Published in final edited form as:

Curr Protoc Hum Genet. ; 88: 6.15.1–6.15.17. doi:10.1002/0471142905.hg0615s88.

***In Silico* Functional Annotation of Genomic Variation**

Mariusz Butkiewicz¹ and William S. Bush¹

¹Institute for Computational Biology, Case Western Reserve University, Cleveland, Ohio

Abstract

This unit describes the concepts and practical techniques for annotating genomic variants in the human genome to estimate their functional significance. With the rapid increase of available whole exome and whole genome sequencing information for human studies, annotation techniques have become progressively more important for highlighting and prioritizing nucleotide variants and their potential impact on genes and other genetic constructs. Here, we present an overview of different types of variant annotation approaches and elaborate on their foundations, assumptions, and the downstream consequences of their use. Computational approaches and tools to assign annotations and to identify variants are reviewed. Further, the general philosophy of assigning potential function to a genetic change within the biological contexts of a disease is discussed.

Keywords

variant annotation

INTRODUCTION

The human genome consists of over 3 billion DNA base pairs and constitutes the genetic blueprint of every human individual (Collins et al., 1998). Individuals vary in their genomic sequence most commonly from single nucleotide variants (SNVs). These changes may alter the expression or products of genes, and thus can influence human phenotypes. Some of these changes have severe consequences, disrupting protein function and leading to either Mendelian or complex forms of disease. Identifying these variants is a major focus of human genetics.

The goal of variant annotation is to identify and prioritize variants, both newly discovered and existing, based on expected consequence severity (Cutting, 2014). In early genomic studies, “annotation” of influential variants was often accomplished via literature search of the most statistically robust findings. While this approach was somewhat subjective and often time consuming, the scope of this process was typically limited to a few genetic variants and/or their nearest genes. In contrast, modern sequencing studies generate vast numbers of genomic variants, and automated strategies are now required to provide an efficient and somewhat objective process for annotation.

Another role of the annotation process is the assessment of potential clinical impact of a variant (Ashley et al., 2010). Clinical sequencing of patient samples is becoming a cost-effective alternative to individual gene testing. When applied to families, it may elucidate new highly penetrant genetic variants for Mendelian disorders. In the latter circumstance

especially, functional annotation of genomic variants may help identify the pathogenic variant. Since variants can represent changes in amino acid sequence of an encoded protein, disease-related biological targets can be more readily identified and evaluated for the protein of interest. Finding a suitable biological target is the first step in a drug discovery process for new disease-related therapeutics.

For research studies, the ultimate goal is to identify genetic changes that directly influence a phenotype of interest. The motivation is to improve our understanding of human biology; for diseases the subsequent goal is to develop new interventions and therapies to reduce disease risk. Since 2005, genome-wide association studies have been the primary method for capturing common genetic variation, relying on patterns of linkage disequilibrium within a population to identify indirect associations between single nucleotide polymorphisms (SNPs) and phenotypes (Korte and Farlow, 2013). With the development of next-generation sequencing technologies, human genomic studies can now capture a broad spectrum of allelic diversity, including low frequency (rare) variants and small-scale insertions and deletions (Metzker, 2010). This technological shift has resulted in a dramatic shift in the scale and scope of generated data (Marx, 2013). Typical studies identify millions of new genetic variants, many of which may only be seen in one or two individuals of the study. As a result, the vast majority of these variants are useless when used in traditional statistical hypothesis testing approaches applied in GWAS. This has led to two critical developments in genomic study design; 1) new statistical approaches that aggregate low frequency variants by functional groupings to improve statistical power, and 2) an increased need to assign a putative biological impact to the variants identified. Both developments rely on the concepts of genome annotation. In this Unit, we discuss the data resources available for functional genomics annotation, the software tools used to assign annotations to identified variants, and the general philosophy of assigning potential function to a genetic change within the biological context of the disease being studied.

GOALS OF ANNOTATION

The primary goal and challenge of annotating genomic variants is to process and condense the vast amounts of information into a useful form for genomic studies. Knowledge about the structure, function, and impact of genomic sequences can be assigned to variants, ultimately resulting in a better understanding of the molecular mechanisms by which a genetic variant may mediate its phenotypic impact. There are two primary types of annotation processes, based on when it is performed relative to genomic analysis: *post-hoc*, and *ad-hoc*.

Post-hoc annotation of risk variants

A typical, albeit non-structured form of annotation, is the post-hoc assignment of variant function following a statistical analysis. This approach is common to publications of genome-wide association studies that perform agnostic tests of association across common variants in the genome. Variants from this analysis that surpass multiple testing corrections and replication studies are considered true positives, and often potential functional consequences of these variants are reported to inform future molecular studies. Post-hoc annotation is often based on literature searches of a given variant or in some cases the

biological function of the nearest gene. The assignment of function based on sequence proximity to a gene can be perilous – an early GWAS for type II diabetes revealed a strongly associated SNP intronic to the *FTO* gene (Frayling et al., 2007). Later studies of the associated variant, however, indicate that its function may be mediated by altering the expression of a downstream gene, *IRX3* (Smemo et al., 2014).

More recently, GWAS have expanded to directly genotype millions of genetic variants, and because the likelihood of directly associating a variant to a phenotype is increased relative to earlier studies which used linkage disequilibrium to identify associations (Bush and Moore, 2012). Because more variants are directly genotyped, the functional variant (rather than simply a variant in strong linkage disequilibrium) can be identified and annotated. Furthermore, next-generation sequencing studies in many cases may perform single-variant tests of association to identify signals that earlier GWAS may have missed. These analyses generally directly identify the causal variant. The overwhelming number of variants identified in sequencing studies may lead to a much larger number of results than GWAS, and as a consequence, post-hoc annotation is sometimes used to prioritize candidates for follow-up based on their putative biological consequences.

The motivations for post-hoc annotation are clear given the typical goals of genomic studies; translating a genetic finding into a therapeutic target first requires translating the finding into a targetable biological mechanism. This has been a sorely overlooked need for GWAS especially, as the majority of publications provide little insight into *how* their reported SNP associations are influencing biological systems to result in a phenotype. While detailed molecular studies are the gold standard for confirming variant function, variant annotation can provide insights about GWAS associations that can guide these costly studies.

Ad-hoc annotation to guide the analysis of genetic variation

Performing ad-hoc annotation prior to statistical analysis of genetic data has grown in importance for a variety of reasons. In the past, variant annotation has been used for computationally intensive analyses, such as the exploration of multi-locus interactions. Tools such as INTERSNP (Herold et al., 2009) and Biofilter (Bush et al., 2009; Pendergrass et al., 2013) use gene and pathway annotations to group variants into pair-wise tests, reducing the multiple testing burden and computational intensity of genome-wide interaction scans. Similar types of annotation data have been used to construct Bayesian networks, and even to weight false discovery rate adjustments. Despite much bioinformatics development in this area, annotation-prioritized association analyses are generally not primary analyses of genomic data; these techniques build upon existing data and potentially bias discoveries toward biological mechanisms that are already established, and this is contrary to the agnostic design of GWAS.

With a shift from genotyping common variants to next-generation sequencing technologies (Metzker, 2010) for capturing lower frequency variants, the importance of functional annotation has grown. Rare variants by definition are infrequent in a population, and traditional statistics for performing association tests are not well-powered (in available sample sizes) to detect the low to moderate effect sizes expected from these variants (Bansal et al., 2010; Neale et al., 2011). As a result, multiple statistical approaches have been

developed to aggregate rare variants across a defined genomic region, typically based on gene boundaries. These algorithms, databases, and tools used to define genomic annotations are much more influential on the successful conduct of a study (Pabinger et al., 2014).

Philosophical issues arising from genomic annotation

With the growing importance and influence of variant annotation on genomic studies, multiple considerations must be addressed before proceeding. Just as early GWAS identified new analytical issues like multiple testing corrections and population stratification (Johnson et al., 2010; Tian et al., 2008), sequencing studies will undoubtedly reveal new issues with variant annotation. While genome builds and functional genomics studies continually improve our understanding of genome structure and function, the quality and depth of understanding is not uniform across the human genome. Furthermore, the algorithms and methods used to assign annotations to variants make their own assumptions and choices that may not always be congruent with the biological hypothesis or phenotype being tested. Newer approaches attempt to quantify the importance of a base-pair change rather than giving a qualitative annotation; it is perhaps overly simplistic to assume that a single score can capture the complexity of a genomic change. Finally, the scope of genomic sequencing can impact the type and possible consequences of genetic variants.

ALGORITHMIC APPROACHES TO GENOMIC ANNOTATION

Multiple approaches have been explored for annotating variant consequence, but these typically fall into two categories: qualitative assignment and quantitative impact scoring. Qualitative assignment of variant consequence is performed by examining the sequence context in which the variant occurs; variants may be considered in an untranslated region (or UTR) if they fall within the non-coding portions of a transcript. These qualitative assignments are made in reference to a catalog of genomic features, such as the exon/intron structure of a gene region, or the locations of regulatory elements, and often multiple assignments can be made to the same variant. Quantitative impact scores are generated using statistical or machine learning algorithms that use multiple features of a variant to predict their impact or deleteriousness. Depending on the algorithm used, there can be a single score per variant, or multiple scores depending upon its genomic/transcriptomic context.

Qualitative assignment of variant consequence

Since the development of Sanger sequencing (Sanger et al., 1977), algorithms have been developed to process DNA sequences to identify functional elements. After the completion of the human genome project (Lander et al., 2001), both the National Center for Biotechnology Information (NCBI) and the European Molecular Biology Laboratory (EMBL) undertook initiatives to build a comprehensive catalog of human transcripts and map these transcripts back to their gene positions on the genome. Since then, multiple efforts have captured transcriptome variation and transcript expression levels across multiple cell lines and human tissues, and these experiments have refined gene and transcript databases. Most notably, the GENCODE project (Harrow et al., 2012) has leveraged multiple lines of experimental data about functional elements of the genome to build a high

quality catalog of transcripts, and these have subsequently been incorporated in the ENSEMBL database (Flicek et al., 2012).

Multiple approaches (Pabinger et al., 2014) have been developed to map variants onto the basic features of a gene region using either ENSEMBL or RefSeq catalogs (Pruitt et al., 2007). These basic features (exons, introns, and UTRs) are supplemented by examining sequence motifs to identify splice sites and additional features. Variants are then assigned consequences based on where they fall within the gene region. Should they fall within coding regions, the specific codon change is assessed to determine the resulting amino acid or stop codon. Short indels or insertions can result in a change in the translational reading frame (frameshift), with potentially drastic consequences for the protein product (Rockah-Shmuel et al., 2013). Commonly used software tools for performing these annotations are shown in Table 1. The following reviews (Castellana and Mazza, 2013; Ritchie and Flicek, 2014) provide a more extensive evaluation regarding strengths and limitations of these tools.

The qualitative annotations produced by these software tools can vary, and may have dramatic differences in interpretation. For example, McCarthy et al. showed that there is rather low concordance (65%) for so-called Loss of Function (LoF) variants between ANNOVAR and variant effect predictor (VEP), even when the two tools use the same transcript set (McCarthy et al., 2014). The Sequence Ontology Project (Mungall et al., 2011) is one resource for attempting to standardize annotations to genomes, and both VEP and ANNOVAR are integrating options for producing annotations referencing sequence ontology (SO) terms.

The SO terminology, however, can be somewhat terse for those not accustomed to annotation. SO terms can be for variants (generally a single base-pair change), insertions, deletions, and duplications. Annotations are generally in terms of *features* which is a higher-level generic term with more specific child terms like *regulatory region*, *transcription factor binding site*, or *transcript*. Variants are also further sub-classified by their impact, with terms such as *stop gained*, *frameshift*, or *missense*. The term *ablation* is used to describe the complete or majority deletion of a genomic feature. As the sophistication of annotation tools progresses, more specific sub-types of variants and features will be identified.

A number of molecular genetics concepts are crucial for understanding SO terms. Transcript splicing and the complex mechanisms of the spliceosome remove introns to produce mature messenger RNAs. The “GU” splice donor sequence at the 5' end of an intron and the “AG” splice acceptor sequence at the 3' end of an intron can be altered with consequences to the transcriptional isoforms produced by a gene. Synonymous variants near the exon boundaries may also impact splicing. An excellent review by Matera and Wang provides detailed information (Matera and Wang, 2014).

While there are numerous transcripts typically identified for any gene, many of these are subject to nonsense-mediated decay (NMD), a post-transcriptional mRNA surveillance mechanism used to degrade non-functional transcripts. These transcripts are generally due to the presence of premature stop codons. More detailed information on the molecular processes involved in NMD can be found in a review by Maquat (Maquat, 2004).

As our molecular understanding of gene expression and its regulation expands, annotation tools will develop to encompass those features as well. With numerous high-profile projects collecting this information, expect many additional variant annotations to become available.

Detailed Examples of Variant Consequences

As examples, we detail the consequences produced by the VEP pipeline, along with their sequence ontology terms and sequence ontology identifier.

transcript_ablation—SO Description: A feature ablation whereby the deleted region includes a transcript feature - SO:0001893. This consequence refers to occasions where a variant (generally a deletion or other structural variant) completely removes a “feature”, typically an exon or an entire gene. **Example:** Chr20:62,550,824 - a 65 base pair insertion abolishes a large segment of a microRNA gene MIR941-2 located on Chromosome 20: 62,550,834-62,550,950.

splice_acceptor_variant—SO Description: A splice variant that changes the 2 base region at the 3' end of an intron - SO:0001574. This consequence refers to variants that disrupt the normal splicing of a transcript during the major splicing pathway (the canonical splicing mechanism responsible for 99% of splicing). The AG region at the 3' end of an intron binds to U2AF1, one of five small nuclear ribonucleoproteins (snRNPs). This can result in a retained intron in the final transcript. **Example:** Chr1:22,453,975 (rs2235527) – a C to G SNP which alters the splice acceptor site on the reverse strand. What should have been a TC sequence (corresponding to the AG acceptor) is now a TG sequence, likely disrupting normal splicing by retaining intron 1 of a WNT4 gene transcript.

splice_donor_variant—SO Description: A splice variant that changes the 2 base region at the 5' end of an intron - SO:0001575. This consequence similarly refers to variants that disrupt the major splicing pathway at the 5' end of an intron where splicing originates. The GU region at the 5' end binds to U1, a snRNP. Disruption of this sequence by a variant can eliminate a splicing event. **Example:** Chr1:1,268,505 (rs141631187) – a G to A SNP which alters the splice donor site on the forward strand. What should have been a GU sequence is now an AU sequence, likely disrupting normal splicing by retaining intron 4 of a *TASIR3* gene transcript.

stop_gained—SO Description: A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript - SO: 0001587. This consequence occurs when one of the three-bases of a codon is changed within a transcript to induce a premature stop codon (either a TAA, TAG, or TGA codon). This annotation is agnostic of the location where this stop codon occurs; stop gained variants near the end of the translated sequence may have no functional impact at all. **Example:** Chr1:1,256,033 (rs11801967) – a C to T SNP which alters a TGG codon #51 (for Leucine) to a TGA stop codon, resulting in a truncated transcript of the *CPSF3L* gene, reducing its size from 55 codons to 51. This transcript undergoes nonsense mediated decay and does not produce a protein product.

frameshift_variant—SO Description: A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three - SO:0001589. This consequence occurs due to insertions or deletions that alter the reading frame of the RNA polymerase, resulting in a radically different translated protein sequence. It is unclear how the location of frameshift variants within a transcript affects their possible function, and it is further unknown how much additional translated sequence is generated by a frameshift without examining the downstream altered sequence for the next stop codon position. Frameshifts could occur near the end of a transcript that induce a stop codon shortly thereafter, and as such may not dramatically alter protein function. **Example:** Chr8:97,243,524 (rs140631053) – a single base-pair (T) insertion alters the reading frame of a protein coding transcript of *UQCRB* gene at codon 159 of 161.

stop_lost—SO Description: A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript - SO:0001578. This consequence occurs when a base-pair change results in a shift from a stop codon (either a TAA, TAG, or TGA codon) to another codon, resulting in an elongated protein product. This elongated product, however, may have another proximal stop codon of its own and thus may not deleteriously affect protein function. **Example:** Chr1: 2,434,523 (rs61751930) is a T to C change part of the seventh of seven exons of the *PLCH2* gene. The resulting change alters the stop codon TAG to non-stop codon CAG, though the distance to the next-most proximal stop codon is unknown without downstream sequence analysis.

start_lost—SO Description: A codon variant that changes at least one base of the canonical start codon - SO:0002012. This consequence occurs when a base-pair change disrupts a start codon (an ATG at the beginning of the translation initiation site). While it is assumed that translation initiation would begin at the next ATG site, molecular studies have shown that translation can initiate at multiple downstream start sites, not necessarily only the next most proximal one (Parsons et al., 2013). **Example:** Chr17:62,893,375 (rs147145975) – a T to C change results in a codon shift from ATG to GTG, abolishing the start site of a protein-coding transcriptional isoform of gene *LRRC37A3*.

inframe_insertion—SO Description: An inframe non synonymous variant that inserts bases into in the coding sequence - SO:0001821. This consequence occurs when a sequence is inserted which adds additional codons without altering the RNA polymerase reading frame (e.g. the sequence length is wholly divisible by 3). Depending upon the length and location of this insertion, the functional consequence of this change may vary; disrupting a critical binding domain or dramatically altering protein stability may result in a highly deleterious change, whereas altering a linker region or other benign structural portion of a protein may result in little functional change. **Example:** Chr1:1147177 – a six base-pair insertion (CCAGGC) results in the inclusion of an additional Glycine and Leucine at residue 224 of a protein coding (canonical) transcript of the *TNFRSF4* gene. This insertion occurs within a transmembrane helix, but does not appear to impact other known domains of the protein.

inframe_deletion—SO Description: An inframe non synonymous variant that deletes bases from the coding sequence - SO:0001822. This consequence occurs when a sequence is deleted which eliminates codons from the transcript without altering the RNA polymerase reading frame (the sequence length is wholly divisible by 3). As with insertions, the length and location of the deletion will determine the functional consequence of this change.

Example: Chr12:102894801 (rs62508727) – an 3 base-pair deletion (TCA) results in the removal of an isoleucine at residue 95 of a protein coding transcript for the PAH gene. This deletion reduces the efficiency of a hepatic enzyme and is associated with autosomal recessive disease Phenylketonuria (Caillaud et al., 1991).

missense_variant—SO Description: A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved - SO:0001583. This consequence occurs when a base-pair change alters the changes the codon of a coding sequence from one amino acid to another based on established codon tables. The position and location of the codon change can have dramatic effects on the functional impact of the variant. Additional information, such post-translational modifications, may be relevant for exploring the impact of these variants. **Example:** Chr15: 85342440 (rs150829393) – a G to A change alters the GGG (glycine) to an AGG (arginine) at residue 1046 of the canonical protein coding transcript of the ZNF592gene. This amino acid change is associate with Spinocerebellar ataxia (Delague et al., 2002; Nicolas et al., 2010).

splice_region_variant—SO Description: A sequence variant in which a change has occurred within the region of the splice site, either within 1–3 bases of the exon or 3–8 bases of the intron - SO:0001630. This consequence occurs within the flanking regions of an intron (that is spliced out) and may influence the splicing process by affecting exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs) (Fairbrother et al., 2002), through the creation of new cryptic splice sites (Eskesen et al., 2004), or by altering splice-site recognition motifs (Fairbrother et al., 2002). **Example:** Chr1:25611065 (rs148114582) – a T to C change in the second base of the second exon is synonymous (GTT to GTC, both coding for Valine), but may also alter splicing as these sequence regions are part of the consensus splicing sequence.

incomplete_terminal_codon_variant—SO Description: A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed - SO: 0001626. This consequence actually reflects uncertainty in the transcript model to which the variant maps. There is insufficient data to properly identify the 3' end of the transcript, and when a variant occurs in a partial codon, a prediction of its impact on the coding sequence cannot be made. This generally occurs in highly complex genes with multiple transcripts. **Example:** Chr11:67,378,053 (rs200417926) – a C to T change alters the last codon of a 136aa protein coding isoform (ENST00000532244) of the gene NDUFV1, which is flagged as having an incomplete 3' CDS sequence.

stop_retained_variant—SO Description: A sequence variant where at least one base in the terminator codon is changed, but the terminator remains - SO:0001567. This consequence occurs when a variant introduces a 'synonymous' change to a stop codon

interchanging a TAA, TAG, or TGA codon. While this is generally thought to be a benign change, it may change the dynamics of protein translation due to codon availability and usage. Codon usage statistics for stop codons are: TAA 0.7 per thousand, TGA 1.3 per thousand, TAG 0.5 per thousand (via GenScript). Example: Chr11:74,717,537 (rs180878536) – an A to G change alter the stop codon of multiple transcripts of the TCIRG1 gene from TAA to TAG.

synonymous_variant—SO Description: A sequence variant where there is no resulting change to the encoded amino acid - SO:0001819. This consequence occurs when a variant alters the codon of a transcript without altering the coded amino acid, possible because some amino acids have redundant codons. The true consequence of these variants can take multiple routes, including changes to RNA processing and post-transcriptional regulation, changes to the global and local stability of an mRNA molecule, changes to codon usage that affects translational efficiency, and possible changes to ribosomal pausing and accessibility. **Example:** Chr1:881,800 – a G to A change in the protein coding transcript of the NOC2L gene changing codon 595 from GTC to GTT, both of which code for the amino acid Valine.

coding_sequence_variant—SO Description: A sequence variant that changes the coding sequence - SO:0001580. This consequence reflects uncertainty in the underlying transcript model, and indicates that a variant alters the coding sequence of a transcript with an incomplete codon. This is generally due to an incomplete 5' sequence. **Example:** Chr1:3,745,870 – a C to T change in a protein coding transcript of the CEP104 gene changing a partial codon nGG to nAG, however the change to this codon cannot be predicted as synonymous or missense due to missing 5' sequence of this transcript (including the first base of the codon).

mature_miRNA_variant—SO Description: A transcript variant located with the sequence of the mature miRNA-SO:0001620. This consequence occurs when a variant falls within the transcribed region of a mature miRNA. miRNAs undergo their own transcriptional process through the Dicer and Drosha mechanisms, and are ultimately bound with proteins. These mature miRNAs have a 6-8nt seed sequence that binds 3'UTR sequences for transcriptional regulation. Alterations to this sequence may disrupt normal miRNA targeting, but it is unknown what effect variants in miRNA transcript have outside the seed region. **Example:** Chr17:46,657,266 – a C to T change in the transcribed sequence of a mature miRNA, MIR10A, located at position 44 of the transcript (outside the seed region).

5_prime_UTR_variant—SO Description: A UTR variant of the 5' UTR - SO:0001623. This consequence occurs when a variant falls within the 5' untranslated region (UTR) of a messenger RNA. Mutations to the 5'UTR of a gene can affect the secondary structure and thermal stability, upstream open reading frames that alter translational efficiency, and internal ribosome entry sites (IRES), all of which have been previously related to disease (Chatterjee and Pal, 2009). Many of the elements altering translational efficiency within the 5' UTR have established sequence motifs. **Example:** Chr1:1,114,579 – a C to T change in the 5' untranslated region of two protein coding transcripts of the TTLL10 gene. It occurs at base 135 of one transcript and base 157 of the other.

3_prime_UTR_variant—SO Description: A UTR variant of the 3' UTR - SO:0001624. This consequence occurs when a variant falls within the 3' untranslated region of a messenger RNA. Mutations to the 3'UTR of a gene can alter the secondary structure and the polyadenylation signal, thus impacting the stability of the transcript. Variants within the 3'UTR may also abolish or create miRNA binding sites, impacting the regulation of transcript pools. **Example:** Chr16:684,591 (rs140632002) – an A to G change in the 3' untranslated region of three protein coding transcripts of the C16orf13 gene (an open reading frame). This variant also occurs in the 3' untranslated region of two transcripts that undergo nonsense mediated decay.

non_coding_transcript_exon_variant—SO Description: A sequence variant that changes non-coding exon sequence in a non-coding transcript - SO:0001792. This consequence occurs when a variant falls within the exon of a non-coding RNA transcript, typically a long non-coding RNA, as other non-coding genes do not have exon/intron structures. **Example:** Chr1:224,918,264 (rs57719973) is a C to T change within the first of three exons of the RP11-3L21.2 gene, a long non-coding RNA validated by the GENCODE project.

intron_variant—SO Description: A transcript variant occurring within an intron - SO:0001627. This consequence occurs when a variant falls within the intron of coding gene. **Example:** Chr1:819,917 (rs6605057) is an A to G change part of the second of two introns of the AL645608.2 gene.

NMD_transcript_variant—SO Description: A variant in a transcript that is the target of NMD - SO:0001621. This consequence occurs when a variant falls within a transcript that undergoes nonsense-mediated decay (NMD). Nonsense mediated decay is a post-transcriptional mRNA surveillance mechanism used to control the quality of transcripts, most notably by degrading transcripts with premature stop/termination codons (PTCs). ENSEMBL uses VEGA classifications (Wilming et al., 2008) for transcripts, many of which are manually curated by the HAVANA project (Hancock et al., 2004), which in this case means that a transcript is labeled as NMD if the coding sequence (following the appropriate reference) of a transcript finishes >50bp from a downstream splice site, or if the variant does not cover the full reference coding sequence then it is annotated as NMD. Importantly, this consequence co-occurs with other consequences (5' UTR, intron, missense, etc) to designate that the transcript this consequence is relative to undergoes NMD. **Example:** Chr1:90,463,679 is a C to T synonymous change at codon 9 of the only NMD transcript for the *ZNF326* gene. This transcript was manually curated by HAVANA as “Known nonsense mediated decay”.

non_coding_transcript_variant—SO Description: A transcript variant of a non-coding RNA gene - SO:0001619. This consequence occurs when a variant falls within the intron of a non-coding gene, typically a long non-coding RNA. **Example:** Chr1:224,918,264 (rs57719973) is a C to T change within the first of three exons of the RP11-3L21.2 gene, a long non-coding RNA validated by the GENCODE project.

upstream_gene_variant—SO Description: A sequence variant located 5' of a gene - SO:0001631. This consequence occurs when a variant falls within 5,000 bases (5KB) before the first exon of a transcript. **Example:** Chr1:1,890,651 is an A to G change that falls upstream of the first exon of a single transcript of the CFAP74 gene by 3,497 bases.

downstream_gene_variant—SO Description: A sequence variant located 3' of a gene - SO:0001632. This consequence occurs when a variant falls within 5,000 bases (5KB) after the last exon of a transcript. **Example:** Chr1:878,661 (rs112422281) is an A to G change that falls downstream of nine transcripts of the NOCL2 gene. Across the different transcripts, it falls between 279 and 3982 bases past the end of the last exon.

TFBS_ablation—SO Description: A feature ablation whereby the deleted region includes a transcription factor binding site - SO:0001895. This consequence refers to occasions where a variant (generally a deletion or other structural variant) completely removes an established transcription factor binding site (TFBS). **Example:** Chr1:113,243,264 is a five base deletion (CCCCCT) that removes an EGR1 binding motif.

TFBS_amplification—SO Description: A feature amplification of a region containing a transcription factor binding site - SO:0001892. This consequence refers to occasions where a variant (generally a short duplication or an insertion) amplifies the effect of an established transcription factor binding site (TFBS). **Example:** Chr8:22,462,393 is an eight base duplication (CGCCCCCTC) that duplicates an EGR1 binding motif.

TF_binding_site_variant—SO Description: A sequence variant located within a transcription factor binding site - SO:0001782. This consequence occurs when a variant falls within an established transcription factor binding site (TFBS). **Example:** Chr5:118,788,316 (rs137853096) is a G to A change that results in a glycine to serine change altering an HSD17B4 binding motif associated with Bifunctional peroxisomal enzyme deficiency (van Grunsven et al., 1998, 1999).

Quantitative scoring of variant impact

Several algorithms have been proposed and developed to score variants based on their potential “deleteriousness”. These algorithms are generally based on classifiers that incorporate multiple features of a variant, such as the evolutionary conservation of the surrounding region or protein domain information. There are two important elements of these classification algorithms: the training set of example deleterious variations, and the context of features used to make predictions.

Classifiers of variant impact require training sets to identify the precise set and/or combination of input features that are associated to a “deleterious” allele; training sets require numerous examples of disease associated variants to establish a frame of reference. These training sets therefore are typically formed from resources such as the Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2005; see UNIT 9.13, Baxevanis, 2012) or the Human Gene Mutation Database (HGMD) (Stenson et al., 2003). The choice of data used as this training set is critically important, as it frames the type of deleterious

variants likely to be predicted; algorithms trained by examining variants of a single type will only be able to accurately predict or identify variants of that same type. An open question in genomics is the extent to which rare-variants associated to common complex disease will be biologically similar to the highly penetrant variants associated to Mendelian forms of disease found in the training catalogs. A detailed comparison of the various catalogs of disease-associated mutations has been published (Peterson et al., 2013).

The choice of features used in these prediction algorithms also frames the context in which functional prediction scores are relevant. Among the first of these types of functional prediction scores was SIFT (Kumar et al., 2009), which was developed specifically to examine the impact of non-synonymous variants that alter the amino acid sequence of a protein. PolyPhen and PolyPhen-2 (Adzhubei et al., 2010; see UNIT 7.20, Adzhubei et al., 2013) are similar approaches that leverage additional information, such as the three-dimensional context of a variant within the protein structure. Another approach is the Combined Annotation Dependent Depletion (CADD) score (Kircher et al., 2014), a support vector machine based approach that integrates multiple annotations into one metric through emphasizing variants that prevailed a simulated selection process. A variety of other algorithms have been similarly developed (Ritchie and Flicek, 2014) all using different feature sets to make predictions, each with their own strengths, weaknesses, and caveats (Castellana and Mazza, 2013). A critical point about all methods for scoring missense variants; they are all made for a variant in the context of a specific protein sequence. Variants that fall within multiple genes, or within genes producing multiple protein isoforms will have multiple (sometimes distinct) scores of variant impact.

Assess the potential clinical impact of a variant

Whole exome and whole genome sequencing are being used in some circumstances to identify pathological variants that influence diseases with a strong family basis or for patients exhibiting symptoms consistent with other known Mendelian syndromes (Veltman and Lupski, 2015). The majority of focus in the literature has been on the quality control procedures, validity, and reproducibility of genomic sequencing data. However, another important consideration is the ability to annotate variants with respect to existing clinical genomics findings.

For over two decades, community-based online resources have organized information about the cosegregation, positional cloning, and phenotypic characteristics of highly-penetrant variants for Mendelian forms of disease (OMIM). Recent efforts have focused on making this information more usable for sequence-based genomic studies, while also augmenting clinical information with measures of confidence in the findings, such as the number of observed cases/families with a given variant. The result is the NCBI resource ClinVar (Landrum et al., 2014), which is now available in multiple formats amenable to variant annotation, such as the Variant Call Format/VCF (Danecek et al., 2011). The precise identification of causal variants for Mendelian diseases is a complex matter, and multiple recent studies have illustrated that OMIM and ClinVar resources may contain some inaccuracies (Freudenberg-Hua et al., 2014; MacArthur et al., 2014).

NON-ALGORITHMIC RESOURCES FOR GENOMIC ANNOTATION

Multiple large-scale collaborative projects have produced a wealth of information about functional elements in the human genome. The ENCODE project has actualized multiple lines of concurrent evidence supporting regulatory elements (ENCODE Consortium, 2004). Chromatin Immuno-Precipitation (ChIP) (Dahl and Collas, 2008; Das et al., 2004) assays have identified the putative binding sites within the genome for a multitude of transcript factors, including POL2 which helps to localize the location of promoter elements in the genome (Corcoran et al., 2009). DNase I Hypersensitivity assays (He et al., 2012, 2013; Madrigal and Krajewski, 2012) are an orthogonal approach, identifying regions of chromatin that are “open” or accessible to digestion by the DNase I enzyme. All these data have been augmented and combined with RNA-sequencing (DNase-Seq) to systematically identify the functional elements and expressed transcripts from a variety of cell lines, all of which has been used by the GENCODE project to refine more complete transcriptional catalogs for all genes (Harrow et al., 2012; Derrien et al., 2012). Alternatives to DNase-Seq are FAIRE-Seq (Giresi et al., 2007) and ATAC-Seq (Buenrostro et al., 2015). In contrast, FAIRE-Seq applies formaldehyde crosslinking, which is considered more efficient than nucleosome-bound DNA, to assess open chromatin states. ATAC-Seq employs an enzyme (transposase Tn5) to capture open chromatin sites. It is considered a rapid and precise approach for epigenomic analyses.

ChIP assays have also been used within both ENCODE and the Epigenomics Roadmap Project (Bernstein et al., 2010) to identify histone methylation and acetylation marks which have been subsequently processed to classify regions of the genome into distinct chromatin states by tools such as Segway (Hoffman et al., 2012a) and ChromHMM (Hoffman et al., 2012b).

The FANTOM consortium (Forrest, 2014) uses an orthogonal technique, Cap Analysis of Gene Expression (CAGE) to identify short genomic sequences bound to active polymerases (Kanamori-Katayama et al., 2011). CAGE data in the most recent FANTOM5 release has been used to pinpoint active promoters (de Hoon et al., 2015) and enhancer elements on over 90 cell types.

Finally, multiple techniques are producing new lines of evidence for the occurrence and frequency of local chromatin looping and trans-chromosomal chromatin contact. These chromatin conformation experiments, such as C3 (Dekker et al., 2002), C4 (Dekker, 2006), C5 (Dostie and Dekker, 2007), Hi-C (de Wit and de Laat, 2012), ChIA-PET (Li et al., 2010), are quite heterogeneous in the resolution and coverage of contact they provide. Moving forward, these approaches will likely yield useful data for the annotation of regulatory regions.

SUMMARY

This unit reviews the annotation of genomic variation. It marks a challenge considering the interpretation of individual variants in the context of disease has not kept pace with its identification or discovery. With steadily declining costs for whole-genome sequencing, clinical annotation and risk assessment will become more time-consuming. Thus, the

interpretation of resulting data will increase in complexity. Here, goals of annotation are discussed covering annotation of risk variants as well as approaches to guide analysis of genetic variation. Further, this unit examines algorithmic approaches, summarizing qualitative and quantitative tools and approaches to annotate genomic variants, as well as non-computational techniques.

Acknowledgments

This work was supported in part by the Consortium for Alzheimer's Sequence Analysis (CASA), funded by the National Institute on Aging (1UF01 AG07133).

LITERATURE CITED

- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010; 7:248–249. [PubMed: 20354512]
- Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr Protoc Hum Genet*. 2013; 76:7.20:7.20.1–7.20.41.
- Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, et al. Clinical assessment incorporating a personal genome. *The Lancet*. 2010; 375:1525–1535.
- Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet*. 2010; 11:773–785. [PubMed: 20940738]
- Baxeavanis AD. Searching Online Mendelian Inheritance in Man (OMIM) for Information on Genetic Loci Involved in Human Disease. *Curr Protoc Hum Genet*. 2012; 73:9.13:9.13.1–9.13.10.
- Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol*. 2010; 28:1045–1048. [PubMed: 20944595]
- Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*. 2015:21–29. [PubMed: 25559105]
- Bush WS, Moore JH. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*. 2012; 8:e1002822. [PubMed: 23300413]
- Bush, WS.; Dudek, SM.; Ritchie, MD. Pacific Symposium on Biocomputing. World Scientific; 2009. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies; p. 368-379.
- Caillaud C, Lyonnet S, Rey F, Melle D, Frebourg T, Berthelon M, Vilarinho L, Osorio RV, Rey J, Munnich A. A 3-base pair in-frame deletion of the phenylalanine hydroxylase gene results in a kinetic variant of phenylketonuria. *J Biol Chem*. 1991; 266:9351–9354. [PubMed: 1709636]
- Castellana S, Mazza T. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Brief Bioinform*. 2013:btt013.
- Chatterjee S, Pal JK. Role of 5'- and 3'-untranslated regions of mRNAs in human diseases. *Biol Cell*. 2009; 101:251–262. [PubMed: 19275763]
- Collins FS, Brooks LD, Chakravarti A. A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. *Genome Res*. 1998; 8:1229–1231. [PubMed: 9872978]
- Corcoran DL, Pandit KV, Gordon B, Bhattacharjee A, Kaminski N, Benos PV. Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PloS One*. 2009; 4:e5279. [PubMed: 19390574]
- Cutting GR. Annotating DNA Variants Is the Next Major Goal for Human Genetics. *Am J Hum Genet*. 2014; 94:5–10. [PubMed: 24387988]
- Dahl JA, Collas P. A rapid micro chromatin immunoprecipitation assay (ChIP). *Nat Protoc*. 2008; 3:1032–1045. [PubMed: 18536650]

- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. *Bioinformatics*. 2011; 27:2156–2158. [PubMed: 21653522]
- Das PM, Ramachandran K, vanWert J, Singal R. Chromatin immunoprecipitation assay. *Biotechniques*. 2004; 37:961–969. [PubMed: 15597545]
- Dekker J. The three “C” s of chromosome conformation capture: controls, controls, controls. *Nat Methods*. 2006; 3:17–21. [PubMed: 16369547]
- Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science*. 2002; 295:1306–1311. [PubMed: 11847345]
- Delague V, Bareil C, Bouvagnet P, Salem N, Chouery E, Loiselet J, Mégarbané A, Claustres M. A new autosomal recessive non-progressive congenital cerebellar ataxia associated with mental retardation, optic atrophy, and skin abnormalities (CAMOS) maps to chromosome 15q24-q26 in a large consanguineous Lebanese Druze Family. *Neurogenetics*. 2002; 4:23–27. [PubMed: 12030328]
- Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*. 2012; 22:1775–1789. [PubMed: 22955988]
- Dostie J, Dekker J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc*. 2007; 2:988–1002. [PubMed: 17446898]
- ENCODE Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science*. 2004; 306:636–640. [PubMed: 15499007]
- Eskenen ST, Eskenen FN, Ruvinsky A. Natural selection affects frequencies of AG and GT dinucleotides at the 5' and 3' ends of exons. *Genetics*. 2004; 167:543–550. [PubMed: 15166176]
- Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. *Science*. 2002; 297:1007–1013. [PubMed: 12114529]
- Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. Ensembl 2013. *Nucleic Acids Res*. 2012:gks1236.
- Forrest, A. PRION. LANDES BIOSCIENCE; 1806 RIO GRANDE ST, AUSTIN, TX 78702 USA: 2014. FANTOM5: A deep sequencing expression atlas of mammalian cell types; p. 8-8.
- Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*. 2007; 316:889–894. [PubMed: 17434869]
- Freudenberg-Hua Y, Freudenberg J, Vacic V, Abhyankar A, Emde AK, Ben-Avraham D, Barzilai N, Oschwald D, Christen E, Koppel J, et al. Disease variants in genomes of 44 centenarians. *Mol Genet Genomic Med*. 2014; 2:438–450. [PubMed: 25333069]
- Giresi PG, Kim J, McDaniel RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res*. 2007; 17:877–885. [PubMed: 17179217]
- van Grunsven EG, van Berkel E, Ijlst L, Vreken P, de Klerk JB, Adamski J, Lemonde H, Clayton PT, Cuebas DA, Wanders RJ. Peroxisomal D-hydroxyacyl-CoA dehydrogenase deficiency: resolution of the enzyme defect and its molecular basis in bifunctional protein deficiency. *Proc Natl Acad Sci U S A*. 1998; 95:2128–2133. [PubMed: 9482850]
- van Grunsven EG, van Berkel E, Mooijer PA, Watkins PA, Moser HW, Suzuki Y, Jiang LL, Hashimoto T, Hoefler G, Adamski J, et al. Peroxisomal bifunctional protein deficiency revisited: resolution of its true enzymatic and molecular basis. *Am J Hum Genet*. 1999; 64:99–107. [PubMed: 9915948]
- Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005; 33:D514–D517. [PubMed: 15608251]
- Hancock, JM.; Hancock, JM.; Zvelebil, MJ. Dictionary of Bioinformatics and Computational Biology. John Wiley & Sons, Ltd; 2004. HAVANA (Human and Vertebrate Analysis and Annotation).

- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 2012; 22:1760–1774. [PubMed: 22955987]
- He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res.* 2012; 22:1015–1025. [PubMed: 22508765]
- He HH, Meyer CA, Long H, Liu XS, Brown M. A closer look into DNase I hypersensitivity. *Epigenetics Chromatin.* 2013; 6:1–1. [PubMed: 23289424]
- Herold C, Steffens M, Brockschmidt FF, Baur MP, Becker T. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics.* 2009; 25:3275–3281. [PubMed: 19837719]
- Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods.* 2012a; 9:473–476. [PubMed: 22426492]
- Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* 2012b:gks1284.
- de Hoon M, Shin JW, Carninci P. Paradigm shifts in genomics through the FANTOM projects. *Mamm Genome.* 2015:1–12. [PubMed: 25398484]
- Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O'Brien SJ. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics.* 2010; 11:724. [PubMed: 21176216]
- Kanamori-Katayama M, Itoh M, Kawaji H, Lassmann T, Katayama S, Kojima M, Bertin N, Kaiho A, Ninomiya N, Daub CO, et al. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* 2011; 21:1150–1159. [PubMed: 21596820]
- Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014; 46:310–315. [PubMed: 24487276]
- Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods.* 2013; 9:29. [PubMed: 23876160]
- Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 2009; 4:1073–1081. [PubMed: 19561590]
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001; 409:860–921. [PubMed: 11237011]
- Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014; 42:D980–D985. [PubMed: 24234437]
- Li G, Fullwood M, Xu H, Mulawadi F, Velkov S, Vega V, Ariyaratne P, Mohamed Y, Ooi H, Tennakoon C, et al. Software ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.* 2010; 11:R22. [PubMed: 20181287]
- MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature.* 2014; 508:469–476. [PubMed: 24759409]
- Madrigal P, Krajewski P. Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Front Genet.* 2012; 3:230. [PubMed: 23118738]
- Maquat LE. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol.* 2004; 5:89–99. [PubMed: 15040442]
- Marx V. Biology: The big challenges of big data. *Nature.* 2013; 498:255–260. [PubMed: 23765498]
- Matera AG, Wang Z. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol.* 2014; 15:108–121. [PubMed: 24452469]
- McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier JB, Donnelly P, Consortium W, et al. Choice of transcripts and software has a large effect on variant annotation. *Genome Med.* 2014; 6:1–16. [PubMed: 24433494]

- Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet.* 2010; 11:31–46. [PubMed: 19997069]
- Mungall CJ, Batchelor C, Eilbeck K. Evolution of the Sequence Ontology terms and relationships. *J Biomed Inform.* 2011; 44:87–93. [PubMed: 20226267]
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ, et al. Testing for an unusual distribution of rare variants. *PLoS Genet.* 2011; 7:e1001322. [PubMed: 21408211]
- Nicolas E, Poitelon Y, Chouery E, Salem N, Levy N, Mégarbané A, Delague V. CAMOS, a nonprogressive, autosomal recessive, congenital cerebellar ataxia, is caused by a mutant zinc-finger protein, ZNF592. *Eur J Hum Genet EJHG.* 2010; 18:1107–1113. [PubMed: 20531441]
- Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform.* 2014; 15:256–278. [PubMed: 23341494]
- Parsons MT, Whitley PJ, Beesley J, Drost M, de Wind N, Thompson BA, Marquart L, Hopper JL, Jenkins MA, Brown MA, et al. Consequences of germline variation disrupting the constitutional translational initiation codon start sites of MLH1 and BRCA2: Use of potential alternative start sites and implications for predicting variant pathogenicity. *Mol Carcinog.* 2013
- Pendergrass SA, Frase AT, Wallace JR, Wolfe D, Katiyar N, Moore C, Ritchie MD. Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. *BioData Min.* 2013; 6
- Peterson TA, Doughty E, Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. *J Mol Biol.* 2013; 425:4047–4063. [PubMed: 23962656]
- Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2007; 35:D61–D65. [PubMed: 17130148]
- Ritchie GR, Flicek P. Computational approaches to interpreting genomic sequence variation. *Genome Med.* 2014; 6:87. [PubMed: 25473426]
- Rockah-Shmuel L, Tóth-Petróczy Á, Sela A, Wurtzel O, Sorek R, Tawfik DS. Correlated Occurrence and Bypass of Frame-Shifting Insertion-Deletions (InDels) to Give Functional Proteins. *PLoS Genet.* 2013; 9:e1003882. [PubMed: 24204297]
- Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci.* 1977; 74:5463–5467. [PubMed: 271968]
- Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gómez-Marín C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature.* 2014; 507:371–375. [PubMed: 24646999]
- Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeyasinghe S, Krawczak M, Cooper DN. Human gene mutation database (HGMD®): 2003 update. *Hum Mutat.* 2003; 21:577–581. [PubMed: 12754702]
- Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet.* 2008; 17:R143–R150. [PubMed: 18852203]
- Veltman JA, Lupski JR. From genes to genomes in the clinic. *Genome Med.* 2015; 7
- Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.* 2008; 36:D753–D760. [PubMed: 18003653]
- de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* 2012; 26:11–24. [PubMed: 22215806]

Table 1

Selection of commonly used annotation tools

Annotation tools	URL
ANNOVAR	http://annovar.openbioinformatics.org/
SNPEff	http://snpeff.sourceforge.net/
VEP	http://www.ensembl.org/info/docs/tools/vep/
GEMINI	http://gemini.readthedocs.org/
SeattleSeq	http://snp.gs.washington.edu/