Computational challenges in genome wide association studies: data processing, variant annotation and epistasis

Pablo Cingolani

PhD.

School of Computer Science

McGill University

Montreal,Quebec

March 2015

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

CHAPTER 1
Introduction

How does one's DNA influence their risk of getting a disease? Contrary to popular belief, your future health is not "hard wired" in your DNA. Only in a few diseases, referred as "Mendelian diseases", are there well known, almost certain, links between genetic mutations and disease susceptibility. For the majority of what are known as "complex traits", such as cancer or diabetes, genomic predisposition is subtle and, so far, not fully understood.

With the rapid decrease in the cost of DNA sequencing, the complete genome sequence of large cohorts of individuals can now be routinely obtained. This wealth of sequencing information is expected to ease the identification of genetic variations linked to complex traits. In this work, I investigate the analysis of genomic data in relation to complex diseases, which offers a number of important computational and statistical challenges. We tackle several steps necessary for the analysis of sequencing data and the identification of links to disease. Each step, which corresponds to a chapter in my thesis, is characterized by very different problems that need to be addressed.

i) The first step is to analyze large amounts of information generated by DNA sequencers to obtain a set of "genomic variants" present n each each individual. To address these big data processing problems, Chapter **??** shows how we designed a programming language (BigDataScript [23]), that simplifies the creation robust, scalable data pipelines.

ii) Once genomic variants are obtained, we need to prioritize and filter them to discern which variants should be considered "important" and which ones are likely to be less relevant. We created the SnpEff & SnpSift

[21, 22] packages that, using optimized algorithms, solve several annotation problems: a) standardizing the annotation process, b) calculating putative genetic effects, c) estimating genetic impact, d) adding several sources of genetic information, and e) facilitating variant filtering.

iii) Finally, we address the problem of finding associations between interacting genetic loci and disease. One of the main problems in GWAS, known as "missing heritability", is that most of the phenotypic variance attributed to genetic causes remains unexplained. Since interacting genetic loci (epistasis) have been pointed out as one of the possible causes of missing heritability, finding links between such interactions and disease has great significance in the field. We propose a methodology to increase the statistical power of this type of approaches by combining population-level genetic information with evolutionary information.

In a nutshell, this thesis addresses computational, analytical, algorithmic and methodological problems of transforming raw sequencing data into biological insight in the aetiology of complex disease. In the rest of this introduction we give the background that provides motivation for our research.

1.1 Genomes and genetic variants

DNA is composed of four basic building blocks, called "bases" or "nucleotides" [4]. These four nucleotides, usually abbreviated $\{A, C, G, T\}$, are Adenine, Cytosine, Guanine, and Thymine. Bases form pairs, either as $A-T$ or $C-G$, that pile-up forming two long polymers, with backbones that run in opposite directions giving rise to a double-helix structure [98]. Arbitrarily, one of the polymers is called the positive strand and the other is called the negative strand.

Proteins are composed by chains of amino acids and, as explained by the central dogma of biology [4], DNA is the template that instructs cellular

machinery how to produce proteins. There are 20 amino acids, which are the building blocks of all proteins. Each of the twenty amino acids is encoded by a group of three DNA bases called "codon" [33]. More than one codon can code for the same amino acid (i.e. $4^3 = 64$ codons > 20 amino acids) allowing for code redundancy. Additionally, there are codons that mark the end of the protein, these are called "STOP" and signal molecular machinery to end the translation process [14].

Proteins compose up to 50% of a cell's dry weight compared to 3% of the DNA [4]. Proteins perform their functions mainly by interacting with other proteins, forming complex pathways that lead to a vast array of cellular functions including catalysis of chemical reactions, cell signaling, and structural conformation of the cell [4]. The 3-dimensional structure of the protein, also called "tertiary structure", is tailored to bind to other proteins in a specific manner to accomplish a specific function.

The human genome has a total of 3 Giga-base-pairs (Gb), and those bases are divided into 22 "autosomal" chromosome pairs (in each pair one chromsome is maternally inherited and the other paternally inherited) and "sex" chromosomes. The longest of the autosomal chromosomes is roughly 250 Mega-bases (Mb) and the shortest one is 50 Mb.

In order to compare DNA from different individuals (or samples), we need a "reference genome". Having a standard reference sequence facilitates comparisons and analysis. For most well studied organisms, "reference genome" sequences are available and current large scale sequencing projects are extending significantly the number of genomes known, e.g. one project seeks to sequence 10,000 mammalian genomes [49], another is targeting all microbes that live within the human gut [93]. The human reference genome (e.g. GRCh37)

does not correspond to the DNA of any particular person, but to a "mosaic" of the genomes of thirteen anonymous volunteers from Buffalo, New York [89].

When the genome of an individual is sequenced, the DNA is compared to the "reference genome". Most of the DNA is the same, but there are differences. These differences, generically known as "genetic variants" (or "variants", for short), describe the particular genetic make-up of each individual. There are several different ways a sample can differ from a reference genome. Each variant is the result of a mutations that happened at some point in the evolutionary history of the individual (or that of the reference genome). Variant types can be roughly categorized in the following way:

Single nucleotide variants (SNV) or Single nucleotide polymorphism (SNP) are the simplest and more common variants produced by single base difference (e.g. a base in the reference genome, at a given coordinate, is an 'A', whereas the sample is 'C'). Depending on whether the variant was identified in an individual or in a population, it is called a Single Nucleotide Variant (SNV) or Single Nucleotide Polymorphism (SNP). It is estimated that there are roughly $3.6M$ SNPs per individual [26]. There are several biological mechanisms responsible for this type of variants: i) replication errors, ii) errors introduced by DNA repair mechanism, iii) deamination (a base is changed by hydrolysis which may not be corrected by DNA repair mechanisms), iv) tautomerism (and alteration on the hydrogen bond that results in an incorrect pairing) [46].

Multiple nucleotide polymorphism (MNP) are sequence differences affecting several consecutive nucleotides and are typically treated as a single variant locus if they are in perfect linkage disequilibrium (e.g. reference is ACG' whereas the sample is TGC'). .

Insertions (INS) refer to a sample having one or more extra base(s) compared to the reference genome (e.g. the reference sequence is AT' and the sample is ACT'). Short insertions and deletions (indels) of a chromosome region range from 1 to 20 bases in length are reported to be 10 to 30 times less frequent than SNV [26]. Small insertions are usually attributed to DNA polymerase slipping and replicating the same bases (this produces a type of insertion known as duplication). Large insertions can be caused by unequal cross-over event (during meiosis) or transposable elements.

Deletions (DEL) are the opposite of insertions, the sample has some base(s) removed with respect to the reference genome (e.g. reference is ACT' and sample is AT'). As in the case of insertions, deletions can also be caused by ribosomal slippage, cross-over events during meiosis. Those include large deletions, which can result in the loss of an exon or one or more whole genes [4]. Short deletions are 10 to 30 times less frequent than SNV [26].

Copy number variations (CNVs) arise when the sample has two or more copies of the same genomic region (e.g. a whole gene that has been duplicated or triplicated) or conversely, when the sample has fewer copies than the reference genome. Copy number variations are often attributed to homologous recombination events [4].

Rearrangements such as inversions and translocations are events that involve two or more genomic breakpoints and a reorganization of genomic segments, possibly resulting in gene fusions or loss of critical regulatory elements. Inversions, a type of rearrangement, result from a whole genomic region being inverted.

As humans have two copies of each autosome, variants could affect zero, one or two of the chromosomes and are called "homozygous reference", "heterozygous", and "homozygous alternative" respectively. Variants are also classified based on how common they are within the population: common ($\geq 5\%$), low frequency ($\leq 5\%$), or rare ($\leq 0.1\%$). How these types of genetic variants influence traits or disease risk is a topic of intense research that is discussed throughout this thesis.

1.2 DNA and disease risk

It would be fair to say that the Garrod family was fascinated by urine. As a physician at King's College, Alfred Baring Garrod, discovered gout related abnormalities in uric acid [55]. His son, Sir Archibald Garrod, was interested in a condition known as alkaptonuria, in which children are mostly asymptomatic except for producing brown or black urine, but by the age of 30 individuals develop pain in joints of the spine, hips and knees. In 1902, Archibald observed that the family inheritance pattern of alkaptonuria resembled Mendel's recessive pattern and postulated that a mutation in a metabolic gene was responsible for the disease. Publishing his finding he gave birth to a new field of study known as "Human biochemical genetics" [55].

Diseases having simple inheritance patterns, such as alkaptonuria, cystic fibrosis, phenylketonuria and Huntington's are also known as Mendelian diseases [55]. The genetic components of several Mendelian diseases have been discovered since the mechanism was first elucidated by Garrod in 1902 and the process has been accelerated in recent years, thanks to the application of DNA sequencing techniques [10].

In complex diseases (or complex traits), such as diabetes or Alzheimer's disease, affected individuals cannot be segregated within pedigrees (i.e. no simple pattern of inheritance can be identified). In contrast to Mendelian

diseases the aetiology of complex traits is complicated due to factors such as: incomplete penetrance (symptoms are not always present in individuals who have the disease-causing mutation) and genetic heterogeneity (caused by any of a large number of alleles). This makes it difficult to pinpoint the genetic variants that increase risk of complex disease.

1.2.1 Heritability and Missing heritability

We all know that "tall parents tend to have tall children", which is an informal way to say that height is a highly heritable trait. It is said that there are 30 cm from the tallest 5% to the shortest 5% of the population and genetics account for 80% to 90% of this variation [100], which means that 27cm of variance are assumed to be "carried" by DNA variants from parents to offspring. Since 2010 the GIANT consortia has been investigating the genetic component of complex traits like height, body mass index (BMI) and waist to hip ratio (WHR). Even though they found many variants associated those traits, their findings only explain 10% of the phenotypic variance which corresponds to only a few centimeters in height [100].

In order to measure heritability we need a formal definition. Heritability is defined as the proportion of phenotypic variance that is attributed to genetic variations. The total phenotypic variation is assumed to be caused by a combination of "environmental" and genetic variations $Var[P] = Var[G] + Var[E] + 2Cov[G, E]$ [1] .

The environmental variance $Var[E]$ is the phenotypic variance attributable only to environment, that is the variance for individuals having the same

[1] Although the referenced paper's notation does not seem absolutely consistent, we quote Emerson *"A foolish consistency is the hobgoblin of little minds"* and proceed...

genome $Var[E] = Var[P|G]$. This can be estimated by studying monozygotic and dizygotic twins.

If the covariance factor $Cov[G, E]$ is assumed to be zero, we can define heritability as $H^2 = \frac{Var[G]}{Var[P]}$. This is called "broad sense heritability" because $Var[G]$ takes into account all possible forms of genetic variance: $Var[G] = Var[G_A] + Var[G_D] + Var[G_I]$, where $Var[G_A]$ is the additive variance, $Var[G_D]$ is the variance form dominant alleles, and $Var[G_I]$ is the variance form interacting alleles (epistasis). Non-additive terms are difficult to estimate, so a simpler form of heritability called "narrow sense heritability" that only takes into account additive variance is defined as $h^2 = \frac{Var[G_A]}{Var[P]}$ [107].

Focusing on narrow sense heritability, the concept of "explained heritability" is defined as the part of heritability due to known variants with respect to phenotypic variation ($\pi_{explained} = h^2_{known}/h^2_{all}$). Similarly, missing heritability is defined as $\pi_{missing} = 1\pi_{explained} = 1h^2_{known}/h^2_{all}$. When all variants associated with traits are known, then $\pi_{missing} = 0$.

Until recently, it was widely assumed by the research community that the problem of missing heritability lied in finding the appropriate genetic variants to account for the numerator of the equation (h^2_{known}) [107]. However, in a series of theorems published recently, it has been proposed that there is a problem in the way the denominator is estimated [107]. The authors created a limiting pathway model ($LP(k)$) that accounts for epistasis (gene-gene interactions) in k biological pathways. They showed that a severe inflation of h^2_{all} estimators occurs even for small values of k (e.g. $k \in [2, 10]$). As a result, genetic variants estimated to account only for 20% of heritability, could actually account for as much as 80% using an appropriate model [107].

Even though this result is encouraging, the problem is now shifted to detecting epistatic interactions, a problem that we discuss in section 1.6 and

Chapter **??**. In the same work [107], the authors show an example of power calculation assuming relatively large genetic effect that would require sequencing roughly $5,000$ individuals to detect links to genetic variants, which is a large but nowadays not uncommon, sample size. Nevertheless other estimates place the sample size requirements as high as $500,000$ individuals [107]. Even though this sounds as an extremely large number of samples, it is quickly becoming possible thanks to large technological advances and cost reductions in sequencing and genotyping technologies.

1.2.2 Conclusions

Although some genetic causes of complex traits, such as type II diabetes, have been found, only a small portion of the phenotypic variance can be explained. This might indicate that many risk variants are yet to be discovered. Recent studies on the topic of missing heritability report that these "difficult to find genetic variants" might be in epistatic interaction (analyzed in section 1.6.6) or rare variants (see section 1.5.6). Analysis of either them requires more complex statistical models and larger sample sizes with the corresponding increase in computational requirements. In Chapter **??** of this thesis, we focus on methods for finding epistatic interactions related to complex disease and develop computationally tractable algorithms that can process data from sequencing experiments involving large number of samples in a reasonable amount of time.

1.3 Identification of genetic variants

Two of the main milestones in genetics were the discovery of the DNA structure in 1953 [98], followed by the first draft of the human genome in 2004 [25]. The cost of sequencing the first human reference genome was around $3 billion (unadjusted US dollars) and it was an endeavor that took around 10 years. Since that time, sequencing technology has evolved substantially

so that a human genome can now be sequenced in three days for a price of less than $1,000, according to prices estimated by Illumina, one of the main genome sequencer manufacturers.

The amount of information delivered by sequencing devices is growing faster than computer speed (Moore's law) and data storage capacity. Just as a crude example, a leading edge sequencing system is advertized to be capable of delivering 18,000 human genomes at 30× coverage per year, yielding over 3.2 PB of information. Having to process huge amounts of sequencing information poses several challenges, a problem informally known as "data deluge". In this section, we explain how sequencing data is generated and how the huge amount of information delivered by a sequencer can be handled in order to make the problem tractable. We want to transform this raw data into knowledge of genomic variants that contribute to disease risk with the ultimate goal to translate these risk variants into biological knowledge. As expected, processing huge datasets consisting of thousands of sample is a complex problem. In Chapter **??** we show how mitigate or solve some of these issues, by designing a computer language specially tailored to tackle what are know as "Big data" problems.

1.3.1 Sequencing data

DNA sequencing machines (or sequencers) are based on different technologies. In a nutshell, a sequencer detects a set of polymers (or chains) of DNA nucleotides and outputs a set of strings of A, C, G, and Ts. Unfortunately, current technological limitations make it impossible to "read" a full chromosome as one long DNA sequence. Instead, modern sequencers produce a large number of "short reads", which range from 100 bases to 20 Kilo-bases (Kb) in length, depending on the technology. Since sequencers are unable to read long DNA chains, preparing the DNA for sequencing involves fragmenting

it into small pieces. These DNA fragments are a random sub-samples of the original chromosomes. Reading each part of the genome several times allows to increase accuracy and ensure that the sequencer reads as much as possible of the original chromosomes. The coverage of a sequencing experiment is defined as the number of times each base of the genome is read on average. For instance, if the sequencing experiment is designed to produce one billion reads, and each read is 150 bases long, then the total number of bases read is 150Gb. Since the human genome is 3Gb, the coverage is said to be 50×.

After sequencing a sample, we have millions of reads but we do not know where these reads originate from in the genome. This is solved by aligning (also called mapping) reads to the reference genome, which is assumed to be very similar to the genome being sequenced. Once the reads are mapped, we can infer if the sample's DNA has any differences with respect to the reference genome, a problem is known as "variant calling".

Although sequencing costs are dropping fast, it is still expensive to sequence thousands of samples and in some cases it makes sense to focus on specific areas of the genome. A popular experimental setup is to focus on coding regions (exons). A technique called "exome sequencing" consists of capturing exons using a DNA chip and then sequencing the captured DNA fragments only. Exons are roughly 1.2% of the genome, thus this technique reduces sequencing costs significantly, for which it has been widely used by many research groups although it has the disadvantage of only analysing coding genomic variation.

1.3.2 Read mapping

Once the samples have been sequenced, we have a set of reads from the sequencer. The first step in the analysis is finding the location in the reference genome where each read is supposed to originate from, a process that is

complicated by a several factors: i) there are differences between the reference genome and the sample genome, ii) sequencing reads may contain errors, iii) several parts of the reference genome are quite similar making reads from those regions indistinguishable, and iv) a typical sequencing experiment generates millions of reads.

Local sequence alignment. We introduce a problem known as *local sequence alignment*: Given two sequences s_1 and s_2 from an alphabet (e.g. $\Sigma = \{A, C, G, T\}$), the alignment problem is to add gap characters ('-') to both sequences, so that a distance, such as Levenshtein distance, $d(s_1, s_2)$ is minimized. This problem has a well known solution, the Smith-Waterman algorithm [91], which is a variation of the global sequence alignment solution from Needleman-Wunsch [76], having an algorithm complexity $O(l_1.l_2)$ where l_1 and l_2 are the length of the sequences. So, Smith-Waterman algorithm is slow since in this case one of the sequences is the entire genome.

In order to speed up sequence alignments, several heuristic approaches emerged. Most notably, BLAST [5], which could be for mapping sequences to a reference genome. BLAST uses an index of the genome to map parts of the query sequence, called seeds, to the reference genome. Once these seeds have been positioned against the reference, BLAST joins the seeds performing an alignment only using a small part of the reference.

Read mapping. Sequence alignment has an exact algorithm solution and several faster heuristic solutions. But even the fastest solutions are too slow to be used with the millions of reads generated in a typical sequencing experiment. Faster algorithms can be used if we relax our requirements in two ways: i) we allow for sub-optimal results, and ii) instead of requiring the output to be a complete local alignment between a read and the genome, we just want to know the region in the reference genome where the read sequence is from.

This relaxed version of the alignment algorithm is called "read mapping" and the reduced complexity is enough to speed up the computations significantly. In a nutshell, a read mapping is regarded as correct if it overlaps the true reference genome region where the read originated. Once the mapping is performed, the read is locally aligned, a strategy similar to BLAST algorithm [65, 60].

Reformulating this as a *mapping* problem allows us to use data structures such as suffix trees to index the reference genome. Using suffix trees we can query for a substring (read) [38] of the indexed string in $O(m)$ time, where m is the length of the query. Alternatively, we can use suffix arrays which are a space optimized alternative to suffix trees [38]. An implicit assumption in this solution, is that the read will be very similar to the reference and that there will be no big gaps. Suffix arrays algorithms are fast but, even though they are memory optimized versions of suffix trees, memory requirements are still high ($O[n\ log(n)]$, where n is the length of the indexed sequence -the genome-) and this becomes the limiting factor. In order to reduce memory footprint of suffix arrays, Ferragina and Manzini [40] created a data structure based on the Burrows-Wheeler transform. This structure, known as an FM-Index, is memory efficient yet fast enough to allow mapping high number of reads. An FM-index for the human genome can be built in only 1Gb of memory, compared to 12Gb required for an equivalent suffix array [65]. Given a genome G and a read R, an FM-index search can find the N_{occ} exact occurrences of R in G in $O(|R| + N_{occ})$ time, where $|R|$ is the length of R [65].

We should keep in mind that suffix trees, suffix arrays and FM-indexes are guaranteed to find all matching substring occurrences, nevertheless a sequencing read may not be an exact substring of the reference genome (due to sample's genome differences with the reference genome, read errors, etc.). So,

even if efficient indexing and heuristic algorithms can decrease mapping time considerably, these algorithms are not guaranteed to find an optimal mapping. Several parameters, such as read length, sequencing error profile, and genome complexity profile can affect performance. The most commonly used implementation of the FM-index mapping algorithms are BWA [65, 66] and Bowtie [60, 59]. Each of them provide optimized versions for the two most common sequencing types: i) short reads with high accuracy [65, 60] or ii) longer reads with lower accuracy [66, 59].

Mapping quality. Sequencers not only provide sequence information, but also provide an error estimate for each base [64]. This is often referred as a quality (Q) value, which is the probability of an error, measured in negative decibels $Q = -10\ log_{10}(\epsilon)$, where ϵ is the error probability. Mapping quality is an estimation of the probability that a read is incorrectly mapped to the reference genome. Mapping algorithms provide estimates of mapping errors. In the MAQ model [67], which is one of the earliest models for calculating mapping quality, three main sources of error are explored: i) the probability that a read does not originate from the reference genome (e.g. sample contamination); ii) the probability that the true position is missed by the algorithm (e.g. mapping error); and iii) the probability that the mapping position is not the true one (e.g. if we have several possible mapping positions). It is assumed that the total error probability can be approximated as $\epsilon \approx max(\epsilon_1, \epsilon_2, \epsilon_3)$.

1.3.3 Variant calling

Genome-wide variant calling has until recently largely been done using genotyping arrays (for SNVs) or Comparative Genomic Hybridization arrays (for CNVs). The inherent limitations of these technologies, particularly their

ability to only assay genotypes at sites that are known in advance to be polymorphic, combined with the declining cost of sequencing, have now made approaches based on high-throughput resequencing the tool of choice for variant calling in clinical studies.

Once the sequencing reads have been mapped to the reference genome, we can try to find the differences between a sequenced sample and the reference genome. This process is called "variant calling" [77]. Several factors complicate this task, the two main ones being sequencing errors and mapping errors, described in 1.3.2. Based on sequencing data and mapping error estimates, tools such as GATK [73] and SamTools/BcfTools [67] use maximum likelihood models can infer when there is a mismatch between a sample and the reference genome and whether the sample is homozygous or heterozygous for the variant. This method works best for differences of a single base (SNV), but it can also work with different degrees of success for short insertions or deletions (InDels) usually consisting of less than 10 bases.

Aligning sequences that contain InDels (gaps) is more difficult than ungapped alignments since finding optimal gap boundary depends on the scoring method being used. This biases variant calling algorithms towards detecting false SNVs near InDels [37]. An approach to reduce this problem is to look for candidate InDels and perform a local realignment in those regions. This local re-alignment process reduces significantly the number of false positive SNVs [37]. Another approach to reduce the number of false positive SNVs calls near InDels involves the "Base Alignment Quality" (BAQ) [63], which is the probability of misalignment for each base. It can be shown that replacing the original base quality with the minimum between base quality and BAQ produces an improvement in SNV calling accuracy. The BAQ can be calculated using a special type of "Hidden Markov Model" (HMM) designed for

sequence alignment [63, 38]. A more sophisticated option for reducing errors consist of performing a local genome re-assembly on each polymorphic region (e.g. HaplotypeCaller algorithm [92]).

Finally, one should note that the error probabilities inferred by the sequencers are far from perfect. Once the variants have been called, empirical error probabilities can be easily calculated [73] by comparing sequenced variants to a set of "gold standard variants" (i.e. variants that have been extensively validated). This allows to re-calibrate or re-estimate the error profile of the reads. This is know as a re-calibration step, and usually improves the number of false positives calls [37].

Due to the nature of short reads, this family of methods does not work for structural genomic variants, such as large insertions, deletions, copy number variations, inversions, or translocations. A different family of algorithms are used to identify structural variants generally making use of pair end reads or split reads, but their accuracy so far has been low compared to SNV calling algorithm [78].

Caveats

i Using current technologies and computational methods for variant calling, detection accuracy varies significantly for different variant types. SNV are by far the most accurately detected. Insertions and deletions, collectively referred as InDels, can be detected less efficiently depending on their sizes. Small InDels consisting of ten bases or less are easier to detect than large InDels consisting of 200 bases or more. The reason being that the most commonly used sequencers reads DNA in stretches roughly 200 bases long. Due to this technological limitations, detection is less reliable for more complex variant types.

1.4 Functional annotations of genomic variants

The development of cost-effective, high-throughput next generation sequencing (NGS) technologies is poised to have a profound impact on our ability to study the effects of individual genetic variants on the pathogenesis and progression of both monogenic and common polygenic diseases. As sequencing costs decrease and throughput increases, it has now become possible to quickly identify a large number of sequence polymorphisms (SNVs, indels, structural) using samples from affected and unaffected subjects and investigate these in epidemiologic studies to identify genomic regions where mutations increase disease risk. However, translating this information into biological or clinical insights is challenging as it is often difficult to determine which specific polymorphisms are the main pathogenetic drivers of disease across a population; and more importantly, how they affect the activity of disease-related molecular pathways in tissues and organism a specific patient. In part, this difficulty results from the large number of genetic variants that are observed in individual genomes (the human population is believed to contain approximately 3.5 million polymorphic sites with minor allele frequency above 5%) combined with the limited ability of computational approaches to distinguish variants with no impact on genome function (the vast majority) from variants affecting gene function or expression that may be associated with disease risk or drug response (the minority). The development of algorithms for automated variant annotation,which link each variant with information that may help predict its molecular and phenotypic impact, is a critical step towards prioritizing variants that may have a functional impact from those that are harmless or have irrelevant functional effects. In this section we review the key concepts and existing approaches in this important field. In Chapter **??** we introduce an approach to collect relevant information that will help answer questions about

genetic variants discovered in next-generation sequencing studies, including: (i) will a given coding variant affect the ability of a protein to carry its functions; (ii) will a given non-coding variant affect the expression or processing of a given gene; and ultimately (iii) will a given coding or non-coding variant have any impact on phenotypes of interest?

Answering these questions is essential for many types of analyses that use large-scale genomics datasets to study quantitative traits and diseases, particularly when only a small number of individuals is studied comprehensively at a genome-wide level. For example, most genome-wide association studies (GWAS) or exome sequencing studies lack the statistical power to identify rare variants or variants with small effects associated with a disease, in part due to the large number of variants assayed. This limitation can be addressed by directing both statistical analysis and subsequent experimental steps to focus on smaller sets of genetic variants that have been prioritized based on external evidence of their putative impact. The common impairment of DNA repair mechanisms and chromatin stability in malignant cells leads to a similar challenge in cancer genomics, where the hundreds or thousands of mutations that distinguish an individual's tumor and germline genomes need to be classified on the basis of their putative phenotypic effects and potential roles in carcinogenesis.

The large number of databases containing potentially helpful information about a given variant make the process of gathering and presenting relevant data challenging, despite excellent tools that already exist to analyze large genomics datasets (including GATK [73] and Galaxy [45]) and visualize the results (such as the UCSC [54] or Ensembl [41] genome browsers). Each of these databases uses its own format and is updated asynchronously, which makes it difficult for any analysis to remain up to date. In addition, the lack

of comprehensive and computationally efficient models that allow integrative analyses using these resources, makes the task of comprehensive variant annotation overwhelming. By efficiently combining information from tens or hundreds of genome-wide databases, the tools described here are designed to greatly facilitate the process of variant annotation, and make it accessible to groups with limited bioinformatics expertise or resources.

1.4.1 Variant types

Although variant calling is a challenging task and remains an important area of research, many high-quality tools exist for calling SNVs and indels. We discuss here the problem of annotating the variants identified by some of these tools. The most common type of variant identified by current technologies and analysis approaches is a single base difference with respect to the reference genome (SNV) followed by multiple base differences (MNP), as well as small insertions and deletions (InDels). Here, we focus on annotating those variants (or combinations of them, called "Mixed" variants), which comprise most of the variants in a typical sequencing experiment. We do not address the annotation or large rearrangements due to the challenges involved in their identification and functional characterization and their relative rarity in the germ line.

1.4.2 Types of genetic annotations

The process of genetic variant annotation consists of the collection, integration, and presentation of experimental and computational evidence that may shed light on the impact of each variant on gene or protein activity and ultimately on disease risk or other phenotypes. Variant annotation has traditionally been divided in two apparently independent but actually interrelated tasks based on the variant's location with respect to known protein-coding genes (see Table 1 for a list of commonly used variant annotations).Coding

variant annotation focuses on variants that are located within coding regions of annotated protein-coding genes and attempts to assess their impact on the function of the encoded protein. In contrast,non-coding variant annotation focuses on variants located outside the coding portion of genes (i.e. in intergenic regions, UTRs, introns, or non-protein-coding genes) and aims to assess their potential impact on transcriptional and post-transcriptional gene regulation. These two categories of variant annotations are not mutually exclusive, as variants located within exons can often have an impact on the gene transcript's processing (splicing). In addition, some transcripts can have both protein-coding and non-coding functions. Despite the intermingling of the notion of coding and non-coding variants, we will consider each type of annotation separately as assessing their impact requires different sources of data and algorithms.

The ultimate goal of variant annotation is to predict the impact of a sequence variant, although this is an ill-defined term. One the one hand, one may be interested in the molecular impact of a variant on the activity of a protein. On the other, others may be interested in a variant's impact on much higher-level phenotypes such as disease risk. Mutations that are predicted to completely abrogate a gene's activity are called loss-of-function (LOF) mutations. Those that are predicted to have less severe consequences are called moderate or low impact mutations. In practice, a variant will be predicted to cause LOF if it has two properties: (i) its molecular impact is reliably predictable by existing computational approaches (e.g. gain of stop-codon); and (ii) its functional impact, reflected by altered protein activity or expression levels, is expected to be large. Many types of variants, including most non-coding variants, may have a large functional impact but lack predictability, and as a consequence are typically not predicted to be LOF variants.

1.4.3 Coding variant annotation

Coding variants occur in translated exons. When a reliable gene annotation is available, their main impact can be classified by determining their effect on the translated amino acid sequence (if any). A synonymous coding variant (also called silent) does not change the sequence of amino acids encoded by the gene, although it may impact aspects of post-transcriptional regulation such as splicing and translation efficiency and can affect the total protein activity through changes in the amount of translated protein that is made in the cell. In contrast, a non-synonymous coding variant changes one or more amino acids encoded by the gene and can directly alter the protein's activity, localization or stability. Non-synonymous variants include missense substitutions that change a single amino acid, nonsense substitutions that lead to the gain of a stop codon, frame-preserving indels that insert or delete one or more amino acids, and frame-shifting indels that may completely alter the protein's amino acid sequence. Primary annotation and assessment of impact, determines whether a variant falls in any of these categories.

Caveats

i *Gene misannotation.* Genomic variants that have a significant effect on a protein's expression or function represent a very small fraction of all variants. Assembly and gene annotation errors or genomic oddities that break classical computational models are also rare, but lead to false positives. This implies that one is likely to find a non-negligible fraction of false-positive high-impact variants among the list of what appear to be the strongest candidates for variants with severe effects. Tools such as SnpEff can anticipate some of the most common causes of misannotation, but the number and diversity of the type of events that can lead to false-positives makes the task very challenging. As a consequence, one should

always manually inspect the top candidates to ensure that they have been assigned to the correct genes and transcripts.

ii *Gene isoforms.* In higher eukaryotes, most genes have more than one transcript (or isoform), due to alternative promoters, splicing, or polyadenylation sites. For example, a human gene has an average of 8.8 annotated messenger RNA (mRNA) isoforms and some genes are believed to have over 4,000 isoforms resulting from complex splicing programs. For these genes, a variant may be coding with respect to one mRNA isoform and non-coding with respect to another. There are two frequent approaches to address this situation: (i) annotate a variant using the most severe functional effect predicted for at least one mRNA isoform; or (ii) use only a single canonical transcript per gene to perform primary annotation.

iii *Variant calling for indels.* Variant annotation relies on knowing the exact genomic coordinates of the variant: this is rarely a problem for isolated SNVs; however, insertions and deletions often cannot be located unambiguously. Consider for example the variant $AA \rightarrow A$. This mutation results in the loss of a single base, but was it the first or second A that was deleted? From the standpoint of the cell, this question is irrelevant and deletion of any A will have the same effect. In contrast, from the standpoint of most variant annotation software, deleting the first A is different from deleting the second. Consider the scenario of a previously annotated transcript where the first A is part of the 5' UTR and the second is the first base of a start codon. If the missing base is assigned to the leftmost position in the motif (as is the current convention), the deletion would be annotated as a low impact 5'UTR variant. However, assigning it to the rightmost A would make it appear (incorrectly) to be

a high-impact start-codon deletion. Similar issues may arise when considering conservation scores or transcription factor binding site (TFBS) predictions.

1.4.4 Loss of function variants

True LOF variants are difficult to predict computationally, but specific types of genetic changes will frequently lead to severely impaired protein activity. These include (i) stop-gains (nonsense mutations) and start-loss; (ii) indels causing frameshifts; (iii) large deletions that remove either the first exon or at least 50% of the protein coding sequence; and (iv) loss of splice acceptor or donor sites that alter the protein-coding sequence. Variants that introduce premature in-frame stop codons (nonsense mutations and most frameshift indels) are expected to abolish protein function, unless the variant is very near the C-terminus of the coding region [102] (effectively, downstream of the last functional domain in the protein). Such mutations may have severe consequences in affected cells, tissues or organism, as is seen for mutations that cause monogenic diseases [88]. In addition, a new stop codon that lies upstream of the last exon will likely trigger nonsense mediated decay (NMD), a process that degrades mRNA before protein synthesis occurs [75]. NMD predictions are not exact and many factors can affect mRNA degradation, including the variant's distance from the last exon-exon junction or poly-A tail, and the possibility that transcription may re-initiate downstream of the LOF variant [15].

A variant that leads to the loss of a stop codon, sometimes called read-through mutation, will result in an elongated protein-coding transcript that terminates at the next in-frame stop codon. While there are no general models that predict how deleterious this may be, such variants can also result in aberrant folding and degradation of the nascent proteins, leading to activation of

cellular stress response pathways, in addition to their direct effects on protein activity and expression levels [88].

The effect of the loss of a start codon depends on the location of a replacement start codon with respect to the translation start site and reading frame of the native protein. If the new start codon maintains the reading frame, the only consequence may be the loss of a few amino acids in the protein transcript; however, in many cases, the new start codon will not be in-frame, thus producing a frame-shifted protein that is later degraded. In addition, the new start codon may lack an appropriate regulatory context (for example, if there is no Kozak sequence nearby or if it disrupts 5' UTR folding) leading to reduced expression of an N-terminally truncated protein. Consequently, losing a start codon is thought to be highly deleterious in most cases, due to the potential that it may reduce both protein production and activity.

Caveats

i *Rare amino acids.* Through a process called translational recoding, a UGA "Stop" codon located in the appropriate mRNA context (determined by both primary mRNA sequence and secondary structure) may be translated to incorporate a selenocysteine amino acid (Sec / U) [4]. In humans, it is known to occur 100 codons located in mRNAs whose 3' UTR contains a Selenocysteine insertion sequence element (SECIS). Since the translation machinery goes so far to encode these special rare amino acids, the expectation is that mutations at those sites would be highly deleterious. This is supported by evidence that reduced efficiency of selenocysteine incorporation is linked to severe clinical outcomes, such as early onset myopathy [72] and progressive cerebral atrophy [3].

ii *False-positives in LOF predictions.* Variants predicted to result in a LOF sometimes actually produce proteins that are partially functional

[71]. In fact, an apparently healthy individual is typically heterozygous for around 100 predicted LOF variants, and homozygous for roughly 10, but many of those are unlikely to completely abolish the protein function. Indeed, these variants are enriched toward the 3' end of the gene, where they are likely to be less deleterious.

1.4.5 Variants with low or moderate impact

Compared to the high impact variants discussed above, where extensive prior biological evidence strongly suggests that a specific type of variant will severely impair protein activity, there are few guidelines that can reliably predict how the majority of nonsynonymous (missense) variants will alter protein function or expression. As a result, the primary annotation performed by SnpEff and most related software packages will broadly categorize missense substitutions and their accompanying amino acid changes (e.g. $K154 \rightarrow L154$) as moderate impact variants. Short indels whose length is a multiple of three are treated similarly, unless they introduce a stop codon, as their effect will usually be localized.

Once missense and frame-preserving indel variants are identified, a more detailed estimation of their impact on protein function can be performed using heuristic and statistical models. The most common approaches are based on conservation, either amongst orthologous or homologous proteins, or protein domains, sometimes adding information of the physio-chemical properties of the reference and variant amino acids (e.g. differences in side chain charge, hydrophobicity, or size). The SIFT algorithm [57] assesses the degree of selection against specific amino acid changes at a given position of a protein sequence by analyzing the substitution process at that site throughout a collection of predicted homologous proteins identified by PSI-BLAST [6]. Based on this multiple sequence alignment and the highly conserved regions it contains,

SIFT calculates a normalized probability of amino acid replacement (called the SIFT score), which estimates the mutation's effect on protein function. Polyphen [2], another commonly used tool, takes the process one step further by searching UniProtKB/Swiss-Prot [28] and the DSSP database of secondary structure assignments [53] to determine if the variant is located in a known active site in the protein. In contrast to other methods that categorize each variant individually, VAAST [85], a commercially available package, computes scores for groups of variants located within a given gene and "collapses" them into a single category, a concept similar to burden testing performed for rare variants identified in exome sequencing studies. For human proteins, SnpEff makes use of the Database for Nonsynonymous SNVs' Functional Predictions [69] (dbNSFP), which collects scores produced by several impact assessment algorithms in a single database. Individually, impact assessment methods usually have an estimated accuracy of 60% to 80% when compared to manually curated databases of human variants, but predictions from several algorithms can be combined to provide a stringent, but more accurate estimate of impact [19].

In most cases these algorithms apply best to SNVs since these are common in populations and there is more genomic sequence and experimental data available to refine the statistical methods. However, some recently developed algorithms are capable of assessing variants other than SNVs, including PROVEAN [19], which extends SIFT to assess the functional impact of indels.

Caveats

i *Imprecise models of protein function.* Accurate impact assessment of coding variants remains an open problem and most computational predictions are riddled with both false positives and false negatives. While both missense variants and frame-preserving indels are broadly cataloged

as having moderate effects, this is mostly due to lack of a comprehensive model and the extremely complex computations that would be required for an in-depth analysis (such as protein structure predictions). In these cases, proteomic information can be revealing. SnpEff adds annotations from curated proteomic databases, such as NextProt [58], which can help to elucidate if a mutation alters a critical protein amino acid or domain (such as amino acids that are post-translationally modified as part of a signaling cascade or that are form the active site of an enzyme) resulting in a protein may no longer function.

ii *Gain of deleterious function.* Computational variant annotation may eventually be able to fairly accurately predict the molecular impact of a variant in terms of the degree to which it translates in a loss of function for the encoded protein. However, gains of function, including the acquired ability to interact with new partners and disrupt their function, remain vastly more difficult to tackle, although several such variants have been linked to disease [99].

iii *Unanticipated effects of synonymous variants.* In most cases, synonymous variants are regarded as non-deleterious (or low impact); however, one needs to seriously consider the possibility that they may have greater functional effects by altering mRNA splicing [32] or secondary structure [86]. Synonymous SNVs may also alter translation efficiency, by changing a frequently used to a rarely used codon and have been linked to changes in protein expression [87].

1.4.6 Non-coding variant annotation

Although coding variants represent less than 2% of variants in the human genome, they make up the vast majority of confirmed disease-related variants

that have been validated at a functional level. This may result from ascertainment bias (since variants in coding regions are straightforward to discover and characterize at a basic level and many studies have largely ignored non-coding variants); or may be explained by the increased complexity of computational approaches and lab assays required to predict and validate the impact of non-coding variants; or by their potentially more subtle impact on gene expression or cell function. Nonetheless, in a compendium of current GWAS studies, roughly 40% of the variants are intergenic and 30% intronic. Functional studies of these variants are increasingly emphasizing the importance of non-coding genetic variation at risk loci for complex genetic diseases and traits [51].

Functional non-coding regions of the genome encompass a wide variety of regulatory elements contained in DNA and RNA molecules that are involved in transcriptional and post-transcriptional regulation. Cis-regulatory elements include (i) binding sites for DNA-binding proteins such as transcription factors and chromatin remodelers; (ii) binding sites for RNA-binding proteins involved in splicing, mRNA localization, or translational regulation; (iii) micro RNA (miRNA) target sites; and (iv) long non-coding RNA (lncRNA) targets on DNA, RNA and proteins. Non-coding transcripts include well-characterized regulatory RNAs (e.g. miRNA, snoRNA, snRNA, piRNA and lncRNAs) as well as RNAs involved directly in protein synthesis (e.g. tRNA and rRNA). The annotation and impact assessment of non-coding variants presents a significant challenge for several reasons: (i) reliable technologies to study transcriptional regulatory regions on a genome-wide basis are only just reaching maturity and provide limited resolution of binding sites for individual transcription factors and regulatory RNA molecules; (ii) non-coding functional regions of most genomes remain incompletely mapped as they vary widely among different cell types and cell states (for example, in diseased and

healthy tissues); (iii) non-coding regulatory elements often are part of complex transcriptional programs that are time-dependent, contain many redundant linkages or reciprocal connections between genes and respond to a wide range of intraand extracellular signals; and (iv) genomic regulatory elements rarely have a strict consensus sequence (for example, compare the position weight matrices used to identify transcription factor or miRNA binding sites with the amino acid triplet code) making the effect of a mutation on gene regulatory programs difficult to predict. As a result, high-quality annotation of non-coding variants relies more heavily on experimental data than is the case for coding variants: since many of these experimental techniques did not study the effects of SNVs on gene regulatory programs, they can only be used to annotate variants and not to predict their effects on gene transcription. In the few cases where the effects of SNVs have been studied (for example, the effects of SNVs that are common in a population and located in genetic loci associated with complex diseases), experimental approaches provide highly accurate functional assessment at a cost of reduced coverage compared to computational approaches.

Large-scale projects such as ENCODE [27] and modENCODE [18] have made major steps toward mapping gene transcription and transcriptional regulatory regions in many tissues and cell types, but similar studies in diseased tissues remain at an early stage (for example, the growing collection of disease-related epigenomes from the Epigenome Roadmap [12]). The base-by-base resolution and number of cell states studied for different types of regulatory elements and non-coding transcripts varies widely among datasets; in part due to the lack of sensitive, comprehensive and high-resolution technologies to study the different molecular species and modes of interaction that can be

altered by non-coding variants. Efficient technologies for genome-wide, high-throughput mapping of binding sites for RNA-binding proteins (PAR-CLiP [8]), miRNAs (PAR-CLiP [47] and CLASH [50]) are starting to be applied on a broad scale as are protocols to map transcription factor binding sites (TFBS) which can improve resolution to a single base (Chip-exo [83]). However, in most cases, DNA and RNA binding sites are only imprecisely located within Chip-Seq peaks that span genomic regions hundreds of base pairs in length, with computational approaches being used to pinpoint the bases most likely mediating the interaction. In the absence of more precise localization data, de novo computational prediction of binding sites for DNA and RNA binding proteins remains insufficiently accurate to be of much use in annotating single noncoding variants.

This limitation is particularly critical for functional predictions of putative target sites for microRNAs and other regulatory RNA species. MicroRNAs are short RNA molecules that regulate gene expression post-transcriptionally by binding the messenger RNA of a gene through complementary, usually in the 3' region of the transcript, which leads to mRNA degradation or inhibits translation. Sequence variants that cause the loss or gain of a miRNA target site would lead to dysregulation of the gene, with likely deleterious effects. Although miRNAs are relatively well documented in most model organisms including human, their binding sites are only starting to be mapped experimentally, and computational predictions has very low specificity. Meaningful information regarding the possible role of a variant in disrupting a miRNA target site is starting to emerge [68], although variants that create new miRNA binding sites remain under the radar.

Even if the position of a functional element could be perfectly determined, predicting a variant's impact on chromatin conformation, promoter activity,

gene expression, or transcript processing remains challenging. For transcription factors, this involves predicting whether the protein will still be able to recognize its mutated site (and with what affinity), as well as predicting the impact of these changes on gene expression levels. The latter is particularly hard to predict as a result of interactions, competition, and redundancy contained in regulatory networks of transcription factors or RNA binding proteins. As a consequence, computational prediction of the functional impact of non-coding variants remains a very active area of research and there is no broad consensus on the best methodology to use [97]. One significant exception is the identification of variants affecting canonical splice sites, defined as two bases on the 3' end on the intron (splice site acceptor) and 5' end of the intron (splice site donor). Variants that affect canonical splice sites are easily detected and typically lead to abnormal mRNA processing, involving exon loss or extension that leads to loss of function of the encoded protein.

1.4.7 Impact assessment of non-coding variants

Two broad classes of publicly available genome-wide datasets are commonly combined to assess the functional impact of non-coding genetic variants: (i) computational predictions of sequence conservation and sites involved in molecular interactions such as transcription factor and RBP binding, as well as miRNA-mRNA target interactions; and (ii) experimental genome-wide localization assays for DNA binding proteins, histone modifications, and chromatin accessibility.

Computational sources of evidence:. Interspecies sequence conservation plays a key role in scoring and prioritizing non-coding variants. This is based on the assumption is that sites or regions that have been more conserved across species than expected under a neutral model of evolution are likely to be functional; suggesting that mutations contained in them are likely

to be deleterious. In the absence of strong experimental data, sequence conservation measures calculated from whole genome multiple alignments, (for example using PhastCons [90], SciPhy [43], PhyloP [80] , and GERP [35]), have been developed to provide a generic indicator of function for non-coding variants. Although high conservation scores generally mean that a genomic region may be functional, the converse is not true and many experimentally proven functional noncoding regions show only modest sequence conservation (for example due to binding site turnover events). Finally, some regulatory regions (e.g. specific elements regulating immune response [81]) are under positive selection and may thus show less conservation than surrounding neutral regions.

In human, genome-wide computational predictions of transcription factor binding sites based on matching to publicly available position weight matrices are available from variety of sources, including Ensembl [41] and Jaspar [16]. Because of the low information content of most binding affinity profiles, the specificity of the predictions is generally very low. Related approaches exist to predict splicing regulatory regions [39] and miRNA target sites [106], some of which are precomputed for whole genomes and available from the UCSC or Ensembl genome browsers. Recent efforts to determine RNA-binding protein sequence affinities can also be used to identify putative binding sites for these proteins in mRNA [82].

Experimental sources of evidence:. To investigate the potential impact of variants on transcriptional regulation, many published experimental data sets produced by large-scale projects such as ENCODE [27], modENCODE [18] and Roadmap Epigenomics [12], can be used directly by annotation packages. These include: (i) ChIP-seq or ChIP-exo experiments that identify

TFBSs on a genome-wide basis; (ii) DNAseI hypersensitivity or Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) assays that identify regions with open chromatin; and (iii) ChIP-seq studies to identify the presence of specific promoter or enhancer-associated histone post-translational modifications, which can be combined to identify active, poised, and inactive enhancers and promoters [82]. Most of these data sets are easily available through Galaxy [45] (as tracks from the UCSC Genome Browser) or through SnpEff (as downloadable pre-computed datasets). In parallel with the types of studies described above, expression quantitative trait loci (eQTLs) represent an agnostic way to map putative regulatory regions. An increasing number of such loci are available through the GTEX database [70]. Experimental data that may support assessment of the impact of variants on post-transcriptional regulation remain sparser, although databases such as doRiNa [7] or starBase [103] contain genome-wide datasets obtained by CLIP-Seq and degradome sequencing. To our knowledge, these data have yet to be used in the context of variant annotation studies.

Combining sources of evidence:. Despite the variety of computational and experimental sources of evidence available, impact assessment for non-coding variants remains relatively crude, due to the fact that biological models of gene regulation remain fairly simple. Nonetheless, significant steps forward have been made recently, and two web-based tools, HaploReg [96] and RegulomeDb [13], perform SNV and indel impact assessment for variants from dbSNV on the basis of a broad body of computational and experimental evidence. Both use pre-computed scores for variants from dbSnp and therefore cannot be used for rare variants, but they are extremely valuable for exploration by associating the variant of interest with a variant in dbSnp via linkage disequilibrium.

Caveats

i *Sparseness of functional sites within ChIP-seq peaks.* Even if a noncoding variant is located in a region that contains a ChIP-seq peak for a given TF and has all the hallmark signatures of regulatory chromatin, the likelihood that it is deleterious remains low, because most DNA bases contained within a peak are non-functional.

ii *Gain of function mutations.* While this section has focused on variants causing the loss of a functional regulatory element, genetic variants may also create new or more effective transcription factor binding sites. These are substantially harder to detect as they can occur in regions that show no evidence of function in individuals possessing the reference allele, and show little conservation across species. Furthermore, computational methods to predict gain of affinity for a given TF caused by a variant have insufficient specificity to be of much use on their own.

1.4.8 Clinical effect of variants

One of the most revealing types of annotation of both coding and non-coding variants reports whether the variant has previously been implicated in a phenotype or disease. Although such information is available for only a small minority of all deleterious variants, their number is growing and should be the first type of annotation one seeks out. Clinical annotations, until recently, have been scattered in a large number of specialized databases of medical conditions with a genetic basis, including the comprehensive, manually curated collection of genetic loci, variants and phenotypes in the Online Mendelian Inheritance in Man database (OMIM, www.omim.org); web pages containing detailed clinical and genetic information about uncommon disorders in the Swedish National Board of Health and Welfare Database for Rare Diseases (www.socialstyrelsen.se/rarediseases) and the peer-reviewed

NIH GeneReviews collection [16] (www.ncbi.nlm.nih.gov/books/NBK1116); and a curated collection of over 140,000 mutations associated with common and rare genetic disorders in the commercial Human Gene Mutation Database (HGMD, www.hgmd.org/). In most cases, these datasets do not use standardized data collection or reporting formats; are designed to primarily provide information to patients and health professionals through a web interface; and rely on heterogeneous criteria to describe disease phenotypes and clinical outcomes; pathological and other clinical laboratory data; as well as the genetic and biologic experiments that have been used to demonstrate disease mechanisms at a molecular or cellular level. These shortcomings are being addressed by initiatives that provide centralized, evidence-based, comprehensive collections of known relationships between human genetic variants and their phenotype that are suitable for computational analysis, such as the NIH effort to aggregate records from OMIM, GeneReviews and locus-specific databases in ClinVar (www.ncbi.nlm.nih.gov/clinvar).

Another important application of variant detection and annotation is in the study of cancer genomes, which is occurring increasingly in clinical settings to support treatment decisions for advanced tumors. Annotation of variants detected in tumor sequences can be analyzed for clinical cohorts, using similar techniques as other complex traits, as well as for individual patients, using techniques to identify differences between somatic (tumor) and germline (healthy) tissues. In the latter case, one looks for cancer-associated mutations that distinguish the somatic genome of cancer cells of an individual from the germline genome in order to find the driving mutations that pinpoint the specific mechanisms underlying tumorigenesis or metastasis. Ideally, these mutations can be used to select a treatment for the patient, establish prognosis, or to identify causative mutations that have led to the cancer's progression.

In such a setting, given that sequence differences between the cancer and germline genomes are of greater interest than the background genetic changes between the germline and a reference genome, variant calling is performed using specialized algorithms, such as MuTect [20] and SomaticSniper [61].

Caveats

i *Annotation accuracy.* Biological knowledge, as well as molecular and phenotypic evidence supports the identification of certain groups of high impact variants based on simple criteria (such as premature stops, frameshifts, start lost and rare amino acid mutations); however, it is often hard to predict whether non-synonymous variants will have equally large effects on an organism's health. Even when the accepted "rules of thumb" used in the primary annotation indicate that protein function is impaired, we should consider that these predictions may be based on a small number of model genes and will require appropriate wet-lab validation or confirmatory studies in cohorts. In addition, as more human genomes are sequenced, it is likely that some genetic variants that have been linked to Mendelian diseases will be found in healthy individuals [84]; and in many cases, may not actually be disease-causing mutations [11].

1.4.9 Data structures and computational efficiency

Most computational pipelines for genomic variant annotation and primary impact assessment are relatively efficient and can annotate variants obtained from large resequencing projects involving thousands of samples within a few minutes or hours even using a moderately powered laptop. This is typically achieved through two key optimizations: (i) creation of reference annotation databases and (ii) implementation of efficient search algorithms. Reference database creation refers to the process of creating and storing precomputed

genomic data from the reference genome, which can be searched quickly to extract information relevant to each variant. This process needs to be performed only once per reference genome and most annotation tools have pre-computed databases for many organisms available for users to download.

Since these databases are typically quite large, efficient search algorithms are used together with appropriate data structures to optimize the search process. In ANNOVAR [95], each chromosome is subdivided in a set of intervals of size k and genomic features for a given chromosome are stored in a hash table of size L/k, where L is the length of the chromosome. Another approach, used by SnpEff, is to use an "interval forest", which is a hash of interval trees [31] indexed by chromosome. Querying an interval tree requires $O[log(n) + m]$ time, where n is the number of features in the tree and m is the number of features in the result.

1.4.10 Conclusions

In Chapter **??** we introduce SnpEff & SnpSift, two approaches we designed for efficiently performing functional annotations of sequencing variants. These packages allow annotating, prioritizing, filtering and manipulating variant annotations as well as combining several public or custom-created databases. It should be noted SnpEff was one of the first annotation packages and has become one of the most widely used annotation software in both research and clinical environments.

1.5 Genome wide association studies

A genome wide association study aims at identifying genetic variants associated to a particular phenotype. First, the genomes (or exome, depending on the study design) of affected individuals (cases) and healthy individuals (controls) need to be sequenced, variants called, annotated and filtered. Then, the goal is to find variants that exhibit some statistical association with the

trait or phenotype of interest, which could be a disease status (e.g. diabetic vs healthy), a biomedical measurement (e.g. cholesterol level), or any measurable characteristic (e.g. height). Since the genome is so large, patterns of mutations that suggest correlation may be encountered by chance, so we need to establish statistical significance in order to distinguish true association from spurious ones. Like most studies, we will focus our discussion on SNVs, but most methods can be extended to other genomic variants.

1.5.1 Single variant tests and models

Consider a simple situation where there is only one variant in the whole genome for the cohort we are analysing. Since each individual has two sets of chromosomes, the variant can be present in one, both, or neither chromosomes, so the number of times a non-reference allele is present in an individual, is $N_{nr} = \{0, 1, 2\}$.

When the trait of interest is binary (e.g healthy vs disease), a cohort can be divided into cases and controls and we can build a 3 by 2 contingency table:

	$HomozygousReference$ ($N_{variant} = 0$)	$Heterozygous$ ($N_{nr} = 1$)	$Homozygousnon-reference$ ($N_{nr} = 2$)
$Cases$	$N_{ca,ref}$	$N_{ca,het}$	$N_{ca,hom}$
$Controls$	$N_{co,ref}$	$N_{co,het}$	$N_{co,hom}$

Further assumptions about how many variants are required to increase disease risk can reduce this 3×2 table to a 2×2 table. In the "dominant model", the effect of a mutated gene dominates over the healthy one, so one variant is enough to increase risk. The opposite, called "recessive model", is when both chromosomes have to be mutated in order to increase risk [9, 24]. In these models, we can count how many cases and controls have at least one variant (dominant model) or two variants (recessive model). This simplifies

the previous table, yielding a 2×2 contingency table, than can be tested using either a χ^2 test or a Fisher exact test [9].

Two other commonly used models, are the "multiplicative" and the "additive" models [9, 24]. In these models, a disease risk is assumed to be multiplied (or increased) by a factor γ with every variant present. We cannot simplify the contingency table, so we assess significance using a Cochran-Armitrage test [24].

1.5.2 Multiple variant tests

In a real case scenario there are thousands or millions of variants. We can extend the concept shown in the previous section by performing individual tests for each variant present in the cohort. Multiple testing can be addressed either by performing a correction, such as False Discovery Rate [9, 24], or using a stricter genome wide significance level. There are 3×10^9 bases in the genome, but taking into account the correlation between nearby variants (linkage disequilibrium), the genome wide significance level is generally accepted to be $p_{value} \leq 10^{-8}$.

In order to check if the null hypothesis of a significance tests is adequate, a QQ-plot is used (i.e. plotting the $y = -log(p_{value})$ vs $x = -log[rank(p_{value})/(N+1)]$, where N is the total number of variants). Adherence of the p-values to a 45 degree line on most of the range implies few systematic sources of association [9, 24]. If the p-values have a higher slope than the $y = x$ line, there might be "inflation", possibly due to co-factors, such as population structure (see section 1.5.4). If the inflation is not too high (e.g. less than 5%), this bias can be corrected by shifting the p-values towards the 45 degree slope. More sophisticated methods are explained in section 1.5.4.

1.5.3 Continuous traits and correcting for co-factors

Methods described so far are suitable for binary "traits" or "phenotypes". Statistical methods that link genetic information to traits can also be used on continuous or "quantitative" traits (e.g. weight, height, cholesterol level, etc.). A linear regression can be used assuming the traits are approximately normally distributed [9, 24]. A significance test (p_{value}) for linear models can be calculated using an F statistic, but more sophisticated methods are also available [9, 24].

Using linear models, it is easy to include known co-factors to correct for biases or inflation. For instance, if it is known that a risk increases with age or that males are more susceptible than females, age and sex can be added to the linear equation in order to correct for these effects [9, 24]. In a similar manner, we can add co-factors to binary traits using logistic regression.

1.5.4 Population structure

It is widely accepted that humans started in Africa and migrated to Europe, then to Asia and later to America [48]. Out of an initial population, a few individuals migrate and colonize a new territory. This implies that the genetic variety of the new colony is significantly reduced, compared to the previous population, since the genetic pool is only a small "founder population". The "Out of Africa" hypothesis implies that each new migration produced a reduction in genetic variety, also known as a "population bottleneck" [48].

As we previously mentioned, each individual inherits two chromosome sets, a maternal and a paternal one. Through recombination a chromosome is formed by a crossover combining maternal and paternal chromosomes and then passed down, thus the offspring has two sets of chromosomes (one from each parent) that on average have half chromosome from each grandparent.

This breaking and shuffling of chromosomes every generation, increases genetic diversity. Nevertheless if variants are located nearby in the chromosome, the chances that they are broken apart by recombination event are smaller than if they are further away from each other. This produces a correlation of close variants or "linkage disequilibrium" (LD). Nearby highly correlated variants are said to be in the same "LD-block" [48]. If a population has low genetic variety, the LD-blocks are large. So African population has more variety (smaller LD-blocks) and conversely, European, Asian and Amerindian populations have less variety (larger LD-blocks) [48].

1.5.5 Population as confounding variable

Imagine that we have a cohort of individuals drawn from two populations (P_A and P_B) and that individuals in P_A have much higher risk of diabetes than individuals from P_B. Now imagine that individuals from P_A have a variant v_A more often, but v_A is actually neutral and has no health effects whatsoever. If we do not take into account population factors, our study would conclude that v_A is the cause of diabetes, just because we see v_A more often in affected individuals. In this case it is clear that population structure is a confounding variable. We could avoid this problem by analyzing each population separately [79], but this would cause a loss of statistical power since we have fewer samples.

A population that is a mixture of two or more population is known as an "admixed population". For instance the "African-American" population is a mixture of, roughly, 80% African and 20% European genomes [48, 9]. This means that analyzing a cohort of African-American individuals, we would get population structure as a confounding variable because of population admixture [48]. Obviously, in this case we cannot analyze each population separately, because each individual in the sample is a mixture of two populations.

The admixed population problem can be studied by performing a correction using the eigen-structure of the sample covariance matrix [79]. Samples can be arranged as a matrix C where each row is a sample and each column represents a position in the genome where there is a variant. The numbers $C_{i,j}$ in the matrix indicate the number of non-reference alleles in a sample (row) at a genomic position (column j). Since the allele can be present in zero, one, or two chromosomes in each individual, the possible values for $C_{i,j}$ are $\{0, 1, 2\}$. The covariance matrix is calculated as $M = \hat{C}^T.\hat{C}$, where $\hat{C}$ is the matrix C corrected to have zero mean columns. Usually, the first two to ten principal components of M are used as factors in linear models (see section 1.5.3) to correct for population structure [79].

Whether a cohort has any population structure and needs correction or not can be tested using two methods: a) plotting the projections of the first two principal components and empirically observing the number of clusters in the chart, or b) using a statistic of the eigenvalues of M based on Tracy-Widom's distribution [79].

1.5.6 Common and Rare variants

The "allele frequency" (AF) is defined as the frequency a variant appears in a population. Variants are usually categorized according to AF into three groups: i) Common variants ($AF \geq 5\%$), "low frequency" ($1\% < AF < 5\%$), and iii) "rare variants" ($AF < 1\%$). Common variants originated earlier in the population while rare variants are either relatively recent or selected against.

There are three main models for disease susceptibility [48, 44]:i) the Common-Disease-Common-Variant hypothesis (CDCV) assumes that if disease is common, it must be caused by a common variant; ii) the "infinitesimal hypothesis" proposes that there are many common variants each having small

risk effects; and iii) the Common-Disease-Rare-Variant hypothesis proposes that there exists many rare variants, each one having large risk effects.

1.5.7 Rare variants test

The "rare variant model" assumes that multiple rare variants have large effects on a trait. The problem is that, since these variants are rare, huge sample sizes are required for tests to identify statistically significant associations. To overcome this problem, methods known as "burden tests" collapse groups of rare variants that are hypothesised to have similar effect and perform statistical significance tests on the group [62]. An example of collapsing technique is to count the number or rare variant in a given window and apply a Fisher exact test, as shown in section 1.5.1. A limitation of some burden tests is that they implicitly assume that all rare variants have the same direction of effect, although in reality they might have no effect, be deleterious, or protective [62, 101].

Several techniques allow weighting rare variants by collapsing them using a kernel matrix. This allows to incorporate other information, such as allele frequency and functional annotations. It can be shown that the statistic induced by kernel weighting functions follows a mixture of χ^2 distributions and there is an efficient way to approximate it [62, 101], avoiding computationally expensive permutations tests.

1.6 Epistasis

In this section we introduced the basic concepts and methodologies used in GWAS. Although fairly mature, there is still heavy research and continuous improvement on GWAS statistical methods. Not only it is well known that traditional (i.e. single marker) GWAS methods fail under non-additive models [34], but also variants so far discovered using these methods do not account for all the expected phenotypic variance attributed to genetic causes (i.e. missing

heritability). As other authors pointed out [30, 107, 108], this might be because we need to look for epistatic variants which are not taken into account using these methods. In the next section, and in Chapter **??**, we cover the topic of epistatic GWAS analysis.

1.6.1 What is epistasis and why it is important

[Historical perspective] The term 'epistasis' was initially used in the context of Mendelian inheritance; environmental effects are relatively unimportant for Mendelian traits, so Ii individuals can be clearly assigned to one of a limited number of classes according to their phenotype. Here, epistasis was used to describe the situation in which the actions of one locus mask the allelic effects of another locus, in the same way that completely dominant alleles mask the effects of the recessive allele at the same locus. [17]

The term 'epistatic' was first used in 1909 by Bateson (1) to describe a masking effect whereby a variant or allele at one locus (denoted at that time as an 'allelomorphic pair') prevents the variant at another locus from manifesting its effect. [29] This was seen as an extension of the concept of dominance. There are, however, some problems with this definition, particularly when applied to binary traits. In human genetics, the phenotype of interest is often qualitative and usually dichotomous, indicating presence or absence of disease. [29] Mathematical models for the joint action of two or more loci usually focus on the penetrance, the probability of developing disease given genotype. [29] Suppose that a predisposing allele is required at both loci in order to exhibit the trait, i.e. one or more copies of both allele A and allele B are required. Then, when the effects of both loci are considered, we obtain the penetrance table shown in Table 2. In this table, the effect of allele A can only be observed when allele B is also present: without the presence of B, the effect of A is not observable. The effect at locus A would appear to be 'masked' by that at

locus B. [29] This leads to a situation that is not precisely analogous to that described by Bateson (1). In Bateson's (1) definition, it is clear that if factor B is epistatic to factor A, we do not expect factor A to also be epistatic to factor B. [29] Table 3 is usually assumed to correspond to a situation in which the biological pathways involved in disease influenced by the two loci are at some level separate or independent (5). [29]

[Epistasis definition] In this review, we provide a historical background to the study of epistatic interaction effects and point out the differences between a number of commonly used definitions of epistasis [29]

[Epistasis & Quantitative traits] In the case of QUANTITATIVE TRAITS, epistasis describes the general situation in which the phenotype of a given genotype cannot be predicted by the sum of its component single-locus effects1 [17] Epistatic QTL-mapping studies in model organisms have detected many new interactions and have therefore concluded that epistasis makes a large contribution to the genetic regulation of complex traits. [17]

[Epistasis examples: Non-human] Extensive work on the control of qualitative genetic variation has highlighted the biological importance of epistasis at a locus-by-locus' level. On the basis of this work, several classic genotype-phenotype patterns that are caused by epistasis such as comb type in chickens, coat colour in various animals, the BOMBAY PHENOTYPE in the ABO blood-group system in humans and kernel colour in wheat [17] In the case of quantitative genetic variation, several or many genes of largely unknown function combine with environmental influences to control trait variation. This is the case for many complex traits that are of medical relevance in humans or of economic importance in plants and livestock. [17] A clear example of this can be seen [in Fig A] which the dominant allele (I) at the KIT locus, which confers white-coat colour in the pig, is dominant over all alleles at the MC1R locus

(E), which confer a darker coat colour. The effects of the various alleles at the E locus can only be determined in individuals with the recessive genotype ii at the I locus. This example was classically termed 'dominant epistasis', which gives a segregation ratio of 12:3:1 for white:black:brown, respectively [17]

[Epistasis examples: Human] D-allele of the angiotensin I converting enzyme (ACE) gene and the C-allele of the angiotensin II type 1 receptor (AGTR1) gene3. The risk of myocardial infarction is significantly increased by the ACE D-allele in patients who carry that particular AGTR1 allele. [17]

[Epistasis & disease / GWAS] The extent to which epistasis is involved in regulating complex traits is not known, and so we cannot assume that epistasis will be found for every trait in every population. [17] However, we argue that epistasis has been overlooked for too long and that it now needs to be routinely explored in complex trait studies. [17] For complex traits such as diabetes, asthma, hypertension and multiple sclerosis, the search for susceptibility loci has, to date, been less successful than for simple Mendelian disorders. This is probably due to complicating factors such as an increased number of contributing loci and susceptibility alleles, incomplete penetrance, and contributing environmental effects [29] The presence of epistasis is a particular cause for concern, since, if the effect of one locus is altered or masked by effects at another locus, power to detect the first locus is likely to be reduced and elucidation of the joint effects at the two loci will be hindered by their interaction. [29]

[Epistasis and networks] Epistasis-nonlinear genetic interactions between polymorphic loci-is the genetic basis of canalization and speciation, and epistatic interactions can be used to infer genetic networks affecting quantitative traits. [52] DATASET: Here, we compared the genetic architecture of three Drosophila life history traits in the sequenced inbred lines of the Drosophila melanogaster

Genetic Reference Panel (DGRP) and a large outbred, advanced intercross population derived from 40 DGRP lines (Flyland)[52] Surprisingly, none of the SNPs associated with the traits in Flyland replicated in the DGRP and vice versa. However, the majority of these SNPs participated in at least one epistatic interaction in the DGRP.[52] Our analysis underscores the importance of epistasis as a principal factor that determines variation for quantitative traits and provides a means to uncover genetic networks affecting these traits. [52]

1.6.2 Epistasis in complex traits

1.6.3 Detecting Epistasis / interactions

1.6.4 Epistatic Evolution and CoEvolution

1.6.5 CoEvolutionary models

1.6.6 Epistatic GWAS

Genome wide association studies have traditionally focused on single variants or nearby groups of variants. An often cited reason for the lack of discovery of high impact risk factors in complex disease is that these models ignore loci interactions [30] which have recently been pointed out as a potential solution for the "missing heritability" problem [107, 108]. With interactions being so ubiquitous in cell function, one may wonder why they have been so neglected by GWAS. There are several reasons: i) models using interactions are much more complex [42] and by definition non-linear, ii) information on which proteins interacts with which other proteins is incomplete [94], iii) in the cases where there protein-protein interaction information is available, precise interacting sites are rarely known [94]. Taking into account the last two items, we need to explore all possible loci combinations, thus the number of N order interactions grows as $O(M^N)$ where M is the number of variants [36]. This requires exponentially more computational power than single loci models. This

also severely reduces statistical power, which translates into requiring larger cohort, thus increasing sample collection and sequencing costs [36].

In Chapter **??** we develop a computationally tractable model for analysing putative interaction of pairs of variants from GWAS involving large case / control cohorts of complex disease. Our model is based on analysing cross-species multiple sequence alignments using a co-evolutionary model in order to obtain informative interaction prior probabilities that can be combined to perform GWAS analysis of pairs of non-synonymous variants that may interact.

The definition of epistasis from a statistical perspective is a "departure from a linear model" [30]. This means that in a logistic regression model the input for sample s includes terms with each of the genotypes at loci i and j), as well as an "interaction term" $g_{s,i} \cdot g_{s,j}$ [29].

$$\begin{aligned} P(d_s|g_{s,i}, g_{s,j}) &= \phi[\theta_0 + \theta_1 g_{s,i} + \theta_2 g_{s,j} + \theta_3 (g_{s,i} g_{s,j}) \\ &\quad ... + \theta_4 c_{s,1} + ... + \theta_m c_{s,N_{cov}}] \end{aligned}$$

where d_s is disease status, $\phi(\cdot)$ is the sigmoid function, $c_{s,1}, c_{s,2}, ...$ are covariates for sample s.

Models involving interactions between more than two variants can be defined similarly, but require more parameters and extremely large samples are required to accurately fit them.

Several families of approaches for epistatic GWAS exist. Here we mention a few:

- Allele frequency: In [1], an analysis of imbalanced allele pair frequencies is performed under the assumption that an implicit test for fitness can

be achieved looking for over/under-represented allele pairs in a given population. In another study [105] the authors infer that interactions can create LD in disease population under two-loci model, then they show how LD-based p-values can uncover interaction and sometimes (in their simulations) outperform logistic regression tests.

- Bayesian model: In [104], a "Bayesian partitioning model" is used by providing Dirichlet prior distributions for each partition and computing posterior probabilities using Markov chain Monte Carlo (MCMC) algorithms. The methodology first test individual makers and picks only the top 10% to further investigate for epistasis, because it is prohibitive to test all loci.
- Machine learning: From a machine learning point of view, finding interacting variants is simply an *"optimisation procedure is to find a set of parameters that allows the machine-learning model to most accurately predict class membership (e.g. affected vs unaffected)"* [74]. Several approaches have emerged to tackle the "interaction problem" and used a variety of different techniques [56, 74] , such as neural networks, cellular automata, random forests, multifactor dimensionality reduction, support vector machines, etc.

Although all these models have advantages under some assumptions, none of them seems to be a "clear winner" over the rest [30]. All of these models suffer from the increase in number of tests that need to be performed, which raises two issues: i) multiple testing, which is often resolved by stringent significance threshold, and ii) computational feasibility, which is solved by efficient algorithms, parallelization, and heuristic approaches to quickly discard uninformative loci combinations. So far, no method for epistatic GWAS has been widely adopted and there is need of different approaches to be explored. In

Chapter **??** we propose an approach to combine co-evolutionary models and GWAS epistasis of pairs of putatively interacting loci.

1.7 Thesis roadmap and Contributions

The original research presented in this thesis covers topics related to the computational and statistical methodologies related to the analysis of sequencing variants to unveil genetic links to complex disease. Broadly speaking, we address three types of problems: (i) Data processing of large datasets from high throughput biological experiments such as resequencing in the context of a GWAS (Chapter **??**); (ii) functional annotations, i.e. calculating variant's impact at the molecular, cellular or even clinical level (Chapter **??**); (iii) identification of genetic risk factors for complex disease using models that combine population-level and evolutionary-level data to detect putative epistatic interactions (Chapter **??**). When applicable, background material specific to each chapter is presented in a preface, together with an explanation of how that chapter ties in with the rest of the thesis.

This thesis comprises text and figures of articles that have either been published, submitted for publication, or ready to be submitted (waiting upon data embargo restrictions):

Chapter ??

1. **P. Cingolani**, R. Sladek, and M. Blanchette. "BigDataScript: a scripting language for data pipelines." Bioinformatics 31.1 (2015): 10-16.

   For this paper, PC conceptualized the idea and performed the language design and implementation. RS & MB helped in designing robustness testing procedures. PC, RS & MB wrote the manuscript.

Chapter ??

2. **P. Cingolani**, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain $w^{1118}; iso-2; iso-3$". Fly, 6(2), 2012.

For this paper, PC conceptualized the idea, implemented the program and performed testing. AP contributed several feature ideas, software testing and suggested improvements. XL, DR, SL, LW, TN, MC, LW performed mutagenesis and sequencing experiments. XL and DR performed the biological interpretation of the data. All authors contributed to the manuscript.

SnpEff's accompanying publication (SnpSift):

3. **P. Cingolani**, V. M. Patel, M. Coon, T. Nguyen, S. Land, D. M. Ruden, and X. Lu." Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift". Toxicogenomics in non-mammalian species, page 92, 2012.

We used SnpEff & SnpSift and developed a number of new functionalities in the context of two collaborative GWAS projects on type II diabetes:

4. M. McCarthy, T2D Genes Consortia. "Variation in protein-coding sequence and predisposition to type 2 diabetes", Ready for submission.

5. A. Mahajan, X. Sim, H. Ng, A. Manning, M. Rivas, H. Heather, A. Locke, N. Grarup, H. K. Im, **P. Cingolani**, et. al. "Identification and Functional Characterization of G6PC2 Coding Variants Influencing Glycemic Traits Define an Effector Transcript at the G6PC2-ABCB11 Locus." PLoS genetics 11.1 (2015): e1004876-e1004876.

Chapter ??

6. **P. Cingolani**, R. Sladek, and M. Blanchette. "A co-evolutionary approach for detecting epistatic interactions in genome-wide association studies". Ready for submission (data embargo restrictions).

For this paper, PC designed the methodology under the supervision of MB and RS. PC implemented the algorithms. PC, RS & MB wrote the manuscript. This work uses data from the T2D consortia, thus it cannot be published until the main T2D paper is accepted for publication (according to T2D data embargo).

Other contributions

During my thesis I have co-authored several other scientific articles (grouped by topic) published, submitted for publication, or ready to be submitted, not mentioned in this thesis:

Epigenetics

7. **P. Cingolani**, X. Cao, R. Khetani, C.C. Chen, M. Coon, A. Bollig-Fischer, S. Land, Y. Huang, M. Hudson, M. Garfinkel, and others. "Intronic Non-CG DNA hydroxymethylation and alternative mRNA splicing in honey bees." BMC genomics 14.1 (2013): 666.
8. M. Senut, A. Sen, **P. Cingolani**, A. Shaik, S. Land, Susan J and D. M. Ruden. "Lead exposure disrupts global DNA methylation in human embryonic stem cells and alters their neuronal differentiation." Toxicological Sciences (2014).
9. D. M. Ruden, **P. Cingolani**, A. Sen, W. Qu, L. Wang, M. Senut, M. Garfinkel, V. Sollars, X. Lu, "Epigenetics as an answer to Darwin's 'special difficulty' Part 2: Natural selection of metastable epialleles in honeybee castes", Frontiers in Genetics (2015).
10. M. Senut, A. Sen, **P. Cingolani**, A. Shaik, S. Land, Susan J and D. M. Ruden. "Lead exposure induces changes in 5-hydroxymethylcytosine clusters in CpG islands in human embryonic stem cells and umbilical cord blood", Submitted to 'Epigenomics.
11. M. Senut, **P. Cingolani**, A. Sen, Arko, A. Kruger, A. Shaik, H. Hirsch, S. Suhr, D. Ruden. "Epigenetics of early-life lead exposure and effects on brain development." Epigenomics 4.6 (2012): 665-674.

GWAS & Disease

12. K. Oualkacha, Z. Dastani, R. Li, **P. Cingolani**, T. Spector, C. Hammond, J. Richards, A. Ciampi, C. Greenwood. "Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness." Genetic epidemiology 37.4 (2013): 366-376.

13. S. Bongfen, I. Rodrigue-Gervais, J. Berghout, S. Torre, **P. Cingolani**, S. Wiltshire, G. Leiva-Torres, L. Letourneau, R. Sladek, M. Blanchette, and others. "An N-ethyl-N-nitrosourea (ENU)-induced dominant negative mutation in the JAK3 kinase protects against cerebral malaria." PloS one 7.2 (2012): e31012.

14. C. Meunier, L. Van Der Kraak, C. Turbide, N. Groulx, I. Labouba, Ingrid, **P. Cingolani**, M. Blanchette, G. Yeretssian, A. Mes-Masson, M. Saleh, and others. "Positional mapping and candidate gene analysis of the mouse Ccs3 locus that regulates differential susceptibility to carcinogen-induced colorectal cancer." PloS one 8.3 (2013): e58733.

15. G. Caignard, G. Leiva-Torres, M. Leney-Greene, B. Charbonneau, A. Dumaine, N. Fodil-Cornu, M. Pyzik, **P. Cingolani**, J. Schwartzentruber, J. Dupaul-Chicoine, and others. "Genome-wide mouse mutagenesis reveals CD45-mediated T cell function as critical in protective immunity to HSV-1." PLoS pathogens 9.9 (2013): e1003637.

16. M. Bouttier, D. Laperriere, M. Babak Memari, M. Verway, E. Mitchell, **P. Cingolani**, T. Wang, M. Behr, R. Sladek, M. Blanchette, S. Mader and J. White. "Genomics analysis reveals elevated LXR signaling reduces M. tuberculosis viability", Submitted to Journal of Clinical Investigation.

17. M. Bouttier, D. Laperriere, M. Babak Memari, M. Verway, E. Mitchell, **P. Cingolani**, T. Wang, M. Behr, R. Sladek, M. Blanchette, S. Mader and J. White. "Genomic analysis of enhancers engaged in M. tuberculosis-infected macrophages reveals that LXR signaling reduces mycobacterial burden", Submitted to PLOS Pathogens.

Fuzzy logic

18. **P. Cingolani** and Jesus Alcala-Fdez. "jFuzzyLogic: a robust and flexible Fuzzy-Logic inference system language implementation." FUZZ-IEEE. 2012.
19. **P. Cingolani** and Jesus Alcala-Fdez. "jFuzzyLogic: a java library to design fuzzy logic controllers according to the standard for fuzzy control programming." International Journal of Computational Intelligence Systems (2013), vol 6, pages 65-75.

References

[1] Marit Ackermann and Andreas Beyer. Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS genetics*, 8(2):e1002463, 2012.

[2] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, and S.R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[3] Orly Agamy, Bruria Ben Zeev, Dorit Lev, Barak Marcus, Dina Fine, Dan Su, Ginat Narkis, Rivka Ofir, Chen Hoffmann, Esther Leshinsky-Silver, et al. Mutations disrupting selenocysteine formation cause progressive cerebello-cerebral atrophy. *The American Journal of Human Genetics*, 87(4):538–544, 2010.

[4] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D Watson, and AV Grimstone. Molecular biology of the cell (3rd edn). *Trends in Biochemical Sciences*, 1995.

[5] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, et al. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[6] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[7] Gerd Anders, Sebastian D Mackowiak, Marvin Jens, Jonas Maaskola, Andreas Kuntzagk, Nikolaus Rajewsky, Markus Landthaler, and Christoph Dieterich. dorina: a database of rna interactions in post-transcriptional regulation. *Nucleic acids research*, page gkr1007, 2011.

[8] Manuel Ascano, Markus Hafner, Pavol Cekan, Stefanie Gerstberger, and Thomas Tuschl. Identification of rna–protein interaction networks using par-clip. *Wiley Interdisciplinary Reviews: RNA*, 3(2):159–177, 2012.

[9] D.J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[10] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. Exome sequencing as a tool for mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, 2011.

[11] Callum J Bell, Darrell L Dinwiddie, Neil A Miller, Shannon L Hateley, Elena E Ganusova, Joann Mudge, Ray J Langley, Lu Zhang, Clarence C Lee, Faye D Schilkey, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science translational medicine*, 3(65):65ra4–65ra4, 2011.

[12] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045–1048, 2010.

[13] Alan P Boyle, Eurie L Hong, Manoj Hariharan, Yong Cheng, Marc A Schaub, Maya Kasowski, Konrad J Karczewski, Julie Park, Benjamin C Hitz, Shuai Weng, et al. Annotation of functional variation in personal genomes using regulomedb. *Genome research*, 22(9):1790–1797, 2012.

[14] Sydney Brenner, AOW Stretton, and S Kaplan. Genetic code: the nonsensetriplets for chain termination and their suppression. 1965.

[15] Saverio Brogna and Jikai Wen. Nonsense-mediated mrna decay (nmd) mechanisms. *Nature structural & molecular biology*, 16(2):107–113, 2009.

[16] Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research*, 36(suppl 1):D102–D106, 2008.

[17] Örjan Carlborg and Chris S Haley. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625, 2004.

[18] Susan E Celniker, Laura AL Dillon, Mark B Gerstein, Kristin C Gunsalus, Steven Henikoff, Gary H Karpen, Manolis Kellis, Eric C Lai, Jason D Lieb, David M MacAlpine, et al. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.

[19] Yongwook Choi, Gregory E Sims, Sean Murphy, Jason R Miller, and Agnes P Chan. Predicting the functional effect of amino acid substitutions and indels. *PloS one*, 7(10):e46688, 2012.

[20] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–219, 2013.

[21] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):0–1, 2012.

[22] Pablo Cingolani, Viral M Patel, Melissa Coon, Tung Nguyen, Susan J Land, Douglas M Ruden, and Xiangyi Lu. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Toxicogenomics in non-mammalian species*, page 92, 2012.

[23] Pablo Cingolani, Rob Sladek, and Mathieu Blanchette. Bigdatascript: a scripting language for data pipelines. *Bioinformatics*, 31(1):10–16, 2015.

[24] G.M. Clarke, C.A. Anderson, F.H. Pettersson, L.R. Cardon, A.P. Morris, and K.T. Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133, 2011.

[25] FS Collins, ES Lander, J. Rogers, RH Waterston, and I. Conso. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[26] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[27] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[28] UniProt Consortium et al. Update on activities at the universal protein resource (uniprot) in 2013. *Nucleic acids research*, 41(D1):D43–D47, 2013.

[29] Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.

[30] Heather J Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[31] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, Clifford Stein, et al. *Introduction to algorithms*, volume 2. MIT press Cambridge, 2001.

[32] Jasmin Coulombe-Huntington, Kevin CL Lam, Christel Dias, and Jacek Majewski. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS genetics*, 5(12):e1000766, 2009.

[33] Francis Crick, Leslie Barnett, Sydney Brenner, and Richard J Watts-Tobin. *General nature of the genetic code for proteins*. Macmillan Journals Limited, 1961.

[34] Robert Culverhouse, Brian K Suarez, Jennifer Lin, and Theodore Reich. A perspective on epistasis: limits of models displaying no main effect. *The American Journal of Human Genetics*, 70(2):461–471, 2002.

[35] E.V. Davydov, D.L. Goode, M. Sirota, G.M. Cooper, A. Sidow, and S. Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS computational biology*, 6(12):e1001025, 2010.

[36] David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.

[37] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. Del Angel, M.A. Rivas, M. Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[38] R. Durbin. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.

[39] William G Fairbrother, Ru-Fang Yeh, Phillip A Sharp, and Christopher B Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–1013, 2002.

[40] P. Ferragina and G. Manzini. Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 390–398. IEEE, 2000.

[41] Paul Flicek, Ikhlak Ahmed, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, et al. Ensembl 2013. *Nucleic acids research*, page gks1236, 2012.

[42] Hong Gao, Julie M Granka, and Marcus W Feldman. On the classification of epistatic interactions. *Genetics*, 184(3):827–837, 2010.

[43] M. Garber, M. Guttman, M. Clamp, M.C. Zody, N. Friedman, and X. Xie. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25(12):i54–i62, 2009.

[44] G. Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145, 2012.

[45] Jeremy Goecks, Anton Nekrutenko, James Taylor, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.

[46] Anthony JF Griffiths. *An introduction to genetic analysis*. Macmillan, 2005.

[47] Markus Hafner, Steve Lianoglou, Thomas Tuschl, and Doron Betel. Genome-wide identification of mirna targets by par-clip. *Methods*, 58(2):94–105, 2012.

[48] D.L. Hartl and A.G. Clark. *Principles of population genetics*. Sinauer associates Sunderland, Massachusetts, 2007.

[49] David Haussler, Stephen J O'Brien, Oliver A Ryder, F Keith Barker, Michele Clamp, Andrew J Crawford, Robert Hanner, Olivier Hanotte, Warren E Johnson, Jimmy A McGuire, et al. Genome 10k: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, 100(6):659–674, 2009.

[50] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human mirna interactome by clash reveals frequent noncanonical binding. *Cell*, 153(3):654–665, 2013.

[51] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[52] Wen Huang, Stephen Richards, Mary Anna Carbone, Dianhui Zhu, Robert RH Anholt, Julien F Ayroles, Laura Duncan, Katherine W Jordan, Faye Lawrence, Michael M Magwire, et al. Epistasis dominates the genetic architecture of drosophila quantitative traits. *Proceedings of the National Academy of Sciences*, 109(39):15553–15559, 2012.

[53] Robbie P Joosten, Tim AH Te Beek, Elmar Krieger, Maarten L Hekkelman, Rob WW Hooft, Reinhard Schneider, Chris Sander, and Gert Vriend. A series of pdb related databases for everyday needs. *Nucleic acids research*, 39(suppl 1):D411–D419, 2011.

[54] Donna Karolchik, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The ucsc genome

browser database: 2014 update. *Nucleic acids research*, 42(D1):D764–D770, 2014.

[55] Martin Alexander Kennedy. Mendelian genetic disorders. *Encyclopedia of Life Sciences*, 2001.

[56] Ching Lee Koo, Mei Jing Liew, Mohd Saberi Mohamad, and Abdul Hakim Mohamed Salleh. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *BioMed research international*, 2013, 2013.

[57] P. Kumar, S. Henikoff, and P.C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073–1081, 2009.

[58] Lydie Lane, Ghislaine Argoud-Puy, Aurore Britan, Isabelle Cusin, Paula D Duek, Olivier Evalet, Alain Gateau, Pascale Gaudet, Anne Gleizes, Alexandre Masselot, et al. nextprot: a knowledge platform for human proteins. *Nucleic acids research*, 40(D1):D76–D83, 2012.

[59] B. Langmead and S.L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 2012.

[60] B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[61] David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, 2012.

[62] B. Li and S.M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[63] H. Li. Improving snp discovery by base alignment quality. *Bioinformatics*, 27(8):1157–1158, 2011.

[64] H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[65] H. Li and R. Durbin. Fast and accurate short-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(5), 2009.

[66] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589, 2010.

[67] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851, 2008.

[68] Chenxing Liu, Fuquan Zhang, Tingting Li, Ming Lu, Lifang Wang, Weihua Yue, and Dai Zhang. Mirsnp, a database of polymorphisms altering mirna target sites, identifies mirna-related snps in gwas snps and eqtls. *BMC genomics*, 13(1):661, 2012.

[69] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbnsfp: a lightweight database of human nonsynonymous snps and their functional predictions. *Human mutation*, 32(8):894–899, 2011.

[70] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[71] D.G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J.K. Pickrell, S.B. Montgomery, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, 2012.

[72] Baijayanta Maiti, Sandrine Arbogast, Valérie Allamand, Mark W Moyle, Christine B Anderson, Pascale Richard, Pascale Guicheney, Ana Ferreiro, Kevin M Flanigan, and Michael T Howard. A mutation in the sepn1 selenocysteine redefinition element (sre) reduces selenocysteine incorporation and leads to sepn1-related myopathy. *Human mutation*, 30(3):411–416, 2009.

[73] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[74] Brett A McKinney, David M Reif, Marylyn D Ritchie, and Jason H Moore. Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88, 2006.

[75] Eszter Nagy and Lynne E Maquat. A rule for termination-codon position within intron-containing genes: when nonsense affects rna abundance. *Trends in biochemical sciences*, 23(6):198–199, 1998.

[76] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[77] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

[78] Jason ORawe, Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, Lifeng Tian, Hakon Hakonarson, W Evan Johnson, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome med*, 5(3):28, 2013.

[79] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[80] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.

[81] Towfique Raj, Manik Kuchroo, Joseph M Replogle, Soumya Raychaudhuri, Barbara E Stranger, and Philip L De Jager. Common risk alleles for inflammatory diseases are targets of recent positive selection. *The American Journal of Human Genetics*, 92(4):517–529, 2013.

[82] Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, et al. A compendium of rna-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 2013.

[83] Ho Sung Rhee and B Franklin Pugh. Chip-exo method for identifying genomic location of dna-binding proteins with near-single-nucleotide accuracy. *Current Protocols in Molecular Biology*, pages 21–24, 2012.

[84] Erin Rooney Riggs, Karen E Wain, Darlene Riethmaier, Melissa Savage, Bethanny Smith-Packard, Erin B Kaminsky, Heidi L Rehm, Christa Lese Martin, David H Ledbetter, and W Andrew Faucett. Towards a universal clinical genomics database: the 2012 international standards for cytogenomic arrays consortium meeting. *Human mutation*, 34(6):915–919, 2013.

[85] A.F. Rope, K. Wang, R. Evjenth, J. Xing, J.J. Johnston, J.J. Swensen, B. Moore, C.D. Huff, L.M. Bird, J.C. Carey, et al. Using vaast to identify an x-linked disorder resulting in lethality in male infants due to n-terminal acetyltransferase deficiency. *The American Journal of Human Genetics*, 2011.

[86] Radhakrishnan Sabarinathan, Hakim Tafer, Stefan E Seemann, Ivo L Hofacker, Peter F Stadler, and Jan Gorodkin. The rnasnp web server: predicting snp effects on local rna secondary structure. *Nucleic acids research*, 41(W1):W475–W479, 2013.

[87] Zuben E Sauna and Chava Kimchi-Sarfaty. Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, 12(10):683–691, 2011.

[88] Gert C Scheper, Marjo S van der Knaap, and Christopher G Proud. Translation matters: protein synthesis defects in inherited disease. *Nature Reviews Genetics*, 8(9):711–723, 2007.

[89] Valerie Schneider and Deanna Church. Genome reference consortium. 2013.

[90] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

[91] Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[92] GATK team. The genome analysis toolkit. Accessed: 2015.

[93] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[94] Kavitha Venkatesan, Jean-Francois Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, et al. An empirical framework for binary interactome mapping. *Nature methods*, 6(1):83–90, 2009.

[95] K. Wang, M. Li, and H. Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[96] Lucas D Ward and Manolis Kellis. Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, 40(D1):D930–D934, 2012.

[97] Lucas D Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, 30(11):1095–1106, 2012.

[98] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[99] David C Whitcomb, Michael C Gorry, Robert A Preston, William Furey, Michael J Sossenheimer, Charles D Ulrich, Stephen P Martin, Lawrence K Gates, Stephen T Amann, Phillip P Toskes, et al. Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. *Nature genetics*, 14(2):141–145, 1996.

[100] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[101] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 2011.

[102] Yumi Yamaguchi-Kabata, Makoto K Shimada, Yosuke Hayakawa, Shinsei Minoshima, Ranajit Chakraborty, Takashi Gojobori, and Tadashi Imanishi. Distribution and effects of nonsense polymorphisms in human genes. *PloS one*, 3(10):e3393, 2008.

[103] Jian-Hua Yang, Jun-Hao Li, Peng Shao, Hui Zhou, Yue-Qin Chen, and Liang-Hu Qu. starbase: a database for exploring microrna–mrna interaction maps from argonaute clip-seq and degradome-seq data. *Nucleic acids research*, 39(suppl 1):D202–D209, 2011.

[104] Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39(9):1167–1173, 2007.

[105] Jinying Zhao, Li Jin, and Momiao Xiong. Test for interaction between two unlinked loci. *The American Journal of Human Genetics*, 79(5):831–845, 2006.

[106] Jesse D Ziebarth, Anindya Bhattacharya, Anlong Chen, and Yan Cui. Polymirts database 2.0: linking polymorphisms in microrna target sites with human diseases and complex traits. *Nucleic acids research*, page gkr1026, 2011.

[107] O. Zuk, E. Hechter, S.R. Sunyaev, and E.S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[108] Or Zuk, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander. Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.