Chapter 8

Exome Sequencing Analysis: A Guide to Disease Variant Detection

Ofer Isakov, Marie Perrone, and Noam Shomron

Abstract

Whole exome sequencing presents a powerful tool to study rare genetic disorders. The most challenging part of using exome sequencing for the purpose of disease-causing variant detection is analyzing, interpreting, and filtering the large number of detected variants. In this chapter we provide a comprehensive description of the various steps required for such an analysis. We address strategies in selecting samples to sequence, and technical considerations involved in exome sequencing. We then discuss how to identify variants, and methods for first annotating detected variants using characteristics such as allele frequency, location in the genome, and predicted severity, and then classifying and prioritizing the detected variants based on those annotations. Finally, we review possible gene annotations that may help to establish a relationship between genes carrying high-priority variants and the phenotype in question, in order to identify the most likely causative mutations.

Key words Exome, Exome sequencing, Variant detection, Exome analysis, Disease variants, Variant annotation, Variant prioritization, Gene annotation

1 Introduction

Whole exome sequencing (WES) presents a powerful tool to study rare genetic disorders [1]. The exome represents the 1–2 % of the human genome that is translated into proteins. By analyzing the exome sequence of a single individual who displays an unusual phenotype or is affected by a rare genetic disease, the mutation that causes that disease or phenotype can be discovered. This method yields significant benefits. Not only can it help to identify the exact cause of a particular condition [2], but it can also identify genes not previously known to be associated with a biological pathway, or inspire new therapeutics for the condition [3].

Prior to the advent of deep-sequencing technologies, the discovery of mutations causing rare diseases was challenging. Traditional methods, such as Sanger sequencing of specific genes, linkage analysis, genome-wide association studies, karyotyping, and homozygosity

Noam Shomron (ed.), *Deep Sequencing Data Analysis*, Methods in Molecular Biology, vol. 1038,
DOI 10.1007/978-1-62703-514-9_8, © Springer Science+Business Media New York 2013

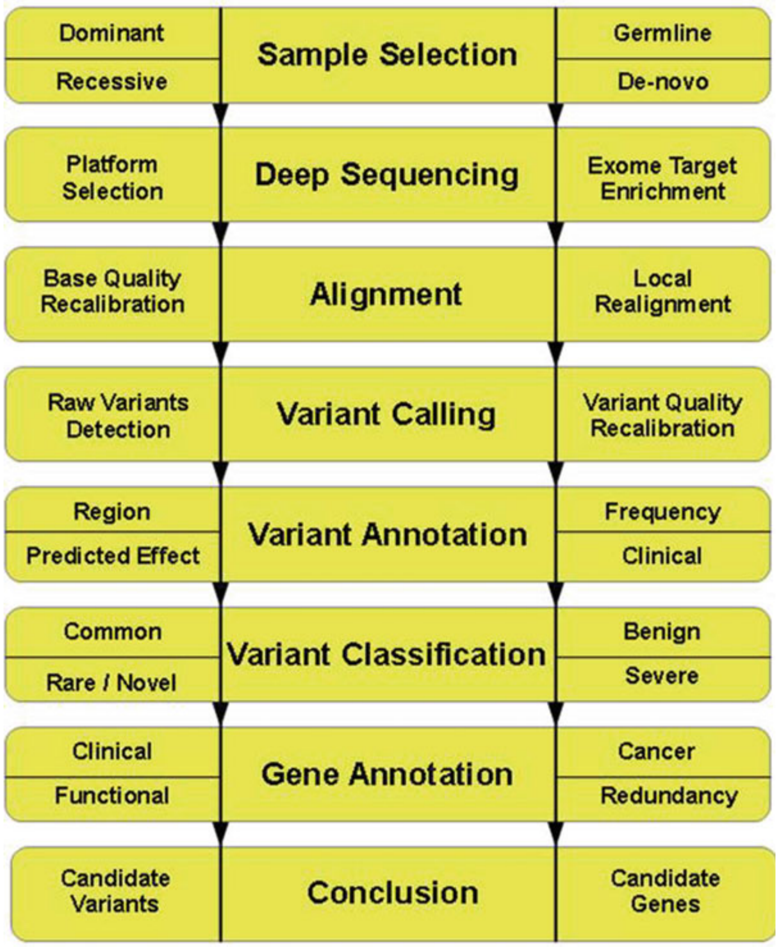

Fig. 1 Deep sequencing disease variant detection work flow

mapping, have many limitations. These methods are slow, less sensitive, or discerning enough to pinpoint actual causative mutations, or rely on large samples of affected individuals and prior knowledge about the genes involved in the disease. In order to detect rare mutations, a more powerful and sophisticated approach is needed. As the throughput of genetic sequencing has increased through the use of massively parallel sequencing [4], it has become feasible to discover specific rare, novel disease-causing variants with only the exomes of a few affected individuals.

1.1 Whole Exome Sequencing vs. Whole Genome Sequencing (Table 1)

The fundamental limitation to sequencing only the exome is that, of course, the causative mutation will only be found if it is in a known functional region of the genome. Nearly 99 % of the genome is ignored in exome sequencing. Exome sequencing generally includes exons and certain select functional elements, such as microRNAs, but genes that have not yet been discovered will not be sequenced, nor will mitochondrial DNA, introns, or any other part of the genome that has not yet been found to be functional (also see further explanation below). Annotation of the human genome still has a long way to go, and many elements of the genome are yet to be discovered [5]. Also, structural variants and copy number variations are harder to identify using targeted exome capture [6].

However, choosing to sequence only the exome has many advantages. For starters, WES is both faster than and one-sixth the cost of whole genome sequencing (WGS) [7]. Whereas sequencing more than one genome would more likely be cost-prohibitive, sequencing more than one exome is usually feasible, and, as we will discuss later, can greatly improve the chances of identifying the causative mutation. Furthermore, WES inherently enriches the data for only the most relevant variations. This enrichment occurs for a few reasons. First, mutations in noncoding regions of the genome are less likely to cause severe phenotypes. Studies of genetic diseases demonstrate that genetic diseases are much more likely to be caused by mutations in the coding regions of the genome. Missense mutations in coding regions are the most common known type of disease-causing mutations [8]. Thus, by looking only at the mutations in the exome, researchers efficiently focus only on the most likely candidate mutations, and they eliminate a significant portion of benign mutations.

As the cost of genome sequencing continues to drop, it may soon be practical to sequence the entire genome rather than just the exome and detect more mutations, which will mean that variations outside the known functional regions will not be missed. However, even if it is viable to sequence the whole genome, it may not make sense to. While the effects of variations in protein-coding regions on transcription and translation are fairly predictable, the effects of mutations in noncoding regions may be unclear. In protein-coding regions, mutations can be classified based on the effect they have on the protein-coding sequence. Mutations in the noncoding region, on the other hand, are more difficult to interpret. Further complicating the matter is the fact that sequencing the whole genome will result in exponentially more variations, since variations in nonfunctional regions are not subject to the same negative selection pressure as variations in the coding regions [9]. Thus, WGS carries with it the risk that the true causative mutation will be buried under a pile of uninformative nonfunctional mutations. Having more information is not always better, if the information cannot be interpreted. For now, sequencing only the exome provides the most efficient means to identify specific variations causing genetic disease.

1.2 Successful Exome Analysis

One of the first reported successful uses of exome sequencing to discover a novel causative mutation was in 2009 [2, 10]. The exomes of four individuals with Miller syndrome, a rare

Mendelian disease that causes craniofacial and limb development abnormalities, were sequenced and analyzed to discover a mutation in a single gene, DHODH, which was not previously associated to the disease. In the following 2 years, exome sequencing was successfully used to implicate over 30 novel variants in a wide variety of genetic diseases, ranging from autism to hearing loss [10]. However, exome sequencing is not an infallible method for causative variant detection [11]. Assuming that the causative mutation is in the exome, it can still be missed if analysis of the found variants is not done properly.

1.3 Chapter Outline

The most challenging part of using exome sequencing for the purpose of variant detection is analyzing, interpreting, and filtering the large number of detected variants. A typical exome sequencing run result in tens of thousands of small single-nucleotide variants (SNVs), insertions, and deletions [12]. Most successful studies use a stepwise filtering process to extract the most likely variants from the pool of thousands. The steps in this process may include searching for variations in known variant databases, such as dbSNP, or comparing variations to variations in other relevant exomes. Ideally, the filtering steps would identify one variant, or a small number of variants, that can be verified by Sanger sequencing or some other method. Unfortunately, researchers are often left with too many possible variants, or none at all. In order to establish exome analysis as the standard tool for variant detection in genetic disease, a robust, methodical analysis pipeline is essential. This chapter describes the process of detecting disease-causing variations in exomes. First, we address strategies in selecting samples to sequence, and technical considerations involved in exome sequencing. Next, we discuss how to identify variants, and methods for first annotating detected variants using characteristics such as allele frequency, location in the genome, and predicted severity, and then classifying and prioritizing the detected variants based on those annotations. Finally, we review possible gene annotations that may help to establish a relationship between genes carrying high-priority variants and the phenotype in question, in order to identify the most likely causative mutations.

2 Materials

Throughout this chapter we describe various tools that may be employed in the different analysis steps. However, other established tools and software are either already available or being developed. We therefore urge readers to always keep up to date regarding novel and useful tools relevant for their analysis (*see* **Notes 6** and 7). The tools described throughout the chapter include the following:

Table 1
Available tools for exome variant detection and evaluation

Tool/resource	Function	Reference	Link
OMIM	Clinical association	–	http://omim.org/
NHGRI-GWAS catalog	Clinical association	[47]	http://www.genome.gov/gwastudies/
Genetic Association Database (GAD)	Clinical association	[58]	http://geneticassociationdb.nih.gov/
Phenopedia and Genopedia	Clinical association	[59]	http://www.hugenavigator.net/
Human Phenotype Ontology	Clinical association	[60]	http://www.human-phenotype-ontology.org/
Mouse Genome Database (MGD)	Clinical association	[61]	http://www.informatics.jax.org/
GERP++	Conservation	[39]	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html
PhyloP	Conservation	[40]	http://compgen.bscb.cornell.edu/phast/help-pages/
PhastCons	Conservation	[41]	http://compgen.bscb.cornell.edu/phast/
Duplicated Genes Database (DGD)	Duplicate genomic regions	–	http://dgd.genouest.org/
dbDNV	Duplicate genomic regions	[65]	http://goods.ibms.sinica.edu.tw/DNVs/
NHLBI Exome Sequencing Project	Known variant database	–	http://evs.gs.washington.edu/EVS/
dbSNP	Known variant database	[20]	http://www.ncbi.nlm.nih.gov/SNP/
Complete Genomics	Known variant database	[35]	http://www.completegenomics.com/
1000 Genomes Project	Known variant database	[9]	http://www.1000genomes.org/
miRBase	miRNA sequence database	[29]	http://www.mirbase.org/
TargetScan	miRNA target database	[31]	http://www.targetscan.org/
miRNA.org	miRNA target database	[32]	http://www.microrna.org/microrna/home.do
Rfam	ncRNA sequence database	[30]	http://rfam.sanger.ac.uk/

(continued)

Table 1
(continued)

Tool/resource	Function	Reference	Link
Biocarta	Pathway annotation	–	http://www.biocarta.com
KEGG	Pathway annotation	[55]	http://www.genome.jp/kegg/
REACTOM	Pathway annotation	[56]	http://www.reactome.org/ReactomeGWT/entrypoint.html
BioGRID interaction database	Protein interactions	[62]	http://thebiogrid.org/
STRING protein functional interactions database	Protein interactions	[63]	http://string-db.org/
Human protein–protein interaction prediction database (PIPs)	Protein interactions	[64]	http://www.compbio.dundee.ac.uk/www-pips/
Catalog of Somatic Mutations in Cancer (COSMIC)	Somatic mutations	[67]	http://www.sanger.ac.uk/genetics/CGP/cosmic/
Annovar	Variant annotation	[33]	http://www.openbioinformatics.org/annovar/
Ensembl SNP effect predictor	Variant annotation	[34]	http://ensembl.org/Homo_sapiens/UserData/UploadVariations
SnpEff	Variant annotation	[48]	http://snpeff.sourceforge.net/
VariantClassifier	Variant annotation	[49]	http://www.jcvi.org/cms/research/projects/variantclassifier
SNPnexus	Variant annotation	[50]	http://www.snp-nexus.org/
MutationAssessor	Variant annotation	[51]	http://mutationassessor.org/
UCSC genome browser	Variant annotation and visualization	[27]	http://genome.ucsc.edu/
Genome Analysis Toolkit (GATK)	Variant discovery	[19]	http://www.broadinstitute.org/gatk/
SIFT	Variant severity	[42]	http://sift.jcvi.org/
PolyPhen2	Variant severity	[43]	http://genetics.bwh.harvard.edu/pph2/
MutationTaster	Variant severity	[44]	http://www.mutationtaster.org/
LRT	Variant severity	[45]	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2752137/

3 Methods

3.1 Sample Selection

Although there are cases in which sequencing of a single individual's whole exome revealed the causative variation [13], such a detection strategy is limited. In order to increase the probability of finding a single novel causative variation, it is recommended to, if possible, sequence a few additional relevant exomes [11]. In this section we describe strategies for choosing which additional individuals' exomes should be sequenced. The optimal strategy will depend on the mode of inheritance of the mutation in question. Information about the rarity or the prevalence of the disease and knowledge of the affected individual of interest's genetic pedigree will help in surmising the mode of inheritance (*see* **Note 1**).

3.1.1 Sample Selection for a Recessive Mode of Inheritance

Recessive mutations are easier to identify by filtering for homozygosity, so it is less crucial to gather exomes from other individuals. If the only exome available is that of the affected individual, having a recessive mutation is to the researcher's advantage, as homozygosity can be used to help narrow down the pool of possible mutations. Additionally, if there is suspected consanguinity in the affected family, the variant will most likely come from a region of the genome that is identical by descent.

If additional exomes are available, a useful strategy to use is to compare exomes from affected and unaffected individuals in the same kindred. The strategy here is to search for variations homozygous only to the affected individuals and to none of the unaffected individuals. It is helpful to choose affected individuals who are as distantly related as possible in order to minimize the amount of similar benign variations.

3.1.2 Sample Selection for a Dominant Mode of Inheritance

If the disease is suspected to follow a dominant mode of inheritance and neither parent of the affected individual of interest is affected, then there are two possible scenarios: either there is partial penetration or the mutation has arisen de novo. While partial penetration is more likely, the discovery of a causative mutation will be more difficult. Having a de novo mutation is much rarer; the typical exome contains zero to three de novo mutations. However, the sample selection strategy is simple: sequence the exome of the affected individual and those of his or her healthy parents, and then look for the variant that is present in the child but not in the parents. It is important to note that this method requires both a high sensitivity and high specificity so as not to miss mutations or mistake sequencing artifacts for de novo mutations.

More often, the dominant genetic mutation will be inherited from one of the parents. Here, there are two possible sample selection strategies, which can be used in tandem. The first strategy is to sequence exomes from as many unrelated affected individuals as possible and search for a common affected gene. It is useful to choose affected individuals from similar geographical ancestry so as to minimize the amount of benign variance and narrow the pool of candidate genes [2]. The exomes should be compared for variations found in the same gene (though not necessarily the same loci) across all exomes.

The second strategy, similar to the one mentioned in the recessive mode of inheritance section, is to sequence exomes from individuals in the same kindred. By comparing unaffected and affected individuals from the same kindred, benign, private mutations can be eliminated. In the case of a dominant mutation, variants should be present in a heterozygous state in the affected population and should not be found at all in the unaffected population.

This last strategy of sequencing affected and unaffected family members is useful for both dominant and recessive mutations, so it could be the strategy of choice if the mode of inheritance is unclear.

3.1.3 Combining Strategies

These selection strategies, of course, are not mutually exclusive. Studies often use a combination of the above strategies. Both unrelated and related individuals can be included in the study. Combining selection strategies may be useful if no assumptions can be made about the mode of inheritance. For example, two related individuals' exomes, one affected, one unaffected, could be used to filter out common variants. If this method alone does not narrow down the variant pool enough, an exome from unrelated, affected individual could be used to further filter the pool of variants.

For a comprehensive review of exome sequencing strategies, *see* ref. 12.

3.2 Technical Considerations

When sequencing an exome for variant detection, technical aspects of the sequencing process should be taken into consideration. There are many different deep-sequencing platforms to choose from. There are a few commercial exome-capturing kits available (e.g., Agilent, Illumina, and Nimblegen), each with slightly different sample preparation and selection methods. Different kits have different methods of capturing or enriching the desired regions of the genome. There are three main enrichment strategies: hybridization, circularization, or PCR. Circularization and PCR are mostly used for enriching a very small amount of genetic material, less than that of an exome. The most optimal and commonly used method for targeted exome capture is hybridization [14]. In this technique, small DNA or RNA probe matching regions of the exome act as bait for matching exomic sequences in a fragment library. There are two ways to do hybrid capture: on-array or in-solution [15]. In on-array hybridization, the probes are immobilized on an array, and a surfeit of target DNA is used as compared to the number of probes. In-solution hybridization, on the other hand, uses an excess of probes as compared to target, and probes are not immobilized on a surface. In-solution hybridization may be the preferred method if there is a very limited amount of sample DNA or if the target size is small (on the order of 3 Mb). Otherwise, the two methods of hybrid capture provide comparable results [15]. Whichever method of hybridization is chosen, there are inherent errors to consider. Due to faulty hybridization, some desired regions may be missed, and there may be noncoding regions of the genome that are captured through off-target effects. If the fragment library has long fragments, the sequences that are captured might extend into the non-targeted regions of the genome. These are called "near target" sequences. Using longer fragments means that more near target sequences will be captured; using shorter fragments means that more target sequences will be missed. Different kits have different library preparation guidelines and use different size selection methods [16]. Generally, fragment size is not a determining factor when choosing a kit, but it is good to be aware of the potential for error. For more information on the errors inherent in targeted exome capture with hybridization, *see* ref. 15. One final important consideration for targeted exome capture is the definition of "exome." Different commercial kits have different definitions of "exome." It can include only the known protein-coding regions, or known coding regions and other known functional elements, such as untranslated regions (UTRs), certain microRNAs, or noncoding exons. The size of the region captured ranges from around 34,000 kb to over 50,000 kb. The choice of kit may depend on which parts of the genome a researcher wishes to capture. References 16–18 each provides detailed comparisons of specific kits (*see* **Note 2**).

3.3 Variant Calling (Figs. 1 and 2)

This section gives a general overview of the process of variant calling. The basic steps in the pipeline are initial mapping of the reads, improvement of alignments and quality scores, variant identification, and recalibration of the variants' quality scores. An in-depth review of variant calling can be found in [19].

The first step in variant calling is mapping the reads. To ensure that variants are properly identified, it is important to have good-quality data. Having adequate coverage of each base is one way to improve data quality. Depending on the source, the necessary minimum coverage for reliably detecting single-nucleotide variations is anywhere from 8× to 200×. In general, a coverage of 20× to 50× at each nucleotide is considered acceptable when identifying variations [14]. The mapping quality and accuracy can then be further improved by marking duplicate reads, local realignment, and base quality recalibration.

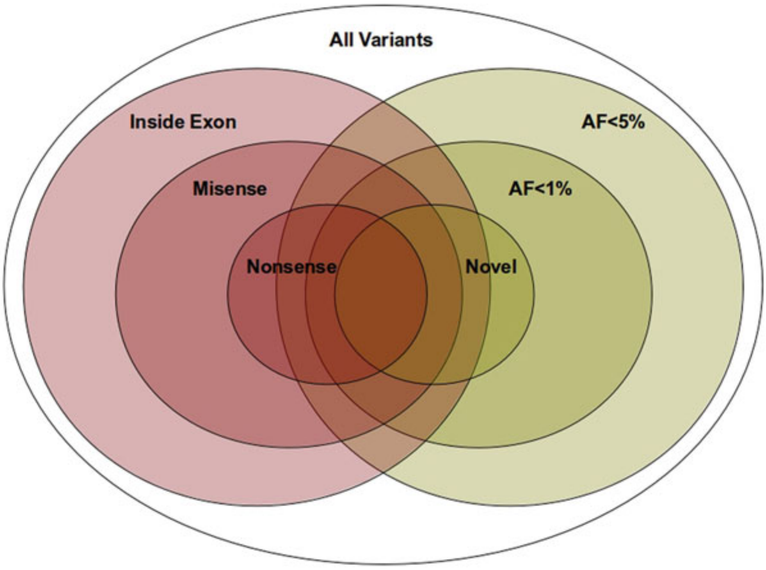

Fig. 2 Variants should be classified into several priority levels and gradually reviewed. Classifications should be designed to include higher priority classes inside lower ones in order to increase the search space without removing possibly relevant variants. In this figure, we chose two classifications (frequency and gene-effect) reducing them into only three subcategories (although more such subcategories can be added). The darker the color shade the higher the priority, with novel nonsense mutations representing the highest priority variants

Having duplicate reads may result in overestimation of coverage and quality of variants. When the sequence library is being created, PCR amplification of the DNA can result in duplicate sequences. This would make it seem like a variant has higher read coverage, and, thus, a higher quality, than it actually should, and can result in false SNVs. Therefore, it is recommended to remove all but one of the reads whose ends match to the same places on the reference genome.

Next, local realignment helps to verify that reads are mapped correctly. Initial alignment algorithms often misalign reads that begin or end with insertions or deletions (indels), mistakenly calling false nucleotide changes. Local realignment tools take into consideration other reads that map to the same region and can provide evidence for indels rather than SNVs.

Base quality recalibration adjusts the quality scores associated with each base pair, based on a sample of tentatively called variants. To determine the likelihood that the SNVs found in the reads are accurate, a subsample of reads is aligned to the reference genome, and differences between the reads and the reference genome are called and cross-referenced to a database of known SNVs (such as dbSNP [20]). Only about 1–10 % of the SNVs called should be novel (not found in the reference database). If there is a greater proportion of novel SNVs than this, the accuracy of the entire exome should be called into question. The quality of the bases is then recalibrated based on the number of unknown SNVs and a number of other factors, including the bases around the unknown SNVs.

Finally, the actual variants are ready to be called. A common file format for called variants is variant call format (VCF). SNVs are determined by simply identifying statistically significant differences between the mapped reads and the reference genome. Indel calling is based on gapped alignment methods or inferences from mapping of paired-end reads. Structural variants require paired-end sequencing to be discovered. The distance between the mapped locations of the paired reads provides a clue as to whether an inversion, translocation, or copy number variation is present.

One final step to ensure good-quality data is to recalibrate the variant quality scores. The likelihood that the called variants are accurate can be estimated by, again, comparing the discovered variants to "true" variants from a database such as dbSNP. Additionally, if exomes from related individuals were sequenced, variant genotypes can be compared between exomes for consistent inheritance patterns.

3.4 Variant Annotation

The variant calling pipeline, described in the previous section, results in a list of high-quality SNVs, insertions, and deletions. The number of discovered variants is usually determined by the size and conservation of the sequenced region. Larger less conserved regions will carry more variants and smaller, functional, conserved targets will carry fewer variants. For example, when performing WGS, millions of variants are usually detected [21–23]. WES, targeting only known functional regions in the genome, that are more conserved, will usually result in thousands of detected variants [2, 24, 25]. Although the list of variants resulting from WES is much shorter, deciphering which of the detected variants have functional significance and which are irrelevant to the phenotype in question remains a challenging task. In order to facilitate the task of pinpointing the most relevant variants, a comprehensive annotation of the different variants is necessary. The more information gathered on each variant, the easier it is to classify and prioritize it. Such prioritization is crucial when dealing with a large data set of candidate variants where the goal is to reduce the list of relevant variants considerably, and retain only a handful of suspected mutations which are later validated and whose association to the phenotype in question confirmed (*see* **Note 3**).

The annotation process involves the collection of as much data as possible on each single detected variant. Most of the annotation data should be gathered from open-access, publicly available, specialized resources. For each of the detected variants, information should be gathered regarding its:

Genomic region: Annotations regarding the region in which each variant is found in the genome are essential for early-stage prioritization since variants found in known functional regions are more likely to have a phenotypic effect than variants found in regions without any known genetic or epigenetic function. Regional information should include whether the variant is up/downstream of a known gene, whether it is located in a gene's intron, exon, splice junction, or UTRs. Different gene annotations (e.g., RefSeq [26], UCSC genes [27], Ensembl [28]) result in different regional annotations with RefSeq being the most curated version (with ~40 K records) and Ensembl the most comprehensive (~180 K records). Using a more comprehensive gene list will produce a more sensitive regional annotation, but the annotation will be less specific, as some of the annotations were not validated and thus could represent false annotations. It is also useful to annotate variants that are found inside functional regions such as miRNA and other noncoding RNA sites (found in miRBase [29] and Rfam [30], respectively). Information about known and predicted microRNA-binding sites can also be retrieved from Web servers such as TargetScan [31] or miRNA.org [32]. Finally, additional regions with known functionally relevant regulatory properties should also be considered (retrieved from the UCSC genome browser [27]).

Exonic effect: In cases where the detected variant is found inside an exon, the expected variant effect on translation should be added to the annotation. If the variant is a SNV it may be synonymous (change in the codon that does not lead to an amino acid (AA) change), non-synonymous (the codon change results in an amino acid change), or stop loss/gain (the codon changes either from or into a stop codon). If the variant is an insertion or a deletion (indel) it may be non-frameshift (an indel that does not cause the translation reading frame to shift) or frameshift (an indel causing a shift in the reading frame). It is important to note that most genes have more than one possible transcript (isoforms); therefore variants may have different effects on different transcripts of the same gene. A comprehensive annotation predicts the effect of each variant on each of the gene's possible isoforms. The exonic effect of a variant can be produced by several available tools such as ANNOVAR [33] and the Ensembl SNP effect predictor [34].

Population frequency: Determining a variant's frequency in the population can facilitate the elucidation of relevance to the phenotype in question. If the phenotype is severe and very rare it is unlikely that the causative allele will have a high frequency in the population. Therefore, each detected variant should be annotated with any available information regarding previous detection. Detected variants in sequencing experiments are recorded in designated public databases such as dbSNP [20], the 1000 genomes project [9], the

NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS/), and the Complete Genomics data [35]. These databases should contain information regarding each variant's allele frequency (the ratio between the number of times the variant allele was observed and the number of times the reference allele was observed). Most variants detected through exome sequencing will have been previously observed and will be found in one of the aforementioned public databases [36, 37]. We emphasize that a variant previously detected and recorded in one of these databases should not be automatically excluded from downstream analysis. A variant does not need to be novel in order for it to be causative and frequency rather than mere novelty should be considered for such prioritization. It is also recommended to incorporate additional allele frequency annotations gathered from previous personal experiments so as to remove common sequencing platform-derived errors and additional variants common in previously sequenced individuals.

Conservation: Variants found inside conserved sites in the genome are more likely to have deleterious effects with phenotypic consequences, since many such regions are conserved as a result of higher functionality and negative selection [38]. Thus, annotating each variant with the conservation levels and inferred constraint levels of its location using tools such as GERP++ [39], PhyloP [40], and PhastCons [41] can facilitate the prediction of its functional and phenotypic effect.

Expected severity: This annotation assigns each variant an expected severity score (a number that indicates how deleterious the variant is expected to be). This score is based primarily on exonic effect and conservation information, and incorporates additional information, such as chemical and physical properties of amino acids and protein structure. Higher severity variants are more likely to alter the transcription or the translation of the gene in which they reside. These alterations may change crucial qualities of the gene such as expression, affinity, and function and are therefore more likely to result in a phenotypic change. Publicly available tools such as SIFT [42], PolyPhen2 [43], Mutation Taster [44], and LRT [45] utilize the aforementioned characteristics in order to predict the deleteriousness level of a given non-synonymous mutation. These tools are used only for non-synonymous mutations because, while synonymous mutations have been previously associated with functional changes that lead to disease [46], non-synonymous variants, which result in an amino acid changes in the translated protein, are much more likely to induce deleterious functional alterations. Annotating non-synonymous variants with severity scores may facilitate the process of filtering mutations predicted to be tolerated, benign, or silent and the prioritization of the remaining mutations.

Clinical associations: As mentioned before, most variants detected in exome sequencing data have already been observed and recorded in one of the major variant databases. Some of these variants may also have clinical information assigned to them. A combination of such clinical information might be highly relevant and shed light on the phenotype of interest. For a comprehensive list of the variants found to be significantly associated with various phenotypes one should use the National Human Genome Research Institute catalog of published Genome-Wide Association Studies (NHGRI-GWAS catalog) [47]. It is also recommended to annotate variants with any recorded phenotypic effect from the Online Mendelian Inheritance in Man database (OMIM; http://omim.org/).

Publicly available tools such as ANNOVAR [33], snpEff [48], VariantClassifier [49], SNPnexus [50], MutationAssessor [51], and the Ensembl SNP effect predictor [34] automatically annotate variant lists supplied by the user with some of the aforementioned annotations and are highly recommended for such a task.

3.5 Variant Classification

Once all the information has been gathered, and annotations complete, the researcher still faces the challenging task of uncovering the most likely candidate variants out of the immense variant list. In order to make this task feasible, it is recommended to group variants that share similar annotations into various classes. Each given classification represents a subgroup of variants that share similar features and annotations, thus simplifying the data and making it much easier to analyze (*see* **Note 4**). It is important to note that the prioritization process may result in filtering out the actual causative variant. Therefore it is recommended to stratify the variant classification into several priority levels and gradually review them from high priority to low. One should refrain from implementing a simple dichotomous classification which may result in missing the causative variant altogether. Common classifications include the following:

Frequency: Based on population allele frequency (AF) annotations (see previous section) it is possible to classify the different variants into different frequency levels which would enable the researcher to filter out common variants and include only a set of variants with appropriate frequency in the downstream analysis. For example, it is highly unlikely that a very rare disease is caused by a common mutation. Therefore only variants classified as rare will be reviewed when searching for the causative variant in a rare disease. A variant is usually considered very common, if it has an AF higher than 5 %, less common for 1 % < AF < 5 %, rare for AF < 1 %, and private for variants detected only in a specific proband [52]. This classification is fairly simple when only one population frequency database is employed. However, it is important to note that different public databases can record different allele frequencies for the same variant. In such cases, the frequency classification may be set by the researcher according to the experimental setting and the various databases employed. For example, if the exome sequencing was performed on individuals from a European ancestry, it is recommended to first review the variant frequency in European populations as it is more informative than the general population frequency [53].

Deleteriousness: Another classification which is highly informative is deleteriousness. In this classification, variants are classified according to their predicted effect in regard to protein translation, viability, and functionality. For this purpose, annotations regarding a variant's type, exonic effect, conservation, and expected severity are taken into account. For example, an insertion that causes a frameshift close to the start of a gene will be considered more deleterious than a synonymous mutation that occurs in a non-conserved region of the gene [54]. In this classification, there are no standard categories, and the researcher may decide how to categorize the different variants. One form of deleteriousness classification can be divided into four groups: the first, which is the most deleterious and likely to affect gene function, includes frameshift insertion–deletion variants, variants that cause a loss or a gain of a stop codon, variants that are found inside a splice junction, and all non-synonymous variants that are considered severe by more than one of the variant effect prediction methods (see previous section). The second class, considered less deleterious, may include all insertions, deletions, and non-synonymous variants. The third classification includes all the variants found in any functionally annotated region of the genome and the forth includes the entire set of detected variants. This classification method, in which all the variants found in one category are also found in the following, less severe, category, allows the user to efficiently review the most likely variant candidates initially, and then gradually expand the variant search space, if necessary.

Sample information: Last but not least, information about the sample of exomes that were sequenced, such as a genetic pedigree and the suspected mode of inheritance of the causative mutation, should be considered. This information may suggest to the researcher the type of the causative variant (e.g., homozygous or heterozygous). If multiple exomes were sequenced, comparisons of variants between exomes will greatly help the filtering process. For example, if the variant is suspected to be recessive, and the affected individual and his or her unaffected parents are sequenced, variants that are homozygous in the affected individual and heterozygous in each of the parents should be pulled out and classified as a high-priority group. Another example of sample information is segregation, in which a variant is expected to appear either in a higher rate or exclusively (depending on the predicted penetrance) in sequenced affected individuals and should not appear in any of the unaffected individuals. Variants that do not segregate as expected could be excluded from downstream analysis.

Combining annotation-based classification with experimental-setting-classifications results in various categories such as "very rare, highly deleterious, found only in cases" or "common mildly deleterious found in both cases and controls." Using these categories the researcher may choose to review only variants falling into the most relevant combination of classifications, hence significantly reducing the number of variants set for downstream analysis.

3.6 Gene Annotation

Once a class or a combination of classes of variants is chosen, it is usually necessary to expand the characterization to the affected gene level. In some cases, variants found in the highest priority level are reviewed and determined to be irrelevant or inconsequential. This is not uncommon as it is has been shown that there are many such loss-of-function mutations in healthy individuals [54]. One way to predict the phenotype caused by a variant and filter out such high-priority, nonfunctional variants is to annotate and review the gene in which each variant resides. The more information gathered on each affected gene the easier it is to assess its functional significance and relevance to the disease in question (*see* **Note 5**). Common informative annotations include the following:

Pathway information: Pathway annotation includes all the pathways in which a given gene takes part in. When studying a disease with known associated pathways and characterized pathogenesis, it is reasonable to first review variants found in genes that participate in its related pathways. This is also true when a phenotypically similar disease exists, and its pathogenesis and related pathways are known. Pathway information can be gathered from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [55], Biocarta (http://www.biocarta.com), and REACTOM [56] (*see* ref. 57 for additional pathway resources).

Clinical associations and phenotype data: There are occasions in which several different variants on the same gene all result in the same phenotype. Gathering as much clinical and phenotype information known for each of the affected genes may help uncover such cases of genetic heterogeneity. Variants found in a gene in which other variants have already been associated with a certain phenotype are more likely to be associated with the same phenotype. Much like in the case of variant clinical associations, the NHGRI-GWAS catalog and the OMIM database can be utilized to annotate gene clinical associations. Further resources for clinical associations include the genetic association database [58], and Phenopedia and Genopedia [59]. Additional phenotype data can be retrieved from the Human Phenotype Ontology [60] and the Mouse Genome Database (MGD) [61].

Gene–gene and protein–protein interactions: Interactions are highly informative in cases in which there are already known genes which associate with the studied trait. In such cases variants detected inside genes that interact with genes associated with the phenotype in questions can be considered as better, more likely candidates. Recently it was shown that loss-of-function mutations predicted to significantly affect protein translation and function are less common in genes with many interactions [54]. This suggests that variants found in genes with many interactions should also be considered better candidates. Interaction annotations can be gathered from databases such as the BioGRID interaction database [62], STRING protein functional interactions database [63], and the human protein–protein interaction prediction database (PIPs) [64].

Duplicate genes and paralogs: Genes that have multiple duplicates, closely related gene family members (paralogs), and other genes have been shown to carry significantly more loss-of-function variants than other genes [54]. Since loss-of-function variants will usually be classified in a high-priority class (see previous section) it is important to annotate genes with information regarding their paralogs and duplicates and consider this information when deciding on the most likely and functionally relevant mutations (e.g., loss-of-function variants found in highly redundant genes should be considered as less likely candidates). Information regarding variants in gene duplicates and gene paralogs and duplicates can be retrieved from dbDNV [65] and the Duplicated Genes Database (DGD; http://dgd.genouest.org/).

Cancer mutations: When the disease in question is a type of cancer, it may be beneficial to gather, for each gene, any available information regarding the amount and type of somatic and germline mutations previously detected in it. A gene commonly mutated in sequenced cancer samples is more likely to play a role in the development of the disease [66]. A comprehensive database containing such information is the Catalogue of Somatic Mutations in Cancer (COSMIC; [67]).

Once all the genes carrying high-priority variants have been sufficiently annotated, it is much easier to prioritize them and select the most probable candidates. In this stage data regarding the phenotype in question should be integrated and compared with the annotations gathered on each gene in order to possibly find matching functionally relevant characteristics and accordingly validate only the most likely subset of variants found in these genes.

4 Notes

Deep sequencing, also known as high-throughput (or next generation) sequencing, has revolutionized the field of human genetics. The immense amount of data produced by deep sequencing has made it the technology of choice when setting out to interrogate genetic phenomenon, specifically when studying human genetic diseases. Although, when introduced, deep sequencing was still very costly and hence carefully implemented, the unprecedented drop in per-base sequencing prices has turned deep sequencing into both the most efficient and cost-effective option when studying human genetics. Today, the bottleneck has shifted from producing sequence data to the actual analysis of the data. In this chapter we reviewed the common considerations, steps, and resources one should implement when utilizing deep sequencing for the purpose of uncovering the genetic cause of a chosen phenotype. Throughout the chapter we emphasized key points that should be implemented in order to both increase sensitivity and facilitate the filtration and prioritization of the massive amount of data produced by deep sequencing, including the following:

1. Observed or predicted modes of inheritance must be considered when selecting samples for exome sequencing. The selection should include the combination of individuals with the highest likelihood of pinpointing the causative mutation.
2. It is beneficial to examine the included genomic regions and limitations of available enrichment kits in order to better focus on regions relevant to specific studies.
3. Detected variants should be annotated with as much information as possible. This information can be gathered from various public databases, and automatically incorporated using several tools.
4. The more annotations a variant has, the better it can be classified and an educated assessment of its relevance to the phenotype may be employed.
5. Information gathered on the genes affected by candidate variants may pinpoint the most likely candidate and associate it with the observed phenotype.
6. The list of resources mentioned throughout the chapter includes highly established tools and databases but does not include the entire arsenal of possible available computational resources.
7. We urge the researchers to incorporate as many resources as possible in downstream analysis as it is an efficient way to stratify variant data and elucidate relevant disease-causing mutations.

References

1. Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 12:227
2. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2009) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30–35
3. Liu P, Morrison C, Wang L, Xiong D, Vedell P, Cui P, Hua X, Ding F, Lu Y, James M, Ebben JD, Xu H, Adjei AA, Head K, Andrae JW, Tschannen MR, Jacob H, Pan J, Zhang Q, Van Den Bergh F, Xiao H, Lo KC, Patel J, Richmond T, Watt M-A, Albert T, Selzer R, Anderson M, Wang J, Wang Y, Starnes S, Yang P, You M (2012) Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. Carcinogenesis 33(7):1270–1276
4. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141
5. Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 470:187–197
6. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276
7. Biesecker LG, Shianna KV, Mullikin JC (2011) Exome sequencing: the expert view. Genome Biol 12:128
8. Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3:391–397
9. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073
10. Gilissen C, Hoischen A, Brunner HG, Veltman JA (2011) Unlocking Mendelian disease using exome sequencing. Genome Biol 12:228
11. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12:745–755
12. Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012) Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20:490–497
13. Brownstein Z, Bhonker Y, Avraham KB (2012) High-throughput sequencing to decipher the genetic heterogeneity of deafness. Genome Biol 13:245
14. Mertes F, ElSharawy A, Sauer S, van Helvoort JM, Van Der Zaag PJ, Franke A, Nilsson M, Lehrach H, Brookes AJ (2011) Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics 10(6):374–386
15. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111–118
16. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR (2011) A comparative analysis of exome capture. Genome Biol 12: R97
17. Asan, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, Wang J, Wu M, Liu X, Tian G, Wang J, Wang J, Yang H, Zhang X (2011) Comprehensive comparison of three commercial human whole-exome capture platforms. Genome Biology 12(9):R95
18. Sulonen A-M, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, Miettinen T, Tyynismaa H, Salo P, Heckman C, Joensuu H, Raivio T, Suomalainen A, Saarela J (2011) Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol 12:R94
19. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498
20. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311
21. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

22. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DCY, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA (2010) Whole-genome sequencing in a patient with Charcot--Marie–Tooth neuropathy. N Engl J Med 362:1181–1191

23. Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636–639

24. He M-L, Chen Y, Chen Q, He Y, Zhao J, Wang J, Yang H, Kung H-F (2011) Multiple gene dysfunctions lead to high cancer-susceptibility: evidences from a whole-exome sequencing study. Am J Cancer Res 1:562–573

25. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van Lier B, Steehouwer M, van Reeuwijk J, Kant SG, Roepman R, Knoers NVAM, Veltman JA, Brunner HG (2010) Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet 87:418–423

26. Pruitt KD, Tatusova T, Klimke W, Maglott DR (2009) NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res 37:D32–D36

27. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler AD (2002) The human genome browser at UCSC. Genome Res 12:996–1006

28. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Harrow J, Herrero J, Hubbard TJP, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SMJ (2011) Ensembl 2012. Nucleic Acids Res 40:D84–D90

29. Kozomara A, Griffiths-Jones S (2010) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39: D152–D157

30. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A (2009) Rfam: updates to the RNA families database. Nucleic Acids Res 37:D136–D140

31. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets. Cell 120:15–20

32. Betel D, Wilson M, Gabow A, Marks DS, Sander C (2007) The microRNA.org resource: targets and expression. Nucleic Acids Res 36: D149–D153

33. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164

34. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26:2069–2070

35. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F, Fernandez A, Staker B, Pant KP, Baccash J, Borcherding AP, Brownley A, Cedeno R, Chen L, Chernikoff D, Cheung A, Chirita R, Curson B, Ebert JC, Hacker CR, Hartlage R, Hauser B, Huang S, Jiang Y, Karpinchyk V, Koenig M, Kong C, Landers T, Le C, Liu J, McBride CE, Morenzoni M, Morey RE, Mutch K, Perazich H, Perry K, Peters BA, Peterson J, Pethiyagoda CL, Pothuraju K, Richter C, Rosenbaum AM, Roy S, Shafto J, Sharanhovich U, Shannon KW, Sheppy CG, Sun M, Thakuria JV, Tran A, Vu D, Zaranek AW, Wu X, Drmanac S, Oliphant AR, Banyai WC, Martin B, Ballinger DG, Church GM, Reid CA (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327:78–81

36. Ng SB, Nickerson DA, Bamshad MJ, Shendure J (2010) Massively parallel sequencing and rare disease. Hum Mol Genet 19(R2):R119–R124

37. Takata A, Kato M, Nakamura M, Yoshikawa T, Kanba S, Sano A, Kato T (2011) Exome sequencing identifies a novel missense variant in RRM2B associated with autosomal recessive progressive external ophthalmoplegia. Genome Biol 12:R92

38. Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ (2009) Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. Genome Res 19:1562–1569

39. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S (2010) Identifying

a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025

40. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110–121
41. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050
42. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814
43. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249
44. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7:575–576
45. Chun S, Fay JC (2009) Identification of deleterious mutations within three human genomes. Genome Res 19:1553–1561
46. Sauna ZE, Kimchi-Sarfaty C (2011) Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12:683–691
47. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367
48. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Ruden DM, Lu X (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3, Fly 6, 0–1. Fly (Austin) 6(2):80–92
49. Li K, Stockwell T (2010) VariantClassifier: a hierarchical variant classifier for annotated genomes. BMC Res Notes 3:191
50. Dayem Ullah AZ, Lemoine NR, Chelala C (2012) SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). Nucleic Acids Res 40(Web Server issue):W65–W70
51. Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39:e118
52. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11:415–425
53. Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ, Whirl-Carrillo M, Wheeler MT, Dudley JT, Byrnes JK, Cornejo OE, Knowles JW, Woon M, Sangkuhl K, Gong L, Thorn CF, Hebert JM, Capriotti E, David SP, Pavlovic A, West A, Thakuria JV, Ball MP, Zaranek AW, Rehm HL, Church GM, West JS, Bustamante CD, Snyder M, Altman RB, Klein TE, Butte AJ, Ashley EA (2011) Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. PLoS Genet 7: e1002280
54. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner M-M, Hunt T, Barnes IHA, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828
55. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40:D109–D114
56. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37: D619–D622
57. Ooi HS, Schneider G, Lim T-T, Chan Y-L, Eisenhaber B, Eisenhaber F (2010) Biomolecular pathway databases. Methods Mol Biol 609:129–144
58. Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nat Genet 36:431–432
59. Yu W, Clyne M, Khoury MJ, Gwinn M (2010) Phenopedia and genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. Bioinformatics 26:145–146
60. Robinson PN, Mundlos S (2010) The human phenotype ontology. Clin Genet 77:525–534

61. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE (2012) The mouse genome database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res 40:D881–D886

62. Stark C, Breitkreutz B-J, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M (2011) The BioGRID interaction database: 2011 update. Nucleic Acids Res 39:D698–D704

63. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39:D561–D568

64. McDowall MD, Scott MS, Barton GJ (2009) PIPs: human protein-protein interaction prediction database. Nucleic Acids Res 37: D651–D656

65. Ho M-R, Tsai K-W, Chen C, Lin W (2011) dbDNV: a resource of duplicated gene nucleotide variants in human genome. Nucleic Acids Res 39:D920–D925

66. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458:719–724

67. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA (2010) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39:D945–D950