# ANNOVAR Variant Annotation and Interpretation



This exercise touches on several functionalities of the program ANNOVAR to annotate and interpret genetic variants identified through next-generation sequencing. More information and a detailed guide on installation can be found here: http://annovar.openbioinformatics.org/en/latest/.

ANNOVAR has three main annotation types:

[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 cDNA or 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 prediction databases (PolyPhen, MutationTaster, FATHMM). Find intergenic variants with GERP++ score < 2, or many other annotations on specific mutations.

Other functionalities of ANNOVAR include the select of user-specific genomic positions in batch, identify a candidate gene list for Mendelian diseases from exome data, and section of dominant/recessive disease models.

In this exercise, we will annotate a .vcf file (DFNB44.vcf). This is from a family with hereditary hearing loss and a pedigree suggestive of a recessive inheritance model. As an example, we have selected a number of homozygous SNPs for an affected family member that were called after exome sequencing. We do encourage always checking every possible inheritance model, even in consanguineous and apparent autosomal recessive families.

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

Annotate variants to genes and indicate 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):

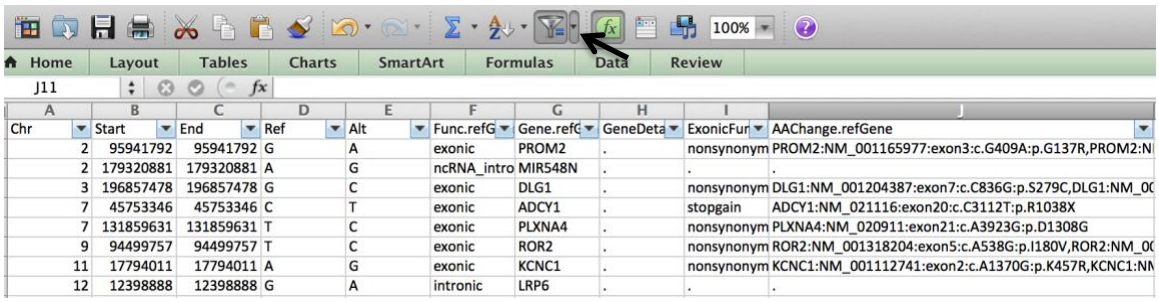
$ table_annovar.pl DFNB44.vcf humandb/ -buildver hg19 -out DFNB44_Gene.vcf -remove -nastring . -protocol refGene -operation g -vcfinput

The annotated output file is written to DFNB44_Gene.vcf.hg19_multianno.txt
Results are also written in VCF format: DFNB44_Gene.vcf.hg19_multianno.vcf

Now look at the resulting table:
$ cat DFNB44_Gene.vcf.hg19_multianno.txt

**Question 1: Which of these variants would you remove based on this annotation?**

The txt file is also easy to view in excel. Open the file in Excel (select "tab-delimited" when opening the file). Click the "DATA" tab at the menu bar, then click the "Filter" button (if you have a large VCF file, you might want to filter out common variants first or select variants based on disease model (see reduction section).




| Chr | Start | End | Ref | Alt | Func.refG | Gene.refG | GeneDeta | ExonicFur | AAChange.refGene |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 95941792 | 95941792 | G | A | exonic | PROM2 | . | nonsynonym | PROM2:NM_001165977:exon3:c.G409A:p.G137R,PROM2:N |
| 2 | 179320881 | 179320881 | A | G | ncRNA_intro | MIR548N | . | . | . |
| 3 | 196857478 | 196857478 | G | C | exonic | DLG1 | . | nonsynonym | DLG1:NM_001204387:exon7:c.C836G:p.S279C,DLG1:NM_00 |
| 7 | 45753346 | 45753346 | C | T | exonic | ADCY1 | . | stopgain | ADCY1:NM_021116:exon20:c.C3112T:p.R1038X |
| 7 | 131859631 | 131859631 | T | C | exonic | PLXNA4 | . | nonsynonym | PLXNA4:NM_020911:exon21:c.A3923G:p.D1308G |
| 9 | 94499757 | 94499757 | T | C | exonic | ROR2 | . | nonsynonym | ROR2:NM_001318204:exon5:c.A538G:p.I180V,ROR2:NM_00 |
| 11 | 17794011 | 17794011 | A | G | exonic | KCNC1 | . | nonsynonym | KCNC1:NM_001112741:exon2:c.A1370G:p.K457R,KCNC1:NM |
| 12 | 12398888 | 12398888 | G | A | intronic | LRP6 | . | . | . |

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 downtream 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 DFNB44.vcf humandb/ -buildver hg19 -out DFNB44_Gene.vcf -remove -nastring . -protocol refGene,knownGene,ensGene -operation g,g,g -arg '-splicing 12 -exonicsplicing','-splicing 12 -exonicsplicing','-splicing 12 -exonicsplicing' -vcfinput

The 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}' DFNB44_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 DFNB44.vcf humandb/ -buildver hg19 -out DFNB44_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.

**Question 3: Which variant is not located in a conserved region?**

This information can be used to evaluate pathogenicity of certain regions, especially non-coding regions.

Here are some highlighted databases:

- 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.

```
$ table_annovar.pl DFNB44.vcf humandb/ -buildver hg19 -out DFNB44_Filter.vcf -remove -nastring . -protocol gnomad_genome,gnomad_exome,popfreq_max_20150413,gme,avsnp147,dbnsfp30a,dbscsnv11,cadd13gt20,clinvar_20170130,intervar_20170202 -operation 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)
- dbSNP147
- gme: Great Middle East allele frequencies from the GME variome project
- dbnsfp30a: whole-exome SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR, VEST, CADD, GERP++, DANN, fitCons, PhyloP and SiPhy scores from dbNSFP version 3.0a
- dbscSNV version 1.1: for splice site prediction by AdaBoost and Random Forest
- Genome-wide CADD version 1.3 score>20
- clinvar_20170130: CLINVAR database with Variant Clinical Significance
- InterVar: Clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline

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**

In Mendelian disease, it is important to evaluate the frequency of a possible causal variant in a control population. This might be different depending on which inheritance model (i.e. dominant alleles should be rarer than recessive in a control cohort). Depending on the disease, reduced penetrance should also be considered. In general, a disease-causing mutations should be rare in any of these control databases:

i. **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.

ii. **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.

iii. **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.

iv. **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).

**v.** **CG46:** CG46 database compiled from unrelated individuals sequenced by the Complete Genomics platform.

**1.b Allele frequencies in disease populations**

vi. **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.

Let us examine if one of our variants we just annotated is in the Clinvar database:
$ awk -F'\t' '{print $1,$2,$71,$72}' DFNB44_Filter.vcf.hg19_multianno.txt

**Question 4: Is one of the variants reported as 'pathogenic' in Clinvar? If yes, which phenotype is associated with this variant?**

Next, look at the gnomAD overall exome frequencies in 123,136 individuals for our variants:
$ awk -F'\t' '{print $1,$2,$14}' DFNB44_Filter.vcf.hg19_multianno.txt

Since this a Middle Eastern family, check the Greater Middle East (GME) Variome:
$ awk -F'\t' '{print $1,$2,$24}' DFNB44_Filter.vcf.hg19_multianno.txt
This is important since certain variants can be very common in some populations but completely absent in others.

**Question 5: Fill in the exome frequencies of all variants in gnomAD and GME in the table below. Would you exclude any variants based on this annotation?**

| Chr | Start | gnomAD_exome_ALL | GME_AF |
|---|---|---|---|
| 2 | 95941792 | 0.0022 | 0.002520 |
| 2 | 179320881 | | |
| 3 | 196857478 | | |
| 7 | 45753346 | | |
| 7 | 131859631 | | |
| 9 | 94499757 | | |
| 11 | 17794011 | | |
| 12 | 12398888 | | |

**1.c All variation**

**vii.** 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. This database includes all variation, including disease-related and controls.

**2. Effect on protein:**

**Missense**

Missense mutations are sometimes more difficult to evaluate compared to loss-of-function mutations. 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 $\geq$10 indicates that the variant is predicted to be within 10% of most deleterious substitutions within the human genome, a score of $\geq$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.
- SIFT: Predicts whether an amino acid substitution affects protein function. The SIFT score ranges from 0.0 (deleterious) to 1.0 (tolerated). SIFT prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences.

- 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.
    - 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 (score 0.909 and 1), possibly damaging (0.447 and 0.908), or benign (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 (0.957 and 1), possibly damaging (0.453 and 0.956), or benign (0 and 0.452).
- LRT: The likelihood ratio test (LRT) of significantly conserved amino acid positions was applied to all codons within the human proteome.
- 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.
- MutationAssessor: Mutation assessor uses a multiple sequence alignment, partitioned to reflect functional specificity, and generates conservation scores for each column to represent the functional impact of a missense variant.
- FATHMM*: Functional Analysis Through Hidden Markov Models. Prediction of the functional consequences of both coding variants and non-coding variants.
- MetaSVM & MetaLR: Deleteriousness prediction for non-synonymous variants. These two ensemble scores (MetaSVM and MetaLR) are based on 10 component scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations.
- 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.
- fitCons*: The fitness consequences of functional annotation, integrates functional assays (such as ChIP-Seq) with selective pressure.
- PhyloP*: Evolutionary conservation at individual alignment sites.

*available genome wide – that means they can be used to evaluate synonymous and non-coding variants as well

Let us evaluate some of these predictions above for our variants
$ awk -F'\t' '{print $1,$2,$37,$42,$51}' DFNB44_Filter.vcf.hg19_multianno.txt
Note that these were loaded from a database here only including the exome. Individual dataset for some of these are available for annotation genome-wide as well.

**Question 6: Can you fill in the other cells, which variants have a prediction to be likely damaging?**

| Chr | Start | Polyphen2_HVAR_score | MutationTaster_pred | CADD_phred |
|---|---|---|---|---|
| 2 | 95941792 | 1 | D | 24.1 |
| 2 | 179320881 | | | |
| 3 | 196857478 | | | |
| 7 | 45753346 | | | |
| 7 | 131859631 | | | |
| 9 | 94499757 | | | |
| 11 | 17794011 | | | |
| 12 | 12398888 | | | |

**Splice mutations:**

- AdaBoost and Random Forest: Adaptive boosting (ADA) and random forest (RF) scores in dbscSNV. dbscSNV 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.

Using the following methods examine the scores for the splice region variant (c.549+3A>G) on chromosome 2 that we found earlier in the exercise:
$ awk -F'\t' '{print $1,$2,$67,$68}' DFNB44_Filter.vcf.hg19_multianno.txt

**Question 7: Can you fill in the ADA and RF scores below for the splice variant on chromosome two. Does this variant affect splicing?**

| Chr | Start | dbscSNV_ADA_SCORE | dbscSNV_RF_SCORE |
|---|---|---|---|
| 2 | 179320881 | | |

**3. ACMG Variant pathogenicity classification for clinical laboratories**

The American College of Medical Genetics and Genomics (ACMG) has published a recommendation to classify variants for use in a clinical setting. Variants according to ACMG/AMP Standards and Guidelines for clinical laboratories are classified as: 'Benign', 'Likely benign', 'Uncertain significance', 'Likely pathogenic' and 'Pathogenic': https://www.acmg.net/docs/Standards_Guidelines_for_the_Interpretation_of_Sequence_Variants.pdf
Please be cautious though with the automated annotation here in ANNOVAR, and double check each candidate variant carefully, especially if you are working in a clinical setting.

Some important recommendations and caveats from ACMG for apparent (loss-of-function) LOF mutations we would like to mention here:

- Beware of genes where LOF is not a known disease mechanism
- Use caution interpreting LOF variants at the extreme 3′ end of a gene, often they survive nonsense mediated decay and a (partial) functional protein might still be expressed
- Use caution with splice variants that are predicted to lead to exon skipping but leave the remainder of the protein intact
- Use caution in the presence of multiple transcripts

**D. Why not combine all annotations?**

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

```
$ table_annovar.pl DFNB44.vcf humandb/ -buildver hg19 -out DFNB44_ANN.vcf -
remove -nastring . -protocol
refGene,knownGene,ensGene,wgRna,targetScanS,phastConsElements46way,tfbsConsSit
es,gnomad_genome,gnomad_exome,popfreq_max_20150413,gme,avsnp147,dbnsfp30a,d
bscsnv11,cadd13gt20,clinvar_20170130,intervar_20170202 -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.

**Question 8: Based on the full annotation table, which of these variant(s) would you consider possibly damaging and for further investigation in this family?**

**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.

**Question 9: Go to http://exac.broadinstitute.org/ and find the ExAC constraint metrics for the genes you considered damaging in the exercise above? What can you conclude from these metrics?**

**F. Useful online annotation tools (mostly limited to a smaller number of variants)**

These webtools are also very useful in annotating variants:
**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: Which of these variants would you remove based on this annotation?**
*We would remove the intronic variant on chromosome 12.*

**Question 2: What has changed compared to the initial annotation (hint: splicing thresholds were changed)?**
*The second variant on chr2, position 179320881 changed to splicing by changing our threshold to 12bp distance from the splice site (c.549+3A>G).*

**Question 3: Which variant is not located in a conserved region?**
*The last intronic variant on chr12 is not located in a conserved region.*

**Question 4: Is one of the variants reported as 'pathogenic' in Clinvar? If yes, which phenotype is associated with this variant?**
*Yes, the nonsense variant on chr7, position 45753346 (ADCY1; c.C3112T:p.R1038X) has previously been reported as pathogenic in recessive hereditary hearing loss.*

**Question 5: Fill in the exome frequencies of all variants in gnomAD and GME in the table below. Would you exclude any variants based on this annotation?**

*Our pathogenic mutation on chr7, position 45753346 (ADCY1; c.C3112T:p.R1038X) is not reported in either database. All of these are pretty rare, we would not exclude any based on frequency.*

| Chr | Start | gnomAD_exome_ALL | GME_AF |
|---|---|---|---|
| 2 | 95941792 | 0.0022 | 0.00252 |
| 2 | 179320881 | . | . |
| 3 | 196857478 | 0.0006 | 0.000504 |
| 7 | 45753346 | . | . |
| 7 | 131859631 | 0.0003 | . |
| 9 | 94499757 | 0.0008 | 0.000504 |
| 11 | 17794011 | 0.0001 | . |
| 12 | 12398888 | . | . |

**Question 6: Can you fill in the other cells, which variants have a prediction to be likely damaging?**

| Chr | Start | Mutation | Polyphen2_HVAR_score | MutationTaster_pred | CADD_phred |
|---|---|---|---|---|---|
| 2 | 95941792 | missense | 1 | D | 24.1 |
| 2 | 179320881 | splice | . | . | . |
| 3 | 196857478 | missense | 0.926 | D | 24.9 |
| 7 | 45753346 | nonsense | . | D | 39 |
| 7 | 131859631 | missense | 0.999 | D | 30 |
| 9 | 94499757 | missense | 0.97 | D | 10.82 |
| 11 | 17794011 | missense | 0.694 | D | 10.52 |
| 12 | 12398888 | intronic | . | . | . |

*This first, 3rd, 4th and 5th variant have an overall predicted damaging effect based on these scores only. Note that for the intronic and splice variants these scores were not calculated because it wasn't possible (Polyphen2) or data was loaded from an exome database (MutationTaster and CADD). The Polyphen2 score is only calculated for missense mutations.*

**Question 7: Can you fill in the ADA and RF scores below for the splice variant on chromosome 2. Is this variant predicted to affect splicing?**

| Chr | Start | dbscSNV_ADA_SCORE | dbscSNV_RF_SCORE |
|---|---|---|---|
| 2 | 179320881 | 0.9998 | 0.938 |

*Both scores are > 0.6, indicating this variant affects splicing in this position. Though this variant is standardly not annotated as splice-altering. Splice region variants can still impact splicing, and annotation with these scores can help you evaluate their effect on splicing. We should consider this variant as possible pathogenic as well.*

**Question 8: Based on the full annotation table, which of these variant(s) would you consider possibly damaging and for further investigation in this family?**
*The nonsense variant on chr7, position 45753346 (ADCY1; c.C3112T:p.R1038X) and splice region variant predicted to affect splicing on chr2, position 179320881 (DFNB59, NM_001042702:exon4:c.549+3A>G) both have been involved in hereditary hearing loss and should be considered for additional validation.*

**Question 9: Go to http://exac.broadinstitute.org/ and find the ExAC constraint metrics for the genes you considered damaging in the exercise above? What can you conclude from these metrics?**

***ADCY1***

| Constraint from ExAC | Expected no. variants | Observed no. variants | Constraint Metric |
|---|---|---|---|
| Missense | 490.4 | 212 | z = 6.15 |
| LoF | 35.0 | 2 | pLI = 1.00 |

**DFNB59**

| Constraint from ExAC | Expected no. variants | Observed no. variants | Constraint Metric |
|---|---|---|---|
| Missense | 101.1 | 109 | z = -0.38 |
| LoF | 15.4 | 10 | pLI = 0.00 |

*The ADCY1 gene is intolerable to loss-of-function and missense mutations according to the pLI and z-scores, the DFNB59 gene is tolerable towards these mutations. We have to note that these metrics are more tailored towards dominant disease, and since both genes are known deafness genes, both variants are still good candidates in this case.*