

## TABLE OF CONTENTS

<b>1 INTRODUCTION.....</b>	<b>1</b>
<b>2 INSTALLATION.....</b>	<b>1</b>
<b>3 USAGE OF ANNOVAR.....</b>	<b>2</b>
<b>3.1 PREPARE DATABASE .....</b>	<b>2</b>
<b>3.2 RUN ANNOVAR ON THE VCF FILE.....</b>	<b>2</b>
<b>3.3 OUTPUT FILE INTERPRETATION .....</b>	<b>3</b>
<b>4.0 WEB-BASED ANNOVAR - WANNOVAR .....</b>	<b>7</b>
<b>4.1 HOW TO USE WANNOVAR .....</b>	<b>8</b>
<b>4.2 RESULT OF WANNOVAR .....</b>	<b>8</b>

## 1 Introduction

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19, hg38, as well as mouse, worm, fly, yeast, and many others). Given a list of variants with chromosome, start position, end position, reference nucleotide and observed nucleotides, ANNOVAR can perform:

- **Gene-based annotation:** identify whether SNPs or CNVs cause protein coding changes and the amino acids that are affected. Users can flexibly use RefSeq genes, UCSC genes, ENSEMBL genes, GENCODE genes, AceView genes, or many other gene definition systems.
- **Region-based annotation:** identify variants in specific genomic regions, for example, conserved regions among 44 species, predicted transcription factor binding sites, segmental duplication regions, GWAS hits, database of genomic variants, DNase I hypersensitivity sites, ENCODE H3K4Me1/H3K4Me3/H3K27Ac/CTCF sites, ChIP-Seq peaks, RNA-Seq peaks, or many other annotations on genomic intervals.
- **Filter-based annotation:** identify variants that are documented in specific databases, for example, whether a variant is reported in dbSNP, what is the allele frequency in the 1000 Genome Project, NHLBI-ESP 6500 exomes or Exome Aggregation Consortium (ExAC) or Genome Aggregation Database (gnomAD), calculate the SIFT/PolyPhen/LRT/MutationTaster/MutationAssessor/FATHMM/MetaSVM/MetaLR scores, find intergenic variants with GERP++ score<2 or CADD>10, or many other annotations on specific mutations.
- **Other functionalities:** Retrieve the nucleotide sequence in any user-specific genomic positions in batch, identify a candidate gene list for Mendelian diseases from exome data, and other utilities.

## 2 Installation

ANNOVAR is written in Perl and can be run as a standalone application on diverse hardware systems where standard Perl modules are installed. It is freely available to personal, academic and non-profit

use only. You cannot redistribute ANNOVAR to other users including lab members. No liability for software usage is assumed.

Download here: [https://www.openbioinformatics.org/annovar/annovar\\_download\\_form.php](https://www.openbioinformatics.org/annovar/annovar_download_form.php)

## 3 Usage of ANNOVAR

One of the functionalities of ANNOVAR is to generate gene-based annotation. For example, from a whole-genome sequencing experiment on a human subject, given a list of 4 million SNVs (single nucleotide variants) and 0.5 million indels (insertions or deletions), it is of interest to identify the genes that are disrupted. For intergenic variants, we are interested in knowing what the two flanking genes are, and what are the distances between the variants and the flanking genes. For exonic variants, we are interested in knowing the amino acid changes.

### 3.1 Prepare Database

Before working on gene-based annotation, a gene definition file and associated FASTA file must be downloaded into a directory if they are not already downloaded. Let's call this directory as humandb/.

Usage:

```
annotate_variation.pl -downdb -buildver hg19 -webfrom annovar refGene humandb/
```

### 3.2 Run ANNOVAR on the VCF file

For beginners, the easiest way to use ANNOVAR is to use the table\_annovar.pl program. This program takes an input variant file (such as a VCF file) and generate a tab-delimited output file with many columns, each representing one set of annotations.

Usage:

```
perl table_annovar.pl VCF_files/example.vcf -buildver hg19 humandb -out example -  
remove -protocol refGene,exac03,avsnp147,dbnsfp30a -operation g,f,f,f -nastring .  
-csvout -polish -vcfinput
```

We can examine the command line in greater detail.

- The -protocol argument indicate the type databases you wish to include in the output.
- 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, r means region-based and f means filter-based (refer to section 1.0). If you do not provide a xref file, then the operation can be g only.
- Sometimes, users want tab-delimited files rather than comma-delimited files. This can be easily done by removing -csvout argument to the above command.
- The -vcfinput argument specify that the input in a VCF file.

The table\_annovar.pl program uses ExAC version 0.3 (referred to as exac03) dbNSFP version 3.0a (referred to as dbnsfp30a), dbSNP version 147 with left-normalization (referred to as avsnp147)

databases and remove all temporary files and generates the output file called example.hg19\_multianno.txt. Database to be included can be modified based on your needs. For more selections, please refer: <https://annovar.openbioinformatics.org/en/latest/user-guide/download/>

### 3.3 Output file interpretation

The output file (example.hg19\_multianno.txt) contains multiple columns. The first a few columns are your input column. Each of the following columns corresponds to one of the "protocol" that user specified in the command line. The Func.refGene, Gene.refGene, GeneDetail.refGene, ExonicFunc.refGene, AChange.refGene columns contain various annotation on how the mutations affect gene structure:

The first column tells whether the variant hit exons or hit intergenic regions, or hit introns, or hit a non-coding RNA gene. If the variant is exonic/intronic/ncRNA, the second column gives the gene name (if multiple genes are hit, comma will be added between gene names); if not, the second column will give the two neighbouring genes and the distance to these neighbouring genes. The possible values of the first column are summarized below:

Value	Default precedence	Explanation	Sequence Ontology
exonic	1	variant overlaps a coding	exon_variant (SO:0001791)
splicing	1	variant is within 2-bp of a splicing junction (use -splicing_threshold to change this)	splicing_variant (SO:0001568)
ncRNA	2	variant overlaps a transcript without coding annotation in the gene definition (see Notes below for more explanation)	non_coding_transcript_variant (SO:0001619)
UTR5	3	variant overlaps a 5' untranslated region	5_prime_UTR_variant (SO:0001623)
UTR3	3	variant overlaps a 3' untranslated region	3_prime_UTR_variant (SO:0001624)
intronic	4	variant overlaps an intron	intron_variant (SO:0001627)
upstream	5	variant overlaps 1-kb region upstream of transcription start site	upstream_gene_variant (SO:0001631)
downstream	5	variant overlaps 1-kb region downstream of transcription end site (use -neargene to change this)	downstream_gene_variant (SO:0001632)
intergenic	6	variant is in intergenic region	intergenic_variant (SO:0001628)

The second field tells the functional consequences of the variant (possible values in this field include: nonsynonymous SNV, synonymous SNV, frameshift insertion, frameshift deletion, nonframeshift insertion, nonframeshift deletion, frameshift block substitution, nonframeshift block substitution).

The third column contains the gene name, the transcript identifier, and the sequence change in the corresponding transcript. A standard nomenclature is used in specifying the sequence changes (you may want to add -hgvs argument so that the cDNA level annotation is compatible with HGVS nomenclature)

Annotation	Precedence	Explanation	Sequence Ontology
frameshift insertion	1	an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence	frameshift_elongation (SO:0001909)
frameshift deletion	2	a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence	frameshift_truncation (SO:0001910)
frameshift block substitution	3	a block substitution of one or more nucleotides that cause frameshift changes in protein coding sequence	frameshift_variant (SO:0001589)
stopgain	4	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate creation of stop codon at the variant site. For frameshift mutations, the creation of stop codon downstream of the variant will not be counted as "stopgain"!	stop_gained (SO:0001587)
stoploss	5	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate elimination of stop codon at the variant site	stop_lost (SO:0001578)
nonframeshift insertion	6	an insertion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence	inframe_insertion (SO:0001821)
nonframeshift deletion	7	a deletion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence	inframe_deletion (SO:0001822)
nonframeshift block substitution	8	a block substitution of one or more nucleotides that do not cause frameshift changes in protein coding sequence	inframe_variant (SO:0001650)
nonsynonymous SNV	9	a single nucleotide change that cause an amino acid change	missense_variant (SO:0001583)
synonymous SNV	10	a single nucleotide change that does not cause an amino acid change	synonymous_variant (SO:0001819)
unknown	11	unknown function (due to various errors in the gene structure definition in the database file)	sequence_variant (SO:0001060)

For more gene-based annotation explanation, please refer:

<https://annovar.openbioinformatics.org/en/latest/user-guide/gene/>

The ExAC\* columns represent allele frequency in all the samples as well as sub-populations in the Exome Aggregation Consortium data sets while the avsnp147 means the SNP identifier in the dbSNP version 147. There are many more version of allele frequency and SNP databases, please refer to: <https://annovar.openbioinformatics.org/en/latest/user-guide/download/>

The other columns contain prediction scores for non-synonymous variants using several widely used tools, including 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. The dbNSFP database is referred to as LJB database, and we used to provide separate files for each individual scores. The description below refers to ljb23. They are helpful for users who only want to infer individual scores for individual prediction method. Detailed information for all the dbNSFP databases are given below:

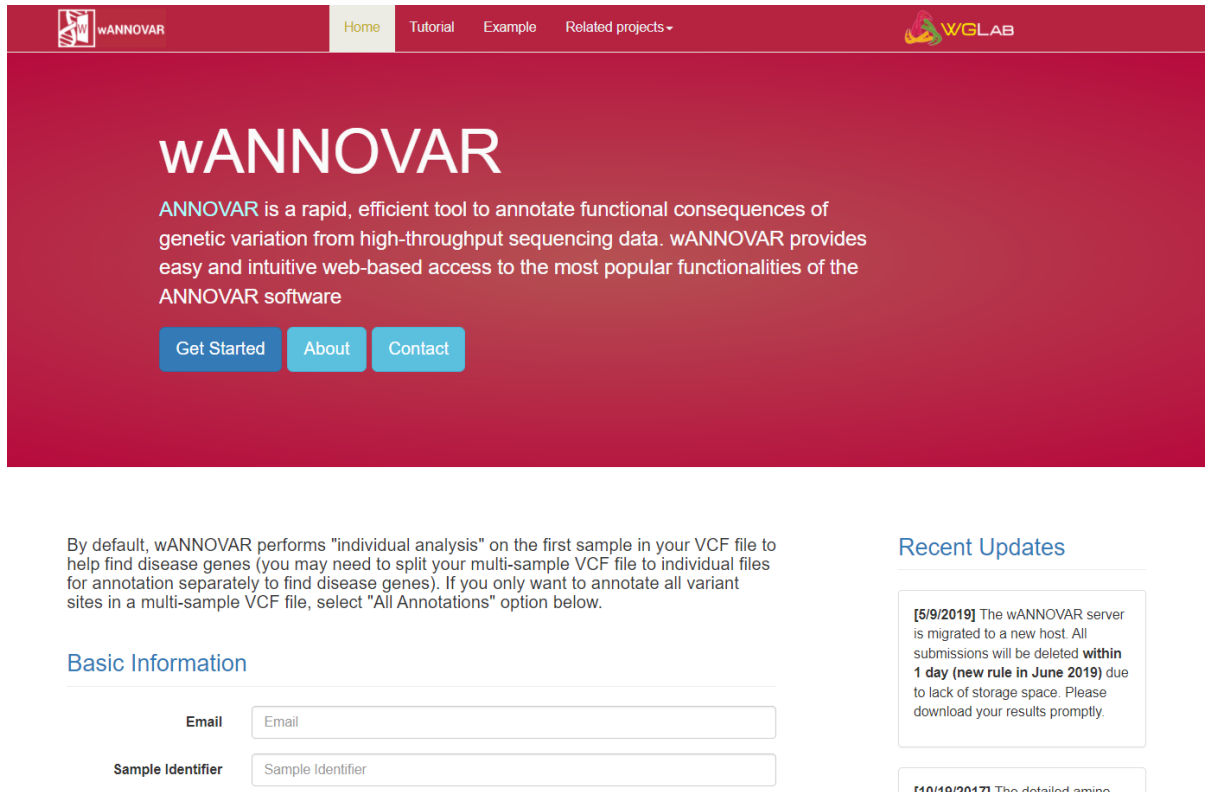
Score (dbtype)	# variants in LJB23 build hg19	Categorical Prediction
SIFT (sift)	77593284	D: Deleterious (sift<=0.05); T: tolerated (sift>0.05)
PolyPhen 2 HDIV (pp2_hdiv)	72533732	D: Probably damaging (>=0.957), P: possibly damaging (0.453<=pp2_hdiv<=0.956); B: benign (pp2_hdiv<=0.452)
PolyPhen 2 HVar (pp2_hvar)	72533732	D: Probably damaging (>=0.909), P: possibly damaging (0.447<=pp2_hdiv<=0.909); B: benign (pp2_hdiv<=0.446)
LRT (lrt)	68069321	D: Deleterious; N: Neutral; U: Unknown
MutationTaster (mt)	88473874	A" ("disease_causing_automatic"); "D" ("disease_causing"); "N" ("polymorphism"); "P" ("polymorphism_automatic")
MutationAssessor (ma)	74631375	H: high; M: medium; L: low; N: neutral. H/M means functional and L/N means non-functional
FATHMM (fathmm)	70274896	D: Deleterious; T: Tolerated
MetaSVM (metasvm)	82098217	D: Deleterious; T: Tolerated
MetaLR (metalr)	82098217	D: Deleterious; T: Tolerated
GERP++ (gerp++)	89076718	higher scores are more deleterious
PhyloP (phylop)	89553090	higher scores are more deleterious
SiPhy (siphy)	88269630	higher scores are more deleterious

For more filter-based annotation explanation, please refer:

<https://annovar.openbioinformatics.org/en/latest/user-guide/filter/>

## 4.0 Web-based ANNOVAR - wANNOVAR

An average biologist who do not want to download and install ANNOVAR software tools can easily submit a list of mutations (even whole-genome variants calls) to the web server (<https://wannovar.wglab.org/>), select the desired annotation categories, and receive functional annotation back by emails.



The screenshot shows the wANNOVAR web interface. At the top is a navigation bar with links: Home, Tutorial, Example, and Related projects. The main heading is "wANNOVAR" with a description: "ANNOVAR is a rapid, efficient tool to annotate functional consequences of genetic variation from high-throughput sequencing data. wANNOVAR provides easy and intuitive web-based access to the most popular functionalities of the ANNOVAR software". Below this are three buttons: "Get Started", "About", and "Contact".

By default, wANNOVAR performs "individual analysis" on the first sample in your VCF file to help find disease genes (you may need to split your multi-sample VCF file to individual files for annotation separately to find disease genes). If you only want to annotate all variant sites in a multi-sample VCF file, select "All Annotations" option below.

**Basic Information**

Email

Sample Identifier

**Recent Updates**

**[5/9/2019]** The wANNOVAR server is migrated to a new host. All submissions will be deleted **within 1 day (new rule in June 2019)** due to lack of storage space. Please download your results promptly.

**[10/19/2017]** The detailed amino...

By default, wANNOVAR performs "individual analysis" on the first sample in your VCF file to help find disease genes (you may need to split your multi-sample VCF file to individual files for annotation separately to find disease genes). If you only want to annotate all variant sites in a multi-sample VCF file, select "All Annotations" option below.

## 4.1 How to use wANNOVAR

### Basic Information

Email

Email Provide email address

Sample Identifier

Sample Identifier Name your sample

Input File

+ Input File Upload VCF file here

or Paste Variant Calls

paste your variant call here

Submit

Reset

Monitor Progress

☒ I agree to the [Terms of Use](#) . Please note that commerical users would need to obtain a license.

### Disease/Phenotype

Enter Disease or Phenotype Terms

please enter your focused disease/phenotype terms

Please use semicolon or enter as separators. Like "alzheimer;brain".  
 Try to use multiple terms instead of a super long term  
 OMIM IDs are also accepted, like 114480 for 'Breast cancer'  
 Better Combined with wANNOVAR's disease model.

### Parameter Settings

Result duration

1 day

Q

Reference Genome

hg19

Select correct version of reference genome

Q

Input Fomat

VCF

Q

Gene Definition

RefSeq Gene

Select preferred database

Q

Individual analysis

Individual analysis

Q

Disease Model

none

Q

## 4.2 Result of wANNOVAR

When done, click Submit and wANNOVAR will the annotation and result to you via email. Interpretation of wANNOVAR output is the same as ANNOVAR. For output example, please refer: <https://wannovar.wglab.org/example.html>

-End-