1 From Genetic Data to Medicine: From DNA Samples to Disease Risk Prediction in Personalized Genetic Tests

Luis G. Leal, Rok Košir, and Nataša Pržulj

1.1 Background

The completion of the 1000 Genomes Project has given us a comprehensive insight into the variability of the human genome. On average, a typical human genome will differ from the reference genome in 4.1 to 5 million sites, the majority of which (86% on average) consist of single nucleotide polymorphisms (SNPs) [1]. SNPs are defined as locations in the DNA sequence where at least two different nucleotides appear in the human population [2]. They have been the focus of many studies, as their presence may have functional consequences: They may affect the transcription factor binding affinity, the mRNA transcript stability, and could produce changes in the amino acid sequences of proteins [3, 4]. These functional changes have effects on the predisposition of individuals to diseases, or the efficacy of drugs on patients.

Given that functional changes could increase predisposition to diseases, SNPs are used as genetic markers to identify genes associated with diseases. According to Szelinger et al. [5], a gene's function may be altered by SNPs at different levels. There are silent, or non-functional SNPs, which do not interfere with the functions of genes, SNPs which increase the risk of a disease, and SNPs having strong functional effects upon disease development (Mendelian disorders); however, only some hundreds of them are likely to contribute to disease risk [1, 6].

Detecting these genetic alterations is fundamental to understanding the development of diseases. With the advent of SNP microarrays, searching for inherited genomic variants was enabled for the first time and it boosted the relationship between computational methodologies and biological understanding [2]. It was not until the rise in next generation sequencing (NGS) and the increase in the density of SNP microarrays, that the SNP identification and genotyping tasks could be executed in mass. Both technologies have shifted the amount of data generated from single SNP studies to whole-genome analyses of multiple individuals at the same time (e.g., the Cancer Genome Atlas Project,[1] the NHLBI Exome Sequencing Project,[2] and the 1000 Genomes Project[3]) [7]. Accurate computational approaches are, however, needed to elucidate heterogeneous disorders from these raw data.

Genome-wide association studies (GWAS) are ideal for detecting novel SNP-disease associations, because disease predisposition can be closely related to the presence of genetic variants. A large number of susceptible loci for common complex diseases (e.g., heart disease, diabetes, obesity, hypertension, cancer) have been found in recent GWAS [8]. For example, genome-wide approaches are important to uncover multiple genetic alterations occurring in cancer development [9]. Different types of cancer, including breast cancer [10] and lung cancer [11], have been studied using this approach. Also, thanks to simultaneous genotyping of SNPs, we have broadened the understanding of diabetes [12], coronary artery disease [13], and hypertension [14], to name a few.

The number of published GWAS increases every year for a wide range of complex traits and different websites gather the data generated from these studies. The full catalog of GWAS is administrated by the National Human Genome Research Institute and the European Bioinformatics Institute (NHGRI-EBI).[4] Other public databases with relevant information are the Single Nucleotide Polymorphism database (dbSNP),[5] the Human Gene Mutation Datbase (HGMD)[6] and the Catalogue of Somatic Mutations in Cancer (COSMIC).[7]

A major purpose of GWAS is to formulate a predictive model based on SNPs for disease diagnostics. GWAS were conceived with the hope of revealing the genetic causes of complex diseases, in the same way that single SNPs driving Mendelian diseases (e.g., cystic fibrosis, hemophilia A, muscular dystrophy) were identified in the past with other approaches [15]. To date, the vast majority of the variants identified by GWAS explain only a fraction of disease heritability, in part because complex diseases have been shown to be the result of multiple interacting SNPs, also known as gene–gene epistatic interactions, and environmental factors [16, 17].

Genetic studies of complex diseases have been approached from two perspectives [18]. First, it is hypothesized that cumulative effects of common variants (i.e., SNPs with allele frequencies higher than 5% in the population) result in a complex disease, which is a focus of many GWAS. Second, it is hypothesized that low frequency variants (0.5%–5%) and rare variants (<0.5%) can have large effects resulting in a complex disease [18, 19]. Thus, the allele's frequency in the population and its effect size on

[1] http://cancergenome.nih.gov
[2] http://evs.gs.washington.edu/EVS
[3] www.1000genomes.org
[4] www.ebi.ac.uk/gwas
[5] www.ncbi.nlm.nih.gov/snp
[6] www.hgmd.cf.ac.uk
[7] http://cancer.sanger.ac.uk/cosmic

the disease are crucial not only to identify the origin of complex diseases, but also to determine the technology (e.g., rare variants can only be determined by NGS, not by microarrays) and the sample size in genetic studies (e.g., large samples are needed to find significant rare variants) [2]. While the amount of studies focused on the effects of low frequency and rare variants is still limited, there are already encouraging results coming from these studies. For example, rare variants associated with osteoporosis, type 2 diabetes, Alzheimer's disease, risk of heart attack, as well as several variants associated with lipid metabolism have been identified [20]. With the ever increasing number of genome projects worldwide based on sequencing we can expect these numbers to rise in the near future. For further information on rare and low frequency variants please refer to an excellent review by Bomba et al. [21].

Traditional univariate statistical methods are used to identify single SNP-disease associations [10, 11, 22]. The association tests examine each SNP independently for association to the disease by means of logistic regression models or contingency table methods when the trait is qualitative (e.g., case/control phenotype), or by means of analysis of variance (ANOVA) when the trait is quantitative (e.g., artery thickness) [3]. Even though these strategies are adequate to study single SNPs, detecting complex genetic architectures demands more sophisticated data-mining approaches [23]. Thus, new algorithms capable of discovering complex multigenic SNPs are being developed for mining data from GWAS studies [24, 25].

Thanks to the completion of the Human Genome Project, the technological advances to genotype SNPs and the detection of markers associated with complex traits via GWAS, new opportunities have appeared for the clinical translation of these discoveries to personalized medicine. In this way, genetic tests have enabled the confirmation or prediction of specific disorders by identifying changes in the chromosomes, DNA sequence, or gene products of individuals [26]. Genetic testing has grown to cover a wide range of variants, including variants associated with adult disease onset, drug dosage, and adverse reactions [27]. As it was envisioned some years ago, the accelerated improvement in genome sequencing techniques has brought GWAS results a step closer to the personal benefit of patients.

Personalized genetic tests (PGTs) have revolutionized our perception of healthcare services under the promise of accurate prediction of disease risk. PGTs are founded on the synergistic relationship between technological advances, medical knowledge, and computational methods, translating the best of them for the benefit of patients. Currently, PGTs can be indicated by health providers, but they also can be accessed through direct-to-consumer (DTC) providers. The DTC genetic testing is offered worldwide via the Internet by various companies; typically, after sending a saliva sample, the consumers receive a report detailing if they carry specific mutations which may increase the disease risk. The idea of DTC services came to life with the availability of GWAS data from different populations; however, the predictive ability of the genetic risk models is a concern [28], especially when inappropriate reference populations are used and the non-genetic factors are omitted [29].

The purpose of this chapter is to summarize a foremost component of PGTs: the methods to transform the raw data from genotyping technologies into disease risk predictions. Because accuracy in risk assessment is essential for personalized medicine, we emphasize the current state and perspectives of the algorithms for SNP identification, as well as the main approaches for predicting SNPs causative of disease. In parallel, we discuss how these components have been implemented in the PGTs market by DTC companies, hence providing the reader with a global picture of the science behind disease risk prediction.

This chapter is structured as follows. First, we introduce the health-related genetic tests and list some companies offering personalized genetic services, including their locations, prices, and types of services they offer. Then, we outline the main platforms for SNP genotyping, along with the algorithms designed for detecting SNPs from their output data. Next, we survey the techniques to predict single-SNP-disease and multiple-SNP-disease associations. We discuss some predictive genetic risk models in DTC services and the factors affecting these approaches. Finally, we discuss perspectives and give recommendations for the improvements of algorithms in personalized genetic testing.

Box 1.1 contains a glossary of terms used in this chapters.

Box 1.1: Glossary of biological concepts

This box presents brief definitions of the biological terms used in the book. Most of these definitions have been adapted from the Genetic Home Reference Glossary.[a]

- **Allele**: Allele represent one of two or more versions of the same gene. Each individual inherits two alleles, one from each parent.
- **Allele frequency**: The measure of an allele's relative frequency (percentage) in a population.
- **Alternative splicing**: The usage of different exons that are all part of the initial transcript, to form the mature mRNA, which will be translated into a protein. Alternative splicing results in the generation of related, but different, proteins from one gene.
- **Coding region/sequence (CDS)**: Represent the region of DNA that will be transcribed into a mature messenger RNA (mRNA) and translated into the amino acid sequence of a protein.
- **Common variants**: Alternative forms of a gene, which are present with a minor allele frequency (MAF) higher than 5%.
- **Contiguous SNPs**: SNPs lying next to each other on the DNA strand.
- **Copy number variants (CNVs)**: A type of structural variation where a section of DNA is present in two or more copies instead of only one.
- **Duplication**: A type of mutation, where a portion of a gene, a whole gene, several genes, or larger regions of the chromosome are copied and are present in duplicate amounts.

(cont.)

[a] http://ghr.nlm.nih.gov/glossary

- **Effect size**: Contribution of a SNP to the genetic component (i.e., heritability) of the disease. This is usually the odds ratio reported in GWAS for the SNP [30, 31].
- **Exons**: Exons represent portions of the DNA sequence of a gene that are transcribed into mRNA and are translated into proteins.
- **Gene**: Genes are the basic physical and functional units of heredity made out of DNA. They make instructions on how to make proteins. The human genome is composed of approximately 19,000 genes [32].
- **Gene–gene epistatic interactions (epistatsis)**: A condition in which the expression of one gene is affected by the expression of one or more independently inherited genes. For example, when the expression of gene B depends on the expression of gene A, then the expression of gene B will not occur if gene A is not expressed. In such a case, gene A is said to be epistatic to gene B.
- **Genotype**: Represents all of the alleles an individual inherited from parents. It can also refer to two specific alleles of a particular gene. At the genomic level, each SNP can have two alleles (e.g., allele *A* and allele *a*); hence, a SNP is linked to one of three possible genotypes, e.g., *AA*, *Aa*, or *aa*.
- **Haplotype**: Describes a combination of alleles or a set of SNPs that are found on the same chromosome and tend to be inherited together. The International HapMap Project collects information of haplotypes.
- **Heritability component**: The heritability component of a disease is the proportion of phenotypic variability in the population explained by genetic factors [24].
- **Heterozygous**: Contrary of the homozygous: an individual inherits two different alleles from parents.
- **Homozygous**: When an individual receives the same alleles from parents, he/she is said to be homozygous.
- **Insertions/deletions (INDELs)**: Types of genetic variation involving the addition (insertion) or loss (deletion) of smaller (single nucleotide) or larger pieces of the DNA strand from a part of a chromosome.
- **Introns**: Introns are portions of the DNA molecule that are transcribed into mRNA, but are not translated into proteins.
- **Inversion**: A type of mutation in which a smaller or larger segment of the DNA molecule is broken away, inverted from end to end and re-inserted back into the chromosome.
- **Linkage disequilibrium (LD)**: Indicates that alleles are physically close to one another on the DNA strand. They occur together more often than accounted by chance alone.
- **Loci**: Particular sites on a chromosome.
- **Minor allele frequency (MAF)**: Refers to the frequency of the least abundant (minor) allele of a SNP in a population.

(cont.)

Box 1.1: Glossary of biological concepts (cont.)

- **Rare variants**: Alternative forms of a gene, which are present with a minor allele frequency (MAF) of less than 1%.
- **SNPs (rSNPs)**: Single nucleotide polymorphisms involve a variation in one single base pair at a specific location in the genome. They represent the main type of single nucleotide variants present in the human genome. SNPs differ from SNVs in that their variation in the population is known. A variation can be said to be a SNP if it is present in at least 1% of the population.
- **Single nucleotide variations (SNV)**: In NGS sequence analysis, variations in a single nucleotide are referred to as SNV, since their population frequency is not known.
- **Structural variants (SV)**: Represent different types of genomic alternations, including duplications, inversions, insertions, deletions etc. To be qualified as SV, the affected region of the DNA has to be 1 kb or larger in size.
- **Untranslated regions (UTRs)**: UTRs represent regions of DNA on either side of the coding regions (CDS) that are not translated into the amino acid sequence of a protein.

1.2 Genetic Tests in Healthcare

Genetic tests are predominantly used to determine whether a patient's DNA sequence has alterations that may result in chromosomal, monogenic, or complex disorders (see Box 1.2) [26, 33]. These alterations in specific genes or chromosomes are important for healthcare in different contexts; for example, they may be responsible for inherited disorders, or they could affect the sensitivity of individuals to a drug therapy. Therefore, types of PGTs have been formulated for a range of applications (see Section 1.2.1) and a number of specialized PGT providers has increased around the world (see Section 1.2.2).

1.2.1 Types of Genetic Tests

While a wide variety of PGTs are available for non-health concerns, including paternity, siblingship, forensic testing, and ancestry, we are interested in health-related genetic tests. Most of the health-related genetic tests evaluate if the patient carries a specific genetic mutation that may increase the disease risk, or a physical trait (Box 1.2). Hence, the test may reveal specific mutations in the DNA, effectiveness of drugs, possibility of drug side effects, or the influence of genetic variants on physical traits [26].

Box 1.2: Types of genetic disorders and PGTs

- **Chromosomal disorders**: Abnormalities such as extra copies, or missing parts of one chromosome.
- **Monogenic disorders or Mendelian diseases**: Mutations in one gene that arise in a severe disorder. The alteration may be linked to one or both alleles, and a person carrying the mutation may have the disorder's symptoms or not (healthy carrier).
- **Complex genetic disorders**: The joint effect of alterations in many genes, lifestyle and environmental factors.
- **Predictive genetic tests**: Detect gene mutations that increase the risk of developing a disorder in adult life. They are thought to be performed in individuals without disease symptoms.
- **Diagnostic genetic tests**: They are thought to be performed in individuals who show disease symptoms. They may confirm the physician's diagnosis and help choose the right treatment.
- **Carrier tests**: These tests find single mutated alleles in asymptomatic individuals. The patient does not show signs of the disease, but their children are at risk of having the genetic condition.
- **Pharmacogenomic tests**: Tests specially designed to evaluate the sensitivity to drug therapy in a patient. They target SNPs associated to drug dosage and risk of adverse effects.

Among the types of health-related PGTs preseted in Box 1.2, we focus on the predictive genetic tests. The results of these tests predict the risk of onset of a particular disease, which depends on the patient's genetic profile and the methodology used to assess the risk. Still the current methodologies do not consider other non-genetic factors of importance (e.g., environmental factors, lifestyle), so the results are highly inaccurate [29]. The probabilistic nature inherent to predictive genetic tests has opened opportunities for improvement, as discussed in Sections 1.4.3 and 1.5.

1.2.2 Genetic Tests Providers

Typically, there are two ways to access the genetic screening services. If a genetic disorder is suspected, a physician orders the test from a laboratory; the laboratory sends the reports back to the healthcare provider and the physician counsels the patient in the interpretation of the results. On the other hand, any person can order a DTC genetic test straight from private companies [34]. The consumer receives a kit to collect a sample of saliva and returns the sample to the company. After the DNA is isolated from the sample and the screening is completed, the reports are sent back to the consumer, or posted online. Despite the variety of tests covered, most of the reports are only for informational purposes, the consumer does not receive a diagnosis and in most cases the companies do not supply medical counselling [28]. Table 1.1 shows

Table 1.1: Some examples of companies offering PGTs for health purposes (prices as per year 2016)

Company	Technologies	Health services
23andme	Illumina Human Omni Express – 24 format chip	• *Health reports*: Personalized information about how the indiv genetics influences susceptibility to diseases. Reports include inherited conditions, drug responses, genetic risk factors and traits. • The price of a personalized saliva-collection kit is 200 USD. T DTC services are available to customers in the USA, Canada, Denmark, Finland, Ireland, Sweeden, the Netherlands, and t United Kingdom.
Counsyl	• Array based ge-notyping. • NGS genotyping test with Illumina HiSeq 2000.	• *Family prep screen*: A screening test for parents. It predicts if p carry genetic diseases that may be inherited to their descenda • *Informed pregnancy screen*: A screening test for pregnant wom predicts if the baby could suffer any chromosomal condition resulting in birth defects. • *Inherited cancer screen:* A screening test to detect genes associa cancer and the chances of suffering different types of cancer. • Costs fluctuate between 150 and 300 USD for individuals wit insurance in the USA. The service can be accessed through p order.
Gene-by-Gene	• Illumina Omni Express array • Illumina MiSeq	• Among the services offered by the company we find: WES[a] (USD), GWAS studies (199 USD) and WGS[b] (10.395 USD). The services are available in the USA through saliva sample colle
FullGenomes	Illumina's HiSeq X	• WES (775 USD) and WGS at different sequencing depths: 30 USD), 10× (745 USD), 4× (375 USD) and 2× (250 USD). DTC available in USA
GenePlanet	Microarrays and sequencing platforms	• *Personal DNA analysis* (560 USD): The service includes inform susceptibility to 20 diseases, predicted response to 6 drugs an traits. This DTC company is based in Slovenia

[a] Whole-exome sequencing

[b] Whole-genome sequencing

some examples of genetic tests for health purposes that can be accessed either through physicians or directly from DTC companies.

The National Institute of Health administers the Genetic Testing Registry (GTR) [35].[8] This database enhances access to details of health-related genetic tests and laboratories worldwide. Although the laboratories voluntarily submit the information and the GTR does not include DTC tests, the database has standardized information for over 32,000 tests from 45 laboratories. The test-specific information is integrated with other NCBI databases in the domains of genomic sequence, sequence variation, genotype-phenotype relationships, and medical literature.

1.3 Common Technologies and Algorithms for SNPs Identification

A number of different technologies are used to assay DNA samples for genetic variants in PGTs. The advent of sequencing technologies has broadened the landscape of variant detection, including SNPs, INDELs, and structural variants. However, the non-sequencing technologies are still crucial for pinpointing specific SNPs and genotyping them in individuals, at low cost. This progress has simultaneously prompted advances in the algorithms for inferring potential genotypes from the raw data (Figure 1.1). The aim of this section is to summarize two common technologies for identifying SNPs, namely microarrays and NGS, and the resulting algorithms that are being used in response to the platforms' evolution (Sections 1.3.1 and 1.3.2).

1.3.1 Microarrays

Microarray technologies provide different alternatives for exploring whole genomes, including gene differential expression identification, copy number estimation, and genotyping [36, 37]. In genotyping, the SNP arrays determine the genotypes of individuals by measuring their relative allele intensities [36]. The first whole-genome sampling method for SNP genotyping was developed by Affymetrix in 2003 [9]. Since then, new generation microarrays have decreased the cost of this technology, improved the coverage and allowed for high throughput genotyping in GWAS [38].

Two main microarray platforms used for the genotyping of SNPs are the Affymetrix GeneChip and the Illumina Bead Array. Despite differences in the physical design and SNP content, both platforms have led to the discovery of hundreds of SNPs related to both complex traits and diseases [39].

1.3.1.1 Affymetrix SNP Microarrays

The Affymetrix SNP microarrays consist of a printed-array format that is produced in parallel by photolithographic manufacturing (see Figure 1.2(a)). For every SNP on the array there are two probes present, each one specific for one SNP allele (see Boxes 1.1 and 1.3 for definitions of biological and technical methods). After fragmenting,

[8] www.ncbi.nlm.nih.gov/gtr/

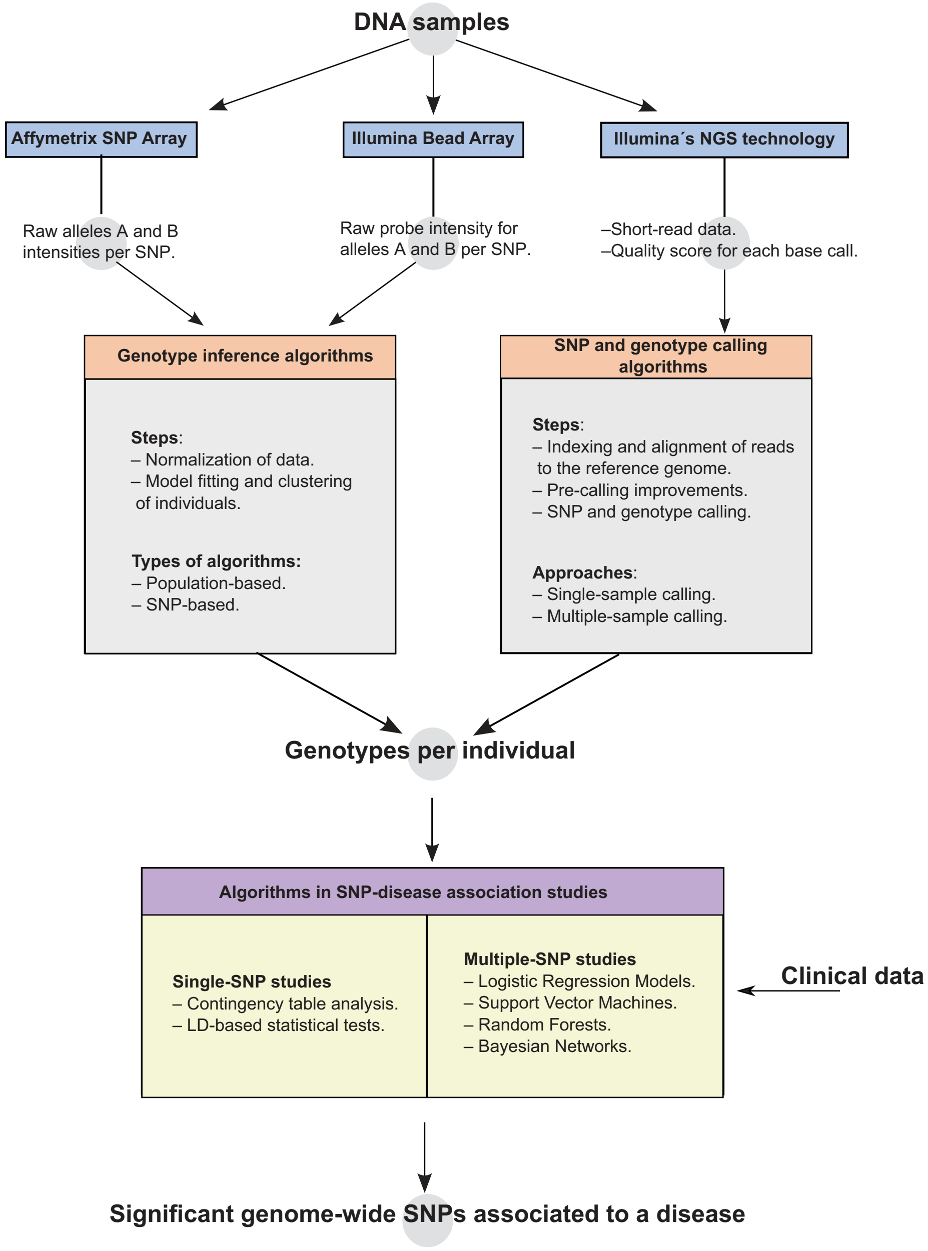
DNA samples
Affymetrix SNP Array
Illumina Bead Array
Illumina´s NGS technology
Raw alleles A and B intensities per SNP.
Raw probe intensity for alleles A and B per SNP.
–Short-read data.
–Quality score for each base call.
Genotype inference algorithms
Steps:
– Normalization of data.
– Model fitting and clustering of individuals.
Types of algorithms:
– Population-based.
– SNP-based.
SNP and genotype calling algorithms
Steps:
– Indexing and alignment of reads to the reference genome.
– Pre-calling improvements.
– SNP and genotype calling.
Approaches:
– Single-sample calling.
– Multiple-sample calling.
Genotypes per individual
Algorithms in SNP-disease association studies
Single-SNP studies
– Contingency table analysis.
– LD-based statistical tests.
Multiple-SNP studies
– Logistic Regression Models.
– Support Vector Machines.
– Random Forests.
– Bayesian Networks.
Clinical data
Significant genome-wide SNPs associated to a disease

Figure 1.1: Workflow of the technologies and algorithms in the discovery of SNP-disease associations.

fluorescence marking and hybridizing of the patient's DNA to the array, the array is scanned and the fluorescence signals (i.e., intensities) are measured. In the initial versions of the Affymetrix GeneChip genotyping microarrays, SNP were detected with the use of five probes that perfectly matched the targeted SNP (perfect match