A useful tutorial

In summer 2020, I taught an online training course that provides some materials for absolute beginners, including those who use personal Windows and Mac laptop computers, rather than Linux servers. The tutorial requires that you install conda in your personal computer, and it can be access here. Some users may find it useful. However, if you are already using a computing cluster and are already familiar with Linux, you do not need to follow this tutorial and can instead just read below.

Download and install

The latest version of ANNOVAR can be downloaded here (registration required). If you have any issue or question about downloading ANNOVAR, plase refer to Download ANNOVAR for more details.

When you have requested the ANNOVAR from the website and downloaded it, you will receive a compressed file `annovar.latest.tar.gz`, you will need to unzip it.

```
tar -xvzf annovar.latest.tar.gz
```

Once you unzip it, the annovar package will show up as a folder `annovar` and it will contains at least these files and folders:

```
annotate_variation.pl
coding_change.pl
convert2annovar.pl
example
humandb
retrieve_seq_from_fasta.pl
table_annovar.pl
variants_reduction.pl
```

In the `annovar` folder, the files end with `.pl` are the perl scripts that we could run. The `example` contains different input file examples. The `humandb` is our warehouse, it stores all the annotation databases that ANNOVAR can directly call and annotate.

Run ANNOVAR

By default, the ANNOVAR provide you with a very small example vcf file and basic annotation for you to run. We will use `ex2.vcf` in `example` as input, and run gene annotation using `table_annovar.pl`. `table_annovar.pl` takes an input variant file (such as a VCF file directly) and generate a tab-delimited output file with many columns, each representing one set of annotations. Additionally, if the input is a VCF file, the program also generates a new output VCF file with the INFO field filled with annotation information. To print the help message for all perl scripts, simply run the script using either `./table_annovar.pl` or `perl table_annovar.pl`.

Let's run our first ANNOVAR.

```
perl table_annovar.pl example/ex2.vcf \
  humandb/ \
  -buildver hg19 \
  -out my_first_anno \
  -protocol refGeneWithVer \
  -operation g \
  -remove -polish -vcfinput -nastring .
```

Results will be in `my_first_anno.hg19_multianno.txt` and `my_first_anno.hg19_multianno.vcf`. This simpliest eample could let you get gene annotation for each variants, but if you want to get more functional annotations, you will need to download additional database.

Download additional database

The `humandb` is our warehouse, it stores all the preprocessed databases of interest so ANNOVAR know how to annotate the variants based on the annotation we required. We need to download appropriate database files using `annotate_variation.pl`. Before download, we need to decide what database we want to use: - genome build (e.g., `hg19` or `hg38`) - annotation (e.g., `gnomad` or `clinvar`) - version (e.g. `clinvar_20240917` or `clinvar_20240611`)

Please check all available database for ANNOVAR in ANNOVAR addional database page.

Example of downloading additional database, and run ANNOVAR using these database (Note that if you already added ANNOVAR path into your system executable path, then typing `annotate_variation.pl` would be okay instead of typing `perl annotate_variation.pl`).

```
annotate_variation.pl -buildver hg19 -downdb -webfrom annovar refGeneWithVer humandb/
annotate_variation.pl -buildver hg19 -downdb cytoBand humandb/
annotate_variation.pl -buildver hg19 -downdb -webfrom annovar gnomad211_exome humandb/
annotate_variation.pl -buildver hg19 -downdb -webfrom annovar avsnp151 humandb/
annotate_variation.pl -buildver hg19 -downdb -webfrom annovar dbnsfp47a humandb/
```

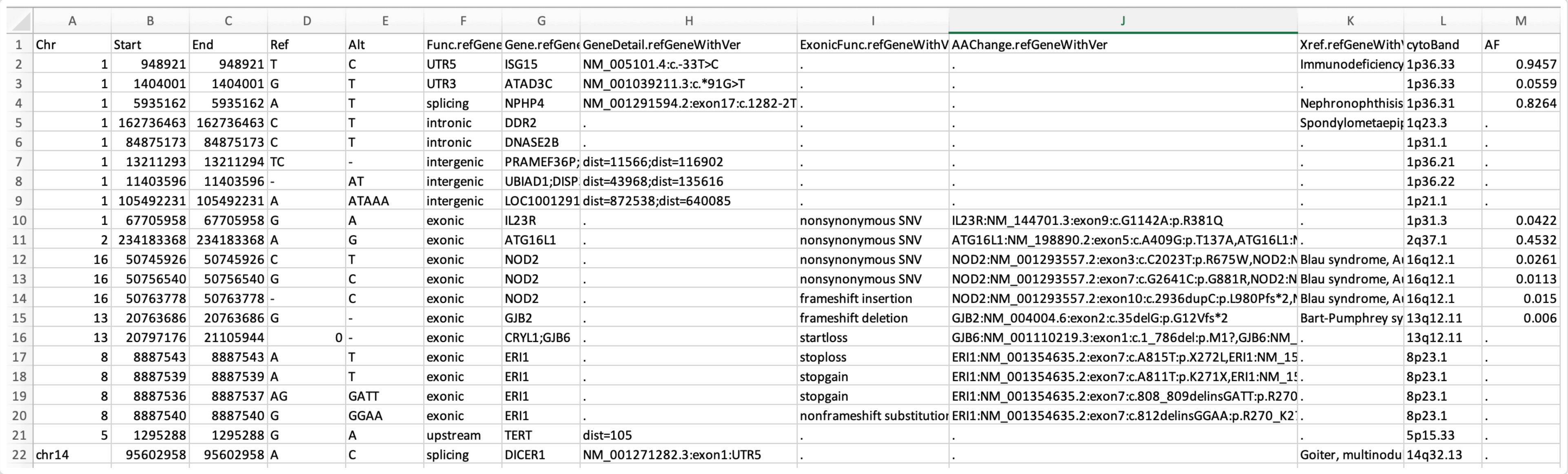
```
table_annovar.pl example/ex1.avinput \
  humandb/ \
  -buildver hg19 \
  -out ex1_anno \
  -protocol refGeneWithVer,cytoBand,gnomad211_exome,avsnp151,dbnsfp47a \
  -operation gx,r,f,f,f \
  -xref example/gene_xref.txt \
  -remove -nastring . -csvout -polish
```

Run the above commands one by one. The first a few commands download appropriate databases into the `humandb/` directory using `annotate_variation.pl`. The final command run TABLE_ANNOVAR, using following databases: - gnomAD exome collection version 2.1.1 (referred to as gnomad211_exome) - dbSNP version 151 (referred to as avsnp151) - dbNFSP version 4.7a (referred to as dbnsfp47a)

We also remove all temporary files (`-remove`), and generate the output file called `myanno.hg19_multianno.csv` (becausse we use `-csvout`). Fields that do not have any annotation will be filled by "." string (`-nastring .`).

We can examine the command line in greater detail. The `-operation` argument tells ANNOVAR which operations to use for each of the protocols: `g` means gene-based, `gx` means gene-based with cross-reference annotation (from `-xref` argument), `r` means region-based and `f` means filter-based. If you do not provide a xref file, then the operation can be `g` only. You will find details on what are gene/region/filter-based annotations in the other web pages. Sometimes, users want tab-delimited files rather than comma-delimited files. This can be easily done by removing `-csvout` argument to the above command.

Open the output file in Excel and see what it contains. The expected output file that I generated can be downloaded here: ex1.hg19_multianno.csv. A screen shot of the first a few columns is shown below:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGeneWithVer	ExonicFunc.refGeneWithV	AAChange.refGeneWithVer	Xref.refGeneWith\	cytoBand	AF
2	1	948921	948921	T	C	UTR5	ISG15	NM_005101.4:c.-33T>C	.	.	Immunodeficiency	1p36.33	0.9457
3	1	1404001	1404001	G	T	UTR3	ATAD3C	NM_001039211.3:c.*91G>T	.	.	.	1p36.33	0.0559
4	1	5935162	5935162	A	T	splicing	NPHP4	NM_001291594.2:exon17:c.1282-2T	.	.	Nephronophthisis	1p36.31	0.8264
5	1	162736463	162736463	C	T	intronic	DDR2	.	.	.	Spondylometaepi	1q23.3	.
6	1	84875173	84875173	C	T	intronic	DNASE2B	.	.	.	.	1p31.1	.
7	1	13211293	13211294	TC	-	intergenic	PRAMEF36P;	dist=11566;dist=116902	.	.	.	1p36.21	.
8	1	11403596	11403596	-	AT	intergenic	UBIAD1;DISP	dist=43968;dist=135616	.	.	.	1p36.22	.
9	1	105492231	105492231	A	ATAAA	intergenic	LOC1001291	dist=872538;dist=640085	.	.	.	1p21.1	.
10	1	67705958	67705958	G	A	exonic	IL23R	.	nonsynonymous SNV	IL23R:NM_144701.3:exon9:c.G1142A:p.R381Q	.	1p31.3	0.0422
11	2	234183368	234183368	A	G	exonic	ATG16L1	.	nonsynonymous SNV	ATG16L1:NM_198890.2:exon5:c.A409G:p.T137A,ATG16L1:N	.	2q37.1	0.4532
12	16	50745926	50745926	C	T	exonic	NOD2	.	nonsynonymous SNV	NOD2:NM_001293557.2:exon3:c.C2023T:p.R675W,NOD2:N	Blau syndrome, A	16q12.1	0.0261
13	16	50756540	50756540	G	C	exonic	NOD2	.	nonsynonymous SNV	NOD2:NM_001293557.2:exon7:c.G2641C:p.G881R,NOD2:N	Blau syndrome, A	16q12.1	0.0113
14	16	50763778	50763778	-	C	exonic	NOD2	.	frameshift insertion	NOD2:NM_001293557.2:exon10:c.2936dupC:p.L980Pfs*2,N	Blau syndrome, A	16q12.1	0.015
15	13	20763686	20763686	G	-	exonic	GJB2	.	frameshift deletion	GJB2:NM_004004.6:exon2:c.35delG:p.G12Vfs*2	Bart-Pumphrey sy	13q12.11	0.006
16	13	20797176	21105944	0	-	exonic	CRYL1;GJB6	.	startloss	GJB6:NM_001110219.3:exon1:c.1_786del:p.M1?,GJB6:NM_	.	13q12.11	.
17	8	8887543	8887543	A	T	exonic	ERI1	.	stoploss	ERI1:NM_001354635.2:exon7:c.A815T:p.X272L,ERI1:NM_15	.	8p23.1	.
18	8	8887539	8887539	A	T	exonic	ERI1	.	stopgain	ERI1:NM_001354635.2:exon7:c.A811T:p.K271X,ERI1:NM_1	.	8p23.1	.
19	8	8887536	8887537	AG	GATT	exonic	ERI1	.	stopgain	ERI1:NM_001354635.2:exon7:c.808_809delinsGATT:p.R270	.	8p23.1	.
20	8	8887540	8887540	G	GGAA	exonic	ERI1	.	nonframeshift substitutio	ERI1:NM_001354635.2:exon7:c.812delinsGGAA:p.R270_K2	.	8p23.1	.
21	5	1295288	1295288	G	A	upstream	TERT	dist=105	.	.	.	5p15.33	.
22	chr14	95602958	95602958	A	C	splicing	DICER1	NM_001271282.3:exon1:UTR5	.	.	Goiter, multinodu	14q32.13	.

The output file contains multiple columns. The first a few columns are your input columns, you could check `example/ex1.avinput` to see what it looks like. Each of the following columns corresponds on one of the "protocol" that user specified in the command line. The *Func.refGene, Gene.refGene, GeneDetail.refGene, ExonicFunc.refGene, AAChange.refGene* columns contain various annotation on how the mutations affect gene structure. The *Xref.refGene* column contains cross-reference for the gene; in this case, whether a known genetic disease is caused by defects in this gene (this information was supplied in the `example/gene_xref.txt` file in the command line). For the next serverals columns, the *AF** columns represent different allele frequency (AF) in gnomAD v2.1.1 database. The column *avsnp151* means the SNP identifier in the dbSNP version 151. The rest of the columns are from `dbnsfp47a` annotation, which contain pathogenic classification (end with `_pred`) or predicted score (end with `_score` or `_rankscore`) from several widely used tools, including AlphaMissense, MetaRNN, SIFT scores, PolyPhen2 HDIV scores, PolyPhen2 HVAR scores, LRT scores, MutationTaster scores, MutationAssessor score, FATHMM scores, GERP++ scores, CADD scores, DANN scores, PhyloP scores and SiPhy scores and so on.

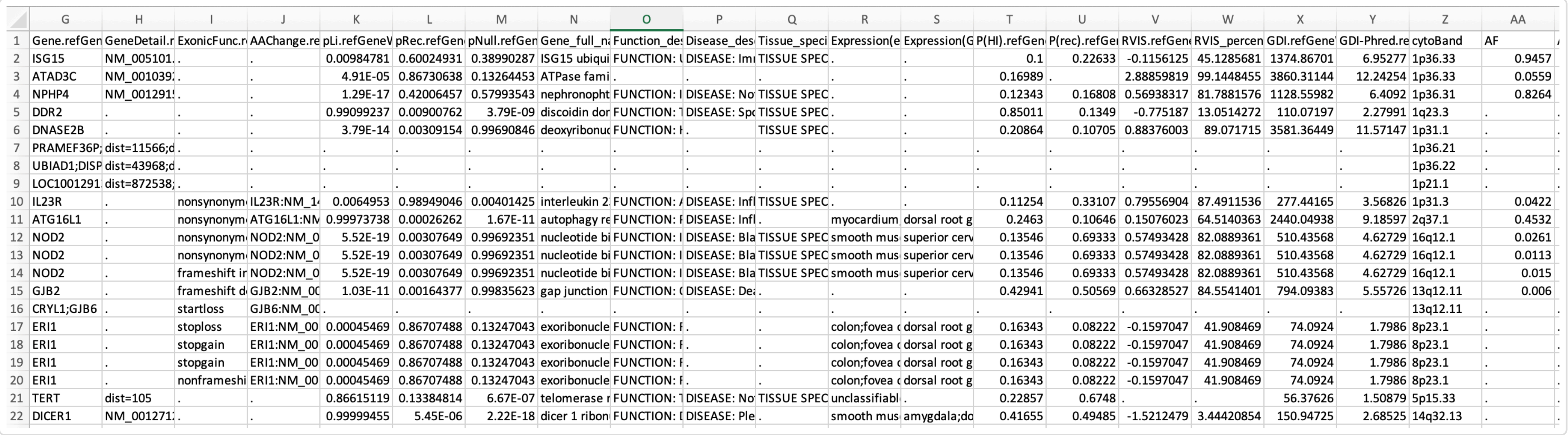
In the command above, we used `-xref` argument to provide annotation to genes. If the file contains header line, it is possible to provide mulitple pieces of annotations to genes (rather than just one single column). To illustrate this, we can check the first two lines (including the header line) of the `example/gene_fullxref.txt` file:

```
head -n 2 example/gene_fullxref.txt
```

```
#Gene_name      pLi      pRec     pNull    Gene_full_name   Function_description     Disease_description      Tissue_specificity(Uniprot)
A1BG    9.0649236354772e-05     0.786086131023045       0.2138232197406 alpha-1-B glycoprotein  .       .       TISSUE SPECIFICITY:
```

The header line starts with `#`. The cross-reference file then contains 15 types of annotations for genes.

You can run the same command above but change `-xref` from `gene_xref.txt` to `gene_fullxref.txt`, and the result file can be downloaded from here. Part of the file is shown below to give users an example:



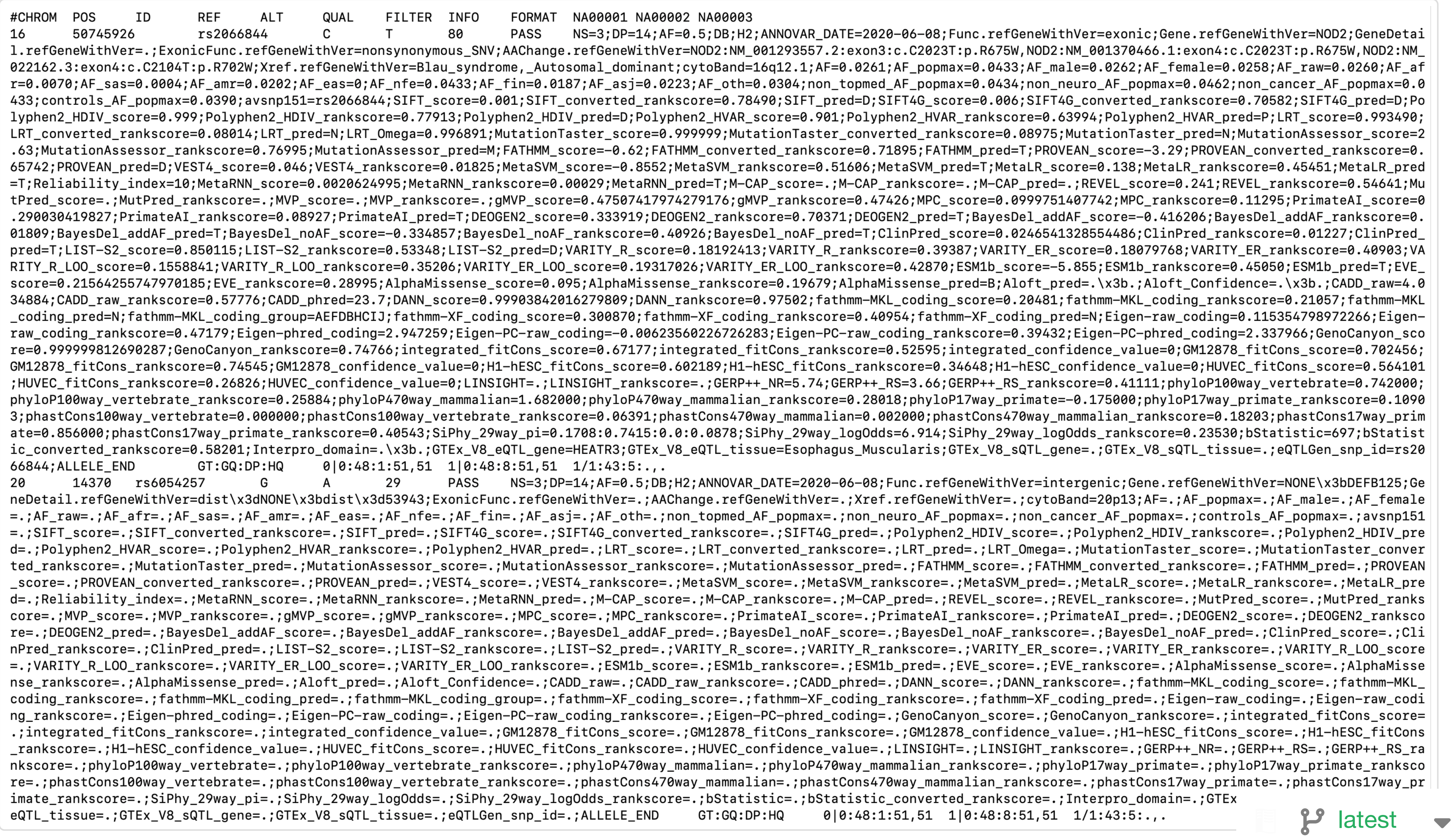
	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	
1	Gene.refGen	GeneDetail.r	ExonicFunc.r	AAChange.re	pLi.refGeneV	pRec.refGen	pNull.refGen	Gene_full_na	Function_des	Disease_des	Tissue_speci	Expression(e	Expression(G	P(HI).refGen	P(rec).refGe	RVIS.refGene	RVIS_percen	GDI.refGene	GDI-Phred.re	cytoBand	AF	/
2	ISG15	NM_005101.	.	.	0.00984781	0.60024931	0.38990287	ISG15 ubiqui	FUNCTION: U	DISEASE: Imi	TISSUE SPEC	.	.	0.1	0.22633	-0.1156125	45.1285681	1374.86701	6.95277	1p36.33	0.9457	
3	ATAD3C	NM_001039	.	.	4.91E-05	0.86730638	0.13264453	ATPase fami	.	.	.	.	.	0.16989	.	2.88859819	99.1448455	3860.31144	12.24254	1p36.33	0.0559	
4	NPHP4	NM_001291	.	.	1.29E-17	0.42006457	0.57993543	nephronopht	FUNCTION: I	DISEASE: No	TISSUE SPEC	.	.	0.12343	0.16808	0.56938317	81.7881576	1128.55982	6.4092	1p36.31	0.8264	
5	DDR2	.	.	.	0.99099237	0.00900762	3.79E-09	discoidin dor	FUNCTION: T	DISEASE: Spc	TISSUE SPEC	.	.	0.85011	0.1349	-0.775187	13.0514272	110.07197	2.27991	1q23.3	.	.
6	DNASE2B	.	.	.	3.79E-14	0.00309154	0.99690846	deoxyribonuc	FUNCTION: H	.	TISSUE SPEC	.	.	0.20864	0.10705	0.88376003	89.071715	3581.36449	11.57147	1p31.1	.	.
7	PRAMEF36P;	dist=11566;d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1p36.21	.	.
8	UBIAD1;DISP	dist=43968;d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1p36.22	.	.
9	LOC1001291	dist=872538;	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1p21.1	.	.
10	IL23R	.	nonsynonym	IL23R:NM_14	0.0064953	0.98949046	0.00401425	interleukin 2	FUNCTION: /	DISEASE: Infl	TISSUE SPEC	.	.	0.11254	0.33107	0.79556904	87.4911536	277.44165	3.56826	1p31.3	0.0422	
11	ATG16L1	.	nonsynonym	ATG16L1:NM	0.99973738	0.00026262	1.67E-11	autophagy re	FUNCTION: F	DISEASE: Infl	.	myocardium,	dorsal root g	0.2463	0.10646	0.15076023	64.5140363	2440.04938	9.18597	2q37.1	0.4532	
12	NOD2	.	nonsynonym	NOD2:NM_0	5.52E-19	0.00307649	0.99692351	nucleotide bi	FUNCTION: I	DISEASE: Bla	TISSUE SPEC	smooth mus	superior cerv	0.13546	0.69333	0.57493428	82.0889361	510.43568	4.62729	16q12.1	0.0261	
13	NOD2	.	nonsynonym	NOD2:NM_0	5.52E-19	0.00307649	0.99692351	nucleotide bi	FUNCTION: I	DISEASE: Bla	TISSUE SPEC	smooth mus	superior cerv	0.13546	0.69333	0.57493428	82.0889361	510.43568	4.62729	16q12.1	0.0113	
14	NOD2	.	frameshift ir	NOD2:NM_0	5.52E-19	0.00307649	0.99692351	nucleotide bi	FUNCTION: I	DISEASE: Bla	TISSUE SPEC	smooth mus	superior cerv	0.13546	0.69333	0.57493428	82.0889361	510.43568	4.62729	16q12.1	0.015	
15	GJB2	.	frameshift d	GJB2:NM_00	1.03E-11	0.00164377	0.99835623	gap junction	FUNCTION: (	DISEASE: Dea	.	.	.	0.42941	0.50569	0.66328527	84.5541401	794.09383	5.55726	13q12.11	0.006	
16	CRYL1;GJB6	.	startloss	GJB6:NM_00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	13q12.11	.	.
17	ERI1	.	stoploss	ERI1:NM_00	0.00045469	0.86707488	0.13247043	exoribonucle	FUNCTION: F	.	.	colon;fovea c	dorsal root g	0.16343	0.08222	-0.1597047	41.908469	74.0924	1.7986	8p23.1	.	.
18	ERI1	.	stopgain	ERI1:NM_00	0.00045469	0.86707488	0.13247043	exoribonucle	FUNCTION: F	.	.	colon;fovea c	dorsal root g	0.16343	0.08222	-0.1597047	41.908469	74.0924	1.7986	8p23.1	.	.
19	ERI1	.	stopgain	ERI1:NM_00	0.00045469	0.86707488	0.13247043	exoribonucle	FUNCTION: F	.	.	colon;fovea c	dorsal root g	0.16343	0.08222	-0.1597047	41.908469	74.0924	1.7986	8p23.1	.	.
20	ERI1	.	nonframeshi	ERI1:NM_00	0.00045469	0.86707488	0.13247043	exoribonucle	FUNCTION: F	.	.	colon;fovea c	dorsal root g	0.16343	0.08222	-0.1597047	41.908469	74.0924	1.7986	8p23.1	.	.
21	TERT	dist=105	.	.	0.86615119	0.13384814	6.67E-07	telomerase r	FUNCTION: T	DISEASE: No	TISSUE SPEC	unclassifiabl	.	0.22857	0.6748	.	.	56.37626	1.50879	5p15.33	.	.
22	DICER1	NM_001271	.	.	0.99999455	5.45E-06	2.22E-18	dicer 1 ribon	FUNCTION: D	DISEASE: Ple	.	smooth mus	amygdala;do	0.41655	0.49485	-1.5212479	3.44420854	150.94725	2.68525	14q32.13	.	.

Similarly, you could run `table_annovar.pl` with the same annotations directly using **VCF file** as input. For example:

```
table_annovar.pl example/ex2.vcf \
  humandb/ \
  -buildver hg19 \
  -out ex2_anno \
  -protocol refGeneWithVer,cytoBand,gnomad211_exome,avsnp151,dbnsfp47a \
  -operation gx,r,f,f,f \
  -xref example/gene_xref.txt \
  -remove -nastring . -polish \
  -vcfinput
```

Result will be written to `myanno.hg19_multianno.txt` (not a csv file because we did not put `-csvout` tag) and `myanno.hg19_multianno.vcf`.

You can download the output file here: ex2.hg19_multianno.vcf. Additionally, a tab-delimited output file is also available as ex2.hg19_multianno.txt, which contains similar information in a different format. You can open the new VCF file in a text editor and check what has been changed in the file: the INFO field in the VCF file now contains annotations that you need, starting with the string ANNOVAR_DATE and ending with the notation ALLELE_END. If multiple alleles are in the same locus, you will see multiple such notations (muleiple "ANNOVAR_DATE ... ALLELE_END" sections) in the INFO field. A screen shot is shown below:



```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NA00001 NA00002 NA00003
16      50745926        rs2066844       C       T       80      PASS    NS=3;DP=14;AF=0.5;DB;H2;ANNOVAR_DATE=2020-06-08;Func.refGeneWithVer=exonic;Gene.refGeneWithVer=NOD2;GeneDetai
l.refGeneWithVer=.;ExonicFunc.refGeneWithVer=nonsynonymous_SNV;AAChange.refGeneWithVer=NOD2:NM_001293557.2:exon3:c.C2023T:p.R675W,NOD2:NM_001370466.1:exon4:c.C2023T:p.R675W,NOD2:NM_
022162.3:exon4:c.C2104T:p.R702W;Xref.refGeneWithVer=Blau_syndrome,_Autosomal_dominant;cytoBand=16q12.1;AF=0.0261;AF_popmax=0.0433;AF_male=0.0262;AF_female=0.0258;AF_raw=0.0260;AF_af
r=0.0070;AF_sas=0.0004;AF_amr=0.0202;AF_eas=0;AF_nfe=0.0433;AF_fin=0.0187;AF_asj=0.0223;AF_oth=0.0304;non_topmed_AF_popmax=0.0434;non_neuro_AF_popmax=0.0462;non_cancer_AF_popmax=0.0
433;controls_AF_popmax=0.0390;avsnp151=rs2066844;SIFT_score=0.001;SIFT_converted_rankscore=0.78490;SIFT_pred=D;SIFT4G_score=0.006;SIFT4G_converted_rankscore=0.70582;SIFT4G_pred=D;Po
lyphen2_HDIV_score=0.999;Polyphen2_HDIV_rankscore=0.77913;Polyphen2_HDIV_pred=D;Polyphen2_HVAR_score=0.901;Polyphen2_HVAR_rankscore=0.63994;Polyphen2_HVAR_pred=P;LRT_score=0.993490;
LRT_converted_rankscore=0.08014;LRT_pred=N;LRT_Omega=0.996891;MutationTaster_score=0.999999;MutationTaster_converted_rankscore=0.08975;MutationTaster_pred=N;MutationAssessor_score=2
.63;MutationAssessor_rankscore=0.76995;MutationAssessor_pred=M;FATHMM_score=-0.62;FATHMM_converted_rankscore=0.71895;FATHMM_pred=T;PROVEAN_score=-3.29;PROVEAN_converted_rankscore=0.
65742;PROVEAN_pred=D;VEST4_score=0.046;VEST4_rankscore=0.01825;MetaSVM_score=-0.8552;MetaSVM_rankscore=0.51606;MetaSVM_pred=T;MetaLR_score=0.138;MetaLR_rankscore=0.45451;MetaLR_pred
=T;Reliability_index=10;MetaRNN_score=0.0020624995;MetaRNN_rankscore=0.00029;MetaRNN_pred=T;M-CAP_score=.;M-CAP_rankscore=.;M-CAP_pred=.;REVEL_score=0.241;REVEL_rankscore=0.54641;Mu
tPred_score=.;MutPred_rankscore=.;MVP_score=.;MVP_rankscore=.;gMVP_score=0.47507417974279176;gMVP_rankscore=0.47426;MPC_score=0.0999751407742;MPC_rankscore=0.11295;PrimateAI_score=0
.290030419827;PrimateAI_rankscore=0.08927;PrimateAI_pred=T;DEOGEN2_score=0.333919;DEOGEN2_rankscore=0.70371;DEOGEN2_pred=T;BayesDel_addAF_score=-0.416206;BayesDel_addAF_rankscore=0.
01809;BayesDel_addAF_pred=T;BayesDel_noAF_score=-0.334857;BayesDel_noAF_rankscore=0.40926;BayesDel_noAF_pred=T;ClinPred_score=0.0246541328554486;ClinPred_rankscore=0.01227;ClinPred_
pred=T;LIST-S2_score=0.850115;LIST-S2_rankscore=0.53348;LIST-S2_pred=D;VARITY_R_score=0.18192413;VARITY_R_rankscore=0.39387;VARITY_ER_score=0.18079768;VARITY_ER_rankscore=0.40903;VA
RITY_R_LOO_score=0.1558841;VARITY_R_LOO_rankscore=0.35206;VARITY_ER_LOO_score=0.19317026;VARITY_ER_LOO_rankscore=0.42870;ESM1b_score=-5.855;ESM1b_rankscore=0.45050;ESM1b_pred=T;EVE_
score=0.21564255747970185;EVE_rankscore=0.28995;AlphaMissense_score=0.095;AlphaMissense_rankscore=0.19679;AlphaMissense_pred=B;Aloft_pred=.\x3b.;Aloft_Confidence=.\x3b.;CADD_raw=4.0
34884;CADD_raw_rankscore=0.57776;CADD_phred=23.7;DANN_score=0.9990384201627980;DANN_rankscore=0.97502;fathmm-MKL_coding_score=0.20481;fathmm-MKL_coding_rankscore=0.21057;fathmm-MKL
_coding_pred=N;fathmm-MKL_coding_group=AEFDBHCIJ;fathmm-XF_coding_score=0.300870;fathmm-XF_coding_rankscore=0.40954;fathmm-XF_coding_pred=N;Eigen-raw_coding=0.115354798972266;Eigen-
raw_coding_rankscore=0.47179;Eigen-phred_coding=2.947259;Eigen-PC-raw_coding=-0.00623560226726283;Eigen-PC-raw_coding_rankscore=0.39432;Eigen-PC-phred_coding=2.337966;GenoCanyon_sco
re=0.999999812690287;GenoCanyon_rankscore=0.74766;integrated_fitCons_score=0.67177;integrated_fitCons_rankscore=0.52595;integrated_confidence_value=0;GM12878_fitCons_score=0.702456;
GM12878_fitCons_rankscore=0.74545;GM12878_confidence_value=0;H1-hESC_fitCons_score=0.602189;H1-hESC_fitCons_rankscore=0.34648;H1-hESC_confidence_value=0;HUVEC_fitCons_score=0.564101
;HUVEC_fitCons_rankscore=0.26826;HUVEC_confidence_value=0;LINSIGHT=.;LINSIGHT_rankscore=.;GERP++_NR=5.74;GERP++_RS=3.66;GERP++_RS_rankscore=0.41111;phyloP100way_vertebrate=0.742000;
phyloP100way_vertebrate_rankscore=0.25884;phyloP470way_mammalian=1.682000;phyloP470way_mammalian_rankscore=0.28018;phyloP17way_primate=-0.175000;phyloP17way_primate_rankscore=0.1090
3;phastCons100way_vertebrate=0.000000;phastCons100way_vertebrate_rankscore=0.06391;phastCons470way_mammalian=0.002000;phastCons470way_mammalian_rankscore=0.18203;phastCons17way_prim
ate=0.856000;phastCons17way_primate_rankscore=0.40543;SiPhy_29way_pi=0.1708:0.7415:0.0:0.0878;SiPhy_29way_logOdds=6.914;SiPhy_29way_logOdds_rankscore=0.23530;bStatistic=697;bStatist
ic_converted_rankscore=0.58201;Interpro_domain=.\x3b.;GTEx_V8_eQTL_gene=HEATR3;GTEx_V8_eQTL_tissue=Esophagus_Muscularis;GTEx_V8_sQTL_gene=.;GTEx_V8_sQTL_tissue=.;eQTLGen_snp_id=rs20
66844;ALLELE_END        GT:GQ:DP:HQ     0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
20      14370   rs6054257       G       A       29      PASS    NS=3;DP=14;AF=0.5;DB;H2;ANNOVAR_DATE=2020-06-08;Func.refGeneWithVer=intergenic;Gene.refGeneWithVer=NONE\x3bDEFB125;Ge
neDetail.refGeneWithVer=dist\x3dNONE\x3bdist\x3d53943;ExonicFunc.refGeneWithVer=.;AAChange.refGeneWithVer=.;Xref.refGeneWithVer=.;cytoBand=20p13;AF=.;AF_popmax=.;AF_male=.;AF_female
=.;AF_raw=.;AF_afr=.;AF_sas=.;AF_amr=.;AF_eas=.;AF_nfe=.;AF_fin=.;AF_asj=.;AF_oth=.;non_topmed_AF_popmax=.;non_neuro_AF_popmax=.;non_cancer_AF_popmax=.;controls_AF_popmax=.;avsnp151
=.;SIFT_score=.;SIFT_converted_rankscore=.;SIFT_pred=.;SIFT4G_score=.;SIFT4G_converted_rankscore=.;SIFT4G_pred=.;Polyphen2_HDIV_score=.;Polyphen2_HDIV_rankscore=.;Polyphen2_HDIV_pre
d=.;Polyphen2_HVAR_score=.;Polyphen2_HVAR_rankscore=.;Polyphen2_HVAR_pred=.;LRT_score=.;LRT_converted_rankscore=.;LRT_pred=.;LRT_Omega=.;MutationTaster_score=.;MutationTaster_conver
ted_rankscore=.;MutationTaster_pred=.;MutationAssessor_score=.;MutationAssessor_rankscore=.;MutationAssessor_pred=.;FATHMM_score=.;FATHMM_converted_rankscore=.;FATHMM_pred=.;PROVEAN
_score=.;PROVEAN_converted_rankscore=.;PROVEAN_pred=.;VEST4_score=.;VEST4_rankscore=.;MetaSVM_score=.;MetaSVM_rankscore=.;MetaSVM_pred=.;MetaLR_score=.;MetaLR_rankscore=.;MetaLR_pre
d=.;Reliability_index=.;MetaRNN_score=.;MetaRNN_rankscore=.;MetaRNN_pred=.;M-CAP_score=.;M-CAP_rankscore=.;M-CAP_pred=.;REVEL_score=.;REVEL_rankscore=.;MutPred_score=.;MutPred_ranks
core=.;MVP_score=.;MVP_rankscore=.;gMVP_score=.;gMVP_rankscore=.;MPC_score=.;MPC_rankscore=.;PrimateAI_score=.;PrimateAI_rankscore=.;PrimateAI_pred=.;DEOGEN2_score=.;DEOGEN2_ranksco
re=.;DEOGEN2_pred=.;BayesDel_addAF_score=.;BayesDel_addAF_rankscore=.;BayesDel_addAF_pred=.;BayesDel_noAF_score=.;BayesDel_noAF_rankscore=.;BayesDel_noAF_pred=.;ClinPred_score=.;Cli
nPred_rankscore=.;ClinPred_pred=.;LIST-S2_score=.;LIST-S2_rankscore=.;LIST-S2_pred=.;VARITY_R_score=.;VARITY_R_rankscore=.;VARITY_ER_score=.;VARITY_ER_rankscore=.;VARITY_R_LOO_score
=.;VARITY_R_LOO_rankscore=.;VARITY_ER_LOO_score=.;VARITY_ER_LOO_rankscore=.;ESM1b_score=.;ESM1b_rankscore=.;ESM1b_pred=.;EVE_score=.;EVE_rankscore=.;AlphaMissense_score=.;AlphaMisse
nse_rankscore=.;AlphaMissense_pred=.;Aloft_pred=.;Aloft_Confidence=.;CADD_raw=.;CADD_raw_rankscore=.;CADD_phred=.;DANN_score=.;DANN_rankscore=.;fathmm-MKL_coding_score=.;fathmm-MKL_
coding_rankscore=.;fathmm-MKL_coding_pred=.;fathmm-MKL_coding_group=.;fathmm-XF_coding_score=.;fathmm-XF_coding_rankscore=.;fathmm-XF_coding_pred=.;Eigen-raw_coding=.;Eigen-raw_codi
ng_rankscore=.;Eigen-phred_coding=.;Eigen-PC-raw_coding=.;Eigen-PC-raw_coding_rankscore=.;Eigen-PC-phred_coding=.;GenoCanyon_score=.;GenoCanyon_rankscore=.;integrated_fitCons_score=
.;integrated_fitCons_rankscore=.;integrated_confidence_value=.;GM12878_fitCons_score=.;GM12878_fitCons_rankscore=.;GM12878_confidence_value=.;H1-hESC_fitCons_score=.;H1-hESC_fitCons
_rankscore=.;H1-hESC_confidence_value=.;HUVEC_fitCons_score=.;HUVEC_fitCons_rankscore=.;HUVEC_confidence_value=.;LINSIGHT=.;LINSIGHT_rankscore=.;GERP++_NR=.;GERP++_RS=.;GERP++_RS_ra
nkscore=.;phyloP100way_vertebrate=.;phyloP100way_vertebrate_rankscore=.;phyloP470way_mammalian=.;phyloP470way_mammalian_rankscore=.;phyloP17way_primate=.;phyloP17way_primate_ranksco
re=.;phastCons100way_vertebrate=.;phastCons100way_vertebrate_rankscore=.;phastCons470way_mammalian=.;phastCons470way_mammalian_rankscore=.;phastCons17way_primate=.;phastCons17way_pr
imate_rankscore=.;SiPhy_29way_pi=.;SiPhy_29way_logOdds=.;SiPhy_29way_logOdds_rankscore=.;bStatistic=.;bStatistic_converted_rankscore=.;Interpro_domain=.;GTEx
eQTL_tissue=.;GTEx_V8_sQTL_gene=.;GTEx_V8_sQTL_tissue=.;eQTLGen_snp_id=.;ALLELE_END     GT:GQ:DP:HQ     0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
```

latest

Additional parameters options

Some people want to have the HGVS formatted strings for not only exonic variant, but also intronic variant that could be say 10bp away from splice site (by default, ANNOVAR only treats variants within 2bp of exon/intron boundary as splice variants, unless a `--slicing_threshold` parameter is set). For `-intronhgvs`, you will need to provide an integer which will then print HGVS notations for intron within this threshold away from exon. In here, we use `-intronhgvs 20`, it means anything within 20bp of intron/exon boundary will have the HGVS notation. for So you can specify this using the command below:

```
table_annovar.pl example/ex2.vcf \
  humandb/ \
  -buildver hg19 \
  -out myanno \
  -remove \
  -protocol refGeneWithVer,cytoBand,gnomad211_exome,avsnp151,dbnsfp47a \
  -operation g,r,f,f,f  \
  -nastring . \
  -vcfinput -polish \
  -intronhgvs 20
```

Finally, for each protocol/operation, you can add extra argument, and it has the same comma-delimited format. For example, you can add `-hgvs` argument to the `refGene` annotation so that the output is in HGVS format (c.122C>T rather than c.C122T). There are the same number of arguments in -arg as in -protocol and -operation.

```
table_annovar.pl example/ex2.vcf \
  humandb/ \
  -buildver hg19 \
  -out myanno \
  -remove \
  -protocol refGeneWithVer,cytoBand,gnomad211_exome,avsnp151,dbnsfp47a \
  -operation g,r,f,f,f \
  -arg '-hgvs',,,, \
  -nastring . -vcfinput -polish
```

Annotate exome VCF file

In this section, we will show how to run ANNOVAR annotation on human exome VCF file, consider both intronic and exonic regions. To make our files more organized, let's create a folder to store our file and result by `mkdir mywork`. We need to download the data we need, we can run this command to download the data into `mywork/`:

```
wget http://molecularcasestudies.cshlp.org/content/suppl/2016/10/11/mcs.a001131.DC1/Supp_File_2_KBG_family_Utah_VCF_files.zip \
  -O mywork/Supp_File_2_KBG_family_Utah_VCF_files.zip
```

To give some background information, this is a zip file as supplementary material of a published paper on exome sequencing of a family with undiagnosed genetic diseases. Through analysis of the exome data, the proband was confirmed to have KBG syndrome, a disease caused by loss of function mutations in ANKRD11 gene. There are several VCF files contained in the zip file, including those for parents, silings and the proband. We will only analyze proband in this exercise, but if you are interested, you may want to check whether this is a de novo mutation by analyzing parental genomes.

Then we can unzip it and take a look what it contains:

```
proband.vcf  Unaffected_brother.vcf  Unaffected_father.vcf  Unaffected_mother.vcf  Unaffected_sister1.vcf  Unaffected_sister2.vcf
```

Because this vcf file used hg19 as reference, we will need to use the databases corresponding to hg19 genome build for proper results. If you have followed our tutorial, you should already have most of the databases already, expect `clinvar_20240611`. Please run command below to download the databases you don't have:

```
annotate_variation.pl -buildver hg38 -downdb -webfrom annovar clinvar_20240611 humandb/
annotate_variation.pl -buildver hg19 -downdb -webfrom annovar refGeneWithVer humandb/
annotate_variation.pl -buildver hg19 -downdb -webfrom annovar gnomad211_exome humandb/
annotate_variation.pl -buildver hg19 -downdb -webfrom annovar dbnsfp47a humandb/
```

Now we have all the input file and datasets we need, let's run `table_annovar.pl` on the exome sequencing of proband `proband.vcf`. We will want to have gene annotation (`refGeneWithVer` operation), ClinVar annotation (`clinvar_20240917` operation), gnomADv2.1.1 exome annotation (`gnomad211_exome` operation), and pathogenicity preditions from various tools (`dbnsfp47a` operation).

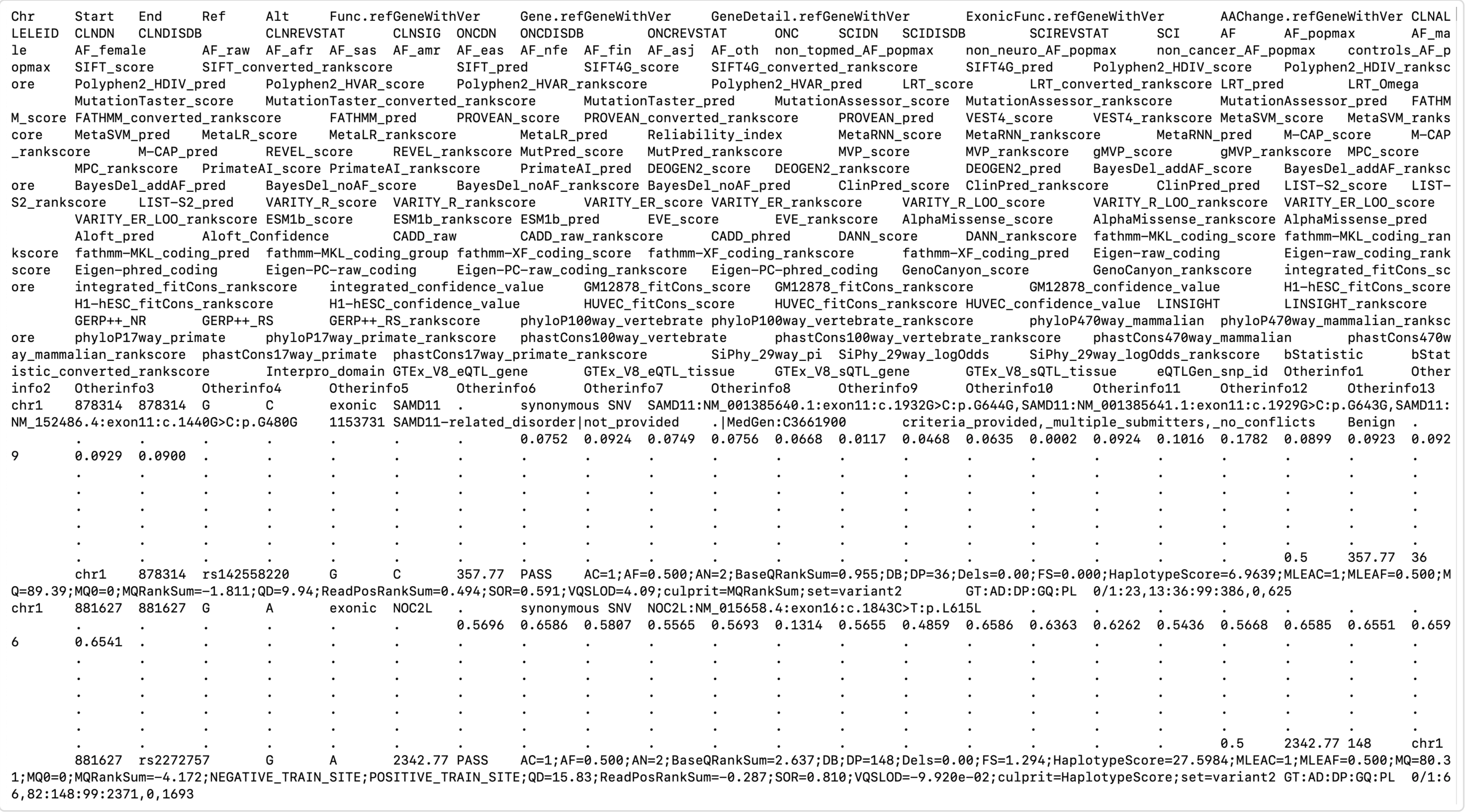
Let's run our command:

```
table_annovar.pl mywork/VCF_files/proband.vcf\
  humandb/ \
  -buildver hg19 \
  -out mywork/proband.annovar \
  -remove \
  -protocol refGeneWithVer,clinvar_20240917,gnomad211_exome,dbnsfp47a \
  -operation g,f,f,f \
  -arg '-hgvs',,, \
  -polish -nastring . \
  -vcfinput \
  -intronhgvs 100
```

Note that we could give arguement for a specific operation, in here we use `-arg '-hgvs',,,` to the `refGeneWithVer` operation. Moreover, we use `-intronhgvs 100` tag seperately and give a range of 100 which means anywhere within 100 bp away from the intron/extron boundary will have HGVS format annotation.

The results will be in `proband.annovar.hg19_multianno.txt` and `proband.annovar.hg19_multianno.vcf` files, which contain annotations for this exome.

We can use `less mywork/proband.annovar.hg19_multianno.txt` to check what the output looks like, you should have a result similar to this:



```
Chr     Start   End     Ref     Alt     Func.refGeneWithVer     Gene.refGeneWithVer     GeneDetail.refGeneWithVer       ExonicFunc.refGeneWithVer       AAChange.refGeneWithVer CLNAL
LELEID  CLNDN   CLNDISDB        CLNREVSTAT      CLNSIG  ONCDN   ONCDISDB        ONCREVSTAT      ONC     SCIDN   SCIDISDB        SCIREVSTAT      SCI     AF      AF_popmax       AF_ma
le      AF_female       AF_raw  AF_afr  AF_sas  AF_amr  AF_eas  AF_nfe  AF_fin  AF_asj  AF_oth  non_topmed_AF_popmax    non_neuro_AF_popmax     non_cancer_AF_popmax    controls_AF_p
opmax   SIFT_score      SIFT_converted_rankscore        SIFT_pred       SIFT4G_score    SIFT4G_converted_rankscore      SIFT4G_pred     Polyphen2_HDIV_score    Polyphen2_HDIV_ranksc
ore     Polyphen2_HDIV_pred     Polyphen2_HVAR_score    Polyphen2_HVAR_rankscore        Polyphen2_HVAR_pred     LRT_score       LRT_converted_rankscore LRT_pred        LRT_Omega
        MutationTaster_score    MutationTaster_converted_rankscore      MutationTaster_pred     MutationAssessor_score  MutationAssessor_rankscore      MutationAssessor_pred   FATHM
M_score FATHMM_converted_rankscore      FATHMM_pred     PROVEAN_score   PROVEAN_converted_rankscore     PROVEAN_pred    VEST4_score     VEST4_rankscore MetaSVM_score   MetaSVM_ranks
core    MetaSVM_pred    MetaLR_score    MetaLR_rankscore        MetaLR_pred     Reliability_index       MetaRNN_score   MetaRNN_rankscore       MetaRNN_pred    M-CAP_score     M-CAP
_rankscore      M-CAP_pred      REVEL_score     REVEL_rankscore MutPred_score   MutPred_rankscore       MVP_score       MVP_rankscore   gMVP_score      gMVP_rankscore  MPC_score
        MPC_rankscore   PrimateAI_score PrimateAI_rankscore     PrimateAI_pred  DEOGEN2_score   DEOGEN2_rankscore       DEOGEN2_pred    BayesDel_addAF_score    BayesDel_addAF_ranksc
ore     BayesDel_addAF_pred     BayesDel_noAF_score     BayesDel_noAF_rankscore BayesDel_noAF_pred      ClinPred_score  ClinPred_rankscore      ClinPred_pred   LIST-S2_score   LIST-
S2_rankscore    LIST-S2_pred    VARITY_R_score  VARITY_R_rankscore      VARITY_ER_score VARITY_ER_rankscore     VARITY_R_LOO_score      VARITY_R_LOO_rankscore  VARITY_ER_LOO_score
        VARITY_ER_LOO_rankscore ESM1b_score     ESM1b_rankscore ESM1b_pred      EVE_score       EVE_rankscore   AlphaMissense_score     AlphaMissense_rankscore AlphaMissense_pred
        Aloft_pred      Aloft_Confidence        CADD_raw        CADD_raw_rankscore      CADD_phred      DANN_score      DANN_rankscore  fathmm-MKL_coding_score fathmm-MKL_coding_ran
kscore  fathmm-MKL_coding_pred  fathmm-MKL_coding_group fathmm-XF_coding_score  fathmm-XF_coding_rankscore      fathmm-XF_coding_pred   Eigen-raw_coding        Eigen-raw_coding_rank
score   Eigen-phred_coding      Eigen-PC-raw_coding     Eigen-PC-raw_coding_rankscore   Eigen-PC-phred_coding   GenoCanyon_score        GenoCanyon_rankscore    integrated_fitCons_sc
ore     integrated_fitCons_rankscore    integrated_confidence_value     GM12878_fitCons_score   GM12878_fitCons_rankscore       GM12878_confidence_value        H1-hESC_fitCons_score
        H1-hESC_fitCons_rankscore       H1-hESC_confidence_value        HUVEC_fitCons_score     HUVEC_fitCons_rankscore HUVEC_confidence_value  LINSIGHT        LINSIGHT_rankscore
        GERP++_NR       GERP++_RS       GERP++_RS_rankscore     phyloP100way_vertebrate phyloP100way_vertebrate_rankscore       phyloP470way_mammalian  phyloP470way_mammalian_ranksc
ore     phyloP17way_primate     phyloP17way_primate_rankscore   phastCons100way_vertebrate      phastCons100way_vertebrate_rankscore    phastCons470way_mammalian       phastCons470w
ay_mammalian_rankscore  phastCons17way_primate  phastCons17way_primate_rankscore        SiPhy_29way_pi  SiPhy_29way_logOdds     SiPhy_29way_logOdds_rankscore   bStatistic      bStat
istic_converted_rankscore       Interpro_domain GTEx_V8_eQTL_gene       GTEx_V8_eQTL_tissue     GTEx_V8_sQTL_gene       GTEx_V8_sQTL_tissue     eQTLGen_snp_id  Otherinfo1      Other
info2   Otherinfo3      Otherinfo4      Otherinfo5      Otherinfo6      Otherinfo7      Otherinfo8      Otherinfo9      Otherinfo10     Otherinfo11     Otherinfo12     Otherinfo13
chr1    878314  878314  G       C       exonic  SAMD11  .       synonymous SNV  SAMD11:NM_001385640.1:exon11:c.1932G>C:p.G644G,SAMD11:NM_001385641.1:exon11:c.1929G>C:p.G643G,SAMD11:
NM_152486.4:exon11:c.1440G>C:p.G480G    1153731 SAMD11-related_disorder|not_provided    .|MedGen:C3661900       criteria_provided,_multiple_submitters,_no_conflicts    Benign  .
        .       .       .       .       .       .       .       0.0752  0.0924  0.0749  0.0756  0.0668  0.0117  0.0468  0.0635  0.0002  0.0924  0.1016  0.1782  0.0899  0.0923  0.092
9       0.0929  0.0900  .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       0.5     357.77  36
        chr1    878314  rs142558220     G       C       357.77  PASS    AC=1;AF=0.500;AN=2;BaseQRankSum=0.955;DB;DP=36;Dels=0.00;FS=0.000;HaplotypeScore=6.9639;MLEAC=1;MLEAF=0.500;M
Q=89.39;MQ0=0;MQRankSum=-1.811;QD=9.94;ReadPosRankSum=0.494;SOR=0.591;VQSLOD=4.09;culprit=MQRankSum;set=variant2        GT:AD:DP:GQ:PL  0/1:23,13:36:99:386,0,625
chr1    881627  881627  G       A       exonic  NOC2L   .       synonymous SNV  NOC2L:NM_015658.4:exon16:c.1843C>T:p.L615L      .       .       .       .       .       .       .
        .       .       .       .       .       .       0.5696  0.6586  0.5807  0.5565  0.5693  0.1314  0.5655  0.4859  0.6586  0.6363  0.6262  0.5436  0.5668  0.6585  0.6551  0.659
6       0.6541  .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
        .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       .       0.5     2342.77 148     chr1
        881627  rs2272757       G       A       2342.77 PASS    AC=1;AF=0.500;AN=2;BaseQRankSum=2.637;DB;DP=148;Dels=0.00;FS=1.294;HaplotypeScore=27.5984;MLEAC=1;MLEAF=0.500;MQ=80.3
1;MQ0=0;MQRankSum=-4.172;NEGATIVE_TRAIN_SITE;POSITIVE_TRAIN_SITE;QD=15.83;ReadPosRankSum=-0.287;SOR=0.810;VQSLOD=-9.920e-02;culprit=HaplotypeScore;set=variant2 GT:AD:DP:GQ:PL  0/1:6
6,82:148:99:2371,0,1693
```

The screenshot showed us the complete columns and the first two variants. We see some familiar columns from our previous exmaple, such as variant basic information (first 5 columns), refGeneWithVer annotation, and gnomad AF columns, and some tool predictions columns from dbnsfp47a. The columns that start with `CLN` are from ClinVar annotation.

At this point, we have our results, and you could choose your own way of downstream analysis of these exome variants. But for demonstration purpose, we have provided a downstream analysis example in advanced use case for exome VCF annotation. Downstream analysis includes chromosome distribution, variant type ditritbution, ClinVar pathogenicity distribution, CADD score, MetaRNN/AlphaMissense pathogenic predictions, etc.

Gene Annotation Example

The purpose of this gene annotation example is to showcase how to perform a correct gene annotation using ANNOVAR, as a respond to this paper (PMID 36268089) which evaluated the ANNOVAR using 298 variants with ground truth of variant annotation. However, the authors might run ANNOVAR in inappropriate way so they had a wrong conclusion about ANNOVAR. Here we used the exact vcf file they provided as a demo to show how to get the proper gene annotation (DNA change, amino acid change), with transcript version provided. Take a look at our vcf file first:

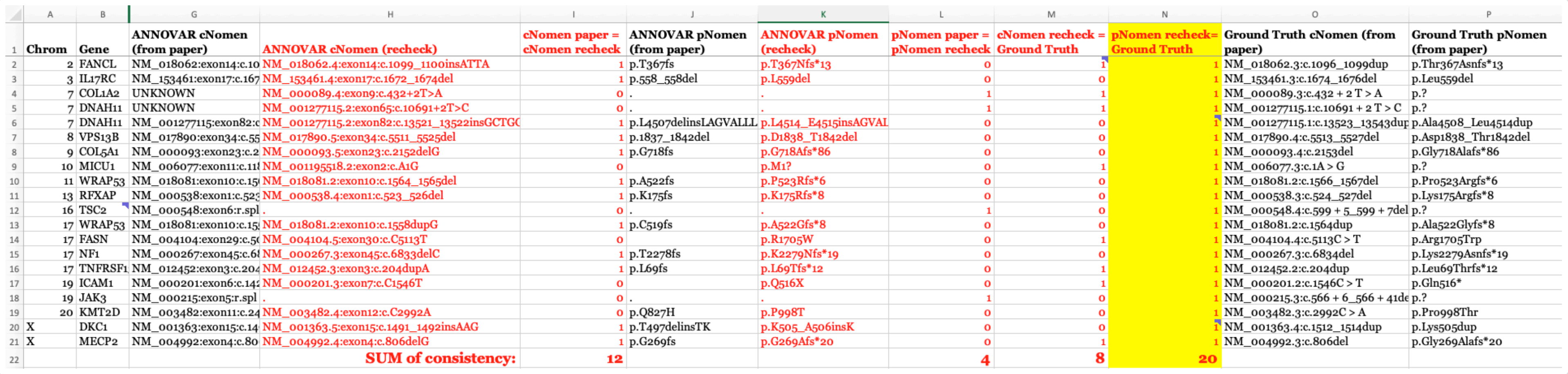
```
##fileformat=VCFv4.0
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO
2   162279995   .   C   G   .   .   .
2   162310909   .   T   C   .   .   .
1   11046609    .   T   C   .   .   .
19  19193983    .   A   T   .   .   .
7   147903589   .   T   C   .   .   .
17  82079248    .   G   A   .   .   .
10  63219963    .   G   C   .   .   .
13  101103286   .   T   A   .   .   .
```

There are 8 columns in a normal vcf file, and in this vcf file there is no quality score, id and other info, it only has the chromosome number, position, reference and alterantive allele, but this will be enough for ANNOVAR to run annotation. Since we only interested in a very simple task: what is the cDNA and amino acid change (if possible) for these variants. We could run the following command:

```
perl table_annovar.pl mywork/PMID_36268089.vcf \
  humandb/ \
  -buildver hg38 \
  -out mywork/myanno_PMID_36268089 \
  -remove \
  -protocol refGeneWithVer \
  -operation g \
  -nastring . \
  -vcfinput \
  -polish
```

The output file of this command is provided here. The first 5 columns describe the chromosome, position, reference allele and alterantive allele for each vairant. The gene name is the 7th column `Gene.refGeneWithVer`, as we can see 'IFIH1', 'MASP2' and 'RFXANK' were shown. For amino acid change of this variant, we could check the 10th column `AAChange.refGeneWithVer`, and it will tell us the amino acid change per transcript. Note that the first variant '2 162279995 162279995 C G' does not have amino acid change becuase it is not in the protein coding region, instead it is in the 'splicing' region. And for the variant '1 11046609 11046609 T C', there are two protein changes 'p.D120G' and 'p.D120G' and this is because there are 2 transcripts (isoforms) for this MASP2 variant, and in this case they are the same amino acid change in the same position, but sometimes you will see different position for amino acid change in different isoforms.

After the annotation, we rechecked our result with the previous paper. The 20 variants provided in the screenshot below are the variants that the paper claimed ANNOVAR had incorrect annotations. The columns in red text are the new columns that we added for rechecking purpose, and the rest of the columns in black text were kept the same from the paper. The cDNA change is called "cNomen" and the amino acid change is called "pNomen" in the paper, so we will keep the same name. At the bottom, we summarize the consistency of ANNOVAR results that we got (cNomen/pNomen recheck), the ANNOVAR results from the paper (cNomen/pNomen paper), and the groud truth annotations based on the paper (Groud Truth). Given that one cDNA change could have various ways of interpretations, amino acid change is more resonable for comparison. At the last columns (highlighted in yellow), we checked our ANNOVAR annotation of amino acid change (pNomen) with the Groud Truth, and ANNOVAR showed 100% accuracy in terms of amino acid change. The version of transcript is provided in our ANNOVAR result as well because we used `refGeneWithVer` database.



Chrom	Gene	ANNOVAR cNomen (from paper)	ANNOVAR cNomen (recheck)	cNomen paper = cNomen recheck	ANNOVAR pNomen (from paper)	ANNOVAR pNomen (recheck)	pNomen paper = pNomen recheck	cNomen recheck = Ground Truth	pNomen recheck= Ground Truth	Ground Truth cNomen (from paper)	Ground Truth pNomen (from paper)
2	FANCL	NM_018062:exon14:c.10	NM_018062.4:exon14:c.1099_1100insATTA	1	p.T367fs	p.T367Nfs*13	0	1	1	NM_018062.3:c.1096_1099dup	p.Thr367Asnfs*13
3	IL17RC	NM_153461:exon17:c.167	NM_153461.4:exon17:c.1672_1674del	1	p.558_558del	p.L559del	0	0	1	NM_153461.3:c.1674_1676del	p.Leu559del
7	COL1A2	UNKNOWN	NM_000089.4:exon9:c.432+2T>A	0	.	.	1	1	1	NM_000089.3:c.432 + 2 T > A	p.?
7	DNAH11	UNKNOWN	NM_001277115.2:exon65:c.10691+2T>C	0	.	.	1	1	1	NM_001277115.1:c.10691 + 2 T > C	p.?
7	DNAH11	NM_001277115:exon82:c	NM_001277115.2:exon82:c.13521_13522insGCTGC	1	p.L4507delinsLAGVALLL	p.L4514_E4515insAGVAL	0	0	1	NM_001277115.1:c.13523_13543dup	p.Ala4508_Leu4514dup
8	VPS13B	NM_017890:exon34:c.55	NM_017890.5:exon34:c.5511_5525del	1	p.1837_1842del	p.D1838_T1842del	0	0	1	NM_017890.4:c.5513_5527del	p.Asp1838_Thr1842del
9	COL5A1	NM_000093:exon23:c.2	NM_000093.5:exon23:c.2152delG	1	p.G718fs	p.G718Afs*86	0	0	1	NM_000093.4:c.2153del	p.Gly718Alafs*86
10	MICU1	NM_006077:exon11:c.11	NM_001195518.2:exon2:c.A1G	0		p.M1?	0	1	1	NM_006077.3:c.1A > G	p.?
11	WRAP53	NM_018081:exon10:c.15	NM_018081.2:exon10:c.1564_1565del	1	p.A522fs	p.P523Rfs*6	0	0	1	NM_018081.2:c.1566_1567del	p.Pro523Argfs*6
13	RFXAP	NM_000538:exon1:c.52	NM_000538.4:exon1:c.523_526del	1	p.K175fs	p.K175Rfs*8	0	0	1	NM_000538.3:c.524_527del	p.Lys175Argfs*8
16	TSC2	NM_000548:exon6:r.spl	.	0	.	.	1	0	1	NM_000548.4:c.599 + 5_599 + 7del	p.?
17	WRAP53	NM_018081:exon10:c.15	NM_018081.2:exon10:c.1558dupG	1	p.C519fs	p.A522Gfs*8	0	0	1	NM_018081.2:c.1564dup	p.Ala522Glyfs*8
17	FASN	NM_004104:exon29:c.5	NM_004104.5:exon30:c.C5113T	0		p.R1705W	0	1	1	NM_004104.4:c.5113C > T	p.Arg1705Trp
17	NF1	NM_000267:exon45:c.6	NM_000267.3:exon45:c.6833delC	1	p.T2278fs	p.K2279Nfs*19	0	0	1	NM_000267.3:c.6834del	p.Lys2279Asnfs*19
17	TNFRSF1	NM_012452:exon3:c.204	NM_012452.3:exon3:c.204dupA	1	p.L69fs	p.L69Tfs*12	0	1	1	NM_012452.2:c.204dup	p.Leu69Thrfs*12
19	ICAM1	NM_000201:exon6:c.14	NM_000201.3:exon7:c.C1546T	0		p.Q516X	0	1	1	NM_000201.2:c.1546C > T	p.Gln516*
19	JAK3	NM_000215:exon5:r.spl	.	0	.	.	1	0	1	NM_000215.3:c.566 + 6_566 + 41de	p.?
20	KMT2D	NM_003482:exon11:c.24	NM_003482.4:exon12:c.C2992A	0	p.Q827H	p.P998T	0	0	1	NM_003482.3:c.2992C > A	p.Pro998Thr
X	DKC1	NM_001363:exon15:c.14	NM_001363.5:exon15:c.1491_1492insAAG	1	p.T497delinsTK	p.K505_A506insK	0	0	1	NM_001363.4:c.1512_1514dup	p.Lys505dup
X	MECP2	NM_004992:exon4:c.80	NM_004992.4:exon4:c.806delG	1	p.G269fs	p.G269Afs*20	0	1	1	NM_004992.3:c.806del	p.Gly269Alafs*20
			SUM of consistency:	12			4	8	20		

Hopefully, after you finish this set of exercises above, you now have a better idea what ANNOVAR is, and can start enjoy the journey of annotating your variants.

If you are interesting in more advanced use of ANNOVAR, please refer to our Advanced Use Case.