Functional Annotation of Rare Genetic Variants

Graham R.S. Ritchie and Paul Flicek

Overview

Genome-wide association studies have successfully identified a growing number of common variants that robustly associate with a wide range of complex diseases and phenotypes. In the majority of cases though, the variants are predicted to have small to modest effect sizes, and, due to the technologies used, many of the signals discovered so far may not be the causal loci. As rare variation studies begin to explore the lower ranges of the allele frequency spectrum, using whole genome or whole exome sequencing to capture a larger proportion of variants, we expect to find variants with a more direct causal role in the phenotype(s) of interest. Interpreting possible functional mechanisms linking variants with phenotypes will become increasingly important.

Experimental investigation is the most direct way to establish if a candidate variant is causally involved in some phenotype, but it is a costly and time-consuming process, and so it is important to try to use as much existing relevant information as possible to prioritise variants for follow-up and to help formulate specific hypotheses

The original version of this chapter was revised. An erratum to this chapter can be found at https://doi.org/10.1007/978-1-4939-2824-8_19

G.R.S. Ritchie (✉)
European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK

Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

University of Edinburgh, Edinburgh, EH16 4UX, UK
e-mail: sigraham.ritchie@ed.au.uk

P. Flicek
European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK

Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK
e-mail: flicek@ebi.ac.uk

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_5

about functional mechanisms to inform subsequent experiments. The genome is complex and different classes of variants may have a wide range of, possibly tissue-specific, effects depending on their genomic context. In this chapter, we review some important classes of genome annotation and highlight some relevant computational tools and databases to help interpret and prioritise candidate variants depending on their genomic context. These resources may also play a role in the discovery of rare variant signals, as association techniques based on collapsing multiple rare variants together (reviewed in Chaps. 13 and 14) may use annotation of genes and regulatory elements to select biologically meaningful groups of variants, and other techniques can use prediction scores to upweight likely functional variants to increase statistical power. In this chapter, we focus on smaller-scale variants such as single nucleotide variants (SNVs) and short sequence insertions and deletions (indels), though some of the approaches we discuss may also be applied to larger structural variants.

Mapping Variants to Annotated Features

An obvious first step in trying to interpret possible functions of sequence variants is to identify overlapping genomic features that may be affected. Features of particular interest include protein-coding and non-coding genes, transcription factor binding sites and other potential regulatory regions. There are a wide range of resources and databases that can be used to identify likely functional genomic features, from very specific resources on a single class of feature such as the miRanda databases of microRNA (miRNA) target sites (Betel et al. 2007) to broad collections of annotations such as the Ensembl (Flicek et al. 2012) and UCSC (Meyer et al. 2013) databases.

For small numbers of variants, looking up the relevant loci in a genome browser, such as Ensembl or UCSC, is a convenient way to find overlapping or nearby features and to visualise variants in their genomic context. Both browsers contain a wealth of information on genes, regulatory regions and informative local genomic properties such as conservation, GC content and co-located or nearby variants (all of which we discuss in more detail later). For larger numbers of variants, automated approaches are clearly required. For simply identifying features overlapping variants, software packages such as BEDTools (Quinlan and Hall 2010) and BEDOPS (Neph et al. 2012a) provide powerful and efficient tools for computing overlaps and proximity (among other useful metrics) between large numbers of genomic loci and can read common variant file formats such as VCF and GVF and annotation files in widely used formats such as BED, GFF, GTF and SAM (more details on these formats are given in the Appendix). More variation specific tools such as the Ensembl Variant Effect Predictor (McLaren et al. 2010) and ANNOVAR (Wang et al. 2010) also identify a wide range of features overlapping variants, but can also make more specific predictions depending on the affected feature.

For many available annotations, especially those in non-coding regions, our understanding of the importance of specific genomic sequences is still in its infancy, and all we can report is that the variant overlaps the relevant annotation. For several classes of feature, such as genes and transcription factor binding sites, we have a more

detailed understanding of the importance of particular nucleotide sequences and so can make reasonably specific predictions about the effect of an allele on the element, as we discuss below. Even when we cannot take this further step, these overlaps provide some indication of the genomic context of the variant locus, and several studies, including the ENCODE consortium (Consortium, The ENCODE Project 2012), have found significant enrichments of trait-associated variants in less well-characterised regions, such as DNase1 hypersensitive sites, suggesting that these variants, or those nearby, affect some as-yet uncharacterised functional elements.

Variants Falling in Protein-Coding Genes

Protein-coding genes are perhaps the best understood genomic features, and given that a variant falls somewhere in an annotated gene structure, there are a number of predictions that can be made about its possible effect on gene function, such as whether it is predicted to change the amino acid sequence of the encoded protein, introduce premature stop codons or affect mRNA splicing. There are several computational tools that are designed to make these predictions that work mainly by first identifying annotated genes overlapping the variants and then applying various biologically informed rules based on both the variant location and allele sequences.

The Ensembl VEP uses a set of standardised consequence terms defined in the sequence ontology (SO) (Eilbeck et al. 2005) to describe the predicted effect of a genetic variant. The use of a standardised term set is important as it allows comparison between the results of different annotation systems, and the ontology structure supports biologically informed grouping and querying of annotation results. The VEP also provides a wide range of ancillary annotation such as cDNA and protein relative coordinates, predicted amino acid substitutions (AASs) and SIFT and PolyPhen predictions for missense variants (discussed below). Several other similar tools such as ANNOVAR and VAT (Habegger et al. 2012) work in a similar way but have different performance characteristics and vary in the amount of ancillary information available.

Variants that are predicted to have the most severe effects on coding genes include those that introduce premature stop codons, disrupt essential mRNA splicing signals and indels that change the translational reading frame. These are collectively termed “loss of function” (LoF) variants and are typically expected to be highly deleterious as they have been implicated in a number of severe diseases (MacArthur et al. 2012). Stop codons introduced early in the transcript mean that the mRNA is likely to undergo a cell surveillance process known as “nonsense-mediated decay” (NMD) (Isken and Maquat 2007) where the aberrant mRNA is degraded to avoid the production of deleterious protein isoforms and so may effectively knock-down the affected transcript. However, stop codons towards the end of the transcript may escape this process and only truncate a few amino acid residues and therefore have minimal effect on protein function, so not all premature stop variants should be considered functionally equivalent.

Frameshifting variants may lead to an entirely different translated sequence and substantial elongation or truncation of the protein product. As with premature stop

codons, the position of the variant in the coding sequence will clearly affect the severity of the variant. Hu and Ng (2012) present a new tool that aims to identify frameshift variants that are likely to be truly deleterious and find that variants that affect fewer and less conserved residues are more likely to be tolerated. Hu and Ng (2012) also find that proximal frameshift variants are frequently compensatory in that a nearby downstream variant restores the reading frame disrupted by an upstream variant, highlighting the importance of considering the haplotype background of a variant.

Variants that disrupt the essential two nucleotide donor and acceptor splice sites at either end of introns are also typically expected to severely disrupt the protein product. While these essential positions are indeed highly conserved, there is also substantial sequence conservation in the flanking nucleotides and in the branch site towards the 3′ end of the intron, so variants in these regions may also affect accurate splicing (indeed, this is one way in which "synonymous" variants in coding sequence might still have functional effects). Desmet et al. (2009) introduce a tool called the Human Splicing Finder which uses position weight matrices to predict the effect of different alleles on splicing motifs in all these relevant regions.

It is important to note that despite the expected severity of loss of function variants, there are still a substantial number of common LoF variants in human populations, and each individual is predicted to carry up to 20 such variants in a homozygous state (MacArthur et al. 2012). This observation implies that we should be cautious about the interpretation of LoF variants without further phenotypic evidence. MacArthur et al. (2012) use their extensive survey of LoF variants found in the 1000 Genomes Project data to develop a classifier that can identify genes that are likely to be tolerant of LoF variants based on conservation and protein network information, and so this approach may be used to filter LoF variants to identify those more likely to have some phenotypic effect.

Other forms of coding variant that have been the subject of substantial research are missense variants predicted to result in a single AAS; these are an interesting class of variant as it appears that some AASs do not have any noticeable effect on protein function and the underlying variants are common in human populations, while others have been implicated in a wide range of diseases—around half of the mutations implicated in human disease from the Human Gene Mutation Database (HGMD) are classified as missense (Stenson et al. 2009). Several computational techniques have been developed to try to discriminate damaging AASs from apparently benign variants. These approaches can be divided into two main classes: those that make predictions based on some biologically informed assumptions about properties of important residues and those that are trained by machine learning methods to discriminate between benign and damaging substitutions. A widely used example of the first class is an algorithm called SIFT (Ng and Henikoff 2001) which makes predictions based entirely on a protein multiple sequence alignment (MSA) by looking for evidence that a substitution at a specific residue might be tolerated because, for example, the mutant residue (or one with similar physico-chemical properties) is found at that position in a related protein from another species, or conversely if a substitution is likely to be damaging because the affected residue is highly conserved. A popular example of the second class of approaches is PolyPhen-2

(Adzhubei et al. 2010) which uses a set of missense variants annotated in the UniProt database (UniProt Consortium 2011) as involved in human disease and trains a naïve Bayes classifier to discriminate between these damaging variants and a control set of common, polymorphic variants. PolyPhen uses a set of 12 predictive features for each variant, including a similar conservation metric from an MSA as used by SIFT, three-dimensional structural data, whether the residue is in a transmembrane region or a protein domain *inter alia*. There are also a number of other tools that take similar approaches but use different sets of annotations. Thusberg et al. (2011) provide a recent review and performance comparison of several AAS prediction tools, and Liu et al. (2011) present a database called dbNSFP which contains precomputed predictions from four tools for all possible AASs in the human genome.

Given the wide variety of these AAS effect prediction tools, a few methods have recently been proposed that combine predictions from a number of different tools to try to improve performance over any single technique. One of the first such methods is known as Condel (González-Pérez and López-Bigas 2011) and integrates scores from five different predictors using a weighted average which the authors show gives a substantial improvement in sensitivity and specificity on some test sets. CAROL (Lopes et al. 2012) integrates predictions from SIFT and PolyPhen using a weighted Z-method, and the authors find that this method can outperform Condel on their test set. There are plug-in modules available for the Ensembl VEP to compute both Condel and CAROL scores for missense variants.

Proteins are typically composed of one or more functional domains, and when considering the effect of any coding variant, it is also useful to check if it might disrupt any important protein domains. There are a number of databases of well-characterised protein domains, such as Pfam (Punta et al. 2011) and InterPro (Hunter et al. 2012), and Ensembl (among other resources) provides a mapping of these domains to gene annotations.

Variation in other gene regions, such as introns and the 5′ and 3′ untranslated regions (UTRs), is typically currently annotated by tools such as the Ensembl VEP and ANNOVAR simply as an overlap. However, these regions are known to contain important signals for gene regulation and may also affect mRNA structural stability. Regulatory features in the UTRs include miRNA target sites found in the 3′ UTRs of many genes. These short sequences are bound by specific miRNAs which typically serve to suppress translation of the mRNA and act as a form of post transcriptional gene regulation. The miRanda algorithm for miRNA target prediction (John et al. 2004) can be used to identify variants that disrupt likely target sites and may also be applied to identify variants that introduce novel target sites. As well as important sequence signals for mRNA splicing, intronic regions may also contain many of the regulatory elements discussed later, such as transcription factor binding sites and enhancers.

An important consideration when interpreting all forms of genetic variants is that many human genes are subject to alternative splicing and may give rise to a number of possible transcripts, frequently depending on tissue or developmental stage. A single variant may therefore be predicted to have a number of different effects depending on which transcripts it falls in—an apparently highly deleterious premature stop codon may have little consequence if it is found in an exon that is

rarely included in any transcript. Rich and detailed annotation of alternatively spliced transcripts is therefore very important for accurate variant interpretation, and the GENCODE gene set (Harrow et al. 2012) represents perhaps the most detailed set of manually annotated transcript models available for human.

Even if a variant is predicted to affect an important transcript, it appears that even severely deleterious genetic variants may be tolerated as long as they are in a heterozygous state and so only disrupt one copy of the gene, although it appears that for some genes (termed haploinsufficient), a single functional copy is not adequate to maintain function (Huang et al. 2010). Huang et al. develop a predictive model of genes that are likely to be haploinsufficient based on a number of gene-level annotations and which can be used to further prioritise variants and highlight the importance of considering variant annotations at the organismal level.

Variants in Non-coding Genes

There is increasing interest in transcribed regions of the genome that do not give rise to protein-coding mRNAs, and a number of different classes of non-coding RNA genes have now been identified and are extensively annotated in the GENCODE resource. There has been less work on interpreting the possible effects of variants in non-coding genes, but some of the approaches described above, such as annotation of variants affecting splicing, may also be applied to these.

The function of many RNA genes depends on the secondary structures formed after the RNA has been transcribed from genomic sequence. Intra-strand base pairing is an important factor in determining this structure, and sequence variants that disrupt base complementarity may thus affect the function of RNA genes. The RNAsnp server (Sabarinathan et al. 2013) uses RNA structure prediction algorithms from the Vienna package (Hofacker 2003) to predict the possible effect of variants on RNA secondary structure.

Some specific classes of RNA genes have other well-characterised functional sequence regions. As discussed above, miRNAs serve an important role in gene regulation, and they do so by binding specific sequences in the UTRs according to base pair interactions. Sequence variants in the binding regions of mature miRNA transcripts may therefore have potentially complex downstream effects on regulatory networks.

Intergenic and Regulatory Variants

Genetic regions remain the most well-characterised regions of the genome, but recent large-scale efforts such as the ENCODE and the NIH Roadmap Epigenomics projects have made available substantial amounts of information about biochemical activity in the ~98 % of the genome that does not encode protein. These data are varied in format and range from specific annotations identifying regions of the

genome bound by transcription factors (TFs) to broad epigenetic marks such as histone modifications and long-range chromatin interactions. Given that the majority of trait-associated variants, 88 % according to a recent survey (Hindorff et al. 2009), do not map to protein-coding loci, the availability of these data provides a promising opportunity to interpret the large numbers of non-genetic variants. It is not, however, currently clear to what extent genetic variation in many of the regions identified in these projects might have phenotypic effects.

Perhaps the most readily interpretable regulatory annotations are TF binding sites. Many TFs bind specific sequence motifs in the genome, and so variants that result in changes in these motifs, particularly at high-information content positions within the motif, might have a direct effect on the binding affinity of the relevant proteins. However, Maurano et al. (2012) find that variants at high-information content, conserved residues of the CTCF TF motifs aligned under regions with experimental evidence of CTCF binding, had no effect on binding intensity, implying there is substantial contextual buffering of variants in TF motifs, and it appears our understanding of the importance of specific sequence variants in these regions is still limited.

As with transcript splicing signals, TF motifs are typically represented as position weight matrices, and so the effect of a variant allele on an aligned motif can be calculated straightforwardly as the difference in alignment score between the two alleles. However, TF motifs are typically short—on the order of 10–20 nucleotides in length—and are found in numerous locations throughout the genome, and so most instances of motifs are unlikely to be functionally important (Pique-Regi et al. 2011). It is therefore important to consider further contextual evidence, such as protein–DNA interaction data for the TF of interest in order to increase prediction accuracy. ChIP-seq data for over 100 TFs in dozens of cell lines and tissues is available from ENCODE and Roadmap Epigenomics projects. The JASPAR database provides the largest open access database of TF motifs, and software such as MOODS (Korhonen et al. 2009) and the MEME suite (Bailey et al. 2009) can be used to align these motifs to sequence of interest and to check the effect of sequence variants. The Ensembl VEP identifies variants that overlap TF motifs lying in matched ChIP-seq peaks and identifies if the variant allele increases or decreases the match to the motif consensus sequence and if the variant lies in a high-information position within the motif.

Active regulatory regions are often recognisable by an accessible chromatin environment, and so assays which identify regions of open chromatin, such as DNase1 hypersensitivity and FAIRE (formaldehyde-assisted identification of regulatory elements), can help identify regulatory elements. DNase1 footprinting (Neph et al. 2012b) can identify specific genomic regions that are likely bound by proteins even when the specific factor cannot be identified and so provide a more specific prediction of a functionally important region. Data from both assays are again available in a wide range of tissues and cell lines. The potential role of variants in establishing accessible chromatin is still not well understood, but Degner et al. (2012) find thousands of variants with significant association with differential chromatin accessibility and argue that variants in these regions may make an important contribution to phenotypic variation.

Other available data include epigenetic marks such as DNA methylation and various histone modifications that mark actively transcribed or repressed genomic

regions and which are associated with regulatory elements such as enhancers and promoters. Two recent software packages, ChromHMM (Ernst and Kellis 2012) and Segway (Hoffman et al. 2012), integrate open chromatin and histone modification data to segment the entire genome into distinct functional regions. They find that these methods identify biologically important regions such as transcription start sites and enhancers. Annotations from these tools may be used to identify the likely functional context of non-coding variants, though we have relatively little understanding of the effect of sequence variation on the elements discovered, and because these tools do not take the sequence into account, it is not possible to compare different predictions for different alleles.

Data from the various techniques discussed here are typically made available in BED (or similar) format (see the Appendix for a description of this file format), and so variants can be annotated as overlapping or lying near these elements as described earlier. There are also Web resources available to identify occupied annotations given variant identifiers or coordinates. RegulomeDB (Boyle et al. 2012) finds overlaps with a wide range of data from the ENCODE project and TF motif alignments and then assigns a rule-based score based on the consistency and specificity of available annotations. HaploReg (Ward and Kellis 2012b) similarly finds overlaps with non-coding annotations but also provides information about linked variants and their associated annotations.

Conservation and Constraint

Genomic regions conserved by natural selection over evolutionary time are likely to be functionally important. By comparing the human sequence to that of other primate and mammalian genomes, we can identify regions and even specific nucleotides that appear to be under constraint. Conservation metrics derived from these sequence alignments provide a powerful means to identify potentially functional sequence features even in the absence of further evidence and can be used to identify and prioritise potentially important variant loci, even within annotation categories. Indeed, several of the quantitative approaches we discussed above make extensive use of conservation information, either at the DNA or protein sequence levels, to derive their scores.

There are several methods that can provide nucleotide resolution conservation scores (important for annotating SNVs), including GERP (Davydov et al. 2010) and phyloP (Siepel et al. 2006), which are based on different algorithmic approaches, but which both use multiple sequence alignments to identify genomic regions with less variation than would be expected under some background model. Nucleotide level conservation scores can also be used to identify runs of especially constrained sequence, which may correspond to functional elements, and these regions can also be used as an informative regional annotation.

Conservation has proven to be an important signal in coding regions, but many regulatory elements appear to have a much faster evolutionary rate, and there is frequently little detectable evolutionary conservation, for example, Schmidt et al. (2010)

find that most binding events for the two transcription factors they study are species specific even among vertebrates. The recent availability of allele frequency data across the genome from projects such as the 1000 Genomes Project (Consortium, The 1000 Genomes Project 2012) offers an alternative approach to estimating constraint on sequence features at potentially shorter timescales than possible using interspecies comparison. Ward and Kellis (2012a) use several metrics of sequence diversity such as variant density, heterozygosity and derived allele frequency computed from the 1000 Genomes Project data to demonstrate that a wide range of non-coding elements demonstrate detectable levels of constraint in human populations. These measures can potentially be used to prioritise variants according to the constraint of overlapping annotations.

Integrative Approaches

Recently, two complementary techniques have been released that integrate a wide variety of the classes of data discussed above with the aim of prioritising candidate functional variants. GWAVA (Ritchie et al. 2014) is a method aimed to identify likely functional regulatory variants and consists of a classifier trained to discriminate between annotated regulatory variants involved in human disease from the HGMD from several different sets of control variants from the 1000 Genomes Project. Features used to differentiate between these classes of variants include genetic context, regulatory annotations, conservation and measures of variation in human populations. The authors demonstrate that the method can identify likely functional variants in a number of contexts relevant to human genetics studies. CADD (Kircher et al. 2014) is also an integrative approach that includes several of the same annotations used in GWAVA, but is also applicable to variants in coding regions as it incorporates transcript-level annotations from the Ensembl VEP and predictions from SIFT and PolyPhen (described earlier). Instead of training on known disease-implicated variants, CADD is trained to discriminate between variants that have become fixed in the human lineage, which presumably represent tolerable variation, from simulated variants unobserved in human populations. This approach is appealing as it can assign a single score to variants falling in any class of genomic element and supports a systematic approach to ranking and prioritising variants across the genome.

Overlap with Known Variants and Associated Loci

While the majority of variants discovered so far in the human genome have not been characterised, an obvious aid to the interpretation of some candidate variant is to check for co-located or nearby variants with some established phenotypic association. These data may take a range of forms, from statistical association with a complex

phenotype such as a GWAS signal to empirical evidence that the variant results in increased expression of some particular gene. Locus-level phenotypic annotation, such as the effect of a gene knockout in a model organism, can also provide useful insight into the possible functional role of a genetic or regulatory variant.

There are a number of useful databases that can be consulted to find known phenotype associations; these can typically be queried either by the variant locus or phenotype of interest. The HGMD (Stenson et al. 2009) aims to collect variants that are "responsible for human inherited disease" and contains thousands of variants curated from the literature that have been implicated in a wide range of human diseases, though with a bias towards monogenic disorders. The Online Mendelian Inheritance in Man (OMIM) resource also includes detailed characterisation of human genes and associated phenotypes and includes some related genetic variants. The NHGRI GWAS catalogue (Hindorff et al. 2009) collects information from GWAS studies and identifies both specific variants and nearby loci associated with the relevant phenotypes.

Even in the absence of any phenotypic data, it is useful to establish if a candidate variant is novel or has been discovered before to find allele frequency information in different populations. A rare variant in one population may be common elsewhere in the world, and as discussed above, allele frequency can be informative about functional constraint. Data from large variant discovery studies such as the HapMap, 1000 Genomes and NHLBI Exome Sequencing Projects can be used to find allele frequencies for several populations around the world. These data are also collated centrally in the Ensembl and dbSNP databases, among other resources.

Summary

Next-generation association studies using sequencing technologies are already exploring the phenotypic consequences of novel variants at lower allele frequencies than previously feasible, and we expect to find variants with direct effects on phenotypic variation. The various resources we have reviewed here can of course be used after an association analysis has been performed to identify candidate functional variants among those linked to the association signals and to inform hypotheses for experimental validation. However, by identifying variants a priori more likely to play a functional role in the trait of interest, annotations may also be used to increase power to discover loci in the first place. This might be especially fruitful for rare variant studies where the sample sizes needed to reliably detect associations using single locus tests are still prohibitive. In a recent study, Schork et al. (2013) find that trait-associated variants are substantially enriched in various functional categories and that annotations can help identify associations that are more likely to replicate in independent samples. We anticipate that careful incorporation of annotation resources into future association studies will yield substantial insights into the contribution of rare variants to human phenotypes.

Appendix

Relevant variant and annotation file formats:

- GFF (General Feature Format): A line-oriented, tab-delimited text file format for describing the location of genomic features. GFF was originally designed to represent gene models but is now used for a wide range of genomic features. The format requires the following eight columns on each line: sequence name, feature source, feature name, start coordinate, end coordinate, score, strand and frame. The ninth column can contain any number of attributes represented as tag-value pairs separated by semicolons.
  - http://www.sequenceontology.org/gff3.shtml
- BED (Browser Extensible Data): BED is also a general format for describing genomic features and again is a line-oriented text file which uses whitespace to delimit data columns. Only three columns are required for a valid BED file: the chromosome (or scaffold) name, the start coordinate and the end coordinate. There are nine further optional fields to include further information such as the name of the feature, associated scores and various display configurations that define how the data is represented in a genome browser. Large BED files can be converted to an efficient binary format known as bigBed.
  - https://genome.ucsc.edu/FAQ/FAQformat.html
- GTF (General Transfer Format): Originally a version of GFF specialised for representing gene models, GTF is now identical to GFF version 2.
- VCF (Variant Call Format): A text file format designed to represent sequence variants (SNVs, indels and structural variants) called against a reference sequence, with a line representing each individual variant. Required tab-delimited columns define the position and alleles of the variant, and further columns can include genotypes, quality scores and QC filters. VCF also supports the inclusion of arbitrary metadata, such as functional annotations for variants, in the INFO column (often identified with a "CSQ" tag).
  - http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41
- GVF (Genome Variation Format): A version of GFF (version 3) specialised for representing genomic variants. The same columns as required for GFF are also required, but there are also a number of required attributes in the ninth column to include variant identifiers and allele sequences, etc. Optional attributes are also available which can represent functional annotations such as genetic consequences.
  - http://www.sequenceontology.org/resources/gvf.html
- SAM (Sequence Alignment/Map Format): A tab-delimited text format for representing sequence reads aligned against some reference sequence (typically a reference genome assembly). Each line represents the alignment of a single read

and has 11 mandatory fields that include details of the alignment sequence, position, quality and a compact representation of the alignment itself in CIGAR format. There is also an efficient binary version of SAM known as BAM. The SAMtools package can be used to convert between SAM and BAM formats.

  – http://samtools.sourceforge.net/

- WIG (Wiggle Track Format): WIG format is used to represent quantitative data across a reference sequence such as conservation scores, GC percentage, etc. It is again a line-oriented format with the value corresponding to each reference position represented on a separate line. Data can be represented with either fixed or variable steps between each data point. Large WIG files can be converted to an efficient indexed binary format called bigWig.

  – https://genome.ucsc.edu/FAQ/FAQformat.html

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P et al (2010) A method and server for predicting damaging missense mutations. Nature 7(4):248–249. doi:10.1038/nmeth0410-248

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L et al (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37(Web Server Issue):W202–W208. doi:10.1093/nar/gkp335

Betel D, Wilson M, Gabow A, Marks DS, Sander C (2007) The microRNA.org resource: targets and expression. Nucleic Acids Res 36(Database):D149–D153. doi:10.1093/nar/gkm995

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M et al (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22(9):1790–1797. doi:10.1101/gr.137323.112

Consortium, The 1000 Genomes Project (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65. doi:10.1038/nature11632

Consortium, The ENCODE Project (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74. doi:10.1038/nature11247

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6(12), e1001025. doi:10.1371/journal.pcbi.1001025

Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK et al (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482(7385):390–394. doi:10.1038/nature10808

Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37(9), e67. doi:10.1093/nar/gkp215

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M (2005) The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6(5):R44. doi:10.1186/gb-2005-6-5-r44

Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. Nature Publishing Group 9(3):215–216. doi:10.1038/nmeth.1906

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S et al (2012) Ensembl 2013. Nucleic Acids Res. doi:10.1093/nar/gks1236

González-Pérez A, López-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88(4):440–449. doi:10.1016/j.ajhg.2011.03.004

Habegger L, Balasubramanian S, Chen DZ, Khurana E, Sboner A, Harmanci A et al (2012) VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics 28(17):2267–2269. doi:10.1093/bioinformatics/bts368

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F et al (2012) GENCODE: the reference human genome annotation for the ENCODE Project. Genome Res 22(9):1760–1774. doi:10.1101/gr.135350.111

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362–9367. doi:10.1073/pnas.0903103106

Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31(13): 3429–3431

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature 9(5):473–476. doi:10.1038/nmeth.1937

Hu J, Ng PC (2012) Predicting the effects of frameshifting indels. Genome Biol 13(2):R9. doi:10.1186/gb-2012-13-2-r9

Huang N, Lee I, Marcotte EM, Hurles ME (2010) Characterising and predicting haploinsufficiency in the human genome. PLoS Genet 6(10), e1001154. doi:10.1371/journal.pgen.1001154

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A et al (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 2012(Database Issue):D306–D312

Isken O, Maquat LE (2007) Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. Genes Dev 21(15):1833–1856. doi:10.1101/gad.1566807

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human microRNA targets. PLoS Biol 2(11), e363. doi:10.1371/journal.pbio.0020363

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310–315. doi:10.1038/ng.2892

Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E (2009) MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics 25(23):3181–3182. doi:10.1093/bioinformatics/btp554

Liu XX, Jian XX, Boerwinkle EE (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32(8):894–899. doi:10.1002/humu.21517

Lopes MC, Joyce C, Ritchie GRS, John SL, Cunningham F, Asimit J, Zeggini E (2012) A combined functional annotation score for non-synonymous variants. Hum Hered 73(1):47–51. doi:10.1159/000334984

MacArthur DG, Balasubramanian S, Frankish A et al (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335(6070):823–828. doi:10.1126/science.1215040

Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA (2012) Widespread site-dependent buffering of human regulatory polymorphism. PLoS Genet 8(3), e1002599

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. Bioinformatics 26(16):2069–2070. doi:10.1093/bioinformatics/btq330

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M et al (2013) The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res 41(Database Issue):D64–D69. doi:10.1093/nar/gks1048

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK et al (2012a) BEDOPS: high-performance genomic feature operations. Bioinformatics 28(14):1919–1920. doi:10.1093/bioinformatics/bts277

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B et al (2012b) An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489(7414):83–90. doi:10.1038/nature11212

Ng P, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11(5): 863–874. doi:10.1101/gr.176601

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res 21(3):447–455. doi:10.1101/gr.112623.110

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C et al (2011) The Pfam protein families database. Nucleic Acids Res 40(D1):D290–D301. doi:10.1093/nar/gkr1065

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842. doi:10.1093/bioinformatics/btq033

Ritchie GRS, Dunham I, Zeggini E, Flicek P (2014) Functional annotation of noncoding sequence variants. Nature Methods 11(3):294–296. doi:10.1038/nmeth.2832

Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J (2013) The RNAsnp web server: predicting SNP effects on local RNA secondary structure. Nucleic Acids Res. doi:10.1093/nar/gkt291

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A et al (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328(5981):1036–1040. doi:10.1126/science.1186176

Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF et al (2013) All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet 9(4), e1003449. doi:10.1371/journal.pgen.1003449

Siepel A, Pollard KS, Haussler D (2006) New methods for detecting lineage-specific selection. Presented at the Proceedings of the 10th International Conference on Research in Computational Molecular Biology, RECOMB 2006: April 2–5, 2006, Venice Lido, Italy, pp 190–205

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN (2009) The Human Gene Mutation Database: 2008 update. Genome Med 1(1):13. doi:10.1186/gm13

Thusberg J, Olatubosun A, Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32(4):358–368. doi:10.1002/humu.21445

UniProt Consortium (2011) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40(Database Issue):D71–D75. doi:10.1093/nar/gkr981

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164. doi:10.1093/nar/gkq603

Ward LD, Kellis M (2012a) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science. doi:10.1126/science.1225057

Ward LD, Kellis M (2012b) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 40(Database Issue):D930–D934. doi:10.1093/nar/gkr917