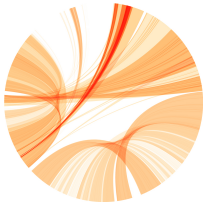
Practical Bioinformatics

Variant Annotation

Part 3

Stefan Wyder
stefan.wyder@uzh.ch
URPP Evolution
www.evolution.uzh.ch

Universität Zürich[UZH]



URPP

Variant Annotation and Prioritization

- What is the effect of your variants (SNPs, insertions, deletions, CNVs, or structural variants)?
- Which variants are the most likely to affect the phenotype?
- Can be used to prioritize variants and to guide experimental validation

Variant Effect Prediction

- which gene/transcript is affected
- location of the variant (e.g. in UTR, CDS, splice site, regulatory region)
- consequence on the protein sequence (e.g. stop gained, frameshift)
- possible impact of AA substitutions
- known previously reported variants

Annotation files

- Variant Effect Prediction depends on annotation and previous knowledge about functional elements
- Preformatted annotation files usually available for model species (Detailedness of Annotation: human >> mouse > fly/yeast/Ecoli ...)
- Gene-based annotation can easily be done for any species where you have i) a genome sequence (even partial) and ii) a gene annotation

Type of Annotation

Gene-based annotation: identify whether SNPs or CNVs cause protein coding changes and the amino acids that are affected.

Region-based annotations: identify variants in specific genomic regions, for example, conserved regions among 44 species, predicted transcription factor binding sites, segmental duplication regions, GWAS hits, database of genomic variants, DNAse I hypersensitivity sites, ENCODE H3K4Me1/H3K4Me3/ H3K27Ac/CTCF sites, ChIP-Seq peaks, RNA-Seq peaks, or many other annotations on genomic intervals.

Filter-based annotation: identify variants that are reported in dbSNP, or identify the subset of common SNPs (MAF>1%) in the 1000 Genome Project, or identify subset of non-synonymous SNPs with SIFT score>0.05, or find intergenic variants with GERP++ score<2, or many other annotations on specific mutations.

Annovar Website

2 main pipelines

ANNOVAR

- http://www.openbioinformatics.org/annovar/
- standalone perl script
- annotation files available for species present in UCSC Genome Browser Annotation Database (http://genome.ucsc.edu/)

ensembl VEP

- http://www.ensembl.org/info/docs/tools/vep/index.html
- available as webtool and standalone perl script
- available for hundreds of species that are present in ensembl (vertebrates, bacteria,insects,plants,fungi)

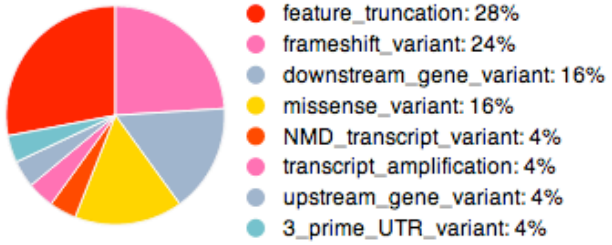
Example ensembl VEP

Variant Effect Predictor results

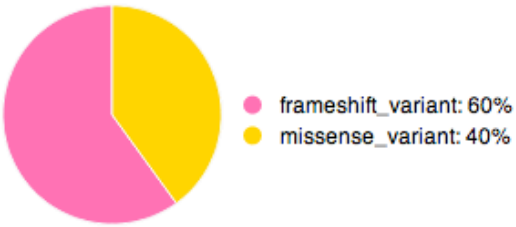
Summary statistics for ticket sRexIRXgOfB7VyM3:

Category	Count
Variants processed	3
Variants remaining after filtering	3
Novel / existing variants	2 (66.7%) / 1 (33.3%)
Overlapped genes	4
Overlapped transcripts	17
Overlapped regulatory features	-

Consequences (all)



Coding consequences



Uploaded variation	Location	Allele	Gene	Feature	Feature type	Consequence	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variation	AA MAF
1_909238_G/C	1:909238	C	ENSG00000187583	ENST00000491024	Transcript	missense_variant	155	155	52	R/P	CGT/CCT	rs3829740	0.219
1_909238_G/C	1:909238	C	ENSG00000187583	ENST00000379407	Transcript	missense_variant	1385	1355	452	R/P	CGT/CCT	rs3829740	0.219
1_909238_G/C	1:909238	C	ENSG00000187583	ENST00000379410	Transcript	missense_variant	1495	1460	487	R/P	CGT/CCT	rs3829740	0.219
1_909238_G/C	1:909238	C	ENSG00000187583	ENST00000379409	Transcript	missense_variant	1646	1616	539	R/P	CGT/CCT	rs3829740	0.219
1_909238_G/C	1:909238	C	ENSG00000187642	ENST00000341290	Transcript	downstream_gene_variant	-	-	-	-	-	rs3829740	0.219
1_909238_G/C	1:909238	C	ENSG00000187642	ENST00000433179	Transcript	downstream_gene_variant	-	-	-	-	-	rs3829740	0.219
1_909238_G/C	1:909238	C	ENSG00000187583	ENST00000480267	Transcript	downstream_gene_variant	-	-	-	-	-	rs3829740	0.219
1_909238_G/C	1:909238	C	ENSG00000187642	ENST00000479361	Transcript	downstream_gene_variant	-	-	-	-	-	rs3829740	0.219
3_361464_A/-	3:361463-361464	-	ENSG00000134121	ENST00000449294	Transcript	frameshift_variant, feature_truncation	345	5	2	-	-	-	-
3_361464_A/-	3:361463-361464	-	ENSG00000134121	ENST00000397491	Transcript	frameshift_variant, feature_truncation	472	5	2	-	-	-	-
3_361464_A/-	3:361463-361464	-	ENSG00000134121	ENST00000421198	Transcript	frameshift_variant, feature_truncation	258	5	2	-	-	-	-
3_361464_A/-	3:361463-361464	-	ENSG00000134121	ENST00000427688	Transcript	frameshift_variant, feature_truncation	380	5	2	-	-	-	-
3_361464_A/-	3:361463-361464	-	ENSG00000134121	ENST00000435603	Transcript	frameshift_variant, feature_truncation	185	5	2	-	-	-	-
3_361464_A/-	3:361463-361464	-	ENSG00000134121	ENST00000453040	Transcript	3_prime_UTR_variant, NMD_transcript_variant, feature_truncation	628	-	-	-	-	-	-
3_361464_A/-	3:361463-361464	-	ENSG00000134121	ENST00000256509	Transcript	frameshift_variant, feature_truncation	647	5	2	-	-	-	-
3_361464_A/-	3:361463-361464	-	ENSG00000134121	ENST00000461289	Transcript	upstream_gene_variant	-	-	-	-	-	-	-
5_121187650_duplication	5:121187650	duplication	ENSG00000181867	ENST00000321339	Transcript	transcript_amplification	1-870	-	-	-	-	-	-

Results depend on the pipeline

- Choice of transcript annotation and software has a large effect on variant annotation (McCarthy, Genome Med 2014)
- For variants considered “loss of function” (LOF: missense, nonsense, nonstop, frameshift, splice site), the concordance between VEP and ANNOVAR was only 44%

Accurate annotation of variants in regulatory and noncoding regions of the genome is still very challenging!

Working with vcf files

VCFtools
http://vcftools.sourceforge.net/

- filter out specific variants
- format conversion
- annotate
- compare vcf files
- intersect
- summarize
- ...

Installation in Ubuntu
sudo apt-get install vcftools

Faster less-documented replacement:
bcftools
http://vcftools.sourceforge.net/htslib.html

Some examples