KANZA BATOOL HAIDER
Independent Study
Supervisor: Prof. Mehmet Baysan
Date: 16.7.2018

WEEK 3- ANNOTATION TYPES

Gene Based Annotation

Important notes for Gene based annotation:

- We can use RefSeq gene definitions
- UCSC Genes definition or Ensemble Gene definition can be used alternatively
- Gene/transcript annotations and FASTA sequences needs to be downloaded from the above data bases and stored in the local disk

Gene based annotation annotates the location of each variant with respect to genes whether each variant is

Exonic

An exon is any part of a gene that will encode a part of the final mature RNA produced by that gene after introns have been removed by RNA splicing and which are responsible for the protein synthesis

Intronic

(An intron is any nucleotide sequence within a gene that is removed by RNA splicing during maturation of the final RNA product which are NOT responsible for the protein synthesis)

Intergenic

An Intergenic region (IGR) is a stretch of DNA sequences located between genes. Intergenic regions are a subset of noncoding DNA. It is one of the DNA sequences sometimes referred to as junk DNA.

Splicing site

Splicing is the editing of the pre-mRNA transcript into a mature messenger RNA (mRNA). After splicing, introns are removed and exons are joined together (ligated).

5'/3' – UTR

The 5' UTR is upstream from the coding sequence. Within the 5' UTR is a sequence that is recognized by the ribosome which allows the ribosome to bind and initiate translation. The 3' UTR is found immediately following the translation stop codon.

Upstream/downstream of genes

Upstream and downstream both refer to relative positions of genetic code in DNA or RNA. Upstream is toward the 5' end of the RNA molecule and downstream is toward the 3' end. When considering double-stranded DNA, upstream is toward the 5' end of the coding strand for the gene in question and downstream is toward the 3' end.

Amino acid changes that may be caused by the mutation

From gathering above information, gene based annotation is able to:

- Identify the genes that are disrupted (SNV and indels)
- For intergenic variants, we are interested in knowing what are the distances between the variants and the flanking genes.
- For exonic variants we are interested in knowing the amino acid changes.

Region Based Annotation

Region based annotations refer to the annotations of variants based in specific genomic elements other than the genes. For example:

Conserved genomic regions

Conserved genes are the gene that has remained essentially unchanged throughout evolution. Conservation of a gene indicates that it is unique and essential: There is not an extra copy of that gene with which evolution can tinker, and changes in the gene are likely to be lethal.

Predicted transcription factor (tf) binding sites

A TF is a protein that can bind to DNA and regulate gene expression. The region of the gene to which TF binds is called a transcription factor binding site. These sites are a subset of DNA binding sites. Overall, these sites can be defined as short segments of DNA that are specifically bound by one or more proteins with various functions.

Predicted Micro RNA sites

Micro RNA is a small non-coding RNA molecule (containing about 22 nucleotides) found in plants, animals and some viruses that functions in RNA silencing and post-transcriptional regulation of gene expression.

Predicted stable RNA secondary structures

Important Note:

These annotations are especially important for whole-genome sequencing data, as the vast majority of variants will be outside of protein coding regions and their functional effects cannot be assessed by gene based annotations.

Filter Based Annotations

For filter based annotations, mutations has to be compared with those detected in the public database for example 1000 Genomes Project and dbSNP.

Pre-computed SIFT scores can be downloaded for all human non-synonymous mutations, to help annotate human exons by filter based procedure

Filtration can be done for specific variants such as SNPs with >1% frequency in the 1000 Genomes Project or non-synonymous SNPS with SIFT scores >0.005

MUTATION PREDICTION ALGORITHMS

For human genomes, following are the two algorithms:

- SIFT
- PolyPhen

These give scores for all possible non-synonymous mutations.

Public databases are in the format of Genetic Feature Format

Next up Plan:

- Study public Data bases and gff3 format file
- Study scoring algorithms SIFT and PolyPhen