PSB 2011 Tutorial: Personal Genomics

CAN ALKAN[1], EMIDIO CAPRIOTTI[2], ELEAZAR ESKIN[3], FEREYDOUN HORMOZDIARI[4], MARICEL G. KANN[5]

[1] Department of Genome Sciences, University of Washington, and Howard Hughes Medical Institute, Seattle, WA, USA

[2] Department of Bioengineering, Stanford University, Stanford, CA, USA

[3] Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA

[4] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada

[5] Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD, USA

1 Introduction

Improvements in sequencing methods and other genotyping assays introduced high throughput, low-cost and more automated technologies. The revolution in DNA sequencing opened many possibilities for researchers working in the fields of genetic variation, diseases of genomic origin, and even personalized medicine [1]. The completion of the pilot phase of the 1000 Genomes Project resulted in the discovery of a vast amount of normal genomic variation including 15 million SNPs, one million indels, and over 20,000 structural variants [2]. The new technologies can also be employed to discover the functional landscape of the human genome as part of the ENCODE Project such as epigenetic variation (methylation patterns and histone modification) and protein-DNA interaction. Further uses of the high throughput sequencing technologies include transcriptome analysis, non-coding RNA discovery, gene expression profiling, rapid testing of genotype-phenotype associations, and identification of pathogens [1, 3].

Our genetic identity not only determines our physical differences, but it also defines our susceptibility against diseases. Several groups are now working on various methods to exploit the power of cost efficient technologies to better perform genotype-phenotype associations, in particular to identify susceptibility to disease and eventually diagnose disease at its early stages. The ultimate goal is to vastly improve the field of pharmacogenomics, which can broadly be defined as the study of the relationship between genotype and drug response and how the drugs affect our metabolism. The wealth of new data gives many opportunities to advance our understanding of how to optimize drug combinations for each individual's genetic makeup. The underlying computational tools for such studies analyze available data to identify differences between a reference genome and sequenced genomes, as well as perform clustering and classification to obtain both normal and disease-related phenotype associations.

This tutorial provides a starting reference to analyze personal genomes (and genomic variation) using various data and techniques such as next generation sequencing (NGS)[1], array comparative genomic hybridization (arrayCGH), and single nucleotide polymorphism (SNP) microarrays.

[1]Parts of this tutorial are taken from review articles written by Michael Brudno, published in *Nature Methods* [4] and *Briefings in Bioinformatics.* [5]

2 Discovery of Human Genome Variation

The genetic variation among human individuals spans a wide range of sizes. The smallest variation is termed single-nucleotide polymorphism (SNP) and is defined as variation occurring when a single nucleotide differs between different human individuals. Small INDELs are larger in size and defined as an insertion or deletion of 1-1,000 basepairs [6, 7]. Structural variants (SVs) are mid-size genomic variation and include insertions, inversions, deletions, and duplications of DNA segments larger than 1000 bp [8–11]. Segmental duplications are also considered to be a type of structural variant, and they are defined as >1000 bp of duplicated genome segments with >90% sequence identity [12–15]. As a more general term, copy number variation (CNV) refers to the duplication or deletion of a segment of DNA sequence compared to a reference genome assembly. The largest types of genome variation are the chromosomal changes such as duplications, deletions, or inversions of large portions of chromosomes and translocation events. Although these genomic variants can be "normal" (i.e. not known to be a cause of disease), many SNPs, INDELs, structural variants, and chromosomal aberrations are associated with disease such as psoriasis [16], HIV susceptibility [17], Crohn disease [18, 19], epilepsy [20, 21], renal disease [20], diabetes [20], autism [22], and more.

In this section, we survey these different types of human genome variation and the tools and methods to detect such variation.

2.1 Single Nucleotide Polymorphism and Small INDEL Polymorphism

SNP microarrays. SNP arrays were introduced to comprehensively and rapidly study single nucleotide polymorphisms in human genomes. The International HapMap Project [23, 24] utilized SNP microarrays to detect and genotype 3.1 million SNPs in 270 individuals from different populations. The SNP arrays contain immobilized oligonucleotide probes specifically designed to test the existence of SNPs common to the human populations. Most commercially available SNP array designs are biased to SNPs that occur more frequently in European populations, thus they are not suitable to study more divergent populations. The first SNP arrays contained approximately 10,000 probes; however, current SNP arrays feature 2 million genetic markers. The test DNA is then labeled using fluorescent molecules and hybridized to the SNP array. Finally, specialized scanners are used to detect the hybridization signals. Computational analyses of SNP arrays are mainly statistical in nature, and most tools are supplied by the microarray design companies such as Affymetrix and Illumina. Non-canonical analyses tools, such as SNPchip [25], are available in the Bioconductor suite [26].

Sequencing-based strategies. Recently many algorithms were developed to discover SNP and small indel variation using high throughput sequencing data sets. Several of such algorithms are listed in Table 1. Since the mapping of a read generated by NGS technologies is only a prediction of its true location, most SNP calling algorithms include a data preparation step in which read mappings are evaluated and filtered. Reads that may be mapping to paralogs or repeat sequences are discarded, or considered only when other reads offer supporting evidence [27–29]. Quality values may also be (re)assigned to the reads based on the basepair traces or various statistics. A re-alignment step [30] may also be employed to better align small indels (1-5 bp), if they are present in the mappings.

In general, a Bayesian approach is applied to the filtered, aligned reads to infer genotypes. These approaches compute the conditional likelihood of the nucleotides at each position using the Bayes rule:

$$P(G|R) = \frac{P(R|G)P(G)}{P(R)}$$

This equation states that one can get the probability of a certain genotype G given the data R (posterior) if one has the overall probability of that genotype (prior) and the probability of observing the given data given from this genotype (likelihood). The denominator can be understood as a normalization factor. Most often, the prior $P(G)$ will be represented by the probability of the variant, for example, the widely used MAQ [28] tool set uses a probability of heterozygosity r. The probability of observing the prepared reads $P(R|G)$ is then estimated for each possible donor genotype. Continuing with the example of MAQ, this probability is computed with a binomial distribution if errors are assumed independent and identical for each base in the read, or otherwise with a weighted product of the observed errors. Finally, a posterior probability $P(G|R)$ is computed, which either estimates the donor nucleotide themselves given the data or the probability of a SNP given the data. Applying a threshold to this probability for SNP discovery offers a sensitivity/specificity trade-off.

Although most methods use a Bayesian approach to SNP discovery, they vary widely in the details, use different interpretation of statistics, and have diverse approaches for small indel discovery. While PolyBayes [27], SOAPsnp [29], and MAQ each assume some prior probability that a site is polymorphic, the rest of the model is different in its implementation. In order to assign a posterior, MAQ estimates a probability of observing the given read errors for each genotype prior via a binomial distribution if errors are correlated or a similar estimating function if they are not. SOAPsnp computes the likelihood estimating a posterior that is based on various features of the reads. PolyBayes assumes knowledge of a probability of error via quality values and uses the product of those directly to compute the posterior. In a recent publication, Hoberman *et al.* [31] present a SNP discovery algorithm with a generally different approach. First, site-specific and more general features are generated from read mappings; and this information is used to train a classifier. Next, this classifier is then used to score the heterozygosity at each position.

For small indel discovery, PolyScan [32] re-evaluated *de novo* signatures, followed by a segment alignment algorithm that is very sensitive to small indels. A statistical model is then presented, but instead of analyzing each column in the multiple alignment, it considers the amount of shift within clusters of re-aligned reads in order to detect small indels. A different approach, mentioned above, was utilized in the MAQ tool [28] and in the Corona Light pipeline [33]. Both of these utilize mate-pair information: at first, all reads are mapped without allowing gaps. Second, mate-pairs with only one end mapped allow the gapped mapping of the second read, in an expected range around the mapped read. This allows for detection of indels, while keeping the computational complexity introduced by gapped alignment limited to a small subset of the reads.

SNP Calling in Color-Space AB SOLiD's di-base sequencing has several properties that present unique challenges for SNP and indel identification. Some tools map the reads by translating the reference and mapping in color-space, but in order to call SNPs, they translate the multiple alignment back to nucleotide space (while correcting likely sequencing errors) and call SNPs as described in the above sections [28,29]. McKernan *et al.* [33] describe Corona Light as a consensus technique where each valid pair of read colors votes for an overall base call. The DiBayes tool implements a Bayesian algorithm that works solely in color-space. Here, the posterior probability is computed for a particular combination of color pairs (dicolors); the prior is based on the expected polymorphism rate, and the likelihood is the probability of seeing a certain dicolor given the error rates. McKernan et al. [33] describe this algorithm as similar to PolyBayes [27], which we discussed

Algorithm	Platform	Strategy	Variation	
			SNP	Indel
MAQ [28]	Illumina	Read pileup	Yes	Yes
SAMtools [35]	Illumina/SOLiD	Read pileup	Yes	Yes
Mosaik	Illumina/454	Hashing reference	Yes	Yes
SOAPsnp [29]	Illumina/454	Likelihood optimization	Yes	Yes
Corona Light [33]	SOLiD	Bayesian framework	Yes	Yes
VARiD [34]	Illumina/454/SOLiD	HMM	Yes	Yes
ProbHD [31]	Illumina/454	Probabilistic framework	Yes	No
SPLINTER [36]	Illumina	Probabilistic framework	Yes	Yes
Dindel [37]	Illumina	Expectation maximization	No	Yes
QCALL [38]	Illumina	Probabilistic framework	Yes	No
Pindel [39]	Illumina	Split read mapping	No	Yes

Table 1: Algorithms to discover SNPs and small indels using sequencing data.

in the previous subsections; however, a detailed description has not yet been published. In addition to AB SOLiD-specific SNP callers, a recent algorithm, VARiD, merges information from both color and letter space data to improve SNP and small indel detection sensitivity [34].

2.2 Structural Variation and Copy Number Variation

SNP microarrays. The SNP genotyping data from various SNP microarray assays, such as Affymetrix and Illumina BeadXpress, can also be used to detect and genotype both common and rare CNVs. The methods that use the SNP microarray data to predict CNVs are usually Hidden Markov Model (HMM) based approaches that make use of the allele frequency of SNPs, the distance between neighboring SNPs, and the signal intensities. Each of these algorithms are fine tuned for different type of SNP microarray assay, and different classes of CNVs. In addition, they can also be used to genotype the copy number of the duplicated DNA segments, however, since the probe density over the duplicated genome intervals are usually poor, these predictions are unreliable. One of the most used CNV detection tools from Illumina SNP genotyping data is called PennCNV [40]. HMMSeg [41] is an HMM based segmentation algorithm that simultaneously analyzes both the normalized total intensity ("LogR ratio") and allelic intensity ratios ("B-allele frequency") [42] to detect regions of homozygous deletion, hemizygous deletion, or amplification. SCIMM [43] uses the same array data and aims to genotype CNVs in a large number of samples using as few as two SNP probes (when the breakpoints of candidate CNVs are known in advance). It was used to identify large CNVs in $\sim$1200 individuals with an emphasis on "hotspots" of human genetic disease [44]. Another recently developed algorithm, named SCOUT [45], is similar to SCIMM in nature, however it performs better in detecting rare CNVs in large cohorts.

Birdseye [46] is similar to SCIMM, but it is developed to use another SNP microarray platform (Affymetrix) and was employed to characterize CNVs in 270 HapMap samples [19]. At the Personal Genomics session in PSB'2010, Yavas et al. also described yet another CNV caller from Affymetrix data, called ÇOKGEN [47].

Array comparative genomic hybridization. The underlying technology of array comparative genomic hybridization (arrayCGH) is similar to SNP microarrays, but the aim is to measure copy number differences between two individuals ("affected" vs. "control"). Oligonucleotides from genomic regions of interest are immobilized in a microarray, DNA samples from two individuals are labeled with different marker molecules, and hybridized to the chip. Finally a specialized scanner compares the signal intensity difference (log_2 score, generated by the two different fluorescent dyes) to measure the copy number difference (Figure 1). Usually a second experiment is performed as a control and to prune "bad" probes using the same samples but with the fluorophores swapped. Selecting the oligonucleotides in array design is particularly important to minimize hybridizations occurring by chance [48]. For each probe a log_2 ratio of signal intensities is calculated. After a normalization procedure based on control regions (known invariant copy number), a genotype is assigned as:

- No difference, $log_2(2/2) = 0$
- Hemizygous deletion in test: $log_2(1/2) = -1$
- Duplication (1 extra copy) in test: $log_2(3/2) = 0.59$
- Homozygous duplication (2 extra copies) in test: $log_2(4/2) = 1$

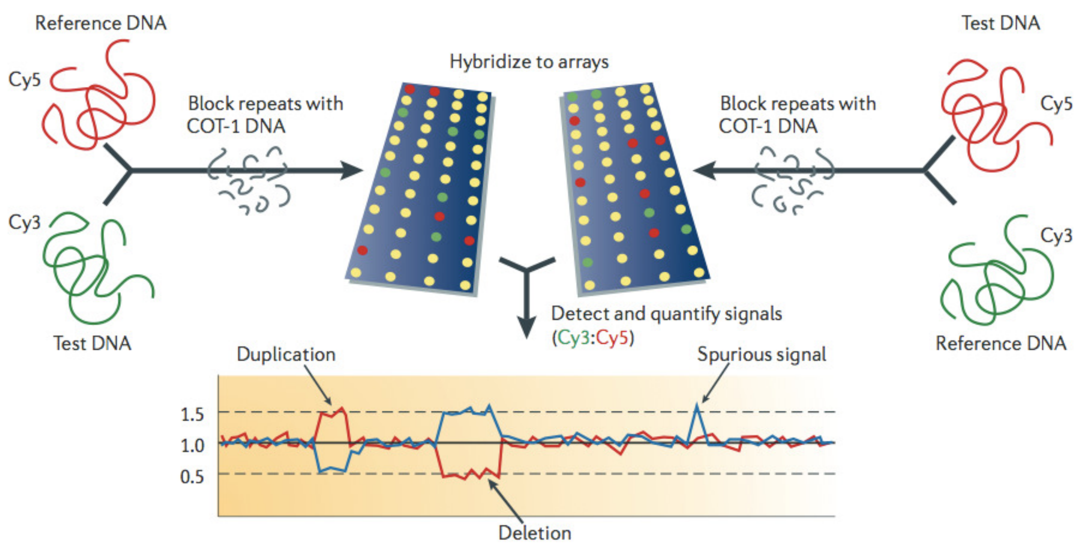

Figure 1: Array comparative genomic hybridization (arrayCGH). Figure adapted from Feuk et al. [9]

ArrayCGH has been utilized in many studies to characterize copy number variation in large cohorts of both normal individuals and patients [49–51] as well as to investigate segmental duplications in primates [52]. Most algorithms to analyze arrayCGH data are based on Hidden Markov Models. Each observed log_2 value reflects an underlying copy number state. Using the observed values, the underlying state for each probe is inferred. Based on some model, the sequence of states most likely to produce the observed values is chosen. ArrayCGH is a powerful and low-cost method to detect CNVs, yet it poses some limitations. First, CNV detection is only possible in the "targeted" regions in the genome, and due to probe uniqueness, the designs are biased against repeats and duplications. Furthermore, although it is possible to assay duplications using arrayCGH, when the copy number differential is low in high-copy regions (for example 10 vs. 11 copies), resolution provided by arrayCGH does not discriminate the copy number difference [14, 15, 52, 53].

Sequencing-based strategies. Detecting SVs between two individuals would be a trivial task if their genomes were already assembled. Since this is currently prohibitive for humans, current methods use only one assembled genome (the reference) and another sequenced genome (the donor). Thus, they are unable to compare the sequences directly and instead rely on detecting variation through signatures—patterns of paired-end mappings that are created by structural variation.

Two of the easiest and most commonly detected signatures are the *basic insertion* and *basic deletion* [8, 54] (Fig 2). A matepair that spans an isolated deletion event maps to the corresponding regions of the reference, but the mapped distance is greater than the insert size. If the event is an insertion, then the distance is smaller. Another variant that leaves a clear signature is an *inversion*. A matepair that spans either (but not both) of its breakpoints will map to the reference with the orientation of the read, lying inside the inversion, flipped. Two such matepairs, respectively spanning each of the two breakpoints, form the basic inversion signature [8, 55, 56] (Fig 2).

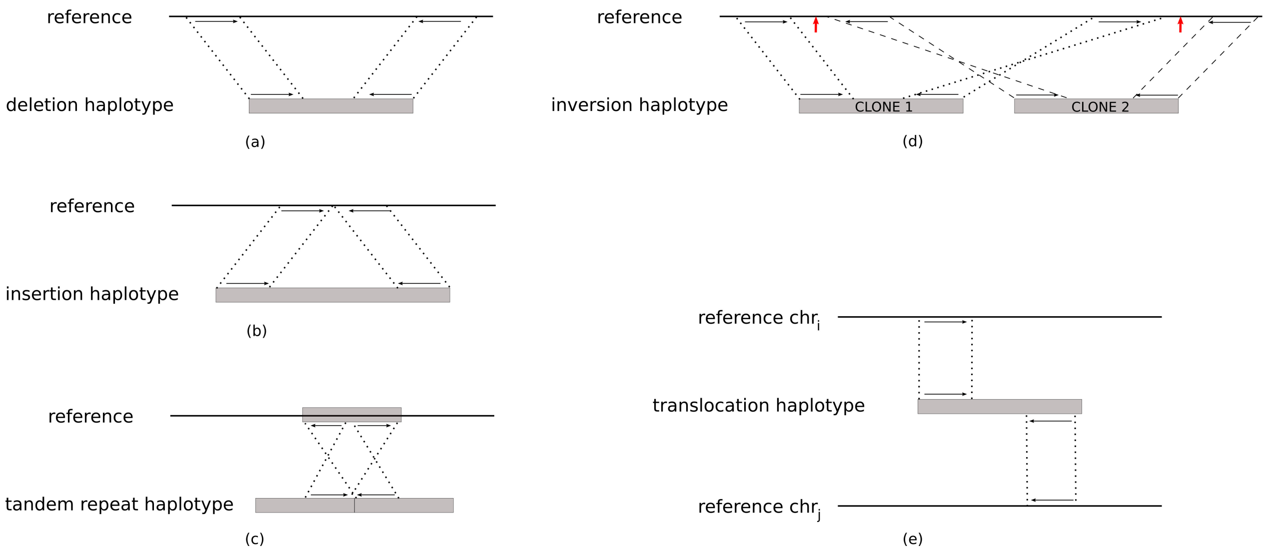

Figure 2: Types of structural variation that can be detected with paired-end sequences: mapped span of paired-end reads appear larger than the expected insert length if there is a (a) deletion and smaller in an (b) insertion haplotype. Disagreement between the mapping orientations and the sequencing library specifications might either report a (c) tandem repeat or an (d) inversion. Also, note that in the case of inversions $CLONE_1$ and $CLONE_2$ predict two different inversion breakpoints (shown with arrows), but by examining the map locations and orientations, one can deduce that both clones predict the same inversion, and both breakpoints of the inversion event can be recovered. If the paired-end reads align confidently on different chromosomes, a (e) translocation event is reported. In this figure, we assumed the expected end-sequence orientation properties in capillary based sequencing and Illumina platforms.

Several methods have been packaged into algorithms and are available to the public, including SegSeq [57], PEMer [58] VariationHunter [56], MoDIL [7], Pindel [39], BreakDancer [59], EWT [60], and AB SOLiD Software Tools [33]. Each one can be characterized in terms of two distinguishing factors—the signatures they detect and the way they cluster/window these signatures. These characterizations are summarized in Table 2. This table can be used to guide a user's decision on which method is most applicable.

Algorithm	Method	Mapping	Capability	Platform	Strategy	SV Classes				
						Del.	Inv.	Novel Ins.	MEI.	Dup.
VariationHunter [56]	RP	mrFAST [14]	Discovery	Illumina	maximal clustering and maximum parsimony	50 bp - 500Kbp	Yes	Yes*	Yes	No
PEMer [58]	RP	BLAT [61]	Discovery	454	Outlier clustering	300 bp - 1 Mbp	Yes	Yes*	No	No
BreakDancer [59]	RP	MAQ [28]	Discovery	Illumina	Outlier clustering and distribution analysis	10 bp - 1 Mbp	Yes	Yes*	No	No
MoDIL [7]	RP	mrFAST [14]	Discovery	Illumina	Outlier fitting distribution analysis	6 bp - 70 Kbp	Yes	Yes*	No	No
CNVer [62]	RP/RD	Bowtie [63]	Discovery Genotyping	Illumina	Outlier fitting and probabilistic graph analysis	1 Kbp - 550 Kbp	No	No	No	Yes
Spanner	RP/RD	Mosaik	Discovery Genotyping	Illumina 454	Outlier clustering	10 bp - 200 Kbp	Yes	Yes	Yes	tandem dup.
SLOPE [64]	RP	MAQ [28]	Discovery	Illumina		10 bp - 1 Mbp	No	Yes*	No	No
HYDRA [65]	RP	BWA [66]	Discovery	Illumina	maximal clustering	50 bp - 1 Mbp	Yes	Yes*	Yes	No
SVDetect [67]	RP	BWA [66]	Discovery	Illumina SOLiD	clustering and Sliding Window	50 bp - 1 Mbp	Yes	Yes*	No	Yes
GASV [68]	RP	BWA [66]	Discovery	Illumina SOLiD	maximal clustering	50 bp - 1 Mbp	Yes	No	No	No
Corona Light [33]	RP	mapreads [33]	Discovery	SOLiD	Outlier clustering and significance testing	5 bp - 85 Kbp	No	Yes*	No	No
mrCaNaVaR [14]	RD	mrFAST [14]	Discovery Genotyping	Illumina 454	Distribution-based outlier clustering	>10 Kbp	No	No	No	Yes
EWT [60]	RD	MAQ	Discovery	Illumina	Significance-testing	200 bp-250 Kbp	No	No	No	Yes
Pindel [39]	SR	MAQ [28]	Discovery	Illumina	Pattern growth	1 bp - 50 Kbp	No	1 bp - 20 bp	No	No
NovelSeq [69]	AS	mrFAST [14]	Discovery	Illumina	local assembly and maximum matching	No	No	> 200 bp	No	No
Cortex	AS	-	Discovery	Illumina 454	de Bruijn graph	30 bp - 40 Kbp	No	> 50bp	Yes	No
SOAPdenovo [70]	AS	-	Discovery	Illumina	de Bruijn graph	50 bp - 4 Kbp	No	> 51 bp	Yes	No
BreakSeq [71]	Breakpoint library	any	Genotyping	any	Read mapping to breakpoints	Yes	Yes	Yes	Yes	Yes

Table 2: Some of the structural variation detection algorithms using next-generation sequencing data. These methods are characterized by the types of signatures (RP: read-pair, RD: read-depth, SR: split-read, AS: assembly), variant classes they can detect and the strategy they use, both of which are shown here. MEI: mobile element insertions. *: RP methods can discover novel sequence insertions of length < insert size only.

In addition to the methods already mentioned in the previous section, there have been more recent approaches that have combined previously developed methodologies into a single framework [33, 59, 72]. For example, BreakDancer combines the standard clustering paradigm (BreakDancerMax) with the distribution-based approach proposed in MoDIL, albeit without hemizygous event detection (BreakDancerMini). AB SOLiD Software Tools combine the standard clustering paradigm with a different distribution-based approach for indel identification, and the binary circular segmentation algorithm to identify regions of gain/loss.

Another prominent tool is PEMer [58], a highly modularized framework for detecting SVs that is specifically tailored to easy modification and development by the user. Some of the PEMer modules include read mapping, filtering of low quality reads, signature detection, and clustering. Such a modularized framework has the potential to facilitate future algorithmic development by allowing algorithmic improvements to particular modules without the need for implementing a whole SV discovery pipeline. However, we note that there is still work to be done to create full-fledged user-friendly tools for biologists.

Segmental duplications are yet another type of structural variants defined as low copy duplications of size >1000 bp and >90% sequence identity [12, 13]. Despite their importance in gene innovation and phenotypic variation, duplicated regions have remained largely intractable due to difficulties in accurately resolving their structure, copy number, and sequence content. Recently, Alkan et al. [14] developed a read mapping tool *mrFAST* that tracks all possible map locations of reads within a given sequence identity threshold. Additional heuristic methods were employed to analyze depth of coverage to detect segmental duplications and predict absolute copy number of the duplicated genes. Furthermore, by inspecting the sequence substitutions and small indels in the duplicated genes this method can distinguish between different copies of highly identical genes, providing a more accurate census of gene content and insight into functional constraint without the limitations of array-based technology [14, 15]. A Hamming distance only version of this read mapping tool, called mrsFAST can also be used for the read depth analysis [73].

3 SNPs and Disease

The interpretation of genomic variation is an active area of research with great impact in molecular biology. SNPs are the major source of human variability, occurring about every 300 base pairs, and are also responsible for the insurgence of human pathologies. Although some progress toward the understanding of disease mechanisms and their association to SNPs has been made, the personalization of medicine is still far away. Meeting that goal will require strong collaborative efforts between health care and academia to assemble larger collections of curated SNPs and to create user friendly, integrated tools that evaluate disease risk associated with genetic variations.

3.1 SNPs Databases and Annotations

Large scale genome-wide association studies and human sequencing projects are producing hundreds of SNPs with putative relevance to cancer [74] and other diseases (see review by Altshuler et al. [75]). Some of these sequence variations in the protein produce changes in the stability, regulation, ability to interact, or to be modified, and are ultimately associated with the disease. The OMIM database [76], manually curated and updated daily, is one of the largest catalogs of human genes and disorders. As part of the NCBI Entrez database, OMIM is freely available and contains over 11,000 genes with known sequence and over 6,000 phenotypes. It should be noted that only a few hundreds of the genes with known sequences currently annotated in OMIM have known phenotypes. Automatic approaches for linking genotype with phenotype information have the potential to overcome the data scarcity problem inherent in manual efforts. To that purpose, several approaches have been developed including PhenoGo [77] that use natural language processed information in combination with Gene Ontology (GO) data to create a collection of over 500,000 phenotype-GO associations, including approximately 33,000 genes from 10 species. Similarly, Gene2Disease automatically assigns priorities to genes related to a disease, and provides a list of candidates based on PubMed Mesh terms and GO. Another resource, Genecards [78], provides a suite of tools that integrate information from over 70 sources including OMIM, constituting a single location to retrieve available information for over 24,000 genes including relationships to diseases when available. The PhenomicDB [79] database uses associated orthology relations to provide multi-species genotype-phenotype mappings across human and several model organisms. The Orthodisease database provides a cluster of more than 3,000 disease genes comprising 26 Eukaryotic organisms. Swissprot is a database of protein sequences that includes disease annotations for about 2,600 of its 270,000 entries (16,600 are for human proteins). PharmaGKB [80] is a catalog of over 300 genes and 400 diseases (with genes involved in drug response), providing a single platform to study relationships between drugs, diseases and genes. Finally, Kann and co-workers have recently mapped all human SNPs and disease mutations (from OMIM [76] and Swiss-Prot [81]) to their corresponding protein domain sites and created a resource for the domain mapping of domain mutations, the DMDM site [82]. DMDM aggregates all the information about human mutations and provides coordinates of all mutations within the human domains. Users will find that most of these databases are freely available (Genecards is limited to nonprofit institutions) and their interface varies in flexibility and convenience. Almost all of them can be easily searched using related words in the query (disease or gene). In addition, the use of standard vocabularies and ontologies within all these databases needs to expand beyond Gene Ontology, so that descriptions of disease phenotypes, cytological changes, and molecular mechanisms can be well-defined and standardized for better discoverability, correlations, and mining. In general, while these databases provide an excellent resource, only a small proportion of the genomic data known to be involved in an inherited disease have both known gene sequence and phenotype. A summary of these resources and others described below can be found in Tables 3 and 4 and in a recent review by Kann [83]. Another major challenge is the integration and organization of phenotypic databases. The NIH, recognizing this need, launched the whole genome association studies. The NCBI's database, dbGaP [84] provides open and controlled access to summary and individual data respectively for several genotype association studies.

3.2 Prediction of deleterious SNPs

Currently, the dbSNP database contains approximately 20 million validated SNPs; yet, their impact on human health is known only for a small fraction of them. The increasing gap between the number of available SNP data and the amount of annotated variants highlights the need for developing computational methods to predict functional SNPs. Recently, several algorithms have been created to predict the effect of non-synonymous coding or missense SNPs [96, 97]. These methods are binary classifiers that use empirical rules [89, 90], machine learning, and statistical algorithms [86–88, 91, 92, 98] to discriminate between disease-related and neutral missense SNPs. The input information for missense SNP predictors is mainly derived from protein sequence, structure, and evolutionary analysis. Sequence information describes the residue composition of the mutated protein and its chemico-physical properties. The structural features provide information about the residue interactions that occur in the mutated region, as well as those that occur non-locally which cannot be detected from sequence analysis alone. A multiple sequence alignment of the protein family provides information about the evolutionary conservation of the mutated residue. A new

Database	URL	Explanation
OMIM [76]	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM	Catalog of human genes and genetic disorders
Genecards [78]	http://www.genecards.org	A compendium of genes, proteins and diseases.
Swissprot [81]	http://www.ebi.ac.uk/swissprot	Database of protein sequences with disease annotation.
DMDM [82]	http://bioinf.umbc.edu/dmdm	Domain mapping of disease mutations, it aggregates all human SNP and disease mutations at the protein domain level.
PhenomicDB [79]	http://www.phenomicDB.de	Phentoytpe-gentotype database integrating data from multiple organisms.
Gene2Disease	http://www.ogic.ca/projects/g2d_2	A database of candidate genes for mapped inherited human diseases.
Orthodisease [85]	http://orthodisease.cgb.ki.se	Eukaryotic Ortholog Groups for Disease Genes.
PhenoGo [77]	http://www.PhenoGO.org	Computed database that provides phenotypic context to existing associations between gene products and Gene Ontology (GO) for multiple organisms.
PharmaGKB [80]	http://www.pharmgkb.org	Pharmacogenetics research database.

Table 3: **Databases with disease annotation.**

class of recently developed methods includes information from functional annotations or functional predictions [29, 92]. Although the algorithms to detect deleterious missense SNPs so far developed can perform quite accurately, they do not provide any information regarding the SNPs' associated pathologies. To overcome this limitation, gene prioritization methods have been developed [99]. Gene prioritization methods are based on the assumption that similar genes are involved in similar biological processes, allowing transferring of disease associations between similar genes. Gene prioritization methods combine different knowledge sources (i.e. functional annotations, protein-protein interactions, biological pathways, and literature information) to rank candidate genes [93–95, 100]. When the genes are poorly annotated in human some of the methods use functional annotations from close homologs. In summary, the methods here discussed predict disease-related missense SNPs and their pathologic effect, but none of them are able to predict the effect of multiple SNPs including the non-coding SNPs. One of the main challenges for the recent future of bioinformatics will be to develop statistical methods that estimate the disease risk associated with a group of SNPs; accomplishing this goal will facilitate the study of disease in the context of complete genomes. Applying these new algorithms to disease prevention and medical diagnosis will have a strong impact on human life style habits, health policies, treatment of diseases, and reduction of health care costs.

Selected tools for disease-related SNPs detection		
Resource	**URL**	**Explanation**
MutPred [86]	http://mutpred.mutdb.org/	Provides structural and functional annotation.
PANTHER [87]	http://www.pantherdb.org/	Hidden Markov model-based tool multiple sequence alignment of protein families.
PhD-SNP [88]	http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi	SVM-based method based on protein sequence information.
PolyPhen [89]	http://genetics.bwh.harvard.edu/pph/	Uses straightforward physical and comparative considerations.
SIFT [90]	http://blocks.fhcrc.org/sift/SIFT.html	Based on sequence homology and the physical properties of amino acids.
SNAP [98]	http://www.rostlab.org/services/SNAP	NN-based method for the detection of functional SNPs
SNPs3D [91]	http://www.snps3d.org/	Based on structure and sequence analysis.
SNPs&GO [92]	http://snps-and-go.biocomp.unibo.it/	SVM-based method including functional annotation.
Selected tools for gene prioritization		
Resource	**URL**	**Explanation**
Endeavour [93]	http://www.esat.kuleuven.be/endeavour	Based on functional annoatation, includes several genomics data.
MedSim [94]	http://www.funsimmat.de/	Functional annotation for genes and proteins in human and mouse.
PhenoPred [95]	http://www.phenopred.org/	Protein–protein interaction gene–disease associations, protein functional information.
ToppGene [59]	http://toppgene.cchmc.org/	mouse phenotype data, human gene annotations and literature.

Table 4: **Selected tools for SNP annotation.**

References

[1] Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141.
Review on high throughput sequencing technologies, and their impact on genetics research.

[2] 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.
Pilot project of the 1000 Genomes Project that aims to catalog normal variation by sequencing 2500 healthy individuals.

[3] Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145.
Review on the technologies behind high throughput sequencing.

[4] Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6:S13–S20.
Review discussing various algorithms to discover structural variation using sequencing data.

[5] Dalca AV, Brudno M (2010) Genome variation discovery with high-throughput sequencing data. Brief Bioinform 11:3–14.
Review discussing various algorithms to discover multitude of classes of genomic variation using sequencing data.

[6] Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, et al. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16:1182–1190.
One of the first resources to catalog normal indel variation in human genomes.

[7] Lee S, Hormozdiari F, Alkan C, Brudno M (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat Methods 6:473–474.
The first method to use a distribution-based clustering approach, allowing the detection of smaller indels and explicitly modeling heterozygosity.

[8] Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727–32.
The first study to systematically discover germline structural variation in a human genome using paired-end sequences.

[9] Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97.
Review on structural variation, its effects, and ways to discover them including experimental methods.

[10] Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426.
One of the first studies to use NGS data to detect structural variants, including using the linking signature for detecting insertions larger than the insert size.

[11] Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64.
The first study to generate high-quality sequences of structural variation breakpoints.

[12] Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001) Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11:1005–17.
The first study to annotate segmental duplications in the human genome assembly.

[13] Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297:1003–7. **The first study to describe the use of read depth to detect segmental duplications from whole-genome shotgun sequence (WGS) data.**

[14] Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41:1061–1067.
The first study to characterize segmental duplications in personal genomes using NGS data, without the limitations of array-based methods.

[15] Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, et al. (2010) Diversity of human copy number variation and multicopy genes. Science 330:641–646.
A catalog and population genetics analysis of copy number polymorphic genes in many populations, and the first study to extensively genotype duplicated paralogs in human genomes.

[16] Hollox EJ, Huffmeier U, Zeeuwen PLJM, Palla R, Lascorz J, et al. (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40:23–25.
Study describing the association of the defensin gene family with psoriasis.

[17] Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, et al. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440.
Study describing the association of the *CCL3L1* gene family with HIV susceptibility.

[18] Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, et al. (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet 79:439–448.
Study describing the association of the defensin gene family with Crohn disease.

[19] McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, et al. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet 40:1107–1112.
Study describing the association of the *IRGM* gene family with Crohn disease.

[20] Mefford HC, Clauin S, Sharp AJ, Moller RS, Ullmann R, et al. (2007) Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. Am J Hum Genet 81:1057–1069.
Study describing the association of the 17q12 locus with renal disease, diabetes, and epilepsy.

[21] Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, et al. (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet 41:160–162.
Study describing the association of the 15q13.3 locus with epilepsy.

[22] Eichler EE, Zimmerman AW (2008) A hot spot of genetic instability in autism. N Engl J Med 358:737–739.
Comment on Weiss et al., NEJM, 2008 that describes the association of the 16p11.2 locus with autism.

[23] International HapMap Consortium (2003) The international HapMap project. Nature 426:789–796.
An extensive catalog of SNPs and small indels in the human genome.

[24] International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–320.
An extensive catalog of SNPs and small indels in the human genome.

[25] Scharpf RB, Ting JC, Pevsner J, Ruczinski I (2007) SNPchip: R classes and methods for SNP array data. Bioinformatics 23:627–628.
R libraries to analyze SNP chips.

[26] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80.
R libraries to analyze biological data,.

[27] Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, et al. (1999) A general approach to single-nucleotide polymorphism discovery. Nat Genet 23:452–456.

[28] Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858.
The study that describes the ubiquitously used NGS mapper, MAQ.

[29] Li R, Li Y, Fang X, Yang H, Wang J, et al. (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res 19:1124–1132.
A SNP detection algorithm using high throughput sequencing data.

[30] Anson EL, Myers EW (1997) ReAligner: a program for refining DNA sequence multi-alignments. J Comput Biol 4:369–383.

[31] Hoberman R, Dias J, Ge B, Harmsen E, Mayhew M, et al. (2009) A probabilistic approach for SNP discovery in high-throughput human resequencing data. Genome Res 19:1542–1552.
Probabilistic SNP detection algorithm using high throughput sequencing data.

[32] Chen K, McLellan MD, Ding L, Wendl MC, Kasai Y, et al. (2007) PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. Genome Res 17:659–666.

[33] McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, et al. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19:1527–1541.
The first whole-genome sequencing of a human individual using two-base encoding.

[34] Dalca AV, Rumble SM, Levy S, Brudno M (2010) VARiD: a variation detection framework for color-space and letter-space platforms. Bioinformatics 26:i343–i349.
First algorithm to incorporate information from different sequencing technologies to improve variant detection.

[35] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The sequence alignment/map format and samtools. Bioinformatics 25:2078–2079.
Describes the commonly used SAM and BAM file formats to store sequence mapping information, and basic SNP and indel detection algorithms.

[36] Vallania FLM, Druley TE, Ramos E, Wang J, Borecki I, et al. (2010) High-throughput discovery of rare insertions and deletions in large cohorts. Genome Res
Algorithm for detection of rare indels and substitutions in pooled-DNA sequencing in Illumina 1G/2G/High-Seq next-gen sequencing using synthetic control information.

[37] Albers CA, Lunter G, Macarthur DG, McVean G, Ouwehand WH, et al. (2010) Dindel: Accurate indel calls from short-read data. Genome Res
Indel algorithm using high throughput sequencing.

[38] Le SQ, Durbin R (2010) Snp detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res
Algorithm to detect SNPs using low-coverage sequencing data.

[39] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics
A method that is able to detect indels with base-pair breakpoint resolution using NGS data, on the basis of the anchored split mapping signature.

[40] Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17:1665–1674.

[41] Day N, Hemmaplardh A, Thurman RE, Stamatoyannopoulos JA, Noble WS (2007) Unsupervised segmentation of continuous genomic data. Bioinformatics 23:1424–1426.
One of the most commonly used HMM segmentation algorithms.

[42] Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. Genome Res 16:1136–1148.

[43] Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet 40:1199–1203.

[44] Itsara A, Cooper GM, Baker C, Girirajan S, Li J, et al. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84:148–161.
A large collection and population analysis of copy number variants, commonly used as control data in medical studies.

[45] Mefford HC, Cooper GM, Zerr T, Smith JD, Baker C, et al. (2009) A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. Genome Res 19:1579–1585.

[46] Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40:1253–1260.

[47] Yavaş G, Koyutürk M, Ozsoyoğlu M, Gould MP, Laframboise T (2010) ÇOKGEN: software for the identification of rare copy number variation from SNP microarrays. Pac Symp Biocomput :371–382.

[48] Sharp AJ, Itsara A, Cheng Z, Alkan C, Schwartz S, et al. (2007) Optimal design of oligonucleotide microarrays for measurement of DNA copy-number. Hum Mol Genet 16:2770–2779.

[49] Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al. (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88.

[50] Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, et al. (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 38:1038–1042.
Study that describes disease associations of multiple genomic loci.

[51] Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–712.
Largest collection of copy number variation in human genomes using array based technology.

[52] Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, et al. (2009) A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457:877–881.
Study describing the accelerated rate of segmental duplications in the genomes of great apes and humans.

[53] Locke DP, Archidiacono N, Misceo D, Cardone MF, Deschamps S, et al. (2003) Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. Genome Biol 4:R50.

[54] Volik S, Zhao S, Chin K, Brebner JH, Herndon DR, et al. (2003) End-sequence profiling: sequence-based analysis of aberrant genomes. Proc Natl Acad Sci U S A 100:7696–701.
The first study describing end-sequence profiling.

[55] Lee S, Cheran E, Brudno M (2008) A robust framework for detecting structural variations in a genome. Bioinformatics 24:i59–i67.

[56] Hormozdiari F, Alkan C, Eichler EE, Sahinal SC (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res 19:1270–1278.
One of the first comprehensive tools for structural variant detection; supports most basic signatures and uses soft clustering.

[57] Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, et al. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods 6:99–103.
First method to discover large somatic CNVs in cancer genomes using sequencing data.

[58] Korbel J, Abyzov A, Mu X, Carriero N, Cayting P, et al. (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10:R23.

[59] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, et al. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6:677–681.

[60] Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19:1586–1592.

[61] Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12:656–664.
BLAT is one of the most commonly used sequence search algorithms developed for longer reads.

[62] Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M (2010) Detecting copy number variation with mated short reads. Genome Res 20:1613–1622.
One of the first algorithms to discover structural variation using sequencing data by incorporating mltiple sequencing signatures.

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

[64] Abel HJ, Duncavage EJ, Becker N, Armstrong JR, Magrini VJ, et al. (2010) SLOPE: a quick and accurate method for locating non-snp structural variation from targeted next-generation sequence data. Bioinformatics 26:2684–2688.

[65] Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, et al. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res 20:623–635.
Describes the HYDRA algorithm for structural variation detection.

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

[67] Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, et al. (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 26:1895–1896.

[68] Sindi S, Helman E, Bashir A, Raphael BJ (2009) A geometric approach for classification and comparison of structural variants. Bioinformatics 25:i222–i230.

[69] Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, et al. (2010) Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 26:1277–1283.
The first algorithm that incorporates local and de novo assembly with one-end-anchored clustering signatures to discover and anchor long novel sequence insertions.

[70] Li R, Zhu H, Ruan J, Qian W, Fang X, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272.

[71] Lam HYK, Mu XJ, Stütz AM, Tanzer A, Cayting PD, et al. (2010) Nucleotide-resolution analysis of structural variants using breakseq and a breakpoint library. Nat Biotechnol 28:47–55.
Collection of validated structural variation breakpoints at the basepair resolution and method for rapid SV genotyping.

[72] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.
The first high coverage NGS dataset of an individual. This data set has been used in many subsequent studies.

[73] Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, et al. (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Methods 7:576–577.
The first cache-oblivious read mapper to increase efficiency.

[74] Collins FS, Barker AD (2007) Mapping the cancer genome. pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. Sci Am 296:50–57.
Overview of the Cancer Genome Atlas project.

[75] Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888.
Review of challenges in the linkage analysis of Mendelian Diseases and GWAS studies.

[76] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33:D514–D517. **A comprehensive knowledgebase of human genes and genetic disorders.**

[77] Lussier Y, Borlawsky T, Rappaport D, Liu Y, Friedman C (2006) PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. Pac Symp Biocomput :64–75 **This computationally-derived resource is primarily intended to provide phenotypic context (cell type, tissue, organ, and disease) for mining existing associations between gene products and GO terms specified in the Gene Ontology databases.**

[78] Safran M, Dalah I, Alexander J, Rosen N, Stein TI, et al. (2010) GeneCards Version 3: the human gene integrator. Database (Oxford) 2010:baq020.
GeneCards provides a suite of tools that integrate information from over 70 sources including OMIM, constituting a single location to retrieve available information for over 29,000 human genes including relationships to diseases when available.

[79] Kahraman A, Avramov A, Nashev LG, Popov D, Ternes R, et al. (2005) PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. Bioinformatics 21:418–420.

[80] Altman RB (2007) PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. Nat Genet 39:426. **A database with manual annotation of drug-gene relationships used in Pharmacogenetics research.**

[81] UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38:D142–D148. **The Uniprot database includes the Swissprot protein database, which provides information about disease mutations.**

[82] Peterson TA, Adadey A, Santana-Cruz I, Sun Y, Winder A, et al. (2010) DMDM: domain mapping of disease mutations. Bioinformatics 26:2458–2459. **DMDM is a database in which each disease mutation can be displayed by its gene, protein or domain location. DMDM provides a unique domain-level view where all human coding mutations are mapped on the protein domain.**

[83] Kann MG (2010) Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Brief Bioinform 11:96–110. **A review of computational approaches for prioritization of disease genes.**

[84] Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, et al. (2007) The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39:1181–1186. **NCBI public repository for individual-level phenotype, exposure, genotype and sequence data and their associations.**

[85] O'Brien KP, Westerlund I, Sonnhammer ELL (2004) OrthoDisease: a database of human disease orthologs. Hum Mutat 24:112–119. **A database of human disease orthologs.**

[86] Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, et al. (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744–2750.
MutPred: Probabilistic method based on Random Forest algorithm to predict the impact of missense SNPs using sevaral machine learning approaches.

[87] Thomas PD, Kejariwal A (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A 101:15398–15403.
PANTHER method for the prediction of deleterious missense SNPs using Hidden Markov Models.

[88] Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734.
Machine learning-based method for the prediction of deleterious missense SNPs using evolutionary information.

[89] Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900. **Method to predict the effect of missense SNP on protein structure and function.**

[90] Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814. **Method for the predicting of deleterious missense SNPs using evolutionary information from multiple sequence alignment.**

[91] Yue P, Moult J (2006) Identification and analysis of deleterious human SNPs. J Mol Biol 356:1263–1274.
Prediction of disease-related missense SNPs using protein structural information.

[92] Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30:1237–1244.
SNPs&GO: Method for prediction of deleterious missense SNPs using functional annotation.

[93] Tranchevent LC, Barriot R, Yu S, Vooren SV, Loo PV, et al. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res 36:W377–W384.
Gene prioritization method using ontologies and annotations, protein-protein interactions, cis-regulatory information, gene expression, sequence information and text-mining.

[94] Schlicker A, Lengauer T, Albrecht M (2010) Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. Bioinformatics 26:i561–i567.
MedSim: ranking of disease candidate genes based on functional comparisons of GO terms. It uses functional annotations of known disease genes for assessing the similarity of diseases and the disease relevance of candidate genes.

[95] Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, et al. (2008) An integrated approach to inferring gene-disease associations in humans. Proteins 72:1030–1037. **Algorithm for the detection of gene-disease associations based on protein-protein interaction network, known gene-disease associations, protein sequence, and protein functional information.**

[96] Karchin R (2009) Next generation tools for the annotation of human SNPs. Brief Bioinform 10:35–52. **Review of the computational approaches for the detection of functional SNPs and suggestions for the improvements of the next generation's methods.**

[97] Mooney S (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform 6:44–56. **Overview of the available tools and resources to analyze functional variation from the perspective of structure, expression, evolution and phenotype.**

[98] Bromberg Y, Yachdav G, Rost B (2008) SNAP predicts effect of mutations on protein function. Bioinformatics 24:2397–2398. **Method for the prediction of functional impact of missense SNPs.**

[99] Tranchevent LC, Capdevila FB, Nitsch D, Moor BD, Causmaecker PD, et al. (2010) A guide to web tools to prioritize candidate genes. Brief Bioinform published online March 21.
Review of 19 computational web tools for human gene prioritization and summary of various biological problems to which they have been successfully applied.

[100] Cheng D, Knox C, Young N, Stothard P, Damaraju S, et al. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res 36:W399–W405. **Text mining based method to infer relationships between human diseases, genes, mutations, drugs and metabolites.**