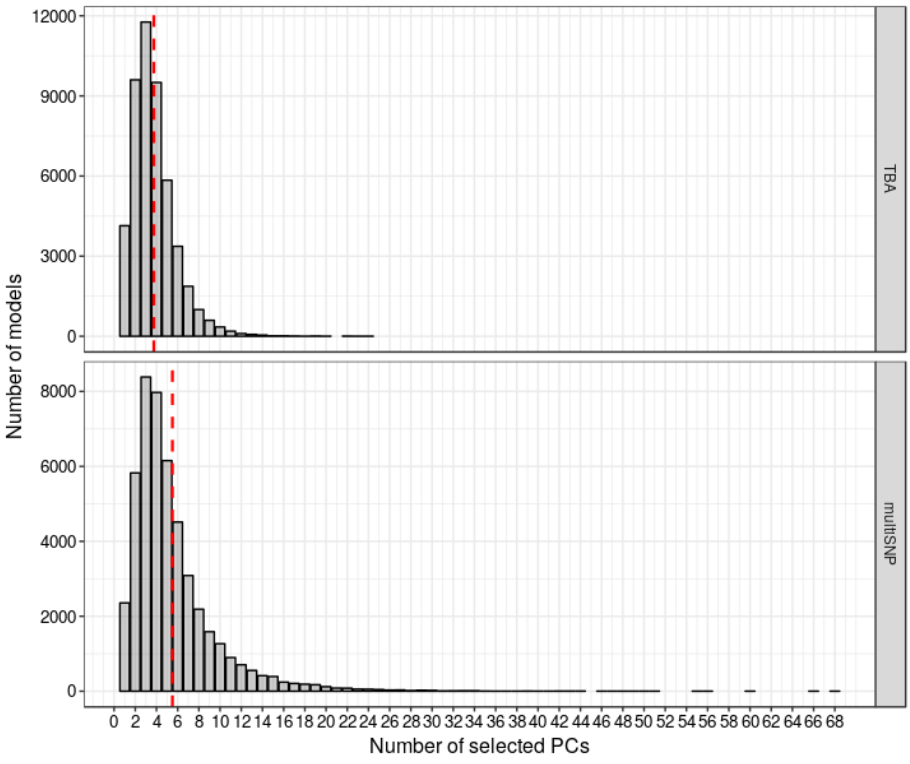
4.2 Concluding remarks

Almost twenty years later, the revolution that was triggered by the release of the first draft of the human genome has not reached its peak yet. Thanks to the development of NGS technologies, we have witnessed an outstanding drop in DNA sequencing costs that has fuelled the advent of several international and national projects that have sequenced, or are sequencing, the whole genome of hundreds of thousands, and soon millions, individuals. WGS will soon became the predominant technology for genetic analysis and this will be another fundamental breakthrough, compared with previous decades that have been dominated by marker genotyping or exome sequencing [61]. In addition, DNA sequencing efforts have been accompanied by numerous projects that have measured molecular phenotypes together with genotypes in large human cohorts (e.g., GEUVADIS and GTEx projects) and, in the most recent years, biobank projects are making available an incredible amount of clinically valuable data. Furthermore, the landscape of DNA sequencing technologies is still evolving: real-time single-molecule sequencing platforms, that do not require template amplification and generate very long reads, are already available and they could occupy important market niches in the next future [64]. Notably, they may contribute to the achievement of an even more comprehensive coverage of human genetic variation, allowing to overcome issues that result from the current usage of the reference genome sequence combined with data obtained through short-read technologies for genotype calling [61].

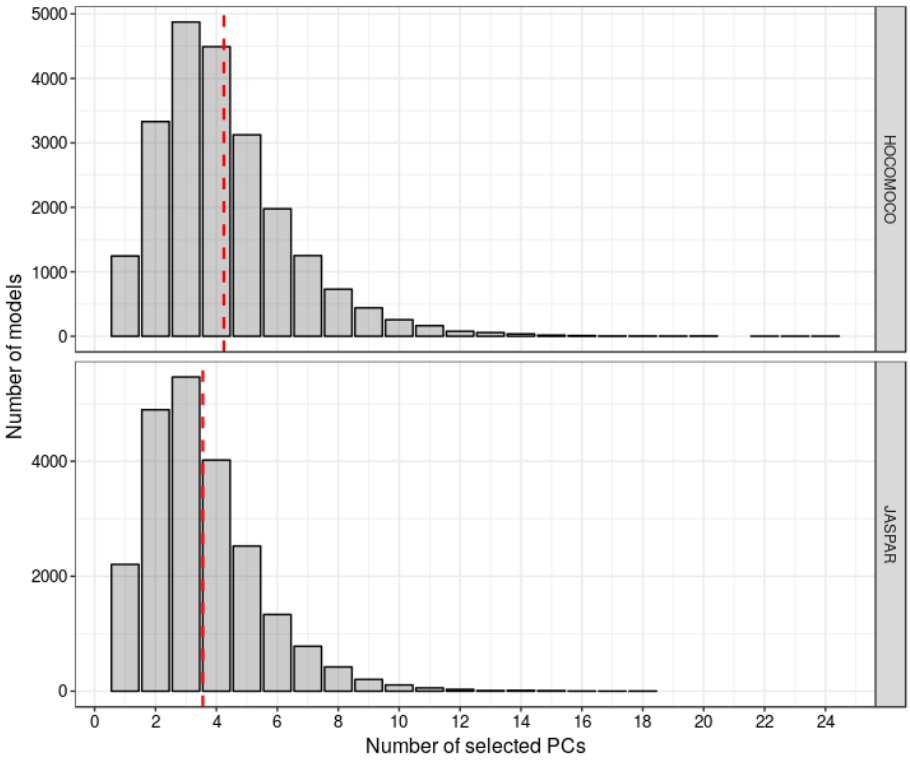
Appendix A

Supplementary material for Chapter 2

Supplementary figures



Supplementary Figure A.1 – Histogram of the number of PCs needed to explain at least 95% of the variance of the TBA values or SNPs calls on all regulatory regions (proximal and distal). The red line is the mean.

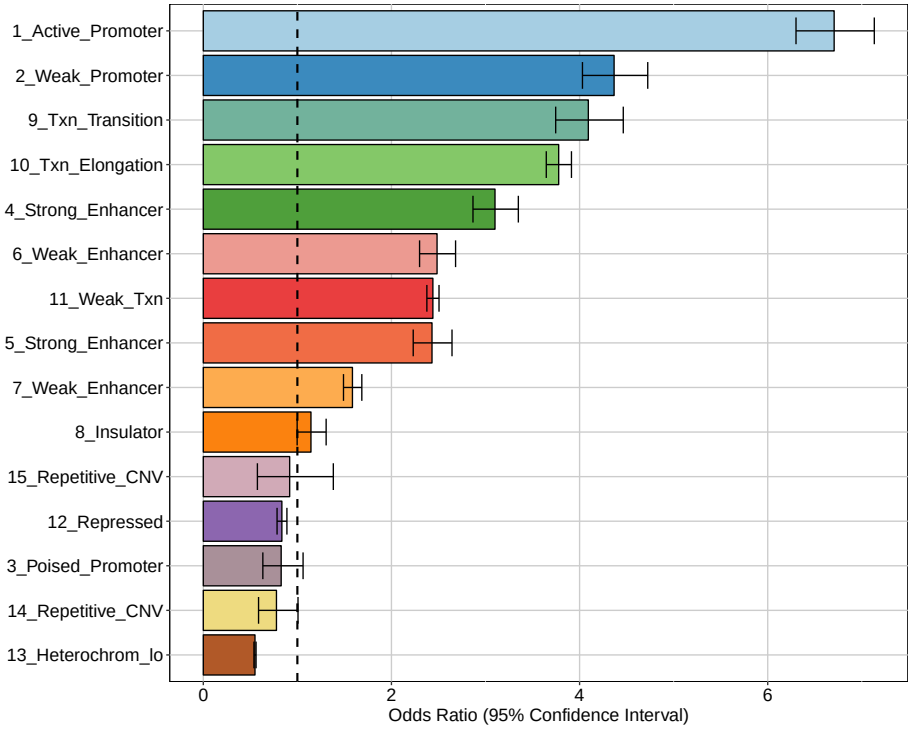


Supplementary Figure A.2 – Histogram of the number of PCs needed to explain at least 95% of the variance of the TBA values for the HOCOMOCO and JASPAR PWM databases, limited to proximal regulatory regions. The red line is the mean.

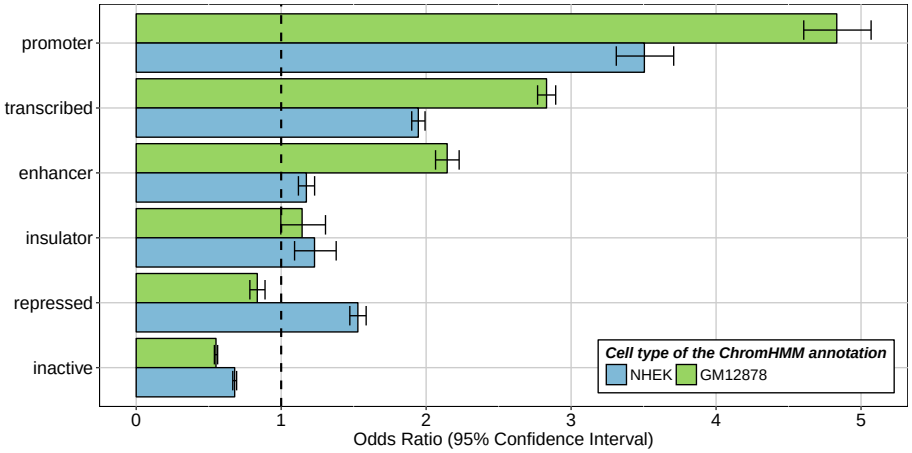
Appendix B

Supplementary material for Chapter 3

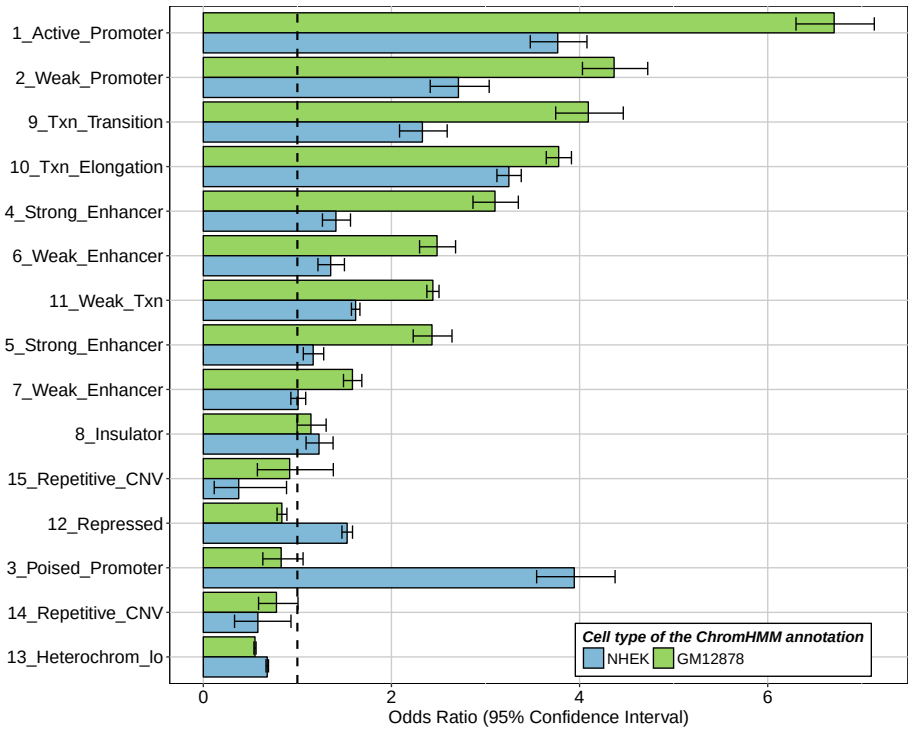
Supplementary figures



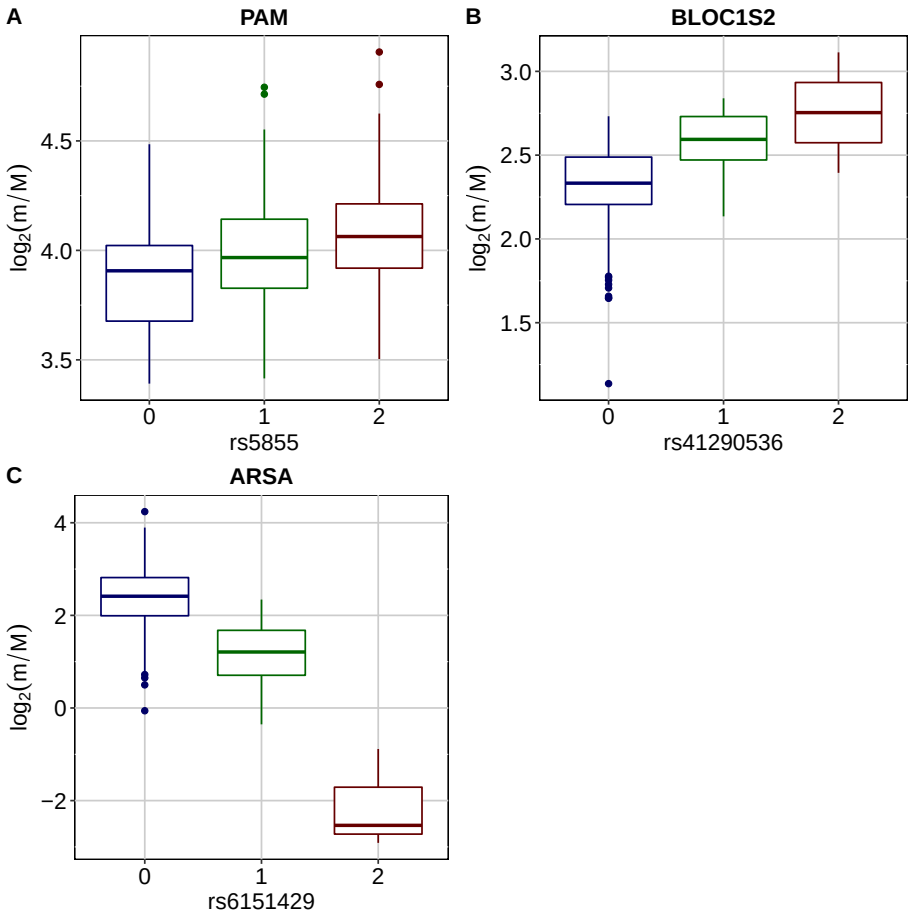
Supplementary Figure B.1 – Enrichment of apaQTLs within chromatin states, taking into account all the 15 chromatin states reported in the ChromHMM annotation. For each of them, the OR obtained by logistic regression and its 95% CI are shown.



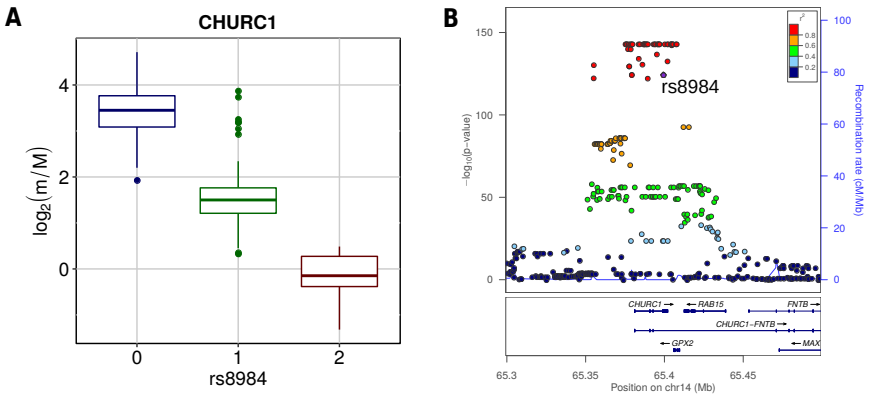
Supplementary Figure B.2 – The results of the enrichment analysis performed with the broad chromatin states of the relevant cell type were compared with those obtained using the ChromHMM annotation of another cell type (NHEK). For each category, the OR obtained by logistic regression and the corresponding 95% CI are shown.



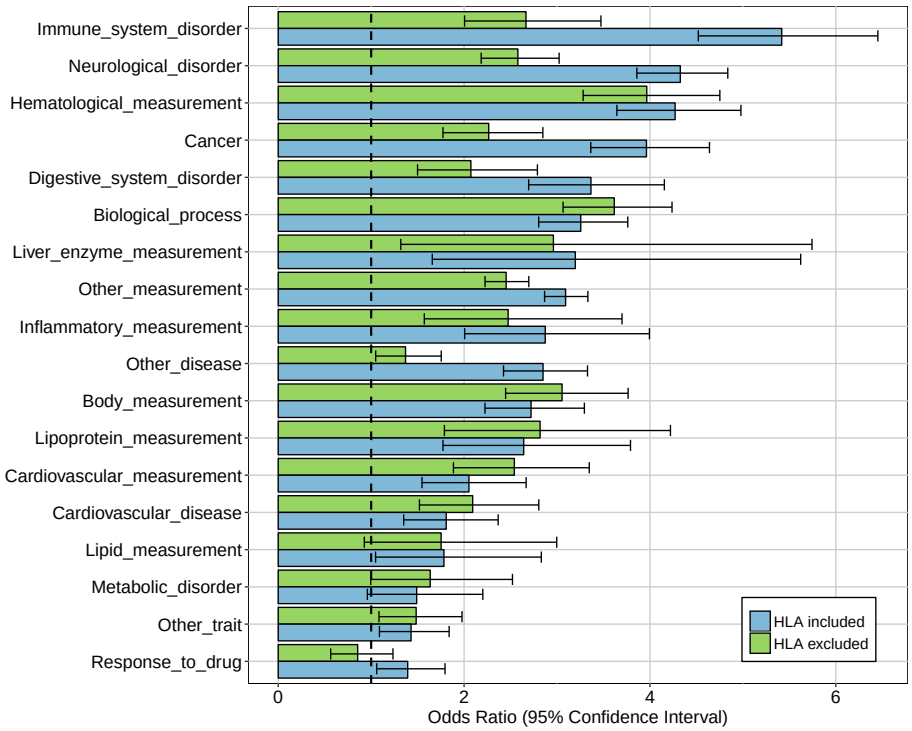
Supplementary Figure B.3 – The results of enrichment analysis done performed with all the chromatin states of the relevant cell type were compared with those obtained using the ChromHMM annotation of another cell type (NHEK). For each category, the OR obtained by logistic regression and the corresponding 95% CIs are shown.



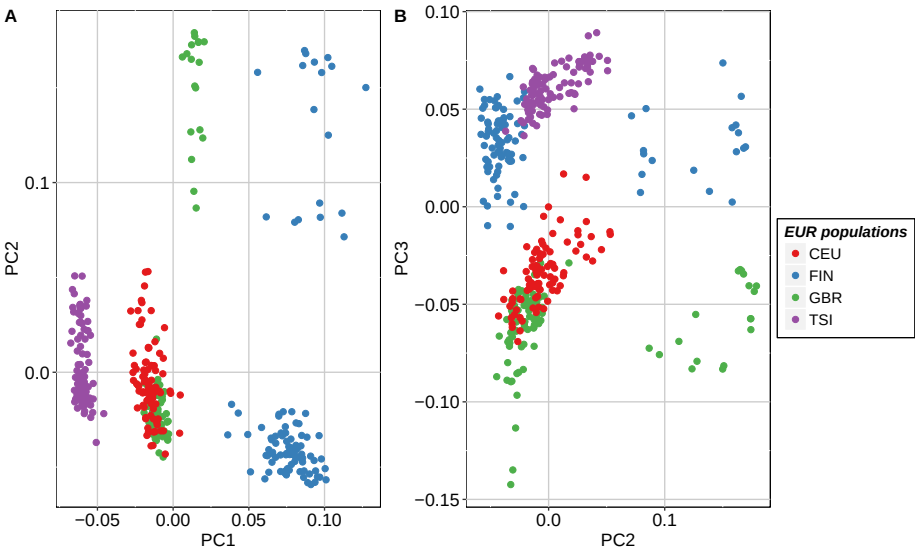
Supplementary Figure B.4 – Boxplots showing the variation of the log2-transformed m/M values obtained for *PAM* (A), *BLOC1S2* (B) and *ARSA* (C), as a function of the genotype of the individuals for a single genetic variant that falls within the cis-window of the tested gene (rs5855, rs41290536 and rs6151429, respectively).



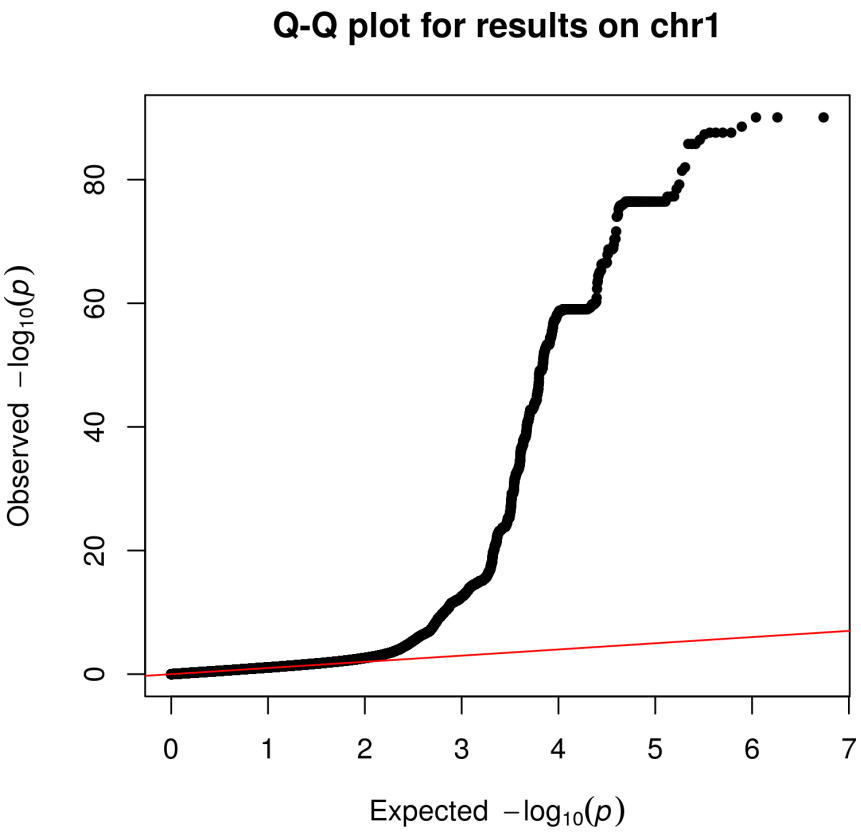
Supplementary Figure B.5 – (A) Boxplot showing the variation of log2-transformed m/M values obtained for *CHURC1* as a function of the genotype of the individuals for rs8984. (B) LocusZoom plot illustrating the results obtained for *CHURC1* in the genomic region around rs8984 (100kb both upstream and downstream its genomic location).



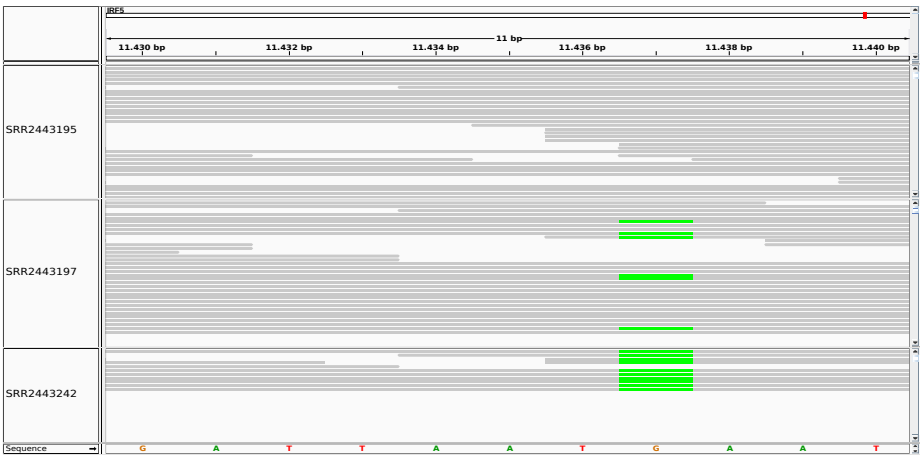
Supplementary Figure B.6 – Comparison of the results of the enrichment analyses performed for multiple categories of complex traits considering all the studied genetic variants (HLA included) or after having excluded those that are located within the HLA locus (HLA excluded). For each category, the OR obtained by logistic regression and the corresponding 95% CIs are shown.



Supplementary Figure B.7 – Principal Component Analysis (PCA) on the genotypic data of EUR individuals. Points are colored according to the subpopulation of origin: Utah Residents (CEPH) with Northern and Western European Ancestry (CEU), Finnish in Finland (FIN), British in England and Scotland (GBR) and Toscani in Italia (TSI).



Supplementary Figure B.8 – Q-Q plot comparing the distribution of P-values obtained fitting apaQTL models for genes on chr1 with the expected uniform distribution. It was generated by the CRAN R package `qqman`.



Supplementary Figure B.9 – Genotypic information was not available for the SLE patients, therefore their genotype in correspondence of the rs10954213 genetic variant was inferred from RNA-Seq data. The figure shows the alignment of RNA-Seq reads in a region around the variant with respect to a reduced genome including only the *IRF5* gene and was generated using the Integrative Genomics Viewer (IGV) software. For example, the SRR2443195 individual (top panel) was considered homozygous for the reference allele, the SRR2443197 individual (central panel) was considered heterozygous and finally the SRR2443242 individual (bottom panel) was considered homozygous for the alternative allele.

Supplementary tables

MOTIF ID	RBP NAMES	OR	95% CI	P-VALUE	FDR
M016_0.6	FMR1	3.72	1.01-11.3	0.0277	0.265
M025_0.6	HNRNPC	1.77	1.06-2.83	0.0213	0.264
M070_0.6	ENSG00000180771;SRSF2	3.83	1.55-8.7	0.00196	0.0518
M075_0.6	TIA1	1.81	0.993-3.1	0.0389	0.326
M081_0.6	CSDA;YB-1	3.9	1.03-12.6	0.0284	0.265
M089_0.6	HNRNPL	3.4	1.08-9.12	0.0216	0.264
M122_0.6	MEX3B;MEX3C;MEX3D	3.85	1.03-12	0.0265	0.265
M140_0.6	ENOX1;ENOX2	3.16	1.36-6.76	0.00431	0.078
M145_0.6	RBM5	4.97	1.51-14.6	0.00441	0.078
M147_0.6	CNOT4	2.11	0.948-4.25	0.0478	0.362
M156_0.6	TIA1	1.81	1.05-2.96	0.0247	0.265
M158_0.6	HNRNPCL1	1.77	1.06-2.83	0.0213	0.264
M160_0.6	KHDRBS1	4.15	1.54-10.3	0.0027	0.0614
M250_0.6	CSDA	2.93	0.939-7.73	0.041	0.326
M256_0.6	ACO1	1.92	1.07-3.25	0.0208	0.264
M291_0.6	EIF4B	10.6	3.5-33.2	2.48e-05	0.00131
M292_0.6	EIF4B	1.98	1.31-2.88	0.000643	0.0256
M320_0.6	MBNL1;MBNL2;MBNL3	1.66	1.2-2.25	0.00163	0.0518
M333_0.6	SRSF9	1.8	0.99-3.07	0.0398	0.326
M344_0.6	RBMX;RBMXL1;RBMXL2	1.77	1.4-2.2	6.38e-07	5.07e-05

Supplementary Table B.1 – Enrichment of RBP-altering SNPs among intragenic apaQTL.

EFO URI	EFO TERM	OR	95% CI	P-VALUE	FDR
EFO_0000540	Immune_system_disorder	5.42	4.52-6.45	1.99e-77	1.19e-76
EFO_0000618	Neurological_disorder	4.33	3.86-4.84	1.89e-142	1.7e-141
EFO_0004503	Hematological_measurement	4.27	3.64-4.98	2.43e-74	1.09e-73
EFO_0000616	Cancer	3.96	3.36-4.64	3.72e-63	1.34e-62
EFO_0000405	Digestive_system_disorder	3.37	2.69-4.15	4e-28	9e-28
GO_0008150	Biological_process	3.26	2.8-3.76	8.22e-56	2.47e-55
EFO_0004582	Liver_enzyme_measurement	3.2	1.66-5.62	0.000167	0.000215
EFO_0001444	Other_measurement	3.09	2.87-3.33	8.45e-190	1.52e-188
EFO_0004872	Inflammatory_measurement	2.87	2.01-3.99	1.71e-09	3.07e-09
EFO_0000408	Other_disease	2.85	2.42-3.33	1.99e-38	5.11e-38
EFO_0004324	Body_measurement	2.72	2.22-3.29	1.46e-23	2.92e-23
EFO_0004732	Lipoprotein_measurement	2.64	1.77-3.79	4.98e-07	7.47e-07
EFO_0004298	Cardiovascular_measurement	2.05	1.55-2.67	2.2e-07	3.61e-07
EFO_0000319	Cardiovascular_disease	1.81	1.35-2.37	3.42e-05	4.73e-05
EFO_0004529	Lipid_measurement	1.78	1.05-2.83	0.0218	0.0231
EFO_0000589	Metabolic_disorder	1.49	0.959-2.2	0.0589	0.0589
EFO_0000001	Other_trait	1.43	1.09-1.84	0.00741	0.00889
GO_0042493	Response_to_drug	1.39	1.06-1.79	0.0135	0.0152

Supplementary Table B.2 – Enrichment of GWAS hits for different trait categories among apaQTL.

EFO URI	EFO TERM	OR	95% CI	P-VALUE	FDR
EFO_0004503	Hematological_measurement	3.97	3.28-4.75	3.16e-48	1.9e-47
GO_0008150	Biological_process	3.62	3.06-4.24	1.27e-54	1.15e-53
EFO_0004324	Body_measurement	3.05	2.45-3.76	2.8e-24	1.01e-23
EFO_0004582	Liver_enzyme_measurement	2.96	1.32-5.74	0.00338	0.00468
EFO_0004732	Lipoprotein_measurement	2.82	1.79-4.22	2.03e-06	3.94e-06
EFO_0000540	Immune_system_disorder	2.67	2.01-3.47	2.32e-12	6.96e-12
EFO_0000618	Neurological_disorder	2.58	2.18-3.02	2.9e-30	1.3e-29
EFO_0004298	Cardiovascular_measurement	2.54	1.89-3.35	1.76e-10	3.95e-10
EFO_0004872	Inflammatory_measurement	2.47	1.57-3.7	3.07e-05	4.61e-05
EFO_0001444	Other_measurement	2.45	2.23-2.7	4.32e-75	7.78e-74
EFO_0000616	Cancer	2.26	1.77-2.85	1.27e-11	3.26e-11
EFO_0000319	Cardiovascular_disease	2.09	1.52-2.8	2.19e-06	3.94e-06
EFO_0000405	Digestive_system_disorder	2.07	1.5-2.79	3.92e-06	6.41e-06
EFO_0004529	Lipid_measurement	1.75	0.926-3	0.0587	0.0622
EFO_0000589	Metabolic_disorder	1.64	0.995-2.52	0.0369	0.0415
EFO_0000001	Other_trait	1.48	1.08-1.98	0.00981	0.0126
EFO_0000408	Other_disease	1.37	1.05-1.75	0.0163	0.0196
GO_0042493	Response_to_drug	0.855	0.564-1.23	0.43	0.43

Supplementary Table B.3 – Enrichment of GWAS hits for different trait categories among apaQTL, after the exclusion of genetic variants within the HLA locus.

Bibliography

[1] Hansen, A. S., Cattoglio, C., Darzacq, X. & Tjian, R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* **9**, 20–32 (2018).

[2] Dahm, R. Friedrich Miescher and the discovery of DNA. *Developmental Biology* **278**, 274–288 (2005).

[3] Roberts, L., Davenport, R., Pennisi, E. & Marshall, E. A History of the Human Genome Project. *Science* **291**, 1195 (2001).

[4] Dulbecco, R. A turning point in cancer research: sequencing the human genome. *Science* **231**, 1055–6 (1986).

[5] Cook-Deegan, R. & McGuire, A. L. Moving beyond Bermuda: sharing data to build a medical information commons. *Genome research* **27**, 897–901 (2017).

[6] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

[7] Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–51 (2001).

[8] Marshall, E. Human genome. Storm erupts over terms for publishing Celera's sequence. *Science* **290**, 2042–3 (2000).

[9] Marris, E. Free genome databases finally defeat Celera. *Nature* **435**, 6 (2005).

[10] Whitfield, J. Human genome pioneer steps down. *Nature* (2002).

[11] Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

[12] Waterston, R. H., Lander, E. S. & Sulston, J. E. On the sequencing of the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 3712–6 (2002).

[13] Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287**, 2185–95 (2000).

[14] Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).

[15] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

[16] Eddy, S. R. The C-value paradox, junk DNA and ENCODE. *Current Biology* **22**, R898–9 (2012).

[17] Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–48 (2012).

[18] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

[19] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).

[20] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

[21] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).

[22] Eddy, S. R. The ENCODE project: Missteps overshadowing a success. *Current Biology* **23**, R259–R261 (2013).

[23] Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences* **111**, 6131–6138 (2014).

[24] Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

[25] Stadhouders, R., Filion, G. J. & Graf, T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* **569**, 345–354 (2019).

[26] Goldberg, A. D., Allis, C. D. & Bernstein, E. Epigenetics: A Landscape Takes Shape. *Cell* **128**, 635–638 (2007).

[27] Liu, G., Mattick, J. & Taft, R. J. A meta-analysis of the genomic and transcriptomic composition of complex life. *Cell Cycle* **12**, 2061–2072 (2013).

[28] Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

[29] Romero, I. G., Ruvinsky, I. & Gilad, Y. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics* **13**, 505–516 (2012).

[30] Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* **5**, 276–287 (2004).

[31] Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166**, 538–554 (2016).

[32] Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research* **46**, D260–D266 (2018).

[33] Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **451**, 535–540 (2008).

[34] Foat, B. C., Morozov, A. V. & Bussemaker, H. J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141–e149 (2006).

[35] Ward, L. D. & Bussemaker, H. J. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics* **24**, i165–i171 (2008).

[36] Molineris, I., Grassi, E., Ala, U., Di Cunto, F. & Provero, P. Evolution of Promoter Affinity for Transcription Factors in the Human Lineage. *Molecular Biology and Evolution* **28**, 2173–2183 (2011).

[37] Grassi, E., Zapparoli, E., Molineris, I. & Provero, P. Total Binding Affinity Profiles of Regulatory Regions Predict Transcription Factor Binding and Gene Expression in Human Cells. *PLOS ONE* **10**, e0143627 (2015).

[38] Gruber, A. J. & Zavolan, M. Alternative cleavage and polyadenylation in health and disease. *Nature Reviews Genetics* (2019).

[39] Takagaki, Y., Seipelt, R. L., Peterson, M. L. & Manley, J. L. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87**, 941–952 (1996).

[40] Mayr, C. & Bartel, D. P. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138**, 673–684 (2009).

[41] Hoffman, Y. *et al.* 3'UTR Shortening Potentiates MicroRNA-Based Repression of Pro-differentiation Genes in Proliferating Human Cells. *PLoS Genetics* **12** (2016).

[42] Gruber, A. R. *et al.* Global 3′ UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nature Communications* **5** (2014).

[43] Berkovits, B. D. & Mayr, C. Alternative 3'UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**, 363–367 (2015).

[44] Taliaferro, J. M. *et al.* Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Molecular Cell* **61**, 821–833 (2016).

[45] Ji, Z., Lee, J. Y., Pan, Z., Jiang, B. & Tian, B. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7028–33 (2009).

[46] Ji, Z. & Tian, B. Reprogramming of 3′ Untranslated Regions of mRNAs by Alternative Polyadenylation in Generation of Pluripotent Stem Cells from Different Cell Types. *PLoS ONE* **4**, e8419 (2009).

[47] Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science (New York, N.Y.)* **320**, 1643–7 (2008).

[48] Berg, M. G. *et al.* U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**, 53–64 (2012).

[49] Elkon, R. *et al.* E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biology* **13** (2012).

[50] Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010).

[51] Stacey, S. N. *et al.* A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nature Genetics* **43**, 1098–1103 (2011).

[52] Graham, R. R. *et al.* Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 6758–63 (2007).

[53] Lee, S. H. *et al.* Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia (2018).

[54] Lembo, A., Di Cunto, F. & Provero, P. Shortening of 3'UTRs Correlates with Poor Prognosis in Breast and Lung Cancer. *PLoS ONE* **7**, e31129 (2012).

[55] Fu, Y. *et al.* Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome research* **21**, 741–7 (2011).

[56] Grassi, E., Mariella, E., Lembo, A., Molineris, I. & Provero, P. Roar: detecting alternative polyadenylation with standard mRNA sequencing libraries. *BMC Bioinformatics* **17**, 423 (2016).

[57] Ha, K. C. H., Blencowe, B. J. & Morris, Q. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biology* **19**, 45 (2018).

[58] Lee, J. Y., Yeh, I., Park, J. Y. & Tian, B. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic acids research* **35**, D165–8 (2007).

[59] You, L. *et al.* APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic acids research* **43**, D59–67 (2015).

[60] Masamha, C. P. *et al.* CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **510**, 412–416 (2014).

[61] Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. Genomic Analysis in the Age of Human Genome Sequencing. *Cell* **177**, 70–84 (2019).

[62] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).

[63] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

[64] Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).

[65] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[66] Birney, E. & Soranzo, N. The end of the start for population sequencing. *Nature* **526**, 52–53 (2015).

[67] Chiang, C. W. K. *et al.* Genomic history of the Sardinian population. *Nature Genetics* **50**, 1426–1434 (2018).

[68] Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).

[69] Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

[70] Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery (2011).

[71] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

[72] Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).

[73] Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019).

[74] Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

[75] Marx, J. L. The cystic fibrosis gene is found. *Science* **245**, 923–5 (1989).

[76] Rommens, J. M. *et al.* Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**, 1059–65 (1989).

[77] Ashley, E. A. Towards precision medicine. *Nature Reviews Genetics* **17**, 507–522 (2016).

[78] Bates, G. P. The molecular genetics of Huntington disease - a history. *Nature Reviews Genetics* **6**, 766–773 (2005).

[79] Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

[80] Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20**, 467–484 (2019).

[81] Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

[82] Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

[83] Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051–1065 (2015).

[84] Banovich, N. E. *et al.* Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. *PLoS Genetics* **10**, e1004663 (2014).

[85] Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).

[86] Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–7 (2015).

[87] Pan, D. Z. *et al.* Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. *Nature Communications* **9**, 1512 (2018).

[88] Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science (New York, N.Y.)* **352**, 600–4 (2016).

[89] Manning, K. S. & Cooper, T. A. The roles of RNA processing in translating genotype to phenotype. *Nature Reviews Molecular Cell Biology* **18**, 102–114 (2017).

[90] Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

[91] Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biology* **16**, 195 (2015).

[92] Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

[93] Samuels, D. C. *et al.* Finding the lost treasures in exome sequencing data (2013).

[94] Kim, Y. C. *et al.* Exome-based Variant Detection in Core Promoters. *Scientific Reports* **6** (2016).

[95] Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

[96] Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).

[97] Strober, B. J. *et al.* Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).

[98] Gaffney, D. J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology* **13**, R7 (2012).

[99] Wen, X., Luca, F. & Pique-Regi, R. Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLoS Genetics* **11**, e1005176 (2015).

[100] Tehranchi, A. K. *et al.* Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell* **165**, 730–41 (2016).

[101] Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genetics* **10**, e1004226 (2014).

[102] Jansen, R. *et al.* Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Human Molecular Genetics* **26**, 1444–1451 (2017).

[103] Foissac, S. & Sammeth, M. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Research* **35**, W297–W299 (2007).

[104] Monlong, J., Calvo, M., Ferreira, P. G. & Guigó, R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature Communications* **5**, 4698 (2014).

[105] Ongen, H. & Dermitzakis, E. Alternative Splicing QTLs in European and African Populations. *The American Journal of Human Genetics* **97**, 567–575 (2015).

[106] Ferreira, P. G. *et al.* Sequence variation between 462 human individuals fine-tunes functional sites of RNA processing. *Scientific Reports* **6**, 32406 (2016).

[107] Yoon, O. K., Hsu, T. Y., Im, J. H. & Brem, R. B. Genetics and Regulatory Impact of Alternative Polyadenylation in Human B-Lymphoblastoid Cells. *PLoS Genetics* **8**, e1002882 (2012).

[108] Thomas, L. F. & Sætrom, P. Single Nucleotide Polymorphisms Can Create Alternative Polyadenylation Signals and Affect Gene Expression through Loss of MicroRNA-Regulation. *PLoS Computational Biology* **8**, e1002621 (2012).

[109] Zhernakova, D. V. *et al.* DeepSAGE Reveals Genetic Variants Associated with Alternative Polyadenylation and Expression of Coding and Noncoding Transcripts. *PLoS Genetics* **9**, e1003594 (2013).

[110] Xiong, H. Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).

[111] Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell* **178**, 91–106.e23 (2019).

[112] Grassi, E., Mariella, E., Forneris, M. *et al.* A functional strategy to characterize expression Quantitative Trait Loci. *Human Genetics* **136**, 1477–1487 (2017).

[113] Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research* **46**, D252–D259 (2018).

[114] Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome research* **24**, 1–13 (2014).

[115] Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research* **42**, D142–D147 (2014).

[116] Hellman, A. & Chess, A. Extensive sequence-influenced DNA methylation polymorphism in the human genome. *Epigenetics & Chromatin* **3**, 11 (2010).

[117] Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487–492 (2013).

[118] Leslie, R., O'Donnell, C. J. & Johnson, A. D. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**, i185–i194 (2014).

[119] Bradfield, J. P. *et al.* A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci. *PLoS Genetics* **7**, e1002293 (2011).

[120] Hernandez, J. B. *et al.* The CREB/CRTC2 pathway modulates autoimmune disease by promoting Th17 differentiation. *Nature Communications* **6**, 7216 (2015).

[121] Han, J.-W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nature Genetics* **41**, 1234–1237 (2009).

[122] Lee, Y.-C. *et al.* Two new susceptibility loci for Kawasaki disease identified through genome-wide association analysis. *Nature Genetics* **44**, 522–525 (2012).

[123] Pamuk, O. N. *et al.* BLK pathway-associated rs13277113 GA genotype is more frequent in SLE patients and associated with low gene expression and increased flares. *Clinical Rheumatology* **36**, 103–109 (2017).

[124] Chang, C.-J. *et al.* Replication and Meta-Analysis of GWAS Identified Susceptibility Loci in Kawasaki Disease Confirm the Importance of B Lymphoid Tyrosine Kinase (BLK) in Disease Susceptibility. *PLoS ONE* **8**, e72037 (2013).

[125] LeBlanc, M. *et al.* Genome-wide study identifies PTPRO and WDR72 and FOXQ1-SUMO1P1 interaction associated with neurocognitive function. *Journal of psychiatric research* **46**, 271–8 (2012).

[126] Beltran, P. J., Bixby, J. L. & Masters, B. A. Expression of PTPRO during mouse development suggests involvement in axonogenesis and differentiation of NT-3 and NGF-dependent neurons. *The Journal of Comparative Neurology* **456**, 384–395 (2003).

[127] Hormozdiari, F. *et al.* Widespread Allelic Heterogeneity in Complex Traits. *American journal of human genetics* **100**, 789–802 (2017).

[128] Li, M.-X., Gui, H.-S., Kwan, J. S. H. & Sham, P. C. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *American journal of human genetics* **88**, 283–93 (2011).

[129] Liu, J. Z. *et al.* A versatile gene-based test for genome-wide association studies. *American journal of human genetics* **87**, 139–45 (2010).

[130] Wang, M. *et al.* COMBAT: A Combined Association Test for Genes Using Summary Statistics. *Genetics* **207**, 883–891 (2017).

[131] Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**, 1091–1098 (2015).

[132] Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS – a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Research* **45**, D139–D144 (2017).

[133] Moyerbrailean, G. A. *et al.* Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? *PLoS Genetics* **12**, e1005875 (2016).

[134] Zuo, C., Shin, S. & Keleş, S. atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* **31**, 3353–3355 (2015).

[135] Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**, 1760–74 (2012).

[136] Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics* **47**, 11.12.1–11.12.34 (2014).

[137] Grassi, E. vcf_rider: library to efficiently compute score on individual genomes starting from vcf files (2017). URL `https://github.com/vodkatad/vcf_rider`.

[138] Mevik, B.-M., Wehrens, R., Liland, K. H. & Hiemstra, P. pls: Partial Least Squares and Principal Component Regression. URL `https://cran.r-project.org/package=pls`.

[139] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909 (2006).

[140] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288 (1996).

[141] Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **42**, 80 (2000).

[142] Mariella, E., Marotta, F., Grassi, E., Gilotto, S. & Provero, P. The Length of the Expressed 3' UTR Is an Intermediate Molecular Phenotype Linking Genetic Variants to Complex Diseases. *Frontiers in Genetics* **10**, 714 (2019).

[143] Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology* **18**, 18–30 (2017).

[144] Spies, N., Burge, C. B. & Bartel, D. P. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome research* **23**, 2078–90 (2013).

[145] Floor, S. N. & Doudna, J. A. Tunable protein synthesis by transcript isoforms in human cells. *eLife* **5**, e10921 (2016).

[146] An, J. J. *et al.* Distinct Role of Long 3' UTR BDNF mRNA in Spine Morphology and Synaptic Plasticity in Hippocampal Neurons. *Cell* **134**, 175–187 (2008).

[147] Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

[148] Tian, B., Pan, Z. & Lee, J. Y. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome research* **17**, 156–65 (2007).

[149] Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–6 (2016).

[150] Gruber, A. J. *et al.* A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome research* **26**, 1145–1159 (2016).

[151] Cunninghame Graham, D. S. *et al.* Association of IRF5 in UK SLE families identifies a variant involved in polyadenylation. *Human molecular genetics* **16**, 579–91 (2007).

[152] Lazzari, E. & Jefferies, C. A. IRF5-mediated signaling and implications for SLE. *Clinical Immunology* **153**, 343–352 (2014).

[153] Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).

[154] Shaw, G. M. *et al.* 118 SNPs of folate-related genes and risks of spina bifida and conotruncal heart defects. *BMC Medical Genetics* **10**, 49 (2009).

[155] Eipper, B. A., Glembotski, C. C. & Mains, R. E. Bovine intermediate pituitary alpha-amidation enzyme: preliminary characterization. *Peptides* **4**, 921–8 (1983).

[156] Czyzyk, T. A. *et al.* Deletion of peptide amidation enzymatic activity leads to edema and embryonic lethality in the mouse. *Developmental biology* **287**, 301–13 (2005).

[157] Gaier, E. D. *et al.* Genetic determinants of amidating enzyme activity and its relationship with metal cofactors in human serum. *BMC endocrine disorders* **14**, 58 (2014).

[158] Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nature genetics* **49**, 1511–1516 (2017).

[159] McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature genetics* **49**, 1126–1132 (2017).

[160] Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).

[161] Shi, Y. & Manley, J. L. The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes & development* **29**, 889–97 (2015).

[162] La Rosa, P. *et al.* Sam68 promotes self-renewal and glycolytic metabolism in mouse neural progenitor cells by modulating Aldh1a3 pre-mRNA 3'-end processing. *eLife* **5** (2016).

[163] Miller, J. E. & Reese, J. C. Ccr4-Not complex: the control freak of eukaryotic cells. *Critical Reviews in Biochemistry and Molecular Biology* **47**, 315–333 (2012).

[164] Millevoi, S. *et al.* An interaction between U2AF 65 and CF Im links the splicing and 3' end processing machineries. *The EMBO Journal* **25**, 4854–4864 (2006).

[165] Millevoi, S. *et al.* A physical and functional link between splicing factors promotes pre-mRNA 3' end processing. *Nucleic acids research* **37**, 4672–83 (2009).

[166] Gunderson, S. I. *et al.* The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. *Cell* **76**, 531–41 (1994).

[167] Lutz, C. S. *et al.* Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro. *Genes & development* **10**, 325–37 (1996).

[168] Liang, S. & Lutz, C. p54nrb is a component of the snRNP-free U1A (SF-A) complex that promotes pre-mRNA cleavage during polyadenylation. *RNA* **12**, 111–121 (2006).

[169] Oktaba, K. *et al.* ELAV links paused Pol II to alternative polyadenylation in the Drosophila nervous system. *Molecular cell* **57**, 341–8 (2015).

[170] Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**, eaat8266 (2019).

[171] Krijger, P. H. L. & de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nature Reviews Molecular Cell Biology* **17**, 771–782 (2016).

[172] MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research* **45**, D896–D901 (2017).

[173] Feng, D. *et al.* Genetic variants and disease-associated factors contribute to enhanced IRF-5 expression in blood cells of systemic lupus erythematosus patients. *Arthritis & Rheumatism* **62**, 562–573 (2010).

[174] Kozyrev, S. *et al.* Structural insertion/deletion variation in IRF5 is associated with a risk haplotype and defines the precise IRF5 isoforms expressed in systemic lupus erythematosus. *Arthritis & Rheumatism* **56**, 1234–1241 (2007).

[175] Hung, T. *et al.* The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression. *Science* **350**, 455–9 (2015).

[176] Kwan, T. *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nature Genetics* **40**, 225–231 (2008).

[177] Elkon, R., Ugalde, A. P. & Agami, R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics* **14**, 496–506 (2013).

[178] Gieselmann, V., Polten, A., Kreysing, J. & von Figura, K. Arylsulfatase A pseudodeficiency: loss of a polyadenylation signal and N-glycosylation site. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 9436–40 (1989).

[179] Steri, M. *et al.* Overexpression of the Cytokine BAFF and Autoimmunity Risk. *New England Journal of Medicine* **376**, 1615–1626 (2017).

[180] Ji, Z. *et al.* Transcriptional activity regulates alternative cleavage and polyadenylation. *Molecular systems biology* **7**, 534 (2011).

[181] Võsa, U., Esko, T., Kasela, S. & Annilo, T. Altered Gene Expression Associated with microRNA Binding Site Polymorphisms. *PLoS ONE* **10**, e0141351 (2015).

[182] Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology* **19**, 327–341 (2018).

[183] Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell* **70**, 854–867.e9 (2018).

[184] Yue, Y. *et al.* VIRMA mediates preferential m6A mRNA methylation in 3'UTR and near stop codon and associates with alternative polyadenylation. *Cell Discovery* **4**, 10 (2018).

[185] Wu, X. & Bartel, D. P. Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability. *Cell* **169**, 905–917.e11 (2017).

[186] Chang, J. W., Yeh, H. S. & Yong, J. Alternative Polyadenylation in Human Diseases. *Endocrinology and metabolism* **32**, 413–421 (2017).

[187] Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nature Communications* **8**, 14519 (2017).

[188] Raj, T. *et al.* Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nature Genetics* **50**, 1584–1592 (2018).

[189] Hoarau, J.-J., Cesari, M., Caillens, H., Cadet, F. & Pabion, M. HLA DQA1 genes generate multiple transcripts by alternative splicing and polyadenylation of the 3' untranslated region. *Tissue Antigens* **63**, 58–71 (2004).

[190] Kulkarni, S. *et al.* Posttranscriptional Regulation of HLA-A Protein Expression by Alternative Polyadenylation Signals Involving the RNA-Binding Protein Syncrip. *The Journal of Immunology* **199**, 3892–3899 (2017).

[191] Misra, M. K., Damotte, V. & Hollenbach, J. A. The immunogenetics of neurological disease. *Immunology* **153**, 399–414 (2018).

[192] Wang, R., Zheng, D., Yehia, G. & Tian, B. A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Research* **28**, 1427–1441 (2018).

[193] Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function (2016).

[194] Casper, J. *et al.* The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research* **46**, D762–D769 (2017).

[195] O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–D745 (2016).

[196] Carlson, M. org.Hs.eg.db: Genome wide annotation for Human (2016). URL `http://bioconductor.org/packages/org.Hs.eg.db/`.

[197] Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research* **39**, D19–21 (2011).

[198] Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).

[199] Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).

[200] Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research* **9**, 677–9 (1999).

[201] Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research* **34**, D590–D598 (2006).

[202] Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

[203] Price, A. L. *et al.* Long-Range LD Can Confound Genome Scans in Admixed Populations. *The American Journal of Human Genetics* **83**, 132–135 (2008).

[204] Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

[205] Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

[206] Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–71 (1994).

[207] Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).

[208] Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *The Journal of Open Source Software* **3**, 731 (2018).

[209] Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Research* **46**, D754–D761 (2018).

[210] Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

[211] Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **4**, 1184–1191 (2009).

[212] Li, L. *et al.* Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Frontiers in genetics* **4**, 103 (2013).

[213] Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

[214] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).

[215] Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

[216] Castelli, E. C. *et al.* HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples. *Human Immunology* **76**, 945–953 (2015).

[217] Lima, T. H. A. *et al.* HLA-F coding and regulatory segments variability determined by massively parallel sequencing procedures in a Brazilian population sample. *Human Immunology* **77**, 841–853 (2016).

[218] Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015).

[219] Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics* **10**, e1004383 (2014).

[220] Wallace, C. *et al.* Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human Molecular Genetics* **21**, 2815–2824 (2012).

[221] Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics* **99**, 1245–1260 (2016).

[222] Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *The American Journal of Human Genetics* **100**, 473–487 (2017).

[223] Lussana, A. AffiXcan. A Functional Approach To Impute Genetically Regulated Expression. (2019). URL `http://bioconductor.org/packages/AffiXcan/`.

[224] Nagpal, S. *et al.* TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *American journal of human genetics* **105**, 258–266 (2019).

[225] Mikhaylova, A. V. & Thornton, T. A. Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations. *Frontiers in Genetics* **10**, 261 (2019).

[226] Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, P650–665 (2018).

[227] McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**, 1279–1283 (2016).

[228] Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499–511 (2010).

[229] Weiner, M. W. *et al.* Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's & Dementia* **13**, e1–e85 (2017).

[230] Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252 (2016).

[231] Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9**, 1825 (2018).

[232] Gusev, A. *et al.* Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature Genetics* **50**, 538–548 (2018).

[233] Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67**, 301–320 (2005).

[234] Robinson, G. That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science* **6**, 15–32 (1991).

[235] Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics* **9**, e1003264 (2013).