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

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

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 ∼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 (log2 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 log2 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 log2 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.
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 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
Databases with disease annotation.

<table>
<thead>
<tr>
<th>Database</th>
<th>URL</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMDM [82]</td>
<td><a href="http://bioinf.umbc.edu/dmdm">http://bioinf.umbc.edu/dmdm</a></td>
<td>Domain mapping of disease mutations, it aggregates all human SNP and disease mutations at the protein domain level.</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Resource</th>
<th>URL</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhD-SNP [88]</td>
<td><a href="http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi">http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi</a></td>
<td>SVM-based method based on protein sequence information.</td>
</tr>
<tr>
<td>SIFT [90]</td>
<td><a href="http://blocks.fhcrc.org/sift/SIFT.html">http://blocks.fhcrc.org/sift/SIFT.html</a></td>
<td>Based on sequence homology and the physical properties of amino acids.</td>
</tr>
<tr>
<td>SNPs3D [91]</td>
<td><a href="http://www.snps3d.org/">http://www.snps3d.org/</a></td>
<td>Based on structure and sequence analysis.</td>
</tr>
</tbody>
</table>

### Selected tools for gene prioritization

<table>
<thead>
<tr>
<th>Resource</th>
<th>URL</th>
<th>Explanation</th>
</tr>
</thead>
</table>

Table 4: Selected tools for SNP annotation.
References


The first study to characterize segmental duplications in personal genomes using NGS data, without the limitations of array-based methods.


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.


Study describing the association of the defensin gene family with psoriasis.


Study describing the association of the CCL3L1 gene family with HIV susceptibility.


Study describing the association of the defensin gene family with Crohn disease.


Study describing the association of the IRGM gene family with Crohn disease.


Study describing the association of the 17q12 locus with renal disease, diabetes, and epilepsy.


Study describing the association of the 15q13.3 locus with epilepsy.


Comment on Weiss et al., NEJM, 2008 that describes the association of the 16p11.2 locus with autism.


An extensive catalog of SNPs and small indels in the human genome.


An extensive catalog of SNPs and small indels in the human genome.


R libraries to analyze SNP chips.


R libraries to analyze biological data.


The study that describes the ubiquitously used NGS mapper, MAQ.


A SNP detection algorithm using high throughput sequencing data.


Probabilistic SNP detection algorithm using high throughput sequencing data.


The first whole-genome sequencing of a human individual using two-base encoding.


First algorithm to incorporate information from different sequencing technologies to improve variant detection.


Describes the commonly used SAM and BAM file formats to store sequence mapping information, and basic SNP and indel detection algorithms.


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.


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.


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.


One of the most commonly used HMM segmentation algorithms.


A large collection and population analysis of copy number variants, commonly used as control data in medical studies.


Study that describes disease associations of multiple genomic loci.


The first study describing end-sequence profiling.


One of the first comprehensive tools for structural variant detection; supports most basic signatures and uses soft clustering.

First method to discover large somatic CNVs in cancer genomes using sequencing data.


[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.


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.


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.


[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.**