Chapter 14

Discovery of Variants Underlying Host Susceptibility to Virus Infection Using Whole-Exome Sequencing

Gabriel A. Leiva-Torres, Nestor Nebesio, and Silvia M. Vidal

Abstract

The clinical course of any viral infection greatly differs in individuals. This variation results from various viral, host, and environmental factors. The identification of host genetic factors influencing inter-individual variation in susceptibility to several pathogenic viruses has tremendously increased our understanding of the mechanisms and pathways required for immunity. Next-generation sequencing of whole exomes represents a powerful tool in biomedical research. In this chapter, we briefly introduce whole-exome sequencing in the context of genetic approaches to identify host susceptibility genes to viral infections. We then describe general aspects of the workflow for whole-exome sequence analysis together with the tools and online resources that can be used to identify and annotate variant calls, and then prioritize them for their potential association to phenotypes of interest.

Key words Host genetics, Antiviral immunity, Exome, Whole-exome sequencing, Sequence alignment, Read depth, Variant calling, Variant annotation, Gene annotation

1 Introduction

1.1 Value and Genetic Approaches to Identify Host Susceptibility Genes to Virus Infection

A characteristic feature of human infections, including virus infections, is that just a proportion of exposed individuals develop clinical disease. Even during the 1918 influenza pandemic, the more recent human immunodeficiency virus (HIV) epidemic or severe acute respiratory syndrome coronavirus (SARS-CoV) pandemic, only a proportion of individuals succumbed to infection [1, 2]. On the contrary, widespread pathogens that are innocuous for the most of the population, such as herpes simplex virus type 1 (HSV-1), can be fatal only to a very few [3]. It is now well established that host genetic variation is an important component of the varied onset, severity, and outcome of infectious disease. Such data have provided important insights into the pathogenesis of virus infections shedding light into antiviral mechanisms required for host defense.

Karen Mossman (ed.), *Innate Antiviral Immunity: Methods and Protocols*, Methods in Molecular Biology, vol. 1656, DOI 10.1007/978-1-4939-7237-1_14,

Several different, yet complementary approaches to the identification of genetic variation important in infectious disease progression have been taken. By far the most common approach has been to look for association in candidate genes using case–control studies. These studies have highlighted a few common, high-penetrance (*see* Table 1 for the definition of the terms) human genetic variants associated with infection and disease resistance due to virus receptor polymorphism. A homozygous 32 base-pair deletion in the chemokine receptor 5 (CCR5 Δ32) gene provides near complete protection against HIV-1 infection [4], whereas homozygous individuals with nonsense mutations in the fucosyltransferase 2, or FUT2, gene are almost completely protected from experimental and natural infections with norovirus [5]. In a second approach, genome-wide linkage analyses paired with candidate-gene approaches have led to the identification of rare large-effect genetic variants in susceptibility to infection against pathogens segregating in families. An excellent example is the dissection of the genetic architecture of childhood herpes simplex encephalitis (HSE), a rare life-threatening complication of primary infection with HSV-1 [6]. A body of elegant studies have revealed that children with mutations in the TLR3-UNC93B-TRIF-TBK1-TRAF3-IRF3 pathway are particularly susceptible to HSE [7], due to impaired CNS-intrinsic TLR3-dependent IFN-α/β and IFN-λ immunity to HSV-1 [8]. Candidate gene approaches, however, have been limited by their reliance on hypothesis based on—often incomplete—biological knowledge.

The sequencing of the human genome and the international HapMap project [9–11] led the way to Genome Wide Association

Table 1
Definition of terms (in alphabetic order)

Term	Meaning
Haplotype	A set of alleles that commonly segregate together and are defined as regions of extended linkage disequilibrium, which in humans is often up to 100 kb in length.
Indel	Insertions and deletions in a genome; the second most common type of variation after SNPs.
Minor allele frequency (MAF)	Refers to the frequency at which the second most common allele occurs in population.
Penetrance	Describes the proportion of individuals with a mutation or risk variant who have the disease. Incomplete penetrance is said when individuals carrying pathogenic mutations manifest no disease phenotype.
Rare allele	Allele present with MAF <1% (PMID: 19293820)
SNP	Single nucleotide polymorphism. Variation of a single nucleotide base, with the minor allele present in at least 1% of alleles in the population.
SNV	Single nucleotide variant. Minor allele frequency undefined.

Studies (GWAS) [12]. This approach does not require a prior hypothesis. Using large well-characterized cohorts of cases and controls, the whole genome is interrogated with a large set of genetic variants to possible association between a variant and the disease trait. One of the most remarkable successes of GWAS in infection diseases was the identification of *IFNL3* variants in association with the clearance of hepatitis C virus (HCV) following treatment (ribavirin and IFN-α) [13–15] or with spontaneous HCV clearance [16, 17], highlighting the importance of IFN-λ3 signaling in innate control of HCV [18].

GWAS applied to other viral infections have confirmed a major role for HLA genes in host susceptibility against HIV, Dengue and hepatitis B viruses and identified several new risk loci [19–21]. However, except for HCV mentioned before, non-HLA loci often span numerous linked genes and have modest effect size challenging their identification. Interestingly, these loci seem to behave in a pathogen-specific fashion, possibly delineating host-pathogen interactions that are specific to a given virus infection.

1.2 Power and Constraints of Whole-Exome Sequencing

In the past few years, the advent of next-generation sequencing technologies (NGS)—such as whole-exome sequencing (WES)—has revolutionized the biomedical field, including the discovery of many new mutations in patients with unexplained infections often seen at the immunodeficiency clinic [22–24]. WES provides a one-step simultaneous interrogation of virtually all exonic and adjacent intronic sequences, which has been remarkably successful both in a diagnostic setting sequencing and as a discovery tool (research exome sequencing) [25, 26].

These studies have been most effective for the discovery of rare, high-penetrance protein-coding variants for presumed monogenic disorders. A recent report counted that out of about 300 primary immuno-deficiencies characterized at the single gene level, close to 1/3 have been identified by NGS in the past 5 years [27]. WES discoveries have provided fresh insights into the mechanisms that control the development, function, and regulation of immune cells during response to infection (recently reviewed in [26, 28]). Notably, they have highlighted (1) pathways that are required for general protection against infection, generally involving genetic block in the T/B-lymphocyte differentiation program or result in absence of specific immune cells, and (2) pathways that are required for response to narrow groups of pathogens, somewhat reminiscent of infection-specific risk loci mapped by GWAS. An example of the latter was the discovery of compound heterozygous mutations in *IRF7* in a child suffering from life-threatening influenza [29]. Each parent was heterozygous for a single mutated allele, indicating autosomal-recessive segregation for the *IRF7*-deficiency. Detailed biochemical analysis indicated that both alleles were loss-of-function mutations, consistent with the mode of inheritance. Mechanistically, *IRF7*-deficiency was linked to both, lack of IFN-α

production in the patient's plamocytoid cells challenged with influenza virus and lack of intrinsic anti-viral immunity in patient-specific fibroblasts or pulmonary epithelial cells derived from induced pluripotent stem cells (iPSC). This study represented the first demonstration of a genetic cause for severe influenza in humans and may well pave the way for the discovery of other influenza susceptibility genes in the IRF7 pathway, akin to mutations in the TLR3-pathways underlying HSE.

The example above illustrates critical requirements for the successful application of WES, including variant prioritization and variant validation. The study design requires a substantial body of previous knowledge about the phenotype including the prevalence in the general population and the penetrance to help in surmising the mode of inheritance [27, 30]. This will dictate the selection samples (*see* **Note 1**). For situations in which there is a single affected case and no family history, sequencing the unaffected parents (as for IRF7-deficiency) permits efficient discovery of de novo mutations and compound heterozygous genotypes. The availability of multiple families with very similar clinical phenotypes substantially increases power for gene discovery.

However, prioritization of disease-causing variants by WES remains one of the main challenges due to the sheer number of variants found in individual exomes. The exome has been defined traditionally as the sequence encompassing all exons of protein-coding genes in the genome and covers between 1 and 2% of the genome [31–33]. Yet this portion houses 85% of the known disease causing variants [34, 35]. An individual exome typically harbors thousands of variants, compared to a reference genome, which are predicted to lead to nonsynonymous amino acid substitutions, alterations of conserved splice site residues, or small insertions or deletions. As presented below, various methods exist to identify which variants deleteriously affect the function of individual proteins. However, each genome is thought to harbor about 100 genuine loss-of-function variants with about 20 genes completely inactivated [36, 37]. Hence, rigorous criteria, including the absence of the candidate variant genotype in individuals without the clinical phenotype together with robust experimental validation, have been proposed to validate disease-causing variants [38]. Whereas study design and experimental approaches need to be developed in a case-by-case situation, below we will present the reagents and methodology for the discovery of and validation of candidate genetic variants in a typical exome-sequencing pipeline.

2 Materials

In addition to DNA samples from cases, their families, and the appropriate controls, the materials required for WES are a well-annotated reference genome, whole-exome capture DNA libraries, and computing facilities.

2.1 Annotated Reference Genome

The human reference assembly defines a standard upon which other whole genome studies are based. The last build of the human reference genome provided by the Genome Reference Consortium reports ~3×10^9 bases having coding and noncoding sequences. The exome is defined as all the exons for the 20,000 protein-coding genes in the human genome and all the exons pertaining to microRNA, small nucleolar RNA, and large intergenic noncoding RNA genes [39]. This information is not static and projects such as GENCODE [40] and RefSeq [41] continue to provide comprehensive annotation of both protein-coding genes and noncoding transcripts. The last assembly of human reference genome (GRCh38) can be accessed via the European Bioinformatics Institute and the Wellcome Trust Sanger Institute (Ensembl) [42] or the University of California Santa Cruz (UCSC) [43] genome browsers.

2.2 Whole-Exome Capture Library

Exome capture essentially consists of the steps of fragmenting a DNA sample, hybridizing the DNA to complementary oligonucleotide baits whose sequence has been designed to hybridize to exon regions. After binding to genomic DNA, these probes are pulled down and PCR amplified through the addition of adapters, allowing exon regions to be selectively sequenced. The most common and efficient strategies are in-solution capture methods offered by Roche/NimbleGen's SeqCap EZ Human Exome Library and Agilent's SureSelect Human All Exon. Several publications have compared the specificity and sensitivity of these platforms [44–46]. The NimbleGen's kit has the greatest bait density of any of the platforms and uses short (55 − 105 bp), overlapping baits to cover the target region [46]. This approach has been found to be an efficient method to cover the target region, sensitively detect variants and has a high level of specificity. Indeed, NimbleGen's kit shows fewer off-target reads than other platforms [46]. Importantly, this bait design has been found to show greater genotype sensitivity than the other platforms in difficult to sequence regions, such as areas of high GC content [44]. The Agilent's kit is the only platform to use RNA probes. The baits are longer than those used in NimbleGen's platform (114 − 126 bp) and the corresponding target sequences are adjacent to one another rather than overlapping. This design has been found to be good at identifying insertions and deletions (indels), because longer baits can tolerate larger mismatches [45].

2.3 High-Performance Computing Facility/Network for Data Storage and Maintenance of Pipelines for WES Analysis

Massively parallel short-read sequencing on NGS platforms typically results in the production of ~50–100 million reads per exome. This large volume of reads needs to be analyzed and stored. Moreover, software packages work best when tools and sequencing data are immediately available in the same network as accessing an external storage location for sequencing data slows down the process. High-performance computing infrastructure (HPC) and IT professionals are needed to access and storage of the generated and analyzed data. The most common infrastructure components

include HPC resources ranging from high-performance computing clusters to cloud computing resources, equipped with batch (queuing) systems, and commonly connected to shared-network-attached storage. Academic researchers have access to these services through national infrastructures, which provide HPC, storage, and ultra-high-speed network connectivity and remote access to research data. These systems are equipped with actively maintained bioinformatics suites for automation of WES analysis. The most widely used variant callers include the Sequence Alignment/Map (SAM) tools [47] and the Genome Analysis Tool Kit (GATK) [48, 49] developed by the Broad Institute. The latter was found the most efficient NGS variant caller in comparative studies [50] (*see* Table 2 for commonly used tools for WES and their weblinks).

3 Methods

A typical pipeline of WES analysis consists of the main following steps: (1) raw data QC and preprocessing, (2) sequence alignment mapping, (3) post-alignment processing, (4) variant analysis, (5) variant prioritization, and (6) variant validation (Fig. 1).

3.1 Raw Data Quality Control (QC) and Preprocessing

An effective QC is critical for a reliable data analysis, since this may affect downstream analysis results. The raw sequence output format for NGS is the FASTQ format (*see* Table 3 for commonly used file formats in WES), which incorporates (1) a text-based representation of sequences (FASTA format) and (2) a per-base quality score of the read provided by the sequencing instrument. The latter is a Phred-like score [51] assigned by an algorithm of the sequencing instrument that estimates the probability that a base is called incorrectly. Several tools have been developed for QC of raw sequence data. The most commonly used is the java script FastQC [52]; it can generate diagnosis plots such as distributions of base quality scores and GC content, N content, and sequence duplication levels. FastQC can also perform standard preprocessing procedure including adapter removal and trimming of low-quality bases at the ends of the reads.

3.2 Sequence Alignment Mapping

After raw data QC and preprocessing, the next step is to map the reads to the reference genome. This is arguably the most crucial step and most time-consuming operation of most WES analysis pipelines. The computational challenge resides in finding an alignment algorithm that is tolerant to imperfect matches, where genomic variations may occur, while being able to align millions of reads at a reasonable speed. To achieve high-speed most alignment algorithms are based on an effective compression algorithm, the Burrows–Wheeler Transformation (BWT) [53]. Many short-read

Table 2
Commonly used tools and weblinks for whole-exome sequence data analysis pipeline

Tool	weblink
Genome browser	
Ensembl	www.ensembl.org
UCSC	http://genome.ucsc.edu
Quality control	
FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Short read mapping	
Bowtie	http://bowtie-bio.sourceforge.net/index.shtml
Bfast	http://bfast.sourceforge.net
Mosaik	https://github.com/wanpinglee/MOSAIK
BWA	http://bio-bwa.sourceforge.net/
Manipulate NGS data (mark duplicates, merge files)	
Picard tools	https://broadinstitute.github.io/picard/index.html
SAMTools	http://www.htslib.org/doc/samtools.html
Variant calling	
GATK	https://software.broadinstitute.org/gatk/
SAMTools	http://www.htslib.org/doc/samtools.html
Variant annotation: (1) Coding effect predictions	
SnpEff	http://snpeff.sourceforge.net/
VEP	http://ensembl.org/info/docs/tools/vep/index.html
SIFT	http://sift.jcvi.org/
PolyPhen2	http://genetics.bwh.harvard.edu/pph2/
Variant annotation: (2) Conservation	
PhyloP	http://compgen.bscb.cornell.edu/phast
GERP++	http://gvs.gs.washington.edu/GVS147/
CADD	http://cadd.gs.washington.edu/
Variant annotation: (3) Gene-level	
MSC	http://lab.rockefeller.edu/casanova/MSC
GAVIN	https://molgenis20.gcc.rug.nl/
Variant annotation: (4) integrative	
ANNOVAR	http://annovar.openbioinformatics.org/en/latest/user-guide/download/
Knowledge-based annotation	
HGPS	http://hgc.rockefeller.edu/
KEGG	www.genome.jp/kegg/
REACTOME	www.reactome.org/
MPO	www.informatics.jax.org/humanDisease.shtml
GEO	www.ncbi.nlm.nih. gov/geoprofiles
GXA	www.ebi.ac.uk/gxa
BioGPS	http://biogps.org
STRING	http://string-db.org
ToppGene	https://toppgene.cchmc.org
GeneMania	http://genemania.org

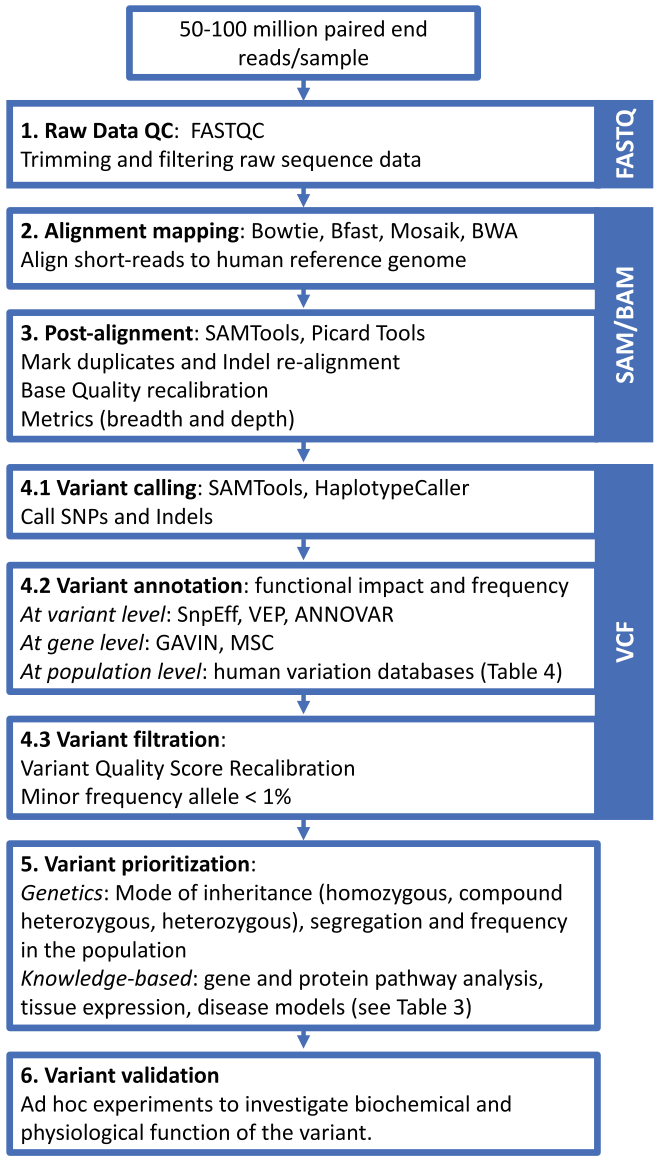

Fig. 1 Basic workflow and tools for whole-exome sequencing project. Following sequencing, reads undergo quality assessment and read alignment against a reference genome, followed by variant identification. The detected variants are annotated to infer their biological relevance. Then, variants are filtered based on quality of the read and frequency on the population. Then variants are prioritized based on the genetic hypothesis for the trait under study and knowledge about the candidate gene/protein. Ultimately, experimental validation is required to ascertain variant discovery. On the right the format outputs are indicated

aligners have been developed using this method: Bowtie [54], Bfast [55], Mosaik [56], and BWA [57]. They vary a lot in speed and accuracy, which are likely to affect the identification of structural variations and influence variant calling. BWA is the most common

Table 3
Description of commonly used file formats in WES workflows

Format	Characteristics
FASTQ file (.fastq)	Text file that stores nucleotide sequence and quality score for downstream analysis. There are typically four lines in a FASTQ file: (1) sequence identifier initialized "@"; (2) biological sequence of nucleotide reads (ACTG); (3) sequence identifier initialized "+"; (4) quality score of corresponding sequencing read, which is coded with ASCII characters.
Sequence alignment/map (SAM) file (.sam)	Text file that stores alignment information of short reads to reference genome. The SAM file contains multiple lines including a header initialized "@" and multiple lines for the sequence alignment.
Binary alignment/map (BAM) file (.bam)	Binary file (stored in a format that is only computer readable) containing the same information as the SAM file, the content of which has been compressed to reduce storage disk space and increase performance.
Browser extensible data (BED) file (.bed)	Tab-delimited text file that consists of several lines each representing a single genomic region, such as an exon. BED files provide the coordinates of those regions including chromosome, start and end positions, and additional fields can be added.
Variant call format (VCF) file (.vcf)	Text file containing meta-information lines (i.e., file format, date, or other information about the overall experiment), a header line naming the columns (chromosome #, position, ID, reference allele, alternative allele, quality, filte, infor), and then data lines each containing information about a position in the genome. It is a standardized text file format for representing SNP, indel, and structural variation calls.

choice of WES alignment [58]. It allows gapped alignment, using very little memory. It performs separated alignment on both strands of a paired-end lane, in multi-threaded execution, unifying results in a single mapping file in the Sequence Alignment Map (SAM) format [47].

3.3 Post-Alignment Processing

To enhance the quality of the alignments for more accurate variant detection, the pipeline carries out three "cleanup" procedures. They consist of read duplicate removal, base quality score recalibration (BQSR), and indel realignment. A final, intermediate step provides important metrics to assess the quality of the data.

3.3.1 Read Duplicate Removal

Many of the reads from massively parallel sequencing instruments are identical—same sequence, start site, and orientation—indicating PCR artefacts [59]. These duplicates may introduce a bias in estimating variant allele frequencies, thus it is advisable that they are removed prior to the variant calling. Programs such as the function rmdup from SAMTools [47] or PicardMarkDuplicates integrated in Picard Tools [49] apply optimal fragment-based duplicate identification and provide unique identifiers for each read group, i.e.,

the set of reads generated from a single run of an instrument. This allows minimizing of experimental noise, reducing the number of false calls and improving the accuracy in the search of the variants.

3.3.2 Indel Re-Alignment

Small insertions or deletions (Indels) in coding regions have been strongly associated with human diseases but accurate Indel calling remains difficult [60, 61]. The local realignment around Indels is an important step. This process searches a consensus alignment among all the reads spanning a deletion or an insertion or both (1) to improve Indel detection sensitivity and accuracy, and (2) to reduce variant false calls due to misalignment of the flanking bases. The alignment is improved by increasing the number of sequences in their local context. The program Haplotype Caller from GATK offers an efficient solution to Indel detection by generating local de novo assembly of aligned reads prior to Indel calling, improving Indel detection [62]. As presented in Subheading 4, the HaplotypeCaller is capable of calling variants and indels simultaneously, which improves Indel detection while producing more accurate variant calls.

3.3.3 BQSR

The per-base quality scores (Phred-score), which convey the probability that the called base in the read is the true sequenced base [51], are quite inaccurate and co-vary with features like sequencing technology, machine cycle, and sequence context. These inaccurate quality scores propagate into faulty SNP discovery [51]. BQSR is a process in which machine learning tools are applied to model these errors empirically and adjust the quality scores accordingly. One of the most commonly used BQSR programs is BaseRecalibrator from the GATK suite, which takes alignment files and for each unknown base, a re-calibrated quality score is calculated to be used for variant calling. Recalibrated scores better reflect the empirical probability of mismatches to the reference genome, and by doing so provide more accurate quality scores [48, 62].

3.3.4 Metrics

Biases in sample preparation, sequencing, genomic alignment, and assembly can result in genomic regions lacking coverage (i.e., gaps) or in regions with much higher coverage than theoretically expected. Hence to evaluate the quality of data to discover variants with reasonable confidence, two important metrics are the breadth and the depth of coverage of a target genome. Breadth of coverage denotes the percentage of bases that are sequenced a given number of times. Depth of coverage represents the number of reads that align at a given position, which is often quoted as average raw or aligned read depth. For example, a genome sequencing study may sequence a genome to 50× average depth and achieve a 95% breadth of coverage of the reference genome at a minimum depth of ten reads. The flagstat command from SAMtools [47] or DepthOfCoverage from GATK [48, 62] provides the calculation

of the fraction of reads that successfully mapped to the reference, with number and percentages of the read mapped and unmapped.

3.4 Variant Analysis

Following these treatment steps of the read, variant analysis consists of three independent steps: variant calling, annotation, and prioritization. Several open source tools are available for variant calling (Table 4).

3.4.1 Variant Calling

Variant calling implies identifying the sites in the sample that statistically differ from the reference genomic sequence. Single nucleotide polymorphisms (SNPs) and Indels are detected where the reads collectively provide evidence of variation (*see* **Note 2**). As with alignment tools, several open source tools are available to identify a high-quality set of variants in WES projects [63]. SAMtools [47] and GATK HaplotypeCaller [48, 62] are widely used in genomic variant analyses. HaplotypeCaller has been found to have high sensitivity for SNP detection and outperform other pipelines for

Table 4
Databases of human genetic variation

Name	Weblink and description
Combined annotation dependent depletion database (CADD)	http://cadd.gs.washington.edu/ Catalog of precomputed scores for all possible SNPs or small Indels of the reference genome and the 1000 Genomes obtained by combining 63 annotations (e.g., SIFT, GERP, others) through a machine-learning framework.
Single nucleotide polymorphism database (dbSNP)	https://www.ncbi.nlm.nih.gov/projects/SNP/ Broad collection of SNPs and Indels submitted by investigators worldwide and curated by NCBI.
Human gene mutation database (HGMD)	http://www.hgmd.org A catalog of all published gene lesions responsible for human inherited disease.
Exome aggregation consortium (ExAC)	http://exac.broadinstitute.org/ Catalogue of exome variation in 60706 individuals some with adult onset diseases (Type 2 Diabetes, schizophrenia) patients presenting severe pediatric diseases have been excluded.
1000 Genomes project	http://www.internationalgenome.org/ Catalogue of genome variation with at least 1% frequency in the population based on whole-genome sequencing of 2504 individuals from 26 populations (including study cohorts for adult onset diseases).
NHLBI exome sequencing project (ESP6500)	http://evs.gs.washington.edu/EVS/ Catalogue of variation within 6500 exomes from well-phenotyped populations from various projects, e.g. Severe Asthma Research Project; Pulmonary Arterial Hypertension population; Acute Lung Injury cohort; Cystic Fibrosis cohort.

Indels [50, 63]. HaplotypeCaller runs a "reading window" along the reference genome, comparing the reference to sequenced reads counting mismatches and Indels. These variations from the reference are used as a measure of entropy, or disorder in the read data. If the level of entropy within the reading window surpasses a cutoff score (default value can be changed), the window is marked as an Active Region, which is inspected to generate the plausible haplotypes. Then, HaplotypeCaller uses a Bayesian statistical model for the calculation of the probability of the genotype, estimating the accuracy of the call with a score of Phred-like quality. The results are reported in a standard Variant Call Format (VCF) file.

3.4.2 Variant Annotation

Annotation of disease-causing variants involves determining (1) the effect they have on the protein-coding sequence, including synonymous and non-synonymous changes, stop-gained or stop-lost, consensus splice site changes for SNPs, frame-shift or other structural impacts on transcript structure for Indels, (2) the frequency of the variant in the population, as disease-causing variants are expected to be rare.

1. Three major tools are used to classify variants functionally: SnpEff (**SNP Eff**ects) [64], VEP (Variant Effect Predictor) [65], and ANNOVAR (**Anno**tate **Var**iation) [66, 67]. SnpEff annotates variants based on their genomic location and predicts coding effects [64], as does VEP, a tool available from the genome browser, Ensembl [65]. Besides annotating functional effects of variants with respect to genes, ANNOVAR has many additional functionalities, such as integrating information from up to 4000 different databases and external resources to annotate the variants [67]. For SNPs, these include (1) calculating their predicted functional importance scores using SIFT (**S**orting **I**ntolerant **F**rom **T**olerant) [68] and PolyPhen2 (**Poly**morphisms **Phen**otyping v**2**) [69] and (2) reporting their conservation levels by PhyloP (**Phylo**genetic ***P***-values) [70, 71] and GERP++ (**G**enomic **E**volutionary **R**ate **P**rofiling) [72]. The CADD (**C**ombined **A**nnotation **D**ependent **D**epletion) database is another useful external linked for deleterious prediction of a variant. The CADD score combines information from several resources to score both protein-altering and regulatory variants [73].

   New tools are being developed for variant annotation that considers gene-level metrics (e.g., conservation at the gene-level, accumulation of mutational load) and provides more sensitive scoring of variants [74]. GAVIN (**G**ene-**A**ware **V**ariant **IN**terpretation for medical sequencing) classifies variants as benign, pathogenic, or a variant of uncertain significance [75]. The MSC (**M**utation **S**ignificance **C**utoff) [76] generates a quantitative score that provides gene-level and gene-specific phenotypic impact cutoff values above which a variant is considered pathogenic with 98% true positive detection rate.

2. To determine variant frequency, ANNOVAR links to external databases such as dbSNP database [77, 78] or the Human Gene Mutation Database [79] to identify the presence or absence of a variant (*see* Table 4 for commonly used databases of human genetic variation). Large-scale genomic studies such as 1000 Genomes Project [36], the US National Institutes of Health–National Heart, Lung, and Blood Institute (NIH-NHLBI), ESP6500 exome-sequencing project [80], and the Exome Aggregation Consortium [37, 81] have catalogued sequence variants from thousands of exomes and genomes, which serve as a valuable resource for allele frequency estimations. These resources are integrated in ANNOVAR, which can find the alternative allele frequency for newly discovered variants in a WES project. The GATK pipeline also integrates ANNOVAR as an external option for variant annotation and can use the tool VariantAnnotator, which is enriched with additional features such as gene set enrichment analysis for downstream analysis.

3.4.3 Variant Filtration

There are two aspects to variant filtration (1) filtering low-quality variants; (2) filtering common variants, which are represent in the general population.

1. Low-quality variants are those including variants with low coverage, low quality, strand biased, as well as those mapping to low-complexity regions or incomplete regions of the reference genome [82]. GATK uses machine learning algorithms (VQSR or variant quality score recalibration) to learn from each dataset what is the annotation profile of "good" and "bad" variants [48, 62]. The tool assigns scores (VQSLOD for variant quality score log-odds) which can be used to set the filtering of "bad" variants. There is tradeoff in the process in which increasing the specificity will decrease the sensitivity of the filtering. VQSR can be applied to SNPs or indels. The availability of in-house databases for WES variants obtained with the same sequencing technology and analysis pipeline is recommended to exclude variants resulting from systematic errors (*see* **Note 3**).
2. Under the assumption that common variants are less likely to cause disease than rare ones, it is important to set a minor allele frequency (MAF) threshold based on disease model of the study. A variant with a MAF greater than 1% is regarded as common; the remainder are considered rare or private to the subject or the kindred studied. Setting the MAF threshold at 1% is recommended, usually filters out over 70% of the variants [83].

3.4.4 Variant Prioritization

At this point the output is a subset of high-quality, low-frequency, predicted pathogenic variants, which require customized filtering process depending on the disease trait. The more information

gathered both on (1) the phenotype and (2) the gene in which the variant resides, the greater the likelihood to accurately assess the functional significance of a variant.

1. A deep knowledge of the clinical and cellular phenotype, the prevalence of the trait in the general population together with an understanding of the familial segregation are essential in the prioritization of gene variants. For example, a recessively inherited disease variant is likely homozygous whereas a dominant disease variant is heterozygous. In general, a dominant allele should be absent in a variant database based on healthy controls or exceedingly rare to allow for reduced penetrance. However, there can be exceptions to these rules. For instance, recessive disease variants can be compound heterozygous. In a cohort, the search for either identical variants or additional rare variants in the same gene can further strengthen the evidence for causality. 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, although this is not always the case.

   Segregation of the variant with disease status is another key criterion for variant prioritization. This requires appropriate WES control data obtained with the same method from healthy subjects, ideally of the same ethnic origin as the patients. In case of complete penetrance, the candidate disease-causing variants found in patients cannot be present in unaffected subjects. In case of incomplete penetrance, the situation is more complex because these hypothetical disease-causing variants can also be present in asymptomatic subjects, including unaffected subjects of the same pedigree.

2. At the gene level, 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 related pathways are known. The HGPS (**H**uman **G**ene **C**onnectome) ranks genes by their biological distance to core genes (known to be associated with phenotype), and provides the distances and all possible biological connections between all pairs of human genes based on protein-protein interaction prediction [74, 84]. Genes can be mapped online to KEGG (**K**yto **E**ncyclopedia of **G**enes and **G**enomes) pathways [85] or REACTOME pathways [86]. It is useful to find information about candidate genes-knockout phenotypes. For this, the Mouse Genome Informatics database enables queries for human–-mouse disease and MPO (**M**ammalian **P**henotype **O**ntology) connections using gene symbols as an input [87]. Expression of candidate gene in the tissues or organs of interest is an important criterion for prioritization. GEO (**G**ene **E**xpression **O**mnibus) profiles [88], the ExA (**Ex**pression **A**tlas [89], and the BioGPS gene annotation portals [90] are excellent resources

for this purpose. Knowledge about protein structure, function, and interactions also can help rank candidate genes. The UniprotKB (**Uniprot k**nowledge**b**ase) collects information from several databases including curated protein sequences and structures with links to annotations of genomic variants [91]. The STRING database and associated search tools [92] are powerful resources for identifying interacting partners of a candidate gene's product or for identifying interactions between the products of a set of genes that bear functional variants. The ToppGene [93] and GeneMania [94] web portals are other resources that perform candidate gene prioritization based on the interactome.

3.4.5 Variant Validation

With all the tools available and new ones emerging monthly, variant filtration and prioritization are becoming more automated. A similar trend is also observed in other parts of variant analysis such as the detection and annotation. Regardless, a deep understanding of the biological questions being asked and the etiology of the disease being studied is crucial for properly choosing tools and parameters that suit a study the best.

Ultimately, variant validation requires experimental confirmation at the level of protein, cell and—if possible—animal model to establish causality. This necessitates solid knowledge of physiology and pathology of the phenotype at the study for the design of appropriate experiments relevant to the nature of the protein. The recent breakthrough of genetic manipulation of human-induced pluripotent stem cells [95], CRISPR genome-editing tools [96, 97] permits establishing the causal relationship between the candidate genotype and the clinical phenotype in relevant cell types [98] or organoids [99] representing relevant tissues, even for isolated cases.

4 Notes

1. Broadly, the mode of inheritance can be recessive, dominant, or X-linked. Recessive mutations are easier to identify by filtering for homozygosity, or compound heterozygous mutations. Dominant inherited mutations will be either inherited from one of the parents or be de novo mutations, in both cases dominant mutations should be absent in unaffected family members or matched unrelated controls.
2. Joint application of variant calling software to multiple samples is recommended to reduce false positive variants. We can also improve variant calling in regions with fewer reads by utilizing reads from multiple samples concurrently. This increases the confidence of any given variant and allele bias and strand bias are much easier to sort.

3. The evaluation of family trios can also eliminate low-quality variants as the majority of variants detected in the child and absent from the parents most likely result from sequence artifacts. Moreover, the accuracy of error detection and variant identification increases with the number of relatives and generations sequenced per family.

References

1. Casanova JL (2015) Human genetic basis of interindividual variability in the course of infection. Proc Natl Acad Sci U S A 112(51): E7118–E7127
2. Herrington CS, Coates PJ, Duprex WP (2015) Viruses and disease: emerging concepts for prevention, diagnosis and treatment. J Pathol 235 (2):149–152
3. Zhang SY, Abel L, Casanova JL (2013) Mendelian predisposition to herpes simplex encephalitis. Handb Clin Neurol 112:1091–1097
4. Dean M et al (1996) Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia growth and development study, multicenter AIDS cohort study, multicenter hemophilia cohort study, San Francisco City cohort, ALIVE study. Science 273 (5283):1856–1862
5. Lindesmith L et al (2003) Human susceptibility and resistance to Norwalk virus infection. Nat Med 9(5):548–553
6. Whitley RJ (2006) Herpes simplex encephalitis: adolescents and adults. Antiviral Res 71(2-3):141–148
7. Rozenberg F (2013) Acute viral encephalitis. Handb Clin Neurol 112:1171–1181
8. Lafaille FG et al (2012) Impaired intrinsic immunity to HSV-1 in human iPSC-derived TLR3-deficient CNS cells. Nature 491 (7426):769–773
9. International HapMap, C (2005) A haplotype map of the human genome. Nature 437 (7063):1299–1320
10. Lander ES et al (2001) Initial sequencing and analysis of the human genome. Nature 409 (6822):860–921
11. Venter JC et al (2001) The sequence of the human genome. Science 291(5507): 1304–1351
12. Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. N Engl J Med 363(2):166–176
13. Ge D et al (2009) Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 461(7262):399–401
14. Suppiah V et al (2009) IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat Genet 41 (10):1100–1104
15. Tanaka Y et al (2009) Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. Nat Genet 41 (10):1105–1109
16. Rauch A et al (2010) Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. Gastroenterology 138 (4):1338–1345. 1345 e1-7
17. Thomas DL et al (2009) Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. Nature 461(7265):798–801
18. Sheahan T et al (2014) Interferon lambda alleles predict innate antiviral immune responses and hepatitis C virus permissiveness. Cell Host Microbe 15(2):190–202
19. Abel L, Alcais A, Schurr E (2014) The dissection of complex susceptibility to infectious disease: bacterial, viral and parasitic infections. Curr Opin Immunol 30:72–78
20. Loeb M (2013) Genetic susceptibility to West Nile virus and dengue. Public Health Genomics 16(1-2):4–8
21. McLaren PJ, Carrington M (2015) The impact of host genetic variation on infection with HIV-1. Nat Immunol 16(6):577–583
22. Conley ME, Casanova JL (2014) Discovery of single-gene inborn errors of immunity by next generation sequencing. Curr Opin Immunol 30:17–23
23. Fodil N, Langlais D, Gros P (2016) Primary Immunodeficiencies and inflammatory disease: a growing genetic intersection. Trends Immunol 37(2):126–140
24. Stoddard JL et al (2014) Targeted NGS: a cost-effective approach to molecular diagnosis of PIDs. Front Immunol 5:531
25. Boycott KM et al (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14 (10):681–691

26. Casanova JL (2015) Severe infectious diseases of childhood as monogenic inborn errors of immunity. Proc Natl Acad Sci U S A 112(51): E7128–E7137
27. Meyts I et al (2016) Exome and genome sequencing for inborn errors of immunity. J Allergy Clin Immunol 138(4):957–969
28. Chou J, Ohsumi TK, Geha RS (2012) Use of whole exome and genome sequencing in the identification of genetic causes of primary immunodeficiencies. Curr Opin Allergy Clin Immunol 12(6):623–628
29. Ciancanelli MJ et al (2015) Infectious disease. Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency. Science 348(6233):448–453
30. Wu L et al (2015) Case-only exome sequencing and complex disease susceptibility gene discovery: study design considerations. J Med Genet 52(1):10–16
31. Ezkurdia I et al (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 23(22):5866–5878
32. Ezkurdia I et al (2014) Analyzing the first drafts of the human proteome. J Proteome Res 13(8):3854–3855
33. Sakharkar MK, Chow VT, Kangueane P (2004) Distributions of exons and introns in the human genome. In Silico Biol 4(4):387–393
34. Majewski J et al (2011) What can exome sequencing do for you? J Med Genet 48 (9):580–589
35. Ng SB et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461(7261):272–276
36. Genomes Project, C et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65
37. Lek M et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536(7616):285–291
38. Casanova JL et al (2014) Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies. J Exp Med 211 (11):2137–2149
39. Consortium GR (2017) https://www.ncbi.nlm.nih.gov/grc/human
40. Pruitt KD et al (2014) RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42(Database issue):D756–D763
41. Harrow J et al (2012) GENCODE: the reference human genome annotation for the ENCODE project. Genome Res 22 (9):1760–1774
42. Aken BL et al (2016) The Ensembl gene annotation system. Database (Oxford) 2016. doi:10.1093/database/baw093
43. Hung JH, Weng Z (2016) Visualizing genomic annotations with the UCSC genome browser. Cold Spring Harb Protoc 2016(11). doi:10.1101/pdb.prot093062. p. pdb prot093062
44. Bodi K et al (2013) Comparison of commercially available target enrichment methods for next-generation sequencing. J Biomol Tech 24 (2):73–86
45. Chilamakuri CS et al (2014) Performance comparison of four exome capture systems for deep sequencing. BMC Genomics 15:449
46. Clark MJ et al (2011) Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29(10):908–914
47. Li H et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25 (16):2078–2079
48. DePristo MA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498
49. McKenna A et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303
50. Liu X et al (2013) Variant callers for next-generation sequencing data: a comparison study. PLoS One 8(9):e75619
51. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8(3):186–194
52. Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
53. Burrows M, Wheeler DJ, (1994) A block-sorting lossless data compression algorithm. Technical report–CaliforniaDigital Equipment Corporation, Palo Alto, 124
54. Langmead B et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25
55. Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS One 4(11):e7767
56. Lee WP et al (2014) MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One 9 (3):e90581
57. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760

58. Shang J et al (2014) Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed Res Int 2014:309650
59. Koboldt DC et al (2010) Challenges of sequencing human genomes. Brief Bioinform 11(5):484–498
60. Mills RE et al (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16 (9):1182–1190
61. Mullaney JM et al (2010) Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet 19(R2):R131–R136
62. Van der Auwera GA et al (2013) From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11 10 1–11 1033
63. Huang HW et al (2015) Evaluation of variant detection software for pooled next-generation sequence data. BMC Bioinformatics 16:235
64. Cingolani P et al (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 (Austin) 6(2):80–92
65. Flicek P et al (2010) Ensembl's 10th year. Nucleic Acids Res 38(Database issue): D557–D562
66. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164
67. Yang H, Wang K (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc 10 (10):1556–1566
68. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4(7):1073–1081
69. Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 76:7.20.1–7.20.41
70. Siepel A, Pollard KS, Haussler D (2006) New methods for detecting lineage-specific selection. In: Apostolico A, Guerra C, Istrail S, Pevzner P, Waterman M (eds) Proceedings of the 10th international conference on research in computational molecular biology. Springer, Germany, pp. 190–205
71. Pollard KS et al (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20(1):110–121
72. Cooper GM et al (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat Methods 7 (4):250–251
73. Kircher M et al (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310–315
74. Itan Y et al (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. Proc Natl Acad Sci U S A 112 (44):13615–13620
75. van der Velde KJ et al (2017) GAVIN: Gene-aware variant INterpretation for medical sequencing. Genome Biol 18(1):6
76. Itan Y et al (2016) The mutation significance cutoff: gene-level thresholds for variant predictions. Nat Methods 13(2):109–110
77. Day IN (2010) dbSNP in the detail and copy number complexities. Hum Mutat 31(1):2–4
78. Sherry ST et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29 (1):308–311
79. Stenson PD et al (2014) The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133(1):1–9
80. Fu W et al (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493 (7431):216–220
81. Karczewski KJ et al (2017) The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 45 (D1):D840–D845
82. Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30(20):2843–2851
83. Bao R et al (2014) Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer Inform 13(Suppl 2):67–82
84. Itan Y et al (2014) HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. BMC Genomics 15:256
85. Kanehisa M et al (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45(D1):D353–D361
86. Croft D et al (2014) The reactome pathway knowledgebase. Nucleic Acids Res 42(Database issue):D472–D477
87. Bello SM, Smith CL, Eppig JT (2015) Allele, phenotype and disease data at mouse genome informatics: improving access and analysis. Mamm Genome 26(7-8):285–294

88. Sayers EW et al (2010) Database resources of the national center for biotechnology information. Nucleic Acids Res 38(Database issue): D5–16
89. Petryszak R et al (2016) Expression atlas update–an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res 44(D1):D746–D752
90. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26(7):873–881
91. The UniProt, C (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45 (D1):D158–D169
92. Szklarczyk D et al (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43(Database issue):D447–D452
93. Chen J et al (2009) ToppGene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 37(Web Server issue):W305–W311
94. Zuberi K et al (2013) GeneMANIA prediction server 2013 update. Nucleic Acids Res 41(Web Server issue):W115–W122
95. Takahashi K et al (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131(5):861–872
96. Cong L et al (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339(6121):819–823
97. Makarova KS et al (2011) Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 9(6):467–477
98. Hamazaki T et al (2017) Concise review: induced pluripotent stem cell research in the era of precision medicine. Stem Cells 35(3):545–550
99. Nie J, Hashino E (2017) Organoid technologies meet genome engineering. EMBO Rep 18 (3):367–376