

Technical Note

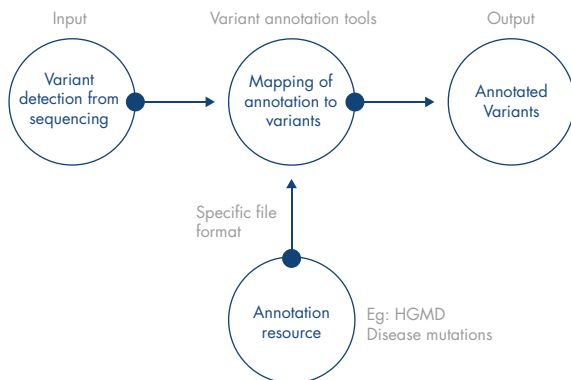
# Use of HGMD mutation data within popular variant annotation tools

Numerous free or open source variant annotation tools are available today to extract, annotate and analyse the many genomes and their identified variants coming from next generation sequencing methods.

There are many different types of information available for annotation of variants with the end goal to use that annotation to define the effect and changes in phenotype that are likely to be caused by the variant. Various information resources can act as a backend database for the annotation tools used within an annotation pipeline where the input file with an undefined collection of variants becomes directly associated with the annotation details (Figure 1).

The value derived from the annotation is directly related to the information resource selected for annotation. Cited in more than 5,000 scientific articles, HGMD is the industry leading database for published, inherited disease mutations.

In this technical note we identify a subset of popular variant annotation tools that are able to work with HGMD data and provide a step-by-step guide for the use of HGMD data by three of the tools: ANNOVAR, snpEff and VariantAnnotation – a Bioconductor package.



**Figure 1.** Variant annotation pipeline

## Open source variant annotation tools

A selection of popular free or open source variant annotation tools are described in Table 1.

Tool	Code source	Annotation format supported	HGMD use described in this application note
ANNOVAR*	Perl	GFF3, VCF	Yes
snpEff	Java	TXT, BED, BigBed, VCF, GFF	Yes
Variant Annotation (Bioconductor package)	R	VCF	Yes
AnnTools	Python, MySQL for data storage	BED	No
CHAOs	Perl	BED, WIG	No
vcfanno	go	BED, BAM, VCF	No
seqminer	R	VCF, BCF, METAL	No

\*ANNOVAR is free for academic use only. Commercial use requires a license from QIAGEN.

## HGMD as an annotation resource

HGMD is a comprehensive database of published inherited disease mutations. Trained genetics experts read the published literature and extract information about germline mutations that have been shown to be associated with a specific disease or phenotype. The database is updated quarterly to ensure that the latest and most relevant information is available. As of the September 2016.3 release HGMD contained information for more than 192,000 mutations.

HGMD data is available by subscription for download in multiple formats supporting variant annotation including BED, GFF and VCF formats. Both hg19 and hg38 reference genomes are supported.

## VCF format

```
##fileformat=VCFv4.1
##Copyright=HGMD. Not for redistribution.
##source=HGMD_PRO_2016.1
##reference=GRCh38
##comment="REF and ALT sequences are both on forward strand of reference assembly"
##INFO=<ID=CLASS,Number=1,Type=String,Description="Mutation Category, https://portal.biobase-international.com/hgmd/pro/global.php#data>
##INFO=<ID=MUT,Number=1,Type=String,Description="HGMD mutant allele">
##INFO=<ID=GENE,Number=1,Type=String,Description="Gene symbol">
##INFO=<ID=STRAND,Number=1,Type=String,Description="Gene strand">
##INFO=<ID=DNA,Number=1,Type=String,Description="DNA annotation">
##INFO=<ID=PROT,Number=1,Type=String,Description="Protein annotation">
##INFO=<ID=DB,Number=1,Type=String,Description="dbSNP identifier, build 137">
##INFO=<ID=PHEH,Number=1,Type=String,Description="HGMD primary phenotype">
#CHROM POS ID REF ALT QUAL FILTER INFO
1 942143 CM1511864 C G . . CLASS=DM7;MUT=ALT;GENE=SAMD11;STRAND=+;DNA=NM_152486.2:c.877C>G;PROT=NF_689699.2:p.F293A;DB=rs200195897;PHEH="Autism_spectrum_disorder"
1 963938 CD142720 CCT C . . CLASS=DM7;MUT=ALT;GENE=KLHL17;STRAND=+;DNA=NM_198317.2:c.1375_1376delCT;PHEH="Schizophrenia"
1 1014143 CM1411641 C T . . CLASS=DM;MUT=ALT;GENE=ISG15;STRAND=+;DNA=NM_005101.3:c.163C>T;PROT=NF_005092.1:p.Q55*;PHEH="Idiopathic_basal_ganglia_calcification"
1 1014316 CI128669 C CG . . CLASS=DM;MUT=ALT;GENE=ISG15;STRAND=+;DNA=NM_005101.3:c.339dupG;PHEH="Mycobacterial_disease_mendelian_susceptibility_to"
```

## GFF3 format

```
##gff-version 3
chr1 hgmd variant_phenotype 942143 942143 . + . ID=1;accession=CM1511864;alt=G;aminoacid_change=P>A;citation_type=Primary;codon_change=CCT-GCT;codon_number=293
chr1 hgmd variant_phenotype 963938 963940 . + . ID=2;accession=CD142720;alt=C;aminoacid_change=W/A;citation_type=Primary;codon_change=W/A;codon_number=458;com
chr1 hgmd variant_phenotype 1014143 1014143 . + . ID=3;accession=CM1411641;alt=T;aminoacid_change=Q>*;citation_type=Primary;codon_change=CAG-TAG;codon_number=55;
chr1 hgmd variant_phenotype 1014316 1014316 . + . ID=4;accession=CI128669;alt=CG;aminoacid_change=W/A;citation_type=Primary;codon_change=W/A;codon_number=113;com
chr1 hgmd variant_phenotype 1014359 1014359 . + . ID=5;accession=CM128668;alt=T;aminoacid_change=E>*;citation_type=Primary;codon_change=GAG-TAG;codon_number=127;
chr1 hgmd variant_phenotype 1022225 1022225 . + . ID=6;accession=CM148517;alt=A;aminoacid_change=G>S;citation_type=Primary,FCR;codon_change=GGC-AGC;codon_number=
chr1 hgmd variant_phenotype 1022313 1022313 . + . ID=7;accession=CM148518;alt=T;aminoacid_change=W>I;citation_type=Primary,FCR;codon_change=AAC-ATC;codon_number=
chr1 hgmd variant_phenotype 1041582 1041582 . + . ID=8;accession=CM126385;alt=T;aminoacid_change=Q>*;citation_type=Primary;codon_change=CAG-TAG;codon_number=353;
```

## BED format

```
track name="hgmd" description="HGMD Mutations" color="176,23,31" visibility=3
chr1 877522 877523 Autism_spectrum_disorder:877C>G 0 +
chr1 899317 899320 Schizophrenia:1375_1376delCT 0 +
chr1 949522 949523 Idiopathic_basal_ganglia_calcification:163C>T 0 +
chr1 949695 949696 Mycobacterial_disease_mendelian_susceptibility_to:339dupG 0 +
chr1 949738 949739 Mycobacterial_disease_mendelian_susceptibility_to:379G>T 0 +
```

## Step-by-step data analysis

Here we demonstrate the steps required to annotate an input sample with HGMD mutation data for three variant analysis tools: ANNOVAR, snpEff and VariantAnnotation.

The dataset used for the analysis is the breast cancer (primary ductal carcinoma TNM stage IIA, grade 3) HCC1187 cell line sample from the Complete Genomics public cancer data set (R. Drmanac et al, Science 327(5961), 78).

## ANNOVAR

**Step 1:** Convert the input VCF file to ANNOVAR's specific file format using the accessory perl script `convert2annovar.pl`. In this example, `HG00731-200-37-ASM.vcf` is the input file and `cgexample` is the name appended to the converted output file

```
$ perl convert2annovar.pl
-format vcf4 vcfBeta-HG00731-
200-37-ASM.vcf -allsample
-outfile cgexample
```

```
kar@sys-mkt108 /cygdrive/i/annovar
$ perl convert2annovar.pl -format vcf4 vcfBeta-HG00731-200-37-ASM.vcf -allsample -outfile cgexample
NOTICE: output files will be written to cgexample.<samplename>.avinput
NOTICE: Finished reading 10344776 lines from VCF file
NOTICE: A total of 10344658 locus in VCF file passed QC threshold, representing 3465464 SNPs (2358709 transitions and 1106755 tr
ansversions) and 6895319 indels/substitutions
NOTICE: Finished writing 3392941 SNPs (2310236 transitions and 1082705 transversions) and 581702 indels/substitutions for 1 samp
les
WARNING: Skipped 4830315 invalid alternative alleles found in input file
WARNING: Found 366 invalid reference alleles in input file
WARNING: Skipped 1658714 invalid genotype records in input file
```

**Step 2:** Annotate the converted VCF file (named cgexample.HG00731-200-37-ASM.avinput in this example) with HGMD annotations using the annotate.variation.pl script. The VCF formatted HGMD file (named HGMD\_PRO\_2016.1\_hg19.vcf in this example) is used as the database file. In this example it is found in the humandb directory.

```
$ perl annotate_variation.pl -infoasscore -buildver hg19 -filter -dbtype vcf -vcfdbfile HGMD_PRO_2016.1_hg19.vcf cgexample.HG00731-200-37-ASM.avinput humandb/
```

```
Kar@MKT /cygdrive/d/annovar
$ perl annotate_variation.pl -infoasscore -buildver hg19 -filter -dbtype vcf -vcfdbfile HGMD_PRO_2016.1_hg19.vcf cgexample.HG00731-200-37-ASM.avinput humandb/
NOTICE: Variants matching filtering criteria are written to cgexample.HG00731-200-37-ASM.avinput.hg19_vcf_dropped, other variants are written to cgexample.HG00731-200-37-ASM.avinput.hg19_vcf_filtered
NOTICE: Processing next batch with 3974643 unique variants in 3974643 input lines
NOTICE: Scanning filter database humandb/HGMD_PRO_2016.1_hg19.vcf...Done
```

**Step 3:** Search the output file (named cgexample.HG00731-200-37-ASM.avinput.hg19\_vcf\_dropped in this example) for annotated variants in the gene of your choice. In this example we have chosen to use BRCA1 since the sample data is taken from a breast cancer cell line.

```
$ egrep -w "hgnc=BRCA1" cgexample.HG00731-200-37-ASM.avinput.hg19_vcf_dropped
```

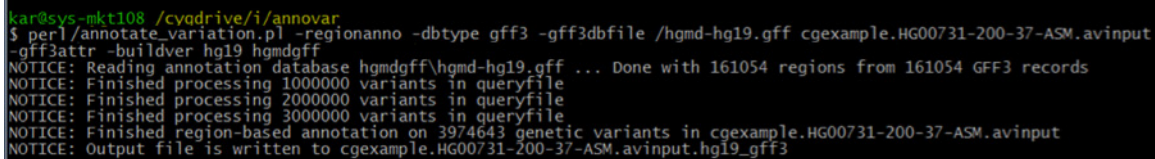
```
Kar@MKT /cygdrive/d/annovar
$ egrep -w "hgnc=BRCA1" cgexample.HG00731-200-37-ASM.avinput.hg19_vcf_dropped
vcf CLASS=DFP;MUT=ALT;GENE=BRCA1;STRAND=-;DB=rs8176318;PHEN="Reduced_activity_association_with" 17 4119
7274 41197274 C A het . 35
vcf CLASS=DM?;MUT=ALT;GENE=BRCA1;STRAND=-;DNA=NM_007294.3:c.5152+66G>A;DB=rs3092994;PHEN="Breast_cancer" 17 4
1215825 41215825 C T het . 26
vcf CLASS=R;MUT=ALT;GENE=BRCA1;STRAND=-;DNA=NM_007294.3:c.4837A>G;PROT=NP_009225.1:p.S1613G;DB=rs1799966;PHEN="Breast_cancer" 17 41223094 41223094 T C het . 12
vcf CLASS=DP;MUT=ALT;GENE=BRCA1;STRAND=-;DNA=NM_007294.3:c.3548A>G;PROT=NP_009225.1:p.K1183R;DB=rs16942;PHEN="Breast_cancer_protection_against_association_with" 17 41244000 41244000 T C het . 43
vcf CLASS=DP;MUT=ALT;GENE=BRCA1;STRAND=-;DNA=NM_007294.3:c.3113A>G;PROT=NP_009225.1:p.E1038G;DB=rs16941;PHEN="Endometriosis_association_with" 17 41244435 41244435 T C het . 43
vcf CLASS=DFP;MUT=ALT;GENE=BRCA1;STRAND=-;DNA=NM_007294.3:c.2612C>T;PROT=NP_009225.1:p.P871L;DB=rs799917;PHEN="Cervical_cancer_decreased_risk_association_with" 17 41244936 41244936 G A het . 4
0
vcf CLASS=DP;MUT=ALT;GENE=BRCA1;STRAND=-;DNA=NM_007294.3:c.1067A>G;PROT=NP_009225.1:p.Q356R;DB=rs1799950;PHEN="Breast_and/or_ovarian_cancer_association_with" 17 41246481 41246481 T C het . 4
6
vcf CLASS=FP;MUT=ALT;GENE=BRCA1;STRAND=-;DB=rs799906;PHEN="Altered_promoter_activity" 17 41278116 4
1278116 T C het . 32
vcf CLASS=DFP;MUT=ALT;GENE=BRCA1;STRAND=-;DB=rs11655505;PHEN="Breast_cancer_decreased_risk_association_with" 1
7 41278377 41278377 G A het . 49
vcf CLASS=FP;MUT=ALT;GENE=BRCA1;STRAND=-;DB=rs799908;PHEN="Altered_promoter_activity" 17 41278916 4
1278916 A G het . 16
vcf CLASS=FP;MUT=ALT;GENE=BRCA1;STRAND=-;DB=rs4793204;PHEN="Reduced_promoter_activity" 17 41279298 4
1279298 A G het . 23
```

Alternatively you can use HGMD gff file as the database file.

**Step 1:** Convert the input VCF file to ANNOVAR's specific file format using the accessory perl script `convert2annovar.pl` (as shown previously)

**Step 2:** Annotate the converted VCF file (named `cgexample.HG00731-200-37-ASM.avinput` in this example) with HGMD annotations using the `annotate.variation.pl` script. The GFF3 formatted HGMD file (named `hgmd-hg19.gff` in this example) is used as the database file. In this example it is found in the `hgmdgff` directory

```
$ perl annotate_variation.pl -regionanno -dbtype gff3 -gff3dbfile /hgmd-hg19.gff cgexample.HG00731-200-37-ASM.avinput --gff3attr -buildver hg19 hgmdgff
```



```
kar@sys-mkt108 /cyadrive/i/annovar
$ perl /annotate_variation.pl -regionanno -dbtype gff3 -gff3dbfile /hgmd-hg19.gff cgexample.HG00731-200-37-ASM.avinput -gff3attr -buildver hg19 hgmdgff
NOTICE: Reading annotation database hgmdgff/hgmd-hg19.gff ... Done with 161054 regions from 161054 GFF3 records
NOTICE: Finished processing 1000000 variants in queryfile
NOTICE: Finished processing 2000000 variants in queryfile
NOTICE: Finished processing 3000000 variants in queryfile
NOTICE: Finished region-based annotation on 3974643 genetic variants in cgexample.HG00731-200-37-ASM.avinput
NOTICE: Output file is written to cgexample.HG00731-200-37-ASM.avinput.hg19_gff3
```

**Step 3:** Search the output file (named `cgexample.HG00731-200-37-ASM.avinput.hg19_gff3` in this example) for annotated variants in the gene of your choice. In this example we have chosen to use `BRCA1` since the sample data is taken from a breast cancer cell line

```
$ egrep -w "hgnc=BRCA1" cgexample.HG00731-200-37-ASM.avinput.hg19_gff3
```

```

C:\cygdrive\i\annovar
1427728,19116388;pmid_notes=Not associated with breast cancer risk in BRCA1/2 mutation carriers,Meta-analysis of disease associa
tion..N/A;ref=G;rsid=rs3213245;snomedct=N/A;uniprot=P18887;variantType=DFP;feature=Increased lung cancer risk association with:
T>C;hyperlink=https://portal.biobase-international.com/hgmd/pro/mut.php?accession%3DCR063419 19 44079687 44079687
G A het 38

kar@sys-ekt108 /cygdrive/i/annovar
$ grep -B -h "hgnc=BRCA1" cgeexample.HG00731-200-37-ASM.avinput.hg19.gff3
gff3 ID=118552;accession=CR095669;alt=A;aminoacid_change=N/A;citation_type=Primary;codon_change=N/A;codon_number=N/A;comments
=[int.5711+421 G>T];confidence=High;disease=Reduced activity%2C association with:ensembl=ENSG00000012048;entrez=672;genomic_sequ
ence=TTTACTTCTCTAAACCTGTGTTACAAA(G/T)GCAGAGAGTCAGACCTTCAATGGAAGAG;hgmdAcc=CR095669;hgnc=BRCA1;hgvs=N/A;icd10=C50-C50.9,C00-
C97.9,C50.9,C76.1,C80,C50,N64.9,C00-D48.9,N63,N60-N64.9;lsdb_source=N/A;mesh=D001943,D009369,D012816,D004194,D001941,D013896,D01
3899;mutationType=R;nucleotideChange=G>T;omim=113705;omim_ref=MTHU019150,MTHU019150,601387,114480,MTHU000126;pmid=19405875;pmid_
notes=N/A;ref=C;rsid=rs8176318;snomedct=N/A;uniprot=P38398;variantType=DFP;feature=Reduced activity association with:G>T;hyperl
ink=https://portal.biobase-international.com/hgmd/pro/mut.php?accession%3DCR095669 17 41197274 41197274
A het 35
gff3 ID=118761;accession=CS045209;alt=T;aminoacid_change=N/A;citation_type=Primary,SAR;codon_change=N/A;codon_number=N/A;comm
ents=polymorphism? not found in 56 controls. familial breast cancer patient from Goa without additional PTC or missense mutation
s.;confidence=Low;disease=Breast cancer;ensembl=ENSG00000012048;entrez=672;genomic_sequence=acacctcagaattgcatattttacacctaac(g/a)tt
taaacacctaagggtttttgctgctgctga;hgmdAcc=CS045209;hgnc=BRCA1;hgvs=N/A;icd10=C50-C50.9,C50.9,C76.1,C50,N60-N64
.9,C80,N63,N64.9,C00-C97.9,C00-D48.9;lsdb_source=N/A;mesh=D013896,D013899,D001943,D009369,D001941,D012816;mutationType=S
;nucleotideChange=5152+66G>A;omim=113705;omim_ref=MTHU000126,114480,MTHU019150,MTHU017027,601387;pmid=15564800,26092435;pmid_not
es=Identified in apparently healthy individuals. Table 1..N/A;ref=C;rsid=rs3092994;snomedct=N/A;uniprot=P38398;variantType=DM;fe
ature=Breast cancer:5152+66G>A;hyperlink=https://portal.biobase-international.com/hgmd/pro/mut.php?accession%3DCS045209 17 4
1215825 41215825
C T het 26
gff3 ID=118899;accession=CM053798;alt=A;aminoacid_change=S>C;citation_type=Primary,FCR,SAR,SAR;codon_change=AGT-TGT;codon_num
ber=1613;comments=N/A;confidence=Low;disease=Ovarian cancer;ensembl=ENSG00000012048;entrez=672;genomic_sequence=CCCCAATTGAAGTTG
CAGAAATCTGCCAG(A/T)GTCCAGCTGCTGCTCATATCTACTGATCTG;hgmdAcc=CM053798;hgnc=BRCA1;hgvs=N/A;icd10=C58.9,C57.4,C56,C76.3,C80,C76.2,C57.9,N00-N99.9,C00-C97.9;lsdb_source=N/A;mesh=D010051,D010386,D005833,D005831,D000
008,D004194,D012816,D010049,D009369;mutationType=M;nucleotideChange=4837A>T;omim=113705;omim_ref=MTHU025028,167000;pmid=15617999
,18992264,21447777,26092435;pmid_notes=Functional analysis indicates likely neutral,computational classification of variants of
uncertain significance. Identified in apparently healthy individuals. Table 1..N/A;ref=T;rsid=rs1799966;snomedct=N/A;uniprot=P38
398;variantType=DM;feature=Ovarian cancer:4837A>T;hyperlink=https://portal.biobase-international.com/hgmd/pro/mut.php?accession%
3DCM053798;ID=118898;accession=CD119485;alt=C;aminoacid_change=N/A;citation_type=Primary;codon_change=N/A;codon_number=1612;com
ments=Descr. in Table S3 (online).;confidence=High;disease=Breast and/or ovarian cancer;ensembl=ENSG00000012048;entrez=672;genom
ic_sequence=N/A;hgmdAcc=CD119485;hgnc=BRCA1;hgvs=N/A;icd10=C58.9,C57.4,C56,C76.3,C80,C76.2,C57.9,N00-N99.9,C00-C97.9;lsdb_source=N/A;mesh=N/A;mutationType=D;nucle
otideChange=4837delA;omim=113705;omim_ref=N/A;pmid=21702907;pmid_notes=N/A;ref=C;rsid=rs397509199;snomedct=N/A;uniprot=P38398;v
ariantType=DM;feature=Breast and/or ovarian cancer:4837delA;hyperlink=https://portal.biobase-international.com/hgmd/pro/mut.php?
accession%3DCD119485;ID=118900;accession=CI045256;alt=TC;aminoacid_change=N/A;citation_type=Primary;codon_change=N/A;codon_num

```

## snpEff

**Step1:** Download the appropriate reference genome. In this example we are using the hg19 reference genome

```
$ java -jar snpEff.jar download -v GRCh37.75
```

```

KarthicL@MKT-KARTHICK /cygdrive/d/snpEff
00:00:00 $ java -jar snpEff.jar download -v GRCh37.75
00:00:00 SnpEff version SnpEff 4.3 (build 2016-06-14 18:42), by Pablo Cin
golani
00:00:00 Command: 'download'
00:00:00 Reading configuration file 'snpEff.config'. Genome: 'GRCh37.75'
00:00:00 Reading config file: D:\snpEff\snpEff.config
00:00:00 done
00:00:00 Downloading database for 'GRCh37.75'
00:00:00 Connecting to http://downloads.sourceforge.net/project/snpEff/data
bases/v4_3/snpEff_v4_3_GRCh37.75.zip
00:29:56 Local file name: 'C:\cygwin64\tmp\snpEff_v4_3_GRCh37.75.zip'
.....
00:30:03 Download finished. Total 662099902 bytes.
00:30:03 Extracting file 'data\GRCh37.75\regulation_CD4.bin'
00:30:03 Creating local directory: 'D:\snpEff\data\GRCh37.75'
00:30:03 Extracting file 'data\GRCh37.75\regulation_GM06990.bin'
00:30:17 Extracting file 'data\GRCh37.75\regulation_GM12878.bin'
00:30:17 Extracting file 'data\GRCh37.75\regulation_H1ESC.bin'

```

**Step 2:** Annotate the input VCF file with HGMD annotations using the `- interval` option in `snpEff` to accept the HGMD file as an annotation file. In this example `sample-hg00731.vcf` is the input file. The BED formatted HGMD file, named `hgmd-hg19.bed` in this example, is used as the database file

```
$ java -Xmx4g -jar snpEff.jar -v -interval hgmd-hg19.bed  
GRCh37.75 sample-hg00731.vcf
```

Input:

```
KarthicL@MKT-KARTHICK /cygdrive/d/snpEff
$ java -Xmx4g -jar snpEff.jar -v -interval hgmd_20161.bed GRCh37.75 test.vcf
00:00:00      SnpEff version SnpEff 4.3 (build 2016-06-14 18:42), by Pablo Cingolani
00:00:00      Command: 'ann'
00:00:00      Reading configuration file 'snpEff.config'. Genome: 'GRCh37.75'
00:00:00      Reading config file: D:\snpEff\snpEff.config
00:00:00      done
00:00:00      Reading database for genome version 'GRCh37.75' from file 'D:\snpEff\./data/GRCh37
00:00:00      .75/snpEffectPredictor.bin' (this might take a while)
00:00:24      done
00:00:24      Reading interval file 'hgmd_20161.bed'
00:00:25      done (161162 intervals loaded).
00:00:25      Loading Motifs and PWMs
00:00:25      Building interval forest
```

Output:

[illegible]

Alternatively, the VCF formatted HGMD file, named HGMD\_PRO\_2016.1\_hg19.vf in this example, can be used as the database file

```
$ java -Xmx4g -jar snpEff.jar -v -interval HGMD_PRO_2016.1_
hg19.vcf GRCh37.75 sample-hq00731.vcf
```

Input:

```
Karthik@MKT-KARTHIK /cydrive/d/snpEff
$ java -Xmx4g -jar snpEff.jar -v -interval HGMD_PRO_2016.1_hg19.vcf GRCh37.75 sample-hg00731.vcf
00:00:00 SnpEff version SnpEff 4.3 (build 2016-06-14 18:42), by Pablo Cingolani
00:00:00 Command: 'ann'
00:00:00 Reading configuration file 'snpEff.config'. Genome: 'GRCh37.75'
00:00:00 Reading config file: D:\snpEff\snpEff.config
00:00:00 done
00:00:00 Reading database for genome version 'GRCh37.75' from file 'D:\snpEff\data\GRCh37.75\snpEffectPredictor.bin' (this might take a while)
00:00:24 done
00:00:24 Reading interval file 'HGMD_PRO_2016.1_hg19.vcf'
00:00:24 done (161162 intervals loaded).
```

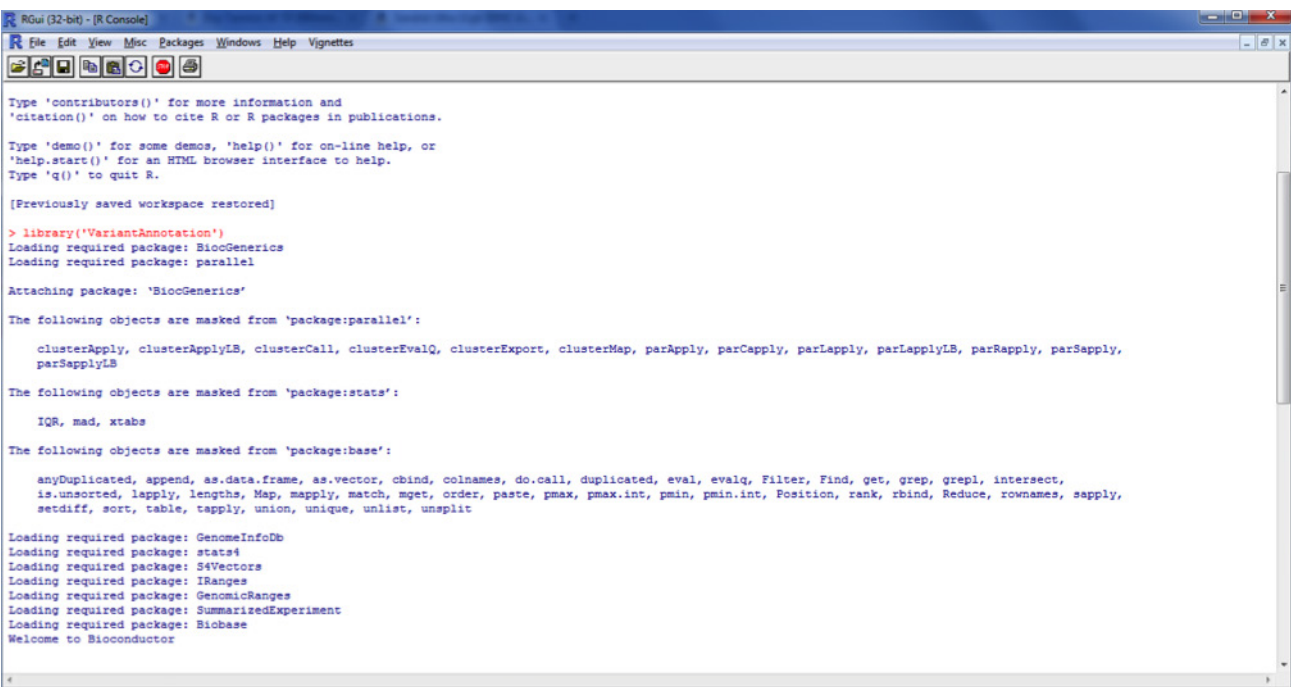
Output:

```
155178611 . C . NS=1;AN=0 GT:PS /./
155178655 . G . NS=1;AN=0 GT:PS /./
155178739 . G . NS=1;AN=0 GT:PS /./
155178764 . T . NS=1;AN=0 GT:PS /./
155178775 . CCGTGACA CCGTGACT NS=1;AN=1;AC=1;CA_XR=dbsnp.86;rs760077;CGA_FI=4580;NM_002455.3|MTX1|CDS|MISSENSE&4580;NM_002455.3|MTX1|CDS|NO-CHANGE&4580
NM_198883.2|MTX1|CDS|MISSENSE&4580;NM_198883.2|MTX1|CDS|NO-CHANGE&4580;NM_007112.3|THS3|TSS-UPSTREAM|UNKNOWN-INC;ANN=CCGTGACT|missense_variant|MODIFIER|MTX1|ENSG00000173171|transcript|ENST0000036837
protein_coding|1|8|c.187A>T|p.Trp63Ser|293/1632|187/1401|63/466|CCGTGACT|missense_variant|MODIFIER|MTX1|ENSG00000173171|transcript|ENST00000316721|protein_coding|1|7|c.187A>T|p.Trp63Ser|199/1441|118
7/1308|63/435|CCGTGACT|upstream_gene_variant|MODIFIER|THS3|ENSG00000169231|transcript|ENST00000368378|protein_coding|c.-1115T>A|11094|CCGTGACT|upstream_gene_variant|MODIFIER|THS3|ENSG000001692
31|transcript|ENST00000457183|protein_coding|c.-1115T>A|11074|CCGTGACT|upstream_gene_variant|MODIFIER|THS3|ENSG00000169231|transcript|ENST00000541990|protein_coding|c.-787T>A|11092|CCGTGACT
upstream_gene_variant|MODIFIER|THS3|ENSG00000169231|transcript|ENST0000418962|nonsense_mediated_decay|c.-1115T>A|11092|CCGTGACT|upstream_gene_variant|MODIFIER|MTX1|ENSG00000173171|transcript|EN
ST00000424959|nonsense_mediated_decay|c.-261A>T|1247|CCGTGACT|upstream_gene_variant|MODIFIER|MTX1|ENSG00000173171|transcript|ENST00000609421|protein_coding|c.-261A>T|1256|CCGTGACT|upstream_g
e_variant|MODIFIER|MTX1|ENSG00000173171|transcript|ENST00000481771|retained_intron|n.-260A>T|1260|CCGTGACT|upstream_gene_variant|MODIFIER|MTX1|ENSG00000173171|transcript|ENST00000495589|processed
transcript|n.-235T>A|111357|CCGTGACT|upstream_gene_variant|MODIFIER|MTX1|ENSG00000173171|transcript|ENST00000495492|retained_intron|n.-346A>T|13464|CCGTGACT|downstream_gene_variant|MODIFIER|
RP11-263K19.4|ENSG00000231064|transcript|ENST00000453136|antisense|n.-3496A>T|13496|CCGTGACT|downstream_gene_variant|MODIFIER|RP11-263K19.4|ENSG00000231064|transcript|ENST00000453788|antisense|n
.-2160T>A|112160|CCGTGACT|downstream_gene_variant|MODIFIER|GBAP1|ENSG00000160766|transcript|ENST00000459805|retained_intron|n.-4834T>A|114834|CCGTGACT|downstream_gene_variant|MODIFIER|RP11-263K1
9.4|ENSG00000231064|transcript|ENST00000422663|antisense|n.-3685A>T|113685|CCGTGACT|downstream_gene_variant|MODIFIER|RP11-263K19.4|ENSG00000231064|transcript|ENST00000430312|antisense|n.-3624A>T|1
13624|CCGTGACT|downstream_gene_variant|MODIFIER|GBAP1|ENSG00000160766|transcript|ENST00000486869|processed_transcript|n.-4834T>A|114834|CCGTGACT|downstream_gene_variant|MODIFIER|GBAP1|ENSG00000160766|transcript|ENST00000486197|retained_intron|n.-4835T>A|114835|CCGTGACT|non_coding_exon_variant|MODIFIER|THS3|ENSG00000169231|tra
nscript|ENST00000486260|processed_transcript|1|14|n.61T>A|11111|CCGTGACT|cutoff|MODIFIER|XUS102|HGMD_PRO_2016|CM111601|n.155178782A>T|11111|GT:PS:FT=HQ;EHQ:CGA_CEQ:GL:CGA_CEQ:DP:AD:CGA_RDP
155178775 VQLOW:23,,122,,1,,1-23,0,0:1,0,0:21,,1,0 NS=1;AN=0 GT:PS /./
155178785 . NS=1;AN=0 GT:PS /./
155178789 . GCGCGCCGCGCC . NS=1;AN=0 GT:PS /./
```

## Variant Annotation – a Bioconductor package

**Step1:** Install the VariantAnnotation package from Bioconductor

```
> library ('VariantAnnotation')
```



```
RGui (32-bit) - (R Console)
File Edit View Misc Packages Windows Help Vignettes

Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> library('VariantAnnotation')
Loading required package: BioGenerics
Loading required package: parallel

Attaching package: 'BioGenerics'

The following objects are masked from 'package:parallel':

  clusterApply, clusterApplyLB, clusterCall, clusterEvalQ, clusterExport, clusterMap, parApply, parCapply, parLapply, parLapplyLB, parRapply, parSapply,
  parSapplyLB

The following objects are masked from 'package:stats':

  IQR, mad, xtabs

The following objects are masked from 'package:base':

  anyDuplicated, append, as.data.frame, as.vector, cbind, colnames, do.call, duplicated, eval, evalq, Filter, Find, get, grep, grepl, intersect,
  is.unsorted, lapply, lengths, Map, mapply, match, mget, order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rownames, sapply,
  setdiff, sort, table, tapply, union, unique, unlist, unsplit

Loading required package: GenomeInfoDb
Loading required package: S4Vectors
Loading required package: IRanges
Loading required package: GenomicRanges
Loading required package: SummarizedExperiment
Loading required package: Biobase
Welcome to Bioconductor
```

**Step 2:** Upload the input vcf file using the “readVcf” function. In this example sample-hg00731.vcf is the input file

```
> vcf <- readVcf("D:/sample-hg00731.vcf", "hg19")
```

```
> vcf <- readVcf("D:/sample-hg00731.vcf", "hg19" )
> vcf
class: CollapsedVCF
dim: 499882 1
rowRanges(vcf):
  GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
info(vcf):
  DataFrame with 21 columns: NS, AN, AC, CGA_XR, CGA_FI, CGA_PFAM, CGA_MIRB, CGA_RPT, CGA_SDO, END, CGA_
info(header(vcf)):
  Number Type Description
  NS 1 Integer Number of Samples With Data
  AN 1 Integer Total number of alleles in called genotypes
  AC A Integer Allele count in genotypes, for each ALT allele
  CGA_XR A String Per-ALT external database reference (dbSNP, COSMIC, etc)
  CGA_FI A String Functional impact annotation
  CGA_PFAM . String PFAM Domain
  CGA_MIRB . String miRBaseId
  CGA_RPT . String repeatMasker overlap information
  CGA_SDO 1 Integer Number of distinct segmental duplications that overlap this locus
  END 1 Integer End position of the variant described in this record
  CGA_WINEND 1 Integer End of coverage window
  CGA_BF 1 Float Frequency in baseline
  CGA_MEDEL 4 String Consistent with deletion of mobile element; type,chromosome,start,end
  MATEID 1 String ID of mate breakend
  SVTYPE 1 String Type of structural variant
  CGA_BNDG A String Transcript name and strand of genes containing breakend
  CGA_BNDGO A String Transcript name and strand of genes containing mate breakend
  CIPOS 2 Integer Confidence interval around POS for imprecise variants
  IMPRECISE 0 Flag Imprecise structural variation
  MEINFO 4 String Mobile element info of the form NAME,START,END,POLARITY
  SVLEN . Integer Difference in length between REF and ALT alleles
geno(vcf):
  SimpleList of length 33: GT, PS, SS, FT, GQ, HQ, EHQ, CGA_CEQ, GL, CGA_CGL, DP, AD, CGA_RDP, CGA_GP,
geno(header(vcf)):
  Number Type Description
  GT 1 String Genotype
  PS 1 Integer Phase Set
  SS 1 String Somatic Status: Germline, Somatic, LOH, or . (Unknown)
  FT 1 String Genotype filters
```

**Step 3:** Upload the HGMD annotations using the “readVcf” function. The VCF formatted HGMD file (named HGMD\_PRO\_2016.1\_hg19.vcf in this example) is used as the database file

```
> hgmd <- readVcf("D:/HGMD_PRO_2016.1_hg19.vcf", "hg19")
```

**Step 4:** Optionally filter the HGMD annotations by their location within or relative to a gene using the locateVariants function and the UCSC HG19 genomic coordinates package specified as txdb. Regions are specified in the region argument and can be one of the following: CodingVariants, IntronVariants, FiveUTRVariants, ThreeUTRVariants,

IntergenicVariants, SpliceSiteVariants or PromoterVariants.

Here we show an example specifying variants located within coding regions

```
> loc <- locateVariants(rowRanges(hgmd), txdb, codingVariants())
```

```
> loc <- locateVariants(rowRanges(hgmd), txdb, CodingVariants())
'select()' returned many:1 mapping between keys and columns
> loc
GRanges object with 443700 ranges and 9 metadata columns:
      seqnames      ranges      strand | LOCATION LOCSTART  LOCEND  QUERYID    TXID      CDSID    GENEID    PRECEDEID    FOLLOWID
      <Rle>      <IRanges> <Rle> | <factor> <integer> <integer> <integer> <character> <IntegerList> <character> <CharacterList> <CharacterList>
1      chr1 [877523, 877523] + | coding      877      877      1      22      28      148398
2      chr1 [877523, 877523] + | coding      832      832      1      23      28      148398
3      chr1 [877523, 877523] + | coding      880      880      1      24      28      148398
4      chr1 [877523, 877523] + | coding      829      829      1      26      28      148398
5      chr1 [877523, 877523] + | coding      274      274      1      29      28      148398
...
443696 chrY [16952726, 16952726] + | coding      1531     1531     161162     78460     226890     22829
443697 chrY [16952726, 16952726] + | coding      2095     2095     161162     78461     226890     22829
443698 chrY [16952726, 16952726] + | coding      1114     1114     161162     78462     226890     22829
443699 chrY [16952726, 16952726] + | coding      2035     2035     161162     78463     226890     22829
443700 chrY [16952726, 16952726] + | coding      2035     2035     161162     78464     226890     22829
-----
seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

And an example specifying variants located within promoter regions

```
> loc <- locateVariants(rowRanges(hgmd), txdb, PromoterVariants())
```

```
> loc <- locateVariants(rowRanges(hgmd), txdb, PromoterVariants())
'select()' returned many:1 mapping between keys and columns
> loc
GRanges object with 38593 ranges and 9 metadata columns:
      seqnames      ranges      strand | LOCATION LOCSTART  LOCEND  QUERYID    TXID      CDSID    GENEID    PRECEDEID    FOLLOWID
      <Rle>      <IRanges> <Rle> | <factor> <integer> <integer> <integer> <character> <IntegerList> <character> <CharacterList> <CharacterList>
(1)    chr1 [1167659, 1167659] - | promoter <NA>      <NA>      16      74      126792
(2)    chr1 [1167659, 1167659] - | promoter <NA>      <NA>      16      4140     51150
(3)    chr1 [1167659, 1167659] - | promoter <NA>      <NA>      16      4141     51150
(4)    chr1 [1167659, 1167659] - | promoter <NA>      <NA>      16      4142     51150
(5)    chr1 [1167674, 1167674] + | promoter <NA>      <NA>      17      74      126792
...
[38589] chrY [2655637, 2655637] - | promoter <NA>      <NA>      161154     78581     6736
[38590] chrY [2655638, 2655639] - | promoter <NA>      <NA>      161155     78581     6736
[38591] chrY [2655641, 2655641] - | promoter <NA>      <NA>      161156     78581     6736
[38592] chrY [2655719, 2655719] - | promoter <NA>      <NA>      161157     78581     6736
[38593] chrY [2655774, 2655774] - | promoter <NA>      <NA>      161158     78581     6736
-----
seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

**Step 5:** Annotate the input VCF file with HGMD annotations using the subsetByOverlaps function. In this example, vcf is the previously uploaded input file and hgmd is the previously uploaded HGMD annotations

```
> out <- subsetByOverlaps(hgmd, vcf)
```

```

> out<-subsetByOverlaps(hgmd,vcf)
> out
class: CollapsedVCF
dim: 200 0
rowRanges(vcf):
  GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
info(vcf):
  DataFrame with 8 columns: CLASS, MUT, GENE, STRAND, DNA, PROT, DB, PHEN
info(header(vcf)):
  Number Type Description
  CLASS 1 String Mutation Category, https://portal.biobase-international.com/hgmd/pro/global.php#cats
  MUT 1 String HGMD mutant allele
  GENE 1 String Gene symbol
  STRAND 1 String Gene strand
  DNA 1 String DNA annotation
  PROT 1 String Protein annotation
  DB 1 String dbSNP identifier, build 137
  PHEN 1 String HGMD primary phenotype
geno(vcf):
  SimpleList of length 0:
> |

```

**Step 6:** View the output. Use the `info(out)` command to view the HGMD annotations

> `info(out)`

```

> info(out)
DataFrame with 200 rows and 8 columns
  CLASS MUT GENE STRAND DNA PROT DB
  <character> <character> <character> <character> <character> <character> <character>
CI148519 DM ALT AGRN + NM_198576.3:c.1362dupC NA NA
CS060109 DP ALT TNFRSF4 - NM_003327.3:c.634+25C>T NA rs2298212
CM134937 DM ALT B3GALT6 + NM_080605.3:c.649G>A NP_542172.2:p.G217S rs397514724
CM1411605 DM ALT B3GALT6 + NM_080605.3:c.766C>T NP_542172.2:p.R256W NA
BM1422338 DM ALT B3GALT6 + NM_080605.3:c.795A>C NP_542172.2:p.E265D rs374677519
... ... ... ... ...
CX941936 DM ALT GBA - NM_001005741.2:c.1447_1466delCTGGACGACGAGTGGCACTGATinsTG NA NA
CM940819 DM ALT GBA - NM_001005741.2:c.1448T>G NP_001005741.1:p.L483R NA
CM870010 DM ALT GBA - NM_001005741.2:c.1448T>C NP_001005741.1:p.L483P rs421016
CM001167 DM ALT GBA - NM_001005741.2:c.685G>A NP_001005741.1:p.A229T NA
CD050144 DM ALT LMNA + NM_170707.3:c.-3_12delGCCATGGAGACCCCG NA rs267607546
  PHEN
  <character>
CI148519 "Congenital_myasthenic_syndrome_with_distal_muscle_weakness_4_atrophy"
CS060109 "Myocardial_infarction_protection_against_association"
CM134937 "Ehlers-Danlos_syndrome-like"
CM1411605 "Spondyloepimetaphyseal_dysplasia_with_joint_laxity"
BM1422338 "Al-Gazali_syndrome"
... ...
CX941936 "Gaucher_disease"
CM940819 "Gaucher_disease"
CM870010 "Gaucher_disease_2"
CM001167 "Gaucher_disease_3"
CD050144 "Muscular_dystrophy_Emer-Dreifuss_neurogenic"
> |

```

Use the `rowRanges(out)` command to show the genomic coordinate information for the mutations

> `rowRanges(out)`

```
> rowRanges(out)
GRanges object with 200 ranges and 5 metadata columns:
      seqnames      ranges strand | paramRangeID      REF      ALT      QUAL      FILTER
      <Rle>        <IRanges> <Rle> | <factor>    <DNAStringSet> <DNAStringSetList> <numeric> <character>
      CI148519      1 [ 977516, 977516] * | <NA>        T          TC          <NA>      .
      CS060109      1 [1147297, 1147297] * | <NA>        G          A          <NA>      .
      CM134937      1 [1168307, 1168307] * | <NA>        G          A          <NA>      .
      CM1411605     1 [1168424, 1168424] * | <NA>        C          T          <NA>      .
      BM1422338     1 [1168453, 1168453] * | <NA>        A          C          <NA>      .
      ...          ...          ...          ...          ...          ...          ...          ...
      CX941936      1 [155205024, 155205044] * | <NA>        CATCAGTGGCCACTGCGTCCAG CCA <NA>      .
      CM940819      1 [155205043, 155205043] * | <NA>        A          C          <NA>      .
      CM870010      1 [155205043, 155205043] * | <NA>        A          G          <NA>      .
      CM001167      1 [155208001, 155208001] * | <NA>        C          T          <NA>      .
      CD050144      1 [156084703, 156084718] * | <NA>        GCCGGCCATGGAGACC      G <NA>      .
      -----
      seqinfo: 24 sequences from hg19 genome; no seqlengths
> rowRanges(out1)
GRanges object with 184 ranges and 5 metadata columns:
      seqnames      ranges strand | paramRangeID      REF      ALT      QUAL      FILTER
      <Rle>        <IRanges> <Rle> | <factor>    <DNAStringSet> <DNAStringSetList> <numeric> <character>
      1:977510_GTGCCAT/. 1 [ 977510, 977516] * | <NA>
      1:1147297_G/A      1 [1147297, 1147297] * | <NA>
      1:1168306_CG/.     1 [1168306, 1168307] * | <NA>
      1:1168406_GCGCCGGTGGAGCTCCAGCGGGAGCAGACCCGCGCTTCGACACCGAATACCG/. 1 [1168406, 1168458] * | <NA>
      1:1265154_T/C      1 [1265154, 1265154] * | <NA>
      ...
      1:155106697_G/A     1 [155106697, 155106697] * | <NA>
      1:155178775_CCGTGACA/CCGTGACT 1 [155178775, 155178782] * | <NA>
      1:155205043_A/.     1 [155205043, 155205043] * | <NA>
      1:155208001_C/<CGA_CNVWIN> 1 [155208001, 155208001] * | <NA>
      1:156084704_C/.     1 [156084704, 156084704] * | <NA>
```

## Obtaining access to HGMD

For more information, or to obtain a quote for a license to HGMD data for use in any of the tools profiled in this technical note, contact [bioinformaticssales@qiagen.com](mailto:bioinformaticssales@qiagen.com).

**EMEA**  
 Silkeborgvej 2 · Prismet  
 8000 Aarhus C  
**Denmark**  
**Phone:** +45 8082 0167  
**E-mail:** [bioinformaticssales@qiagen.com](mailto:bioinformaticssales@qiagen.com)

**Americas**  
 1001 Marshall Street, Suite 200, Redwood City  
 CA 94063  
**USA**  
**Phone:** +1 650 381 5111 or **Toll Free:** +1 866 464 3684  
**E-mail:** [bioinformaticssales@qiagen.com](mailto:bioinformaticssales@qiagen.com)