

Variant Annotation and Functional Prediction

Copyrighted © 2021 Isabelle Schrauwen and Suzanne M. Leal

This exercise touches on several functionalities of the program ANNOVAR to annotate and interpret candidate genetic variants associated with disease, identified through next-generation sequencing methods, imputation or genotyping. When variants are identified to be associated with disease, a common strategy is to perform multiple *in silico* analyses to predict whether they potentially have an impact on gene function.

More information and a detailed guide on installation of ANNOVAR can be found here: <http://annovar.openbioinformatics.org/en/latest/>. ANNOVAR has three main annotation types to help evaluate variants:

[1] **Gene-based annotation:** This annotation annotates variants in respect to their effect on genes (RefSeq genes, UCSC genes, ENSEMBL genes, GENCODE genes, AceView genes) and also outputs the effect of the mutation on the protein in standard HGVS nomenclature (if an effect is predicted).

[2] **Region-based annotation:** With this annotation you can identify variants in specific genomic regions (i.e. conserved regions, 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).

[3] **Filter-based annotation:** Identify variants that are documented in specific frequency databases (dbSNP, Genome Aggregation Consortium, etc) or functional effect prediction databases (PolyPhen, MutationTaster, FATHMM, etc). For example, find intergenic variants with a CADD c-score >20.

In this exercise, we will evaluate *APOC3* variants by annotating a .vcf file (*APOC3.vcf*). Variants in *APOC3* are associated with apoC-III protein levels, triglycerides levels, and coronary heart disease. Previous studies suggested that lifelong deficiency of apoC-III has a cardioprotective effect. When rare variant association tests are performed, variants are often analyzed as a group; and when an association has been found which has been replicated it is not necessarily true that all tested variants are causal. However, for low variant frequencies it is often not possible to test individual variants for an association with a trait. Therefore, bioinformatics tools are often used to predict which variants are likely to be functional and therefore could be involved in trait etiology. For this exercise, six variants in *APOC3* were selected for annotation as an example.

First of all, once you are logged in into dockerhub go to the /work directory where the datafiles for this exercise are located by typing:

```
$ cd work
```

The `table_annovar.pl` in ANNOVAR command accepts VCF files. Type in `table_annovar.pl` to learn about the annotation options (Tip: add Annovar to your PATH to be able to use this command in any directory). More info on VCF processing and left-normalization for indels can be found here:

<http://annovar.openbioinformatics.org/en/latest/articles/VCF/>. Note, ANNOVAR can also accept compressed .vcf.gz files.

`$ table_annovar.pl`

A. Gene-based annotation: Using Ensembl, RefSeq and UCSC Genome Browser

First, we will evaluate the location of these variants in *APOC3*. We will use the Gene-based annotation function, which annotates variants to coding and non-coding genes and indicates the amino acids that are affected. Users can flexibly use RefSeq, UCSC genome browser, ENSEMBL, GENCODE, AceView, or other gene definition databases.

Let us first annotate our variants with the standard refGene database (NCBI):

`$ table_annovar.pl APOC3.vcf humandb/ -buildver hg19 -out APOC3_Gene.vcf -remove -nastring . -protocol refGene -operation g -vcfinput`

Each of the options in the command line is preceded with '-' (again, more information can be found by typing `table_annovar.pl`). The `-operation` option defines the type of annotation, `g`=gene-based; `f`=filter-based and `r`=region-based.

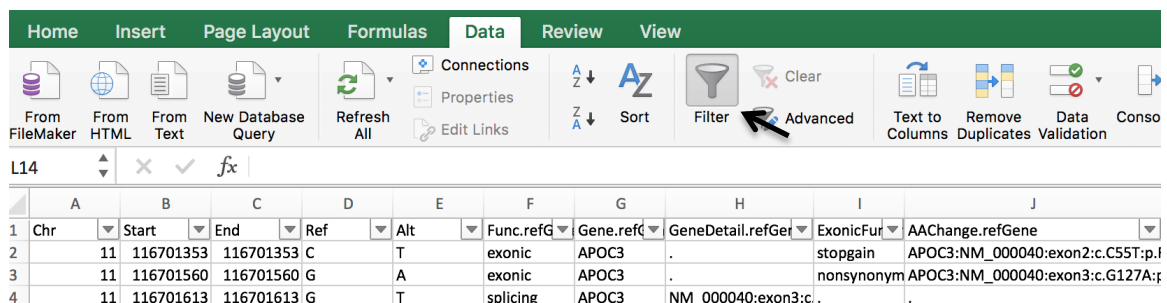
The annotated output file is written to `APOC3_Gene.vcf.hg19_multianno.txt`
Results are also written in VCF format: `APOC3_Gene.vcf.hg19_multianno.vcf`

Now look at the resulting table:

`$ cat APOC3_Gene.vcf.hg19_multianno.txt`

Question 1: Of the six *APOC3* variants that were analyzed how many are exonic__?

The output txt file is also easy to view in excel. Open the file in Excel and select "tab-delimited" when opening the file. To filter data, click the "data" tab at the menu bar, then click the "Filter" button.



	A	B	C	D	E	F	G	H	I	J
1	Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGene	ExonicFunc	AAChange.refGene
2	11	116701353	116701353	C	T	exonic	APOC3	.	stopgain	APOC3:NM_000040:exon2:c.C55T;p.I
3	11	116701560	116701560	G	A	exonic	APOC3	.	nonsynonym	APOC3:NM_000040:exon3:c.G127A;p
4	11	116701613	116701613	G	T	splicing	APOC3	NM_000040:exon3:c.	.	.

Notice all variants are automatically reported following the HGVS nomenclature. Variants are categorized based on these groups:

exonic	variant overlaps a coding region
splicing	variant is within 2-bp of a splicing junction (use -splicing threshold to change this)
ncRNA	variant overlaps a transcript without coding annotation in the gene definition
UTR5	variant overlaps a 5' untranslated region
UTR3	variant overlaps a 3' untranslated region
intronic	variant overlaps an intron
upstream	variant overlaps 1-kb region upstream of transcription start site
downstream	variant overlaps 1-kb region downstream of transcription end site (use -neargene to change this)
intergenic	variant is in intergenic region

Next we will annotate using three main databases: Ensembl, RefSeq and UCSC Known Gene, and change boundaries of splice variants (default is 2 bp from splice site, let's set this to 12 bp):

```
$ table_annovar.pl APOC3.vcf humandb/ -buildver hg19 -out APOC3_Gene.vcf -remove -nastring . -protocol refGene,knownGene,ensGene -operation g,g,g -arg '-splicing 12 -exonicsplicing','-splicing 12 -exonicsplicing','-splicing 12 -exonicsplicing' -vcfinput
```

This file has many columns, view select columns with awk (depending on which columns you are interested in seeing) using the below command or alternatively, you can open the file in excel:

```
$ awk -F'\t' '{print $1,$2,$6,$7,$8,$9,$10}' APOC3_Gene.vcf.hg19_multianno.txt
```

Question 2: What has changed compared to the initial annotation (hint: the splicing thresholds were changed) _____

_____?

B. Region based annotation

Another functionality of ANNOVAR is to annotate regions associated with variants: For example, DNase I hypersensitivity sites, ENCODE regions, predicted transcription factor binding sites, GWAS hits, and phastCons 46-way alignments to annotate variants that fall within conserved genomic regions as shown here:

```
$ table_annovar.pl APOC3.vcf humandb/ -buildver hg19 -out APOC3_Region.vcf -remove -nastring . -protocol phastConsElements46way -operation r -vcfinput
```

Note \$ cat resultingfile.txt here to view your results in the terminal or use awk to print certain columns of interest. Only conserved regions will display a score (maximum 1000) and a name.

Question 3: Which of the *APOC3* variants are within a conserved genomic region__?

We can also identify variants that were previously reported to be associated with diseases or traits in genome-wide association studies:

```
$ table_annoar.pl APOC3.vcf humandb/ -buildver hg19 -out APOC3_Region.vcf -  
remove -nastring . -protocol gwasCatalog -operation r -vcfinput
```

The gwasCatalog track in ANNOVAR is not fully comprehensive, but will point you towards major associations.

Question 4. Which of these variants are reported in the ANNOVAR GWAS catalog, and what has it been associated with _____?

The region-based annotation can be used to evaluate pathogenicity of certain regions, especially non-coding regions. In addition to the examples above, here are some other useful databases in region-annotation:

- wgRna: variants disrupting microRNAs and snoRNAs
- targetScanS: Identify variants disrupting predicted microRNA binding sites
- tfbsConsSites: Transcription factor binding sites
- The Encyclopedia of DNA Elements (ENCODE): A comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. Several annotations are possible depending on your interests and can be found here: <http://annovar.openbioinformatics.org/en/latest/user-guide/region/>

C. Filter based annotation

Filter based annotation includes annotation to certain databases, such as gnomAD, dbSNP, and prediction programs to evaluate pathogenicity. There are many options, but we selected these as particularly helpful for complex diseases:

```
$ table_annoar.pl APOC3.vcf humandb/ -buildver hg19 -out APOC3_Filter.vcf -remove  
-nastring . -protocol  
gnomad_genome,gnomad_exome,popfreq_max_20150413,gme,avsnp150,dbnsfp33a,db  
csnv11,cadd13gt20,clinvar_20170905,gwava -operation f,f,f,f,f,f,f,f,f,f -vcfinput
```

This command will annotate the following:

- gnomAD genome
- gnomad_exome (includes ExAC)
- popfreq_max_20150413: A database containing the maximum allele frequency from 1000G, ESP6500, ExAC and CG46 (use popfreq_all_20150413 to see all allele frequencies)
- dbSNP150
- gme: Great Middle East allele frequencies from the GME variome project
- dbnsfp33a: whole-exome SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, MetaSVM, MetaLR,

- VEST, M-CAP, CADD, GERP++, DANN, fathmm-MKL, Eigen, GenoCanyon, fitCons, PhyloP and SiPhy scores.
- dbSNV version 1.1: for splice site prediction by AdaBoost and Random Forest
 - Genome-wide CADD version 1.3 score > 20
 - clinvar_20170905: CLINVAR database with Variants of Clinical Significance
 - gWAVA: Prioritization of noncoding variants by integrating various genomic and epigenomic annotations.

Build your own filter annotations here:

<http://annovar.openbioinformatics.org/en/latest/user-guide/download/>

We can split these annotations up into several categories what will help to evaluate pathogenicity:

1. Allele frequency databases

1.a Allele frequency in control populations

Evaluating the frequency of a possible causal/associated variant in several control population is important in any disease/trait. The use of these databases might be different depending on the prevalence of your disease of interest, but these databases can provide valuable information on the rarity of variants and population-specific variants. If a variant is rare in your population, it is encouraged to check whether it might be more frequent in other populations, which might alter your conclusions on pathogenicity:

- gnomAD and ExAC databases:** The [Genome Aggregation Database](#) (gnomAD) and the [Exome Aggregation Consortium](#) (ExAC) are a coalition of investigators seeking to aggregate and harmonize genome and exome sequencing data from a wide variety of large-scale sequencing projects. The ExAC dataset contains spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. gnomAD spans 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals and includes ExAC data. For both databases individuals known to be affected by severe pediatric disease are removed, as well as their first-degree relatives, so this data set should aid as a useful reference set of allele frequencies for severe disease studies - however, note that some individuals with severe disease may still be included in the data set.
- BRAVO:** Genome sequencing variants of 62,784 individuals sequenced for NHLBI's TOPMed program, to enhance the understanding of fundamental biological processes that underlie heart, lung, blood and sleep disorders. Currently not implemented in ANNOVAR yet, can be found here: <https://bravo.sph.umich.edu/freeze5/hg38/>.
- GME database:** The Greater Middle East (GME) Variome Project (<http://igm.ucsd.edu/gme/>) is aimed at generating a coding base reference for the countries found in the Greater Middle East. This dataset is especially useful when dealing with Mendelian families from the Middle East. Although these individuals are not a random sample, they were ascertained as a wide variety of distinct phenotypes such that cohort-specific effects are not expected to bias patterns of variation. For the final filtered set, primarily healthy individuals from families were

- selected, and wherever possible, removed from datasets the allele that brought the family to medical attention, leaving 1,111 high-quality unrelated individuals.
- iv. **1000G database:** The [1000 Genomes Project](#) ran between 2008 and 2015, creating a public catalogue of human variation and genotype data. Phase 3 includes 26 different populations, and might be useful when interested in population specific variation.
 - v. **ESP6500:** The [NHLBI GO Exome Sequencing Project \(ESP\)](#) includes 6,503 samples drawn from multiple cohorts and represents all of the ESP exome variant data. In general, ESP samples were selected to contain deeply phenotyped individuals, the extremes of specific traits (LDL and blood pressure), and specific diseases (early onset myocardial infarction and early onset stroke), and lung diseases. This dataset contains a set of 2,203 African-Americans and 4,300 European-Americans unrelated individuals, totaling 6,503 samples (13,006 chromosomes).
 - vi. **CG46:** CG46 database compiled from unrelated individuals sequenced by the Complete Genomics platform.

1.b Allele frequencies in disease populations

- vii. **Clinvar:** ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes hosted by the National Center for Biotechnology Information (NCBI) and funded by intramural National Institutes of Health (NIH) funding. Although this database is mainly used for Mendelian disease variants, several rarer variants with a decent effect size in more complex disorders can be found in here as well, such as *APOC3*.

Let us examine if one of our variants we just annotated is in the Clinvar database:

```
$ awk -F'\t' '{print $1,$2,$103,$104}' APOC3_Filter.vcf.hg19_multianno.txt
```

Question 5: Is one of the variants reported as ‘pathogenic’ in Clinvar? If yes, which variants and which phenotype has been associated with these variants_____

_____?

Next, look at the gnomAD overall exome and genome frequencies in 123,136 individuals for our variants, and specific exome populations:

```
$ awk -F'\t' '{print $1,$2,$6,$14}' APOC3_Filter.vcf.hg19_multianno.txt
```

```
$ awk -F'\t' '{print $1,$2,$15,$16,$17,$18,$19,$20,$21,$22}'
```

```
APOC3_Filter.vcf.hg19_multianno.txt
```

Question 6: Are these variants common or rare, and are some more frequent in a specific population_____

_____?

1.c All variation

- viii. dbSNP: The Single Nucleotide Polymorphism database (dbSNP) or Database of Short Genetic Variations is a public-domain archive for a broad collection of simple genetic polymorphisms.

2. Effect on gene function:

2.a Missense variants

Missense mutations are sometimes more difficult to evaluate compared to loss-of-function mutations. If a variant occurs at a nucleotide or amino acid that is conserved through evolution, it is usually assumed that the specific nucleotide or amino acid is important to function. Whereas conservation scores such as PhyloP use evolution information to measure deleteriousness, there are also tools which combine information on evolution, biochemistry, structure and from public available databases etc., e.g. CADD, Eigen, MutationTaster.

We highlighted a select useful scoring methods that will help evaluate pathogenicity of a missense mutation:

- CADD*: Combined Annotation Dependent Depletion (CADD) is a framework that integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations. A scaled C-score of ≥ 10 indicates that the variant is predicted to be within 10% of most deleterious substitutions within the human genome, a score of ≥ 20 indicates the variant is predicted to be within 1% of the most deleterious variants, and so on. In the annotation above, we added all CADD scores in the exome + all CADD score in the genome > 20 c-scores. This score includes single nucleotide variants as well as insertion/deletions.
- Eigen and Eigen-PC*: Integrates different annotations into one measure of functional importance, a single functional score that can be incorporated in fine-mapping studies. Results for Eigen and Eigen-PC are similar for coding variants, but Eigen-PC has a considerable advantage over Eigen for noncoding variants. A positive Eigen-PC score is considered more damaging than a negative score.
- SIFT: Sorting Tolerant from Intolerant predicts whether an amino acid substitution affects protein function. SIFT prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences. It assumes that important positions in a protein sequence have been conserved throughout evolution and therefore substitutions at these positions may affect protein function. The SIFT score ranges from 0.0 (deleterious or “D”) to 1.0 (tolerated or “T”).
- PolyPhen2. Polymorphism Phenotyping v2. A tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. It obtains information from multiple sources such as variant site (e.g. active, binding, transmembrane, etc), multiple sequence alignment, secondary and 3D structure (if a known model exists), accessible surface area, etc.
 - PolyPhen2 HVAR: This metric is useful for diagnostics of Mendelian diseases, which requires distinguishing mutations with drastic effects from all the remaining human variation, including abundant mildly deleterious

alleles. The variant is considered probably damaging (D; score 0.909 and 1), possibly damaging (P; 0.447 and 0.908), or benign (B; 0 and 0.446).

- PolyPhen2 HDIV: PolyPhen HDIV should be used when evaluating rare variants involved in complex phenotypes and analysis of natural selection from sequence data. Variants can be classified as following: Probably damaging (D; 0.957 and 1), possibly damaging (P; 0.453 and 0.956), or benign (B; 0 and 0.452).
- LRT: The Likelihood Ratio Test. Using a comparative genomics data set of protein-coding sequences from 32 vertebrate species, the LRT was used to compare the null model that each codon is evolving neutrally, with the alternative model that the codon has evolved under negative selection. LRT can accurately identify a subset of deleterious mutations that disrupt highly conserved amino acids within protein-coding sequences, which are likely to be unconditionally deleterious. LRT prediction 'D' stands for 'deleterious' and 'N' stands for 'neutral'.
- MutationTaster*: Mutation taster performs a battery of *in silico* tests to estimate the impact of the variant on the gene product / protein. Tests are made on both, protein and DNA level. MutationTaster is not limited to substitutions of single amino acids but can also handle synonymous or intronic variants. It has four types of prediction outcomes: "disease_causing_automatic", "disease_causing", "polymorphism", and "polymorphism_automatic", which are coded as "A", "D", "N", and "P," respectively. Among them, "D" and "N" are determined by the prediction algorithm, whereas "A" and "P" are determined by external information. "A" and "D" can be regarded as prediction for deleteriousness.
- FATHMM and fathmm-MKL*: Functional Analysis Through Hidden Markov Models can be used for the prediction of the functional consequences of both coding variants and non-coding variants, using different algorithms. The more recent MKL algorithm can be used for all variants, utilizes various genomic annotations, and learns to weight the significance of each component annotation source. Variants are classified as either "damaging" ("D") or "tolerated" ("T").
- GERP++*: Genomic Evolutionary Rate Profiling (GERP) is a method for producing position-specific estimates of evolutionary constraint using maximum likelihood evolutionary rate estimation. GERP++ uses a more rigorous set of algorithms. Positive scores represent a substitution deficit (i.e., fewer substitutions than the average neutral site) and thus indicate that a site may be under evolutionary constraint. Negative scores indicate that a site is probably evolving neutrally. It was suggested that a RS score threshold of 2 provides high sensitivity while still strongly enriching for truly constrained sites; in practice, the threshold depends on the user.
- PhyloP*: (phylogenetic p-values) Evolutionary conservation at individual alignment sites, based on multiple alignments of 100 vertebrate species (100-way) or 20 mammals (20-way) under a null hypothesis of neutral evolution. Positive PhyloP scores indicate conserved sites (slower evolution than expected under neutral drift), the greater the score, the more conserved the site is; negative PhyloP scores indicate fast-evolving site (faster evolution than expected under neutral drift).

*available genome wide – that means they can be used to evaluate synonymous and non-coding variants as well (not all available genome-wide in ANNOVAR for annotation though). These scores are all integrated in dbSNFP, and more information and references can be found here: <https://sites.google.com/site/jpopgen/dbNSFP>

Let us evaluate some of these predictions above for our variants

```
$ awk -F'\t' '{print $1,$2,$36,$86,$70}' APOC3_Filter.vcf.hg19_multianno.txt
```

Note that these were loaded from a database here only including the exome. Individual datasets for some of these are available for annotation genome-wide as well.

Question 7: Can you fill in the other cells, which of the 3 missense variants have a prediction to be likely damaging?

Chr	Position	Ref Allele	Alt Allele	Variant Type	Polyphen2_HDIV	PhyloP_100way	CADD_phred
11	116701560	G	A	missense	1	4.302	23.6
11	116703532	A	G	missense			
11	116703580	A	G	missense			

2.b Splice variants:

- AdaBoost and Random Forest: Adaptive boosting (ADA) and random forest (RF) scores in dbSNV. dbSNV includes all potential human SNVs within splicing consensus regions (−3 to +8 at the 5' splice site and −12 to +2 at the 3' splice site). A score > 0.6 is considered damaging. Changing your splice boundaries to include splice region in combination with these scores can be useful to identify additional splice modifying variants.
- Regsnpintron: For all intronic SNPs including splice variants. See paragraph below. Please note that the current version is not working but should be updated soon.

Using the following methods examine the scores for the splice variants that we found earlier in the exercise:

```
$ awk -F'\t' '{print $1,$2,$99,$100}' APOC3_Filter.vcf.hg19_multianno.txt
```

Question 8: Can you fill in the ADA and RF scores below for the splice variants. Do these variants affect splicing?

Chr	Start	dbSNV_ADA_SCORE	dbSNV_RF_SCORE
11	116701353		
11	116701613		

2.c Intronic & non-coding SNPs:

- gWAVA: Genome-wide annotation of variants (GWAVA) is a tool that supports prioritization of noncoding variants by integrating various genomic and epigenomic annotations. There are different scores based on 3 different versions of the classifier and all are in the range 0-1 with higher scores indicating variants predicted as more likely to be functional.
- Regsnpintron: prioritize the disease-causing probability of intronic SNVs (uses a machine learning algorithm). The columns are "fpr (False positive rate), disease

Disease category (B: benign [FPR > 0.1]; PD: Possibly Damaging [0.05 < FPR <= 0.1]; D: Damaging [FPR <= 0.05]), splicing_site Splicing site (on/off). Splicing sites are defined as -3 to +7 for donor sites, -13 to +1 for acceptor sites.

Note: Most of the prediction scores in this filter-based annotation exercise were loaded through dbSNFP and therefore only exonic variants were annotated. Whole genome scores for the following are available in ANNOVAR as well as separate annotations: FATHMM, Eigen, CADD, GERP, gWAVA, regsnpintron, revel, mcap.

D. Why not combine all annotations?

We can combine all the annotations above into one and single command:

```
$ table_annotar.pl APOC3.vcf humandb/ -buildver hg19 -out APOC3_ANN.vcf -remove
-nastring . -protocol
refGene,knownGene,ensGene,wgRna,targetScanS,phastConsElements46way,tfsConsSit
es,gwasCatalog,gnomad_genome,gnomad_exome,popfreq_max_20150413,gme,avsnp15
0,dbnsfp33a,dbcsnv11,cadd13gt20,clinvar_20170905,gwava -operation
g,g,g,r,r,r,r,f,f,f,f,f,f,f,f,f,f -arg '-splicing 12 -exonicsplicing',''-splicing 12 -
exonicsplicing',''-splicing 12 -exonicsplicing',,,,,,,,,,,,,, -vcfinput
```

This makes it easy for you to make your own, customized annotation table.

Conclusion

The first five variants we studied are all rare variants shown to be associated with low apoC-III protein and triglycerides levels in blood. rs76353203, rs140621530 and rs147210663 were described in the following paper: “Loss-of-function mutations in *APOC3*, triglycerides, and coronary disease” (PMID: 24941081); rs121918381 was described in “Molecular cloning of a human apoC-III variant: Thr 74-Ala 74 variant prevents O-glycosylation” (PMID: 3123586); rs121918382 was described in “Apolipoprotein C-III(Lys58Glu): Identification of an apolipoprotein C-III variant in a family with hyperalphalipoproteinemia” (PMID: 2022742). The last variant (rs4225) is common, and was proposed as a candidate variant involved plasma triglycerides levels and coronary heart disease (PMID: 27624799).

Question 9: Based on the bioinformatics tools predictions, what do you think about the impact of the six variants on the function of the apoC-III protein?

E. Other useful annotations

Mitochondrial annotations

ANNOVAR has a database to annotate the impact of mitochondrial mutations: mitimpact24, use in the filter option

Gene intolerance to mutations scores

These scores can help evaluate whether a gene is tolerable or intolerable to damaging mutations:

- ExAC constraint metrics (pLI and z-scores): can be found on the ExAC website (Gene search).
 - Synonymous and missense: A signed Z score for the deviation of observed counts from the expected number was created. Positive Z scores imply increased constraint (intolerance to variation; i.e. the gene had fewer variants than expected). Negative Z scores indicated that the gene had more variants than expected.
 - LoF: For this metric, three classes of genes with respect to tolerance to LoF variation are assumed: 1) null (where LoF variation is completely tolerated), 2) recessive (where heterozygous LoFs are tolerated), 3) haploinsufficient (i.e. heterozygous LoFs are not tolerated). The observed and expected variants counts were used to determine the probability that a given gene is extremely intolerant of loss-of-function variation (falls into the third category). The closer pLI is to 1, the more LoF intolerant. A pLI ≥ 0.9 is considered as an extremely LoF intolerant set of genes.
- LoFtool score: gene loss-of-function score percentiles. The smaller the percentile, the most intolerant is the gene to functional variation.
- RVIS-ESV score: RVIS score measures genetic intolerance of genes to functional mutations.
- GDI score: the gene damage index (GDI) depicts the accumulated mutational damage for each human gene in the general population. Highly mutated/damaged genes are unlikely to be disease-causing. Yet these genes generate a big proportion of false positive variants harbored in such genes. Removing high GDI genes is a very effective way to remove confidently false positives from WES/WGS data. Damage predictions (low/medium/high) are made for different disease types.

F. Useful online annotation tools

These webtools are also very useful in annotating variants:

WGS Annotator (WGSa) - an annotation pipeline for human genome re-sequencing studies: <https://sites.google.com/site/jpopgen/wgsa/using-wgsa-via-aws>

Web Annovar: <http://wannovar.wglab.org/>

Seattleseq: <http://snp.gs.washington.edu/SeattleSeqAnnotation138/>

Ensembl variant predictor: <http://www.ensembl.org/info/docs/tools/vep/index.html>

Snp-nexus: <http://www.snp-nexus.org/>

Answers

Question 1: Of the six *APOC3* variants that were analyzed how many are exonic variants?

4 variants are exonic (3 nonsynonymous and one stop gain), one splice site variant and one 3'UTR variant.

Chr	Position	Ref	Alt	avsnp150	Func	Gene	ExonicFunc	cDNA/AAchange
11	116701353	C	T	rs76353203	Exonic	APOC3	stopgain	NM_000040:c.C55T;p.R19X
11	116701560	G	A	rs147210663	Exonic	APOC3	nonsynonymous	NM_000040:c.G127A;p.A43T
11	116701613	G	T	rs140621530	Splicing	APOC3	.	NM_000040:exon3:c.179+1G>T
11	116703532	A	G	rs121918382	Exonic	APOC3	nonsynonymous	NM_000040:c.A232G;p.K78E
11	116703580	A	G	rs121918381	Exonic	APOC3	nonsynonymous	NM_000040:c.A280G;p.T94A
11	116703671	G	T	rs4225	UTR3	APOC3	.	NM_000040:c.*71G>T

Question 2: What has changed compared to the initial annotation (hint: splicing thresholds were changed)?

The first variant at position 11:116701353 changed to exonic:splicing by changing our threshold to 12bp distance from the splice site. This variant is located at the -1 position of a 5' donor splice site and could affect splicing as well.

Question 3: Which the *APOC3* variants are within conserved genomic region?

The second and third variant.

Question 4. Which of these variants is reported in the ANNOVAR GWAS catalog, and what has it been associated with?

*The first variant, rs76353203, is indicated to have been associated with Triglyceride levels and high density lipoprotein cholesterol levels. This variant was the first variant in *APOC3* to have been associated with apoC-III deficiency, lower serum triglycerides, and higher levels of HDL cholesterol, and lower levels of LDL cholesterol (Pollin et al, 2008; PubMed: 19074352) and reached genome-wide significance in several GWAS for lower plasma triglyceride levels studies afterwards (PubMed: 24941081; PMID:24343240)*

Question 5: Is one of the variants reported as 'pathogenic' in Clinvar? If yes, which variants and which phenotype has been associated with these variants?

The first 5 variants are in Clinvar and reported as pathogenic, associated with coronary heart disease, Hyperalphalipoproteinemia, and Apolipoprotein_c-iii.

Question 6: Are these variants common or rare, and are some more frequent in a specific population?

The first five variants (exonic and splice) are rare, the last variant in the 3'UTR is common (44% overall prevalence in the genomes). The second variant has a higher frequency in the ASJ population (1.1%; Ashkenazi Jewish) compared to all other populations.

Question 7: Can you fill in the other cells, which of the 3 missense variants have a prediction to be likely damaging?

Chr	Position	Ref Allele	Alt Allele	Variant Type	Polyphen2_HDIV	PhyloP_100way	CADD_phred
11	116701560	G	A	missense	1	4.302	23.6
11	116703532	A	G	missense	0.611	0.719	15.56
11	116703580	A	G	missense	0.123	0.194	0.175

The first missense variant is very likely to be damaging. The second as well, though the last one is not predicted to be damaging by these 3 scoring methods, and more methods should be evaluated.

Question 8: Can you fill in the ADA and RF scores below for the splice variants. Do these variants affect splicing?

Chr	Start	Func.refGene	Effect	dbscSNV_ADA_SCORE	dbscSNV_RF_SCORE
11	116701353	exonic;splicing	c.C55T:p.R19X	0.0001	0.16
11	116701613	splicing	c.179+1G>T	1.000	0.936

The second variant is likely to affect splicing, as both scores are > 0.6. The first variant is located within exon (-1 position) of a 5' donor site, but is unlikely to affect splicing. This variant creates a stop mutation instead.

It is important to note that splice region variants (not standardly annotated unless you change boundaries) can still impact splicing, and annotation with these scores can help you evaluate their effect on splicing.

Question 9: Based on the bioinformatics tools predictions, what do you think about the impact of the six variants on the function of the apoC-III protein?

The first 3 variants, studied in a GWAS of 3734 participants and validated in 110,970 persons (PMID: 24941081), are predicted to be the most impactful on gene function based on all annotations.