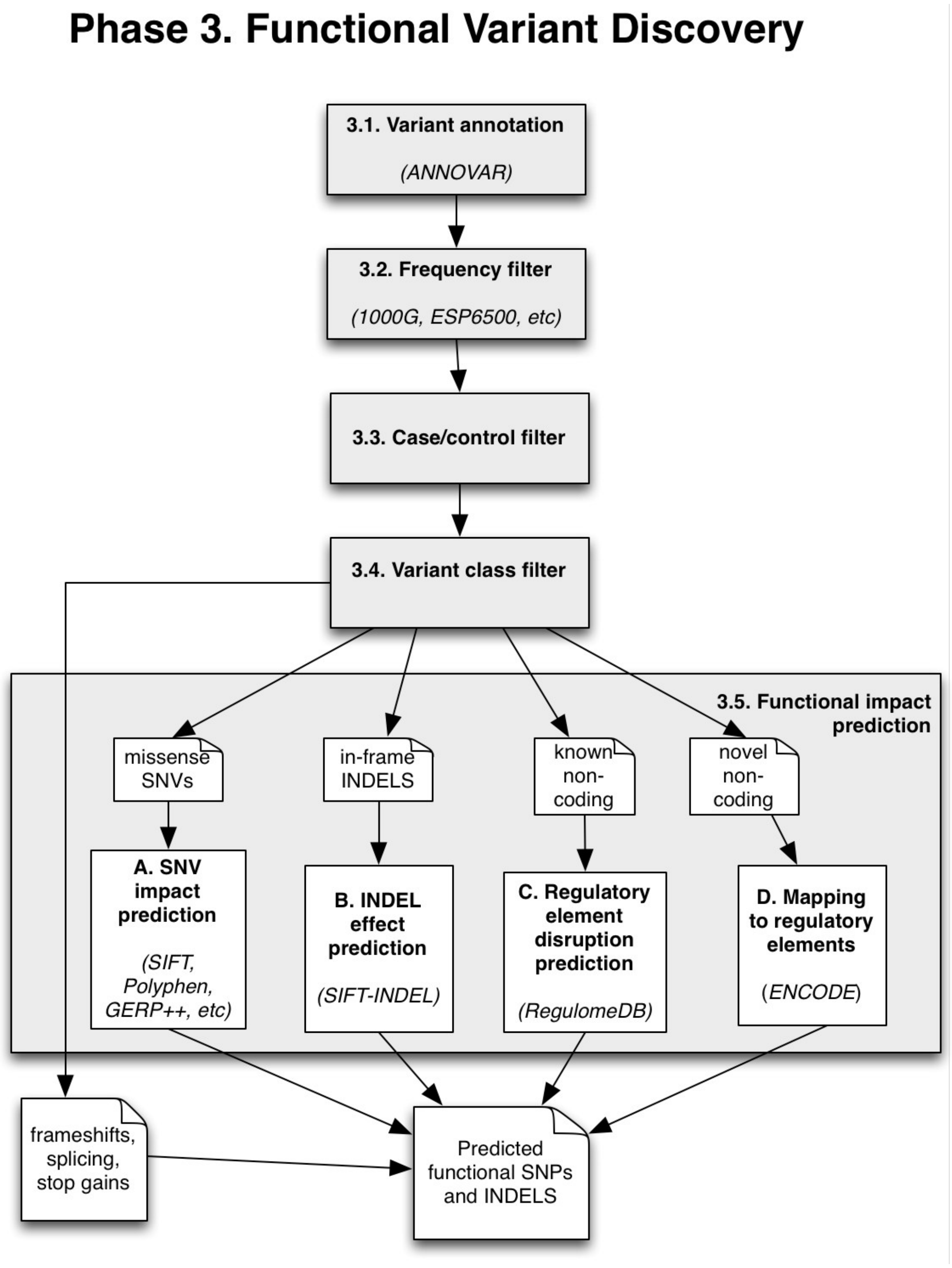
Phase 3. Functional Variant Discovery


Phase 3. Functional Variant Discovery
3.1. Variant annotation
(ANNOVAR)
3.2. Frequency filter
(1000G, ESP6500, etc)
3.3. Case/control filter
3.4. Variant class filter
3.5. Functional impact prediction
missense SNVs
in-frame INDELS
known non-coding
novel non-coding
A. SNV impact prediction
(SIFT, Polyphen, GERP++, etc)
B. INDEL effect prediction
(SIFT-INDEL)
C. Regulatory element disruption prediction
(RegulomeDB)
D. Mapping to regulatory elements
(ENCODE)
frameshifts, splicing, stop gains
Predicted functional SNPs and INDELS

Introduction

There is no single 'recipe' for identifying functional variants. Every filtering strategy inevitably starts with multiple often-logical assumptions about the properties of causal variants that may also discard candidates too early. Novelty is one example, where the variant is not expected to be present in public repositories, even though apparently healthy individuals carrying it may have been sequenced in unrelated studies. High predicted functional impact is another, where a variant is expected to simultaneously occur at an evolutionarily conserved genomic location and also be scored as deleterious by multiple predictive algorithms may even be too stringent a rule in a Mendelian trait. In a similar vein, in a population genetics study, a novel amino acid class-changing variant that is not predicted to have a functional effect by the standard tools may have a molecular effect. An inheritance model may itself be incorrect, e.g. a condition may be autosomal recessive rather than X-linked as hypothesized, or digenic rather than recessive, etc.

It is therefore important not to conclude too early that the variant(s) of interest “are not in the exome” and to exhaust all alternate plausible rules and scenarios before that conclusion is reached. Using *disease variant discovery* as a context, we aim to present here a range of guidelines that can be used to develop an appropriate variant prioritisation strategy.

This phase will be presented as 5 sub-phases:

1. Variant annotation - using ANNOVAR (Wang et al, 2010) as an example
2. Frequency
3. Case-control filtering
4. Variant class filtering
5. Functional impact prediction

3.1. Variant annotation

ANNOVAR adds numerous functional annotations to a set of variants. It is available as a standalone version (http://annovar.openbioinformatics.org/) and as a webserver (http://wannovar.usc.edu). The webserver is intuitive and produces reports in several easy to parse formats, but may only be appropriate for non-sensitive data. ANNOVAR annotates SNVs and indels with: allele frequencies in public databases, gene/transcript effects, site conservation scores, predicted functional effects on the protein. It also simplifies downstream filtering by assigning simple annotations such as zygosity, gene symbols, whether the variant is synonymous or non-synonymous, etc.

Both the standalone and web-versions understand VCF formats produced in Phase II and variants can be submitted in bulk. As outputs are available as tab-delimited text, most of the filters described below can reasonably easily be implemented in a spreadsheet software package such as Excel.

3.2. Frequency filtering

One of the primary criteria for predicting whether a variant is likely to have a functional effect on the protein is rarity (Nelson et al, 2012). For example, a rare nonsense SNP can be expected to have a larger functional effect than a frequently occurring one.

ANNOVAR provides variant frequencies from the 1000 Genomes Project, both as an average across all the populations studied and for individual groups. Similarly, variant frequencies derived from approximately 6500 exomes from the NHLBI Exome Sequencing Project (http://esp.gs.washington.edu) are also provided. The frequency cut-off for rarity has to be decided on a case-by-case basis, but a reasonable starting point would be <1% (‘0.01’ in the ANNOVAR output). Additionally, the absence of a dbSNP identifier could indicate potentially novel variants, but we advise checking final candidates against the latest version of dbSNP.

3.3. Case-control filtering

Working with an inheritance model assists with the filtering out variants that do not fit the expected profile:

1. *Autosomal dominant* – candidates are heterozygous variants seen in all affected individuals in a family and unaffected individuals are homozygous for the reference allele
2. *Autosomal recessive* – homozygous variant seen in all affected individuals in a family where unaffected parents are heterozygous for the allele and other unaffected family members individuals are either heterozygous or homozygous for the reference allele
3. *Spontaneous* – allele in the affected individual not present in either parent (private mutation)
4. *Somatic* – e.g. variants seen in tumour exomes but not in normal tissue from the same individual

The ability to perform this step is, however, most often dependent on the availability of exomes from unaffected family members or from matching normal tissue from the same individual under study. An important consideration to bear in mind is whether the working family pedigree is accurate.

NOTE: *for polygenic disorders or population genetic studies, these criteria cannot be applied. Rather, variants should be assessed for statistical overrepresentation in cases versus controls or between population groups and then the resulting candidates further filtered as described below. Lists of genes predicted to bear a functional variant(s) can then be subjected to further biological contextualisation using techniques such as functional overrepresentation analysis, pathway association, etc.*

3.4. Variant class filter

Further filtering on the *ExonicFunc* column for variants labelled as “nonsynonymous”, “stopgain”, “frameshift” and “splicing” generates a subset of variants with the potential to functionally affect the protein. The latter three classes should automatically be selected as having *probable functional impact*. Also, by this step the nonsynonymous variants are automatically ‘interesting’ based on the fact they are rare and correctly segregate with the target group of interest. This is particularly relevant for population genetic studies where identifying deleterious variants is not the only focus, but positive selection, for example, is. Similarly, in the study of multigenic diseases, finding multiple variants of modest functional impact may be the focus.

3.5. Functional impact prediction

Each variant category has to be assessed with a specific set of tools to predict their functional impact. For this guide, we will assume that synonymous variants have no functional impact. We will also assume that frameshift SNVs/indels, splice site variants and nonsense variants have a functional effect, particularly since at the stage we would be dealing with rare or novel variants.

A. Single nucleotide variants (SNVs)
ANNOVAR annotates missense variants with multiple functional prediction scores, which when used in conjunction with site conservation scores can be used to identify candidates that likely have deleterious effects. Similarly, indels, nonsense and and splicing variants that occur at evolutionarily conserved sites are most likely to affect protein function. Various permutations of these two annotation classes can be produced to identify variants that have the ‘highest’ impact,

Example filters, which can be defined using the appropriate prediction shorthand from ANNOVAR (‘T’ or ‘B’ for tolerated/benign and ‘D’ or ‘P’ for deleterious/pathogenic) and conservation score cut-offs (tool-specific):

- Variant with multiple ‘deleterious’ annotations AND at highly a conserved site
- Variant with multiple ‘deleterious’ annotations
- Variant occurring at a site scored as conserved by all tools AND causing an amino acid class change, yet is not annotated as deleterious

NOTES:

- New variant effect predictors are frequently added to ANNOVAR. While the majority of publications still use a combination of SIFT (Ng and Henikoff, 2003) and PolyPhen (Adzhubei, 2013), we recommend becoming familiar with the other tools and applying them in the filtering pipeline to prevent discarding of candidates due to false negatives.
- There are no standard rules for these filters and have to be defined based on the hypothesized molecular mechanism of the disease. For example, in Mendelian disorders, it is reasonable to assume that the single causal variant will have an extreme functional effect and a very stringent filter is appropriate as a first pass.

- A relaxing of the rules may, however, be appropriate since all functional prediction tools produce false negatives and for example, novelty/rarity + the expected inheritance pattern + site conservation are already substantial evidence to implicate a variant as being causal.
- If ranking of candidate SNPs is desired, multiplying all the functional prediction *scores* (also reported by ANNOVAR) and all the conservation scores and sorting by that result will cause those cases where the majority of tools are in agreement to 'bubble to the top'.

B. In-frame insertions and deletions
Although it is often expected that indels are inevitably deleterious, it is unlikely to be the case for in-frame ones. That said, rare in-frame indels are more likely to be deleterious than common variants. SIFT-indel (Hu and Ng, 2013), can be used to predict the functional impact of non-frameshift indels (http://sift.bii.a-star.edu.sg/www/SIFT_indels2.html).

C. Known non-coding variants
Variants that have been assigned dbSNP identifiers are annotated by the Regulome database project as being located in and having potential disruptive effect on regulatory elements and intergenic regions in the human genome. The known and predicted regulatory DNA elements include regions of DNAase hypersensitivity, transcription factor binding sites and biochemically characterised promoter regions. Rs-IDs can be submitted to a webserver at http://www.regulomedb.org, which will annotate them based on likely functional impact.

D. Novel noncoding variants
Coordinates of these variants can be intersected with coordinates of likely regulatory and noncoding functional genomic elements to predict whether they may affect protein binding. Again, as they are novel and therefore likely rare, their presence at a functional element suggests that they may have a biochemical or phenotypic effect.

Summary

Overall, the described filters will act as a funnel. It is important to note that not all steps may be relevant, that each filter should be customised to the study requirements, and that the 'funnel' will not always comprise of all the possible filters. Variants that fulfil multiple criteria are the mostly likely to have a genotype-to-phenotype association.

References

Hu J, Ng PC. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. PLoS One. 2013 Oct 23;8(10):e77940. doi:

10.1371/journal.pone.0077940. eCollection 2013. PubMed PMID: 24194902; PubMed Central PMCID: PMC3806772.

Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20. doi: 10.1002/0471142905.hg0720s76. PubMed PMID: 23315928.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012 Sep;22(9):1790-7. doi: 10.1101/gr.137323.112. PubMed PMID: 22955989; PubMed Central PMCID: PMC3431494.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, Warren L, Aponte J, Zawistowski M, Liu X, Zhang H, Zhang Y, Li J, Li Y, Li L, Woollard P, Topp S, Hall MD, Nangle K, Wang J, Abecasis G, Cardon LR, Zöllner S, Whittaker JC, Chissoe SL, Novembre J, Mooser V. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012 Jul 6;337(6090):100-4. doi: 10.1126/science.1217876. Epub 2012 May 17. PubMed PMID: 22604722; PubMed Central PMCID: PMC4319976.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM; Broad GO; Seattle GO; NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012 Jul 6;337(6090):64-9. doi: 10.1126/science.1219240. Epub 2012 May 17. PubMed PMID: 22604720; PubMed Central PMCID: PMC3708544.

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

doi: 10.1093/nar/gkq603. Epub 2010 Jul 3. PubMed PMID: 20601685; PubMed Central
PMCID: PMC2938201.

Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812-4. PubMed PMID: 12824425;
PubMed Central PMCID: PMC168916.