

1. How to report a bug or to ask a question?

The best way is: First, do read the website thoroughly especially FAQ. I received too many emails whose answers can easily be found in FAQ. Second, if you cannot find answer in FAQ or do not understand the answer well, then drop me an email as which should contain (1) command line argument (2) error message in screen or the LOG file (3) sometimes example inputfile (4) in case you use Mac or Windows, let me know. The reason is that I can fix something or diagnose something only if I can understand the question and reproduce the results. So do yourself a favor and do me a favor, include details in your email to avoid wasting our mutual time sending multiple emails.

There is no such thing as "ANNOVAR development team", as I am the only person who reply user emails, address user questions, and fix bugs. As of April 2015, I have communicated over 13,399 emails with ANNOVAR users. If you read FAQ #1 before sending me an email, it will save both of us a lot of valuable time.

2. How to annotate variants in a VCF file?

The easiest way is to use `table_annovar.pl`: just add `-vcfinput` argument and supply a VCF file as input file, and your output file will be in VCF format with INFO field populated with ANNOVAR annotations that you have specified in `-protocol` argument. One additional output file called `*multianno.txt` will be in tab-delimited text format for easier manual examination in Excel or other programs.

It is also possible to handle VCF file manually when retrieving a subset of records from VCF file without altering its content. For example, I want to find out all novel variants (not in dbSNP135 and not in 1000G and not in NHLBI-ESP5400) in a VCF file, but without changing the VCF format. This can be done using `convert2annovar.pl` with the `-includeinfo` argument, so that you convert VCF file to ANNOVAR inputfile without losing any VCF-specific information. Then annotate the inputfile by a series of filter operation, then convert the outputfile to VCF file using the `cut -f 3-` command in Linux system.

3. Why I cannot download the databases listed in your download page?

What is your command line? Did you add `"-webfrom annovar"`?

4. How to find frequency information from 1000 Genomes Project data?

The instructions were described in this page ([../user-guide/filter/](#)). But one important thing to emphasize is that due to historical reasons, one must use something like `-dbtype 1000g2015aug_all` (or `-dbtype 1000g2015aug_eur` for European population and `-dbtype 1000g2015aug_afr` for African population), not `-dbtype ALL.sites.2014_10` for annotation.

5. How to annotate copy number variations (CNV)?

The REF and ALT in the input file can be 0. You can then annotate the file by gene-based and region-based annotation.

6. What is the difference between vcf4 and vcf4old format in convert2annovar.pl?

In August 2013, I changed the VCF4 conversion subroutine in `convert2annovar.pl`, but I kept the vcf4old format for users who like the "old-fashion" conversion. The difference is that nowadays people tend to do multi-sample calling or candidate variant calling, so that the variants listed in the VCF4 file do

not necessarily have mutations for a specific sample. This happens when genotype call is 0/0 (reference/reference). I got some complaints from users about the inability to process multi-sample VCF files, so I decided to make this change.

By default "vcf4" will only process the first sample, and will only print out mutations that exist in the first sample. So if you have a multi-sample VCF file, then usually only a subset of lines will exist in the output file. The `-format vcf4` can be combined with `-allsample` argument, which will print out a separate output file for each sample in the VCF4 file (again by default, only the first sample in the VCF4 file will be processed). More importantly, if you use `-format vcf4 -allsample -withfreq`, then all input lines from VCF will be kept in output lines, yet an allele frequency measure is included in each line calculating the frequency of each variant among all the samples in the VCF file.

In general, `-format vcf4old` should be considered as obsolete and should not be used by most users, since `-format vcf4` can now accomplish everything that `-format vcf4old` can do with appropriate combinations of arguments.

7. How to back convert cDNA coordinate such as c.385A>G to genomic coordinate such as chr1:123456A>G?

Read "all variants in a transcript" section from this page ([../user-guide/input/](#)).

8. Why my run of gene-based annotation differ slightly from those shown in website?



UCSC database updates constantly and ANNOVAR executable also updates constantly, so it is expected that ANNOVAR output format or the annotations may change slightly over time.

9. Why the gene name from ANNOVAR output is wrong?

The official gene symbol for human genome is maintained by HGNC, and they change gene name in a constant basis. Every other database tries to synchronize with HGNC, but there is usually a delay. ANNOVAR annotation uses gene name defined in RefSeq (default) or Ensembl or UCSC Gene or GENCODE, so they may differ from the "official" gene symbol in rare occasions. Similarly, OMIM and other clinical databases will also use names that differ from "official" names, depending on how updated they are. For example, if you use early 2016 version of ANNOVAR's RefSeq gene annotation, the CASC5 gene will be there, but in late 2016, this gene was renamed as KNL1 in RefSeq. Similarly, the gene is called CASC5 in OMIM, with an annotation that "HGNC Approved Gene Symbol: KNL1" in OMIM records.

To make sure that you capture all OMIM genes in your results, you will have to maintain a gene name table that has both OMIM gene names and the official HGNC gene names for those OMIM genes, and then search result files generated by ANNOVAR. This is because depending on the date/version/source of ANNOVAR's database, different types of gene names could be in the output file.

10. Why ANNOVAR produced different non-synonymous SNP annotations than another software?

For example, ANNOVAR may report a mutation as W185R mutation, but another software may report the same mutation as R285W mutation. This could be due to a variety of reasons: (1) the use of different gene-definition systems. Depending on your command line argument, ANNOVAR always use the latest refGene, knownGene or ensGene to ensure that the information is up to date. You should check what gene definition system is used by the other annotation software. (2) Even if both software tools are using Ensembl, they could be using different versions of the gene definition. (3) ANNOVAR automatically  latest  excludes any transcript in gene definition file that does not have a complete coding sequence or has a premature stop codon (since this means the protein annotation is wrong). Each gene definition (especially Ensembl) has a lot of such transcripts. (4) ANNOVAR uses precedence rules, so if a variant is

intronic for one transcript but coding for another transcript, it will be reported as coding only. You need to use `-separate` argument to show all annotations if this is of interest to you. (5) This also could be due to the presence of bugs in one software or the other. If there is a potential bug that you find in ANNOVAR, please report to me.

11. How to infer the version number for RefSeq transcripts in ANNOVAR annotation results?

Updated 2017 since UCSC changed their MySQL schema again: Run this command (for human hg19 build): `mysql --user=genome --password=password --host=genome-mysql.cse.ucsc.edu -A -D hg19 -e 'select distinct hg19.refGene.name,hgFixed.gbCdnalInfo.version from hg19.refGene,hgFixed.gbCdnalInfo WHERE hg19.refGene.name=hgFixed.gbCdnalInfo.acc' > refseq_version.txt`

Starting from Nov 2014, when you download refGene for human (hg18/hg19/hg38), the corresponding `refGeneVersion.txt` file will be automatically downloaded to help users who cannot figure out how to run mysql. However, you will need to run the MySQL command manually for other species.

Starting from June 2017, we include `hg19_refGeneWithVer.txt` and `hg19_refGeneWithVerMrna.fa` file into the ANNOVAR package. Therefore, users can directly use `-dbtype refGeneWithVer` to annotate genetic mutations with RefSeq version number.

12. What is the difference between comma and semicolon when they are used to separate gene names in gene annotation?

The semicolon (";") separate different annotations, for example, coding variants for one gene and splice variants for another gene (but these two genes may have the same name, since one gene may have multiple transcripts). The comma (",") separates different genes with the same annotation, for example, multiple genes may have overlapping exons, so a variant may be annotated as exonic in multiple genes.

13. Why I cannot run ANNOVAR in my web browser such as Chrome?


ANNOVAR is a command-line software that requires a Perl interpreter in your system. Typically, Linux systems already include a Perl interpreter by default, yet you need to install one yourself in Windows (use strawberry perl or activeperl). The `table_annovar.pl` is a perl script that you will execute using a perl interpreter, and it is not a URL that you visit in your web browser.

14. Why a very common variant has very low frequency in filter annotation in hg38?

This very rare situation happens for some ANNOVAR filter databases, that were generated by lifting over the corresponding hg19 databases. In some genomic positions, the nucleotide identity differs between hg19 and hg38, resulting in this problem. For example, chrX:152652814 has allele frequencies near 50:50, and from build 37 to build 38, they switched which allele was reference and which allele was alt. So for SNPs where the ref/alt alleles were swapped in build 38, Annovar does not annotate these SNPs's frequency if you use a lift over allele frequency database in ANNOVAR (currently, the list of hg38 databases generated by liftOver is annotated in the download page). This is a very rare event, but extra caution is always good to examine your results if you happen to use one of the liftOver filter databases provided by ANNOVAR.

15. It should be amino acid X in this position but ANNOVAR reports Y in this position!

For example, ANNOVAR reports p.X100Z as the amino acid change, but another web resource shows that position 100 should have wildtype of Y not X.

 v: latest ▼

Whichever website you use, regardless of whether it is swissprot, refseq or whatever, remember that they always have their own way to collect proteome data and compile data, and these ways may result in slight discordance to theoretical protein sequence. Sometimes, these websites may directly translate a

RefSeq transcript such as NM_123456 to a protein sequence and present it, but although ANNOVAR uses NM_123456 as well it uses the "theoretical mRNA sequence" inferred by ANNOVAR as opposed to those provided in RefSeq. By "theoretical", I mean a protein sequence that is translated from the "theoretical" mRNA sequence which is specified by a gene model as well as a whole-genome DNA sequence given a specific genome build. ANNOVAR is a software that produces this "theoretical" protein sequence, so if you want to stick with a specific genome build and a specific gene definition system, then ANNOVAR gives the correct results.

Exceptions exist when the gene model is not annotated correctly. In other word, when the exon start site, end site, splicing site have some slight errors. In this case, the protein sequence produced by ANNOVAR may be wrong and may contain pre-mature stop codons. (There are many many reasons this may happen) If you ever encounter such a variant, just try a different gene model (for example, using `-dbtype knowngene` or `-dbtype ensgene`) to reannotate this variant. If you want to investigate this variant even more closely, considering using the `coding_change.pl` program in ANNOVAR, which will print out the theoretical protein sequence before mutation and after mutation, and will flag any potentially wrong theoretical protein sequence with WARNING messages.

In Nov 2011, I updated ANNOVAR so that any reference transcripts with premature stop codon (potential gene model annotation error or transcript-to-genome mapping error) will no longer be used in `exonic_variant_function` output file.

16. Why ANNOVAR reports a 3-bp deletion as frameshift deletion?

For example, "9 5720612 5720614 AGT -" (hg19 coordinate) is annotated as non-frameshift deletion by CLCbio but ANNOVAR thinks it is a frameshift deletion. Biologically, a 3-bp frameshift deletion is indeed possible: This could happen, for example, when the 3-bp deletion covers only 1 or 2 bp in exons, and indeed this is the case for this deletion. ANNOVAR knows how to handle these types of complicated situations but other software may not.

17. Why ANNOVAR reports T182A,T190A,T300A as the amino acid change but another web server reports only T300A?

Alternative splicing is prevalent in human genome and as a result, it is best to annotate amino acid change with respect to a certain transcript rather than gene. Other servers or software may randomly pick one script as the representative "gene" and gives one single answer. ANNOVAR tries to be comprehensive and always accompany annotation by transcript names, and it is up to the user which representative transcript they want to use or if they want to use all.

There has never been a consensus in the field which transcript should be used to represent a gene when multiple transcripts are available. The most popular approach is to use the longest transcript nowadays. However, in the medical genetics field, for certain specific diseases and specific genes, there are 'canonical' transcripts that everybody uses by default for historical reasons, and you will need to manually select this canonical transcript from ANNOVAR output file to communicate with the rest of the field.

18. Why ANNOVAR reports c.C100T when my input is G to A change?

The c.C100T is a cDNA (actually, mRNA) level change. ANNOVAR input (G to A) has to be in the forward strand, and if the transcript is in the reverse strand, there will be a C to T change in the mRNA.

19. Why ANNOVAR reports c.T5997G when my input is T to C change in chr14:31582550-31582550 in hg19 coordinate?

First, this transcript is in the reverse strand, so the mutation is changed to "G". Second, your input is wrong: this position should be A in hg19, so c.T5997 should be the reference base. Maybe you used a wrong genome build, or your genotype calling software has a bug. ANNOVAR did it correctly. Starting from September 2011, ANNOVAR will try to print out WARNING messages telling user that they used wrong reference alleles in their input file for exonic variants.

20. Why my mutation gets lost by ANNOVAR?

A user reported that the input "17 16256671 16256671 C G" in hg19 coordinate was reported as a "CENPV:NM_181716:exon1:c.C80C" mutation, so the C->G change gets lost by ANNOVAR. Read the FAQ item above: the input is wrong, as this position should be a G wildtype in reference genome, so the C80C mutation is the correct mutation in the opposite strand.

21. Why only one isoform is in exonic_variant_function but two in variant_function?

A user reported that the input "1 14143003 14143003 A G" in hg19 coordinate was reported to hit "NM_001135610,NM_012231" in variant function, but only NM_001135610 in exonic_variant_function file (when `-transcript` argument was used). If you add `-separate` argument, you'll see that the change on NM_012231 is synonymous, so it is not printed out due to precedence rule.

22. Why ANNOVAR reports "unknown" in exonic_variant_function?

"unknown" means that the gene structure is not correctly annotated (complete ORF information is not available). Previous versions of ANNOVAR will always give an answer such as non-synonymous SNVs, etc, but I got too many user emails complaining about "bugs" (even though ANNOVAR is innocent in this case). So after December 2011, if errors exist in gene structure annotation (RefSeq, Ensembl, UCSC, etc), ANNOVAR will just report unknown for exonic_variant_function; in other words, although the variant is clearly within an exon, we cannot say for sure how it affects protein sequence as the ORF annotation is not correct.


23. Why ANNOVAR reports the same function for two different mutations in two sites?

Sometimes different mutations are reported to have the same function in gene-based annotation. For example, these mutations at chromosome 4 at coordinates 8945506, 8950251, 8954996, 8959741, 8964486, 8969231, 8973977 are all reported to be USP17:NM_001105662:exon1:c.A25G:p.R9G. There is nothing wrong: if you check the USP17 gene in genome browser, you'll see that there are at least 9 copies of the gene in each haplotype. So all the mutations (if they are real) all have the same function. In reality, it is likely that these mutations are not real, but are rather artifacts of base-level differences between any random two copies of the same gene.

24. Why ANNOVAR complains "exonic SNPs have WRONG reference alleles" in gene-based annotation?

This happens when ANNOVAR thinks the "reference allele" in your input does not fit the "reference allele" in the mRNA FASTA file in ANNOVAR's database. This could be due to several reasons, (1) wrong `-buildver`, or (2) you did not specify the correct reference allele, or (3) mRNA FASTA file is outdated as the gene model gets updated pretty quickly by UCSC.

To solve this problem, first check (1) and (2) to make sure that you did have the correct input. If you cannot find an error, then update the FASTA file by `retrieve_seq_from_fasta.pl` command, with more details here.

 v: latest ▼

25. Why FASTA sequence in ANNOVAR differ from those in public databases?

For example, the mRNA of the MYBPC3 gene (NM_000256) extracted from ucsc and the other one extracted from annovar `hg18_refGeneMrna.txt` file differ. The reason is simple: FASTA in ANNOVAR is built from ANNOVAR using chr:start-end records, not copied/pasted from any public database. Any errors in chr:start-end will lead to errors in ANNOVAR-compiled FASTA. To avoid future complaints, FASTA sequences with premature stop codon will no longer be used in exonic annotation, although they still exists in the FASTA file.

26. Why ANNOVAR's TFBS annotation differ from what I have from another web server?

There are MANY different transcription binding sites (TFBS) annotations generated by hundreds of research groups in the world. ANNOVAR used a keyword "TFBS" for only one specific type of annotation that have a long history in Genome Browser, but it does not mean that this is the ultimate solution for TFBS prediction. ANNOVAR can certainly take many other types of TFBS annotations for but it won't use the keyword "tfbs" for that. In fact, as you can see from the region-based annotation page, ANNOVAR can also annotate TFBS ChIP-Seq from the ENCODE project.

The take home message is that there are many annotations on TFBS, and they may differ from each other substantially. Use caution when interpreting the data. Ultimately, it is the biologist himself/herself who can decide whether or not the annotation makes sense; ANNOVAR facilitate this process but it cannot make the decision for you.

27. Where are the values for -protocol and -argument come from in table_annovar.pl?

The protocol values corresponds to file names that are stored in the directory specified by the user in command line (with a couple of exceptions such as 1000g-related files). They are generally referred to as database files, and they can either come from ANNOVAR's own repository (via `-downddb -webfrom annovar` argument), or from UCSC's annotation databases (via `-downddb` argument), or provided/compiled by users. Therefore, there are unlimited possibilities for protocols, and there is not a comprehensive list that we can provide.


The argument values correspond to each of the protocols, as optional argument that you would use for `annotate_variation.pl` on this specific protocol. In other words, `-protocol`, `-operation` and `-arg` are all parallel lists of corresponding entries and should have equal comma-delimited number of entries.

28. How to handle huge multi-sample VCF files?

You can just cut the first sample (basically the first ~10 columns), then annotate this file by `table_annovar`. Then just "paste" the annotation with the rest. For example, `cut -f 1-10 input.vcf | grep -v -P '^#' > input1.vcf; cut -f 11- input.vcf | grep -v -P '^#' > genotype`, then `annotate input1.vcf, generate input1.anno.vcf, then paste input1.anno.vcf genotype > input.anno.vcf` to generate the combined output file. You may want to add the VCF header back in.

29. Why the SIFT/PolPhen scores in ANNOVAR differ from those obtained from another website?

The AVSIFT scores (now obsolete!) in ANNOVAR was based on Ensembl55 database, and sometimes there are major differences from those computed from ensembl63 (default in SIFT website). If you select ensembl55 from SIFT website you'll see that the scores are consistent and identical. The LJB_SIFT scores in ANNOVAR was based on the original Liu et al paper, so read the paper for details on how they compile the scores. In most recent version we use the dbnsfp* keyword, and all scores are directly taken from the dbNSFP database.

 v: latest ▼

But in general, calculation of scores depend on version of software, parameters of program, source of data files, definition of gene structure, handling of alternative transcripts and multiple scores, so there are many reasons why there are differences in scores calculated by different people. ANNOVAR now tries to

be synchronized with the ljb* database, so the scores may be different from another web server.

30. Can ANNOVAR identify all SNPs annotated within dbSNP in a given region (say chr1:3751541-3751607)?

In ANNOVAR, filter annotation identifies exact matches including base pair identity, yet region annotation identifies overlapping regions. When you use `--filter`, the program will tell whether the region chr1:3751541-3751607 is a SNP within dbSNP (highly unlikely to be the case). In more recent versions of ANNOVAR, region annotation can handle snp130 now. For example, just try `annotate_variation.pl ex1.human humandb/ -region -dbtype snp130`. However, this command requires about 10GB memory to run.

However, if you are only looking at one single specific region, a simple script can be used to address this question, after using `-downdb snp130` in ANNOVAR: `perl -ne '@a=split(/\t/, $_); $a[1] eq "chr1" and $a[3]>=3751541 and $a[3]<=3751607 and print $a[4], "\n" < hg18_snp130.txt.`

31. How to annotate simple repeat regions in human genome?

Read these pages: <http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=199336701&c=chr1&g=simpleRepeat> (<http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=199336701&c=chr1&g=simpleRepeat>), <http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=199336701&c=chr1&g=rmsk> (<http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=199336701&c=chr1&g=rmsk>), <http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=199336701&c=chr1&g=rmskRM327> (<http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=199336701&c=chr1&g=rmskRM327>), then pick one that matches your goal, then annotate by ANNOVAR.


32. How to handle E. coli, Arabidopsis thaliana and other genomes not in UCSC?

For gene-based annotations (say for example, `-dbtype refGene`), ANNOVAR requires 2 files: a refGene file specifying gene model, and a FASTA file with sequence for each transcript. You can make 3 files for the genome using the following rules:

For refGene file, each line has 16 tab-delimited columns: \$bin, \$name, \$chr, \$dbstrand, \$txstart, \$txend, \$cdsstart, \$cdsend, \$exoncount, \$exonstart, \$exonend, \$id, \$name2, \$cdsstartstat, \$cdsendstat, \$exonframes. The only real important thing is \$name (transcript name), \$chr (chromosome), \$dbstrand (strand of the transcript in reference genome), \$txstart, \$txend (transcription start and end), \$cdsstart, \$cdsend (translation start and end, remember that there are 5'/3'-UTR in each transcript so the \$cdsstart is not the same as \$txstart), \$exoncount (number of exons), \$exonstart \$exonend (comma-delimited exon start and end sites). Remember that all start sites use zero-based coordinates.

For refLink file, you can make anything. The file will be ignored. (It is important for very old genome annotations when name2 field is not present in refGene, but it is not really useful today as people will not use old genome assembly nowadays).

For FASTA file, make sure that the \$name in ">\$name" matches the refGene file, in a case-sensitive manner. You can build the file yourself, or you can directly use `retrieve_seq_from_db.pl` in ANNOVAR to generate this file, given a FASTA file for the genome. Make sure that strand is correct in the cDNA if you build the file yourself.

After you have three files, you can directly run ANNOVAR by specifying `-buildver` argument to match your file prefix.  V. latest ▼

If you have GFF3 files, then convert it to UCSC compatible format first (try the http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/gff3ToGenePred (http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/gff3ToGenePred) tool). This is the easiest thing to do and multiple users have reported success on multiple novel species.

Trouble shooting: If you can generate variant_function annotation but not exonic_variant_function annotation, then double check the GFF file. The gff3ToGenePred requires gene/mRNA/CDS/exon notation, but some GFF3 files use "transcript" rather than "mRNA" resulting in lack of coding information in output files. Manually change "transcript" to "mRNA" in GFF3 will solve this problem.

33. Why the total number of homozygous and heterozygous variants is more than the number of variant site (convert2annovar.pl)?

Suppose we see 30 reads in a site, and 10 are A, 10 are G and 10 are C. This is one site, but may be presented as two heterozygous mutations from the genotype calling algorithm. This could be due to a tri-allelic SNP, or a genomic duplication, or just sequencing error.

34. Can ANNOVAR handle IUPAC code in input?

No. ANNOVAR is a variant annotation program, not a genotype annotation program. It needs to see A, C, G, T, not a IUPAC code representing ambiguity of an allele, or an IUPAC code representing a genotype call.

35. Can ANNOVAR handle genotype calls in input?

No. ANNOVAR is a variant annotation program, not a genotype annotation program. You can only specify the allele of an observed variant (such as A, G, etc), not a genotype on a specific position (such as AG genotype).

36. How to handle two very close SNPs in the same codon?

If two SNPs are separated by only one or two nucleotides, it is best to treat them as a block substitution, rather than two separate variants. Otherwise, the annotation may not be correct if the two SNPs happen to impact the same codon.

37. How to select the X-way phastCons conservation track in ANNOVAR?


This totally depends on the genome build, and you need to check genome browser for the number of tracks. For example, for chicken genome, if you select galGal3 as the --buildver, then you'll see in the genome browser page (by hovering mouse on top of "Most Conserved") that it is 7way.

38. Can ANNOVAR print out translated protein sequence?

`annotate_variation.pl` cannot do that directly, and it is very difficult to modify the existing exonic annotation subroutine to do this. Therefore, in June 2011 version of ANNOVAR, I added the `coding_change.pl` program to infer translated protein sequence before and after mutation occur.

39. Is it possible to add column names to the input file that are carried through the processing?

Some users routinely use extra columns and would like to include the column headers rather than having to edit the resulting ANNOVAR output (usually ANNOVAR will treat the line with column names as "invalid" line and put it into the invalid_input file). This can be done with the `-comment` argument, which treats any input line starting with "#" as the comment line and do not discard it.

 v: latest ▼

40. Can ANNOVAR call genotypes from sequencing data?

ANNOVAR does NOT generate "genotype calling". Dozens of other software tools can perform SNP calling from sequencing data. However, if the user refers to "assigning rs identifiers to SNPs", ANNOVAR can certainly be very helpful (see the example on filtering against dbSNP).

41. How to check if new version of ANNOVAR is available?

Either go to ANNOVAR website to see what's the latest version and compare to your current version (type `annotate_variation.pl` without argument will print out version information). Or use `annotate_variation.pl -downdb null .` to enable automatic web-based checking of new version without downloading any database.

42. How to list all annotation databases in ANNOVAR web server?

You can use `-webfrom annovar -downdb avdblist` to see a list of files, file sizes and time stamp. This only works on human genome though.

43. How to handle OMIM data?

Many people studying Mendelian diseases perhaps are interested in annotating variants against the OMIM database. However, the 16 June 2011 News from UCSC shows that although they released newly re-engineered OMIM tracks for both hg18 and hg19, "the OMIM data are the property of Johns Hopkins University and will not be available for download from UCSC". For now, you can just go to <http://omim.org/downloads> (<http://omim.org/downloads>), fill out the forms and get a copy of the data. I cannot make a derivative database for you, per their guideline. If you only need a gene symbol to OMIM ID mapping, you can get that data from HGNC here: http://www.genenames.org/cgi-bin/hgnc_downloads (http://www.genenames.org/cgi-bin/hgnc_downloads)

44. What is the version of ENSEMBL used in ANNOVAR?


ANNOVAR retrieves ensGene definition from UCSC, so it depends on the version that UCSC has used. For human hg19 build, just go to <http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=211204337&g=ensGene> (<http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=211204337&g=ensGene>), and see what is the latest release for Ensembl gene prediction.

45. How to annotate ENSEMBL gene on the hg38 genome coordinate?

The ensGene file for hg38 is not provided by ANNOVAR because UCSC did not generate this file. However, a user pointed out that UCSC have replaced the ensGene.txt using GENCODEV26 (wgEncodeGencodeCompV26.txt track). Both files contain the same information. Therefore, if you want to annotate Ensemble genes based on hg38, you should use the GENCODE file instead. Detailed instructions are given in the gene-based annotation page.

46. How ANNOVAR handles different coordinate systems for mitochondria?

UCSC's build (for example, hg19) differ from NCBI's build (for example NCBI 37) in a few subtle manners, for example, replacing contigs by chr_random, and the use of different mitochondria assemblies. UCSC's hg19 assembly used the old version mitochondria genome (NC_001807), but 1000 genomes consortium has replace the chrM with the latest Cambridge Reference Sequence version (NC_012920). So if you align your sequence data and call variants against the NC_012920, then you cannot really annotate your variants using UCSC's gene definition. It is necessary to stick with the identical coordinate. For autosomes and chrX/Y, this is not a real issue as they are pretty consistent.

 v: latest ▼

In addition, For most organisms the "stop codons" are "UAA", "UAG", and "UGA". In vertebrate mitochondria "AGA" and "AGG" are also stop codons, but not "UGA", which codes for tryptophan instead. "AUA" codes for isoleucine in most organisms but for methionine in vertebrate mitochondrial

mRNA.

47. How to get -downdb to work if I am behind a proxy server?

The -downdb use `wget` by default without any argument. You can add `-nowget` in the command line, so that Perl HTTP/FTP modules will be used instead which should handle proxy well. Or you can modify the ANNOVAR source code to use wget with proxy functionality.

48. How to download databases not stored in UCSC or ANNOVAR-DB?

In general, you just need to manually download these databases, and reformat them to standard ANNOVAR genericdb format (Chr, Start, End, Ref, Alt, and other information), and use them. Occasionally, you may also automate the process by supplying the URL directly; for example, to download Regulome, you can do `perl annotate_variation.pl --downdb --webfrom http://www.regulomedb.org/downloads/ (http://www.regulomedb.org/downloads/) RegulomeDB.dbSNP141 /Users/user/Desktop/annovar/humandb`.

49. How to handle MAF files from TCGA?

You can use this script (<http://www.openbioinformatics.org/annovar/download/maf2annovar.pl>) to convert MAF to ANNOVAR input format and then annotate the file.

50. How to further speed up ANNOVAR?

You can use the `-thread` argument (if your operating system and your perl build support it), so that multi-threading functionality is used to process the input files in parallel. However, it is extremely important that your database directory (for example, `humandb/` directory) can accommodate random disk access well. Typically, if you use a very large number of threads, you have to use SSD drive to achieve satisfactory performance. Mechanical drives cannot tolerate it for most large databases. Additionally, borrowing ideas from an ANNOVAR user, if you have a machine with large memory, you can also just simply create a RAM disk to treat a portion of the memory as a hard drive and then copy the `humandb` into this RAM disk. For example, do a `mount -t tmpfs -o size=100G tmpfs /tmp/newhumandb/`, followed by `sysctl vm.swappiness=1` to reduce swappiness, and then use the `/tmp/newhumandb` to store databases and perform annotation.

293 Comments ANNOVAR documentation

Login

Recommend 14

Tweet

Share

Sort by Best

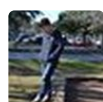


Join the discussion...

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name



Linhai Percy Zhao • a year ago

Hi Dr. Wang,

I found that for gnomAD frequencies, the users can only choose one out of gnomad_exome and gnomad_genome. Is there a combined frequency like those provided on gnomAD browser? Thanks!

Y: latest

1 ^ | v • Reply • Share ›



Kai Wang Mod → Linhai Percy Zhao • a year ago

No, they are two separate data sets with different characteristics. However, you can combine them yourself if needed.

^ | v • Reply • Share ›



David Gómez Sánchez → Kai Wang • a year ago

And another question: If we want to combine them, would their frequencies overlap for variants situated in exonic regions? If that's the case, would you recommend using gnomad_exome's frequencies for this variants?

^ | v • Reply • Share ›



Kai Wang Mod → David Gómez Sánchez • a year ago

These are from different data sets, and the samples do not overlap.

^ | v • Reply • Share ›



David Gómez Sánchez → Kai Wang • a year ago

Dear Dr.Wang, could you please explain us what are gnomad_genome and gnomad_exome made of? what's their size? I'm mostly interested in exonic variants but also in intronic in a much lower degree. Is gnomad_exome bigger than gnomad_genome regarding coding variants?

^ | v • Reply • Share ›



Kai Wang Mod → David Gómez Sánchez • a year ago

it is explained in detail in <http://gnomad.broadinstituit...> 123,136 exome sequences and 15,496 whole-genome sequences

^ | v • Reply • Share ›



Chi-Yu Yen • 3 years ago

Hi , I have a question regarding output format. I ran my annotation analysis using the [table_annovar.pl](#) command with clinvar_20150629 operation filter. The output csv file has a clinvar_20150629 but in that particular field it has a lot of information that is hard to fit in a standard column width. I'm wondering if there is any additional parameter that I have to use to make is separate? Thanks

for example: something like this...could I separate CLINSIG, CLNDBN, CLNREVSTAT, etc?

clinvar_20150629


CLINSIG=non-pathogenic|other|non-pathogenic;CLNDBN=not_specified|Familial_colorectal_cancer|Hereditary_cancer-predisposing_syndrome;CLNREVSTAT=criteria_provided\x2c_single_submitter|no_assertion_cri

1 ^ | v • Reply • Share ›

v: latest ▼



Kai Wang Mod → Chi-Yu Yen • 3 years ago

 please use clinvar_20160302 instead in table_annovar. This has the "column"-based format.

^ | v • Reply • Share ›



Chi-Yu Yen → Kai Wang • 3 years ago

Thanks Kai! I have another similar format question. How about for "AAChange.refGene" column? I believe it's from the gene-based annotated refGene. Is there a tabular version as well? Thank you.

^ | v • Reply • Share ›



Kai Wang Mod → Chi-Yu Yen • 3 years ago

This is gene based annotation, you have to parse it yourself if you need to have separate fields.

^ | v • Reply • Share ›



qiao • 3 years ago

Hi,

I noticed that annotations for a deletion site disaccord among different transcripts, and what troubles us is that this mutation is a 52bp-deletion in chr9, however the annotation shows a 210bp deletion in transcript NM_058195(c.350_559del, details see below) . Is something wrog with this transcript? Thanks!

chr9 21971000 21971051

CCTCAGCCAGGTCCACGGGCAGACGGCCCCAGGCATCGCGCACGTCCAGCCG - exonic
CDKN2A . frameshift deletion

CDKN2A:NM_000077:exon2:c.307_358del:p.R103fs,

CDKN2A:NM_001195132:exon2:c.307_358del:p.R103fs,

CDKN2A:NM_058195:exon2:c.350_559del:p.A117fs

9p21.3 0.5 5818.73 349 chr9

21970999 .

TCCTCAGCCAGGTCCACGGGCAGACGGCCCCAGGCATCGCGCACGTCCAGCCG T 5818.73
PASS

AC=1;AF=0.5;AN=2;BaseQRankSum=6.246;ClippingRankSum=1.086;DP=359;FS=10.593;MLE.
GT:AD:DP:GQ:PL 0/1:187,162:349:99:5856,0,7351


1 ^ | v • Reply • Share ›



Kai Wang Mod → qiao • 3 years ago

Unlike the other transcripts, if you examine NM_058195 carefully, you will see that this deletion actually take out the entire exon2 (i.e., intron/exon boundaries are included in the entire deletion), so in principle the transcript does not splice any more, and as a result, the cDNA portion of the transcript after the deletion site is considered as deleted.

^ | v • Reply • Share ›

 v: latest ▼



Nelson Chuang • 4 years ago

does table_annovar not support -comment anymore? I want to keep the sample names from

the vcf file. I even tried -otherinfo but that didn't carry over the sample names. Thanks!

1 ^ | v • Reply • Share ›



Kai Wang Mod → Nelson Chuang • 4 years ago

what's your command line and error/warning message? Please read FAQ #1.

^ | v • Reply • Share ›



Nelson Chuang → Kai Wang • 4 years ago

I apologize, I hoped it was a quick question/answer. Here is my command line. The program says "comment" is not a valid argument, and I don't see it listed under available arguments in the help either.

```
perl annovar/table_annovar.pl GBR.private.tranche3.pass.vcf annovar/humandb/
-buildver hg19 -out GBR.private.tranche3.pass.anno -remove -comment -
protocol
refGene,cytoBand,genomicSuperDups,esp6500siv2_all,1000g2014oct_all,avsnp1
-operation g,r,r,f,f,f,f -nastring . -vcfinput -otherinfo
```

^ | v • Reply • Share ›



Kai Wang Mod → Nelson Chuang • 4 years ago

Only annotate_variation supports -comment.

You are using table-annovar on a VCF file. Your command should generate a txt file and a VCF file. The VCF file have comment line from original VCF file.

^ | v • Reply • Share ›



Nelson Chuang → Kai Wang • 4 years ago

Yea I have been copying it over from the vcf, but was hoping it was implemented to save me time. It's not a big deal. Thank you for your time.

^ | v • Reply • Share ›



clasguitar → Nelson Chuang • 2 years ago

Hi Nelson,

I am interested in doing the same thing (carrying over sample names from the vcf file while implementing the [table_annovar.pl](#) program). So I assume that the sample genotype information that is maintained in the table_annovar output file is kept in the original order? (you did not run into problems with mis-labeling when you had to copy and paste the sample names from the original vcf file to the table_annovar output?). Thanks.

^ | v • Reply • Share ›

v: latest ▼



Chi-Yu Yen • 3 years ago

Hi Dr Wang I have a question on how to integrate the downloaded OMIM database with

Hi Dr.Wang, I have a question on how to integrate the downloaded OMIM database with ANNOVAR. I downloaded the OMIM database(genemap.txt) and tried to format to the ANNOVAR genericdb format, however; I didn't find any Ref, Alt column in the genemap.txt file and it seems to not able to read it correctly. Would you help me to shed some lights on how I should format it correctly? Thank you.

1 ^ | v • Reply • Share ›



Kai Wang Mod → Chi-Yu Yen • 3 years ago

OMIM only annotates genes (locus) and disease, not variants (despite the presence of allelic series for some genes). You can write your own script to check whether a gene is a OMIM gene and what phenotype it associates, but this is done on results generated from Annovar, not on Annovar itself.

^ | v • Reply • Share ›



Nicky Pan • 2 months ago

Hi Dr.Wang,

Recently, gnomAD released version r2.1.

Is there a plan to update the gnomAD database in ANNOVAR, or how to make my own gnomAD databases for using in ANNOVAR, just like the ClinVar?

^ | v • Reply • Share ›



Kai Wang Mod → Nicky Pan • 2 months ago

I plan to update several databases soon (next a few weeks) including gnomAD. You can also make your own database but may need to write a script to reformat the data depending on the data format in the new release.

^ | v • Reply • Share ›



Nicky Pan → Kai Wang • 2 months ago

thank you for your reply. Looking forward to the updates.

^ | v • Reply • Share ›



Alexandros Pavlaras • 3 months ago

Hello Dr.Wang,

I am currently trying to detect RNA Editing Sites in my project and I'm using ANNOVAR to annotate the Editing Sites I obtained using SPRINT. I met with a problem and I'd like to know if there's a solution to it. My RNA-seq data isn't strand specific. If I have an A to G change occurring in the "+" strand or a T to C change in the "-" strands, I know it's ADAR editing, so I'm good with the gene regions. Though, when it comes to intergenic or ncRNA regions, ANNOVAR doesn't provide a strand annotation, so I can't define the Editing Site. Does it have to do with my RNA-seq data not being strand specific ? Is there a way to get strand annotation for those regions ? I'm new to the bioinformatics field, so I'm sorry for any of my ambiguities. Thank you in advance for your time and precious information.

Best regards,

Alexandros

^ | v • Reply • Share ›

v: latest ▼

**Kai Wang** Mod → Alexandros Pavlaras • 3 months ago

All input needs to be in forward strand in the genome. Whenever you do alignment of RNA-Seq data with reference genome, the mutations in the results should always be the forward strand (because reference genome is forward strand). Any alignment software will know how to flip strand when doing the alignment. So when you see a mutation call from most variant calling software with most alignment software, it should be already in forward strand. (I am not familiar with SPRINT so I cannot tell with 100% certainty).

^ | v • Reply • Share ›

**陆晓凡** • 3 months ago

Hi Dr.Wang,

I met a problem when annotate a nonframeshift substitution but it was supposed to be frameshift.

Here is the example of data (test.avinput) and code.

my data and the confusing annotations are the last two which I thought they were supposed to be frameshift.

```
14 69256505 69256505 G A
14 69256506 69256506 G A
14 69256987 69256987 C -
14 69257040 69257041 -- GG
14 69257144 69257145 -- CC
```

my cmd:

```
annotate_variation.pl -out test -build hg19 example/test.avinput humandb/
```

After annotation I got the following results and I do not know why the last two are nonframeshift.

line1 synonymous SNV

[see more](#)

^ | v • Reply • Share ›

**Kai Wang** Mod → 陆晓凡 • 3 months ago

you should use table_annovar and use the -polish argument. There are instructions in the quick start-up guide page in the website.

^ | v • Reply • Share ›

**Rohan** • 5 months ago

Hi Is there a way by which we can annotate data as per hgvs ? There is python package available but is difficult to integrate.

^ | v • Reply • Share ›

v: latest ▼

**Kai Wang** Mod → Rohan • 5 months ago

depending on your command line. if using annotate-avriation, add -hgvs argument. If using table_annovar add -arg '-hgvs' that corresponds to the same position as

using `table_annovar`, add `myanno` that corresponds to the same position as specified in `-protocol` and `-operation`.

^ | v • Reply • Share ›



Harold Smith • 6 months ago

Hello,

I obtained an unexpected result and hope someone has an explanation.

The experiment: WGS, 20X coverage, *C. elegans* strain with a known mutation

The pipeline: BMap for alignment, FreeBayes for variant calling, ANNOVAR for annotation

The problem: the known mutation (AG -> TA) is a nonsense allele (codon AGG -> TAG). I expected it to be classified as 'stopgain' in the `exonic_variant_function` report, but instead it was classified as 'nonframeshift substitution'. While technically correct, it obscures the impact of the mutation on the coding sequence. At a minimum, I would expect the amino acid change to be reported in the relevant field, but it is not.

Let me know if you require additional information.

Thanks,
Harold

P.S.-I confirmed that the gene sequence in `RefGeneMrna.fa` is correct, so the underlying annotation is not the reason for this behavior.

Also, it appears that the ANNOVAR mailing list/Google group is not working, as I was unable to post this question there.

^ | v • Reply • Share ›



Kai Wang Mod ➔ Harold Smith • 6 months ago

You need to use `table_annovar.pl` with `-polish` argument as shown in quick start-up guide. It should be classified as stopgain by annovar. Also read FAQ #1 since you did not provide any details on your command. Mailing list is only for announcing critical updates, not for addressing questions.

^ | v • Reply • Share ›



Harold Smith ➔ Kai Wang • 5 months ago

I downloaded the latest version of ANNOVAR and used the following command:

```
perl table_annovar.pl sample_9_candidates.vcf WS265db/ --buildver WS265 --
out myanno --polish --remove --protocol refGene --operation g --vcinput --xref
WS265db/WS265.gene_xref.txt
```

The mutation I mentioned in the previous message is now classified as 'stopgain', but classification of some of the more complex alleles is still unexpected. Here's one example:

```
1 20026 20026 TACT CACA exon1c WS265db/WS265.gene_xref.txt nonframeshift
```

130036 30039 TAGT GAGA EXONIC WBGene00022279 . nonframeshift substitution
 WBGene00022279:Y74C9A.5:exon3:c.761_764TCTC:p.N254_Y255delinsIS
 sesn-1 0 7.60564e-07 10 | 30036 . TAGT GAGA 7.60564e-07 .
 AB=0;ABP=0;AC=0;AF=0;AN=2;AO=2;CIGAR=1X2M1X;DP=10;DPB=10;DPRA=0
 GT:DP:DPR:RO:QR:AO:QA:GL 0/0:10:10,2:8:297:2:15:0,-1.59338,-24.7189

AACChange field indicates the protein coding change (N254_Y255delinsIS) but the ExonicFunc field indicates 'nonframeshift substitution' even though the substitutions are nonsynonymous.

Thank you for your assistance.

^ | v • Reply • Share ›



Kai Wang Mod → Harold Smith • 5 months ago

your input is a block substitution, changing 4 base pairs (so it is nonframeshift substitution). From amino acid perspective, it changes NY to IS, so both amino acids are changed. This is not typically regarded as nonsynonymous.

^ | v • Reply • Share ›



Harold Smith → Kai Wang • 5 months ago

Thank you for your prompt reply. Actually, this variant is nonsynonymous; the two amino acid changes just happen to be adjacent to each other. But your response indicates that ANNOVAR is reporting them as you intended. It appears that I will need to decompose the VCF into allelic primitives before using ANNOVAR (or parse the ANNOVAR output to reclassify variants of this type as nonsynonymous). Thanks again.

^ | v • Reply • Share ›



Kai Wang Mod → Harold Smith • 5 months ago

In my opinion, it is more correct to report them as two amino acids changed to another two amino acids. If you decompose the change into two separate mutations (and therefore two separate amino acid changes), I feel that this may be less informative. As you suggested, regardless of how one treats them, the mutations are still non-synonymous mutations as the amino acid sequences are changed. The difference is whether to treat this as one mutation, or two separate mutations adjacent to each other; I feel the former is better.

^ | v • Reply • Share ›



Amina Attia • 7 months ago

Dear Dr Wang,

Thank you for your work on ANNOVAR.

v: latest ▼

I have a question about the "AACChange" column. Some variants show strange results. For instance, chr13 103398261 T > C is described as "nonsynonymous SNV" whereas when I carefully read the reference sequence, this mutation TGT>TGC is synonymous. chr13

Carefully read the reference sequence, this mutation TCG > TGC is synonymous. chr10:114059240 G > A (CCG > CCA) is also synonymous but described by the "AAChange" column as non synonymous. This problem occurs for some other variants. Do you have any idea of where it comes from ? I used the last version of ANNOVAR, hg19 and refGene as a database. Thank you very much for your help !

^ | v • Reply • Share ›



Kai Wang Mod ➔ Amina Attia • 7 months ago

Please read FAQ #1 and provide details. For example, did you do -polish? these are essential information.

^ | v • Reply • Share ›



I. N. Motoike • 8 months ago

Dear Dr. Wang,

We are planning to release whole-genome reference panel by conducting a large-scale whole genome sequencing of DNA samples from participants in our cohort studies, in academic project.

We would like to add basic annotation, gene name, exonic functions etc., by using of annovar, to the reference panel.

Is this possible in annovar's academic license?

If the distribution of annovar's result is allowed in your license, it would be very helpful for us and for data users.

We would appreciate it if you could consider with this matter.

We look forward to hearing back from you!

Thanks

^ | v • Reply • Share ›



Kai Wang Mod ➔ I. N. Motoike • 8 months ago

yes you can do that and release results. thank you.

^ | v • Reply • Share ›



I. N. Motoike ➔ Kai Wang • 7 months ago

Thank you for your kind and quick reply.

I am very appreciate you.

We will try to do that, and if we use annovar, we will cite.

Thanks a lot!

^ | v • Reply • Share ›



Jianxin Wang • 9 months ago

Hi Dr. Wang,

v: latest ▼

Below are some questions and a suggestion after using Annovar in the past few years.

1. I found most (if not all) of the annovar annotations in the vcf header have a type of "String".

Is there any particular reason for this decision? For example, the minor allele frequency from various populations have floating point values but is specified to have a type of "String" rather than "float".

2. All the annotation feature names are specified in the vcf file body using "key=value" pairs. This will inflate the file size after annotation with annovar. Is it possible to use the approach used in VEP? i.e. define the feature mapping in the vcf header and use something like: ANNOVAR=feature1_value|feature2_value|feature3_value|...? to insert Annovar specific annotations in the vcf body?

3. Creating .avinput file from a large vcf file is time consuming. Some annotations may fail due to memory limitations (we use a cluster here that require specify amount of memory before job submission). Is it possible to restart the annotation without regenerating the .avinput file after making memory requirement change in the job scripts?

4. I also have a suggestion: for large vcf files (multiple samples), it would be faster to create an intermediate file by just cutting the first few columns in the vcf and then do the annotation. After annotation is done, merge the annotation results to the original vcf. By doing this way, I found annovar uses much less memory and runs much faster. Can annovar run this way?

^ | v • Reply • Share ›



Kai Wang Mod → Jianxin Wang • 9 months ago

1. You can manually change it to float. The String is default for most annotations to be compatible with all annotations (say for example, an annotation type may be float numbers, but may have many missing values as "NA" or ".", so if this is the case, treating this as float will cause complaints from VCF analysis tools.

2. It is definitely possible, just have not got the time or request to do it, and in a way it makes parsing more difficult for most people.

3. You can change input type from vcf to avinput (delete -vcfinput argument), generate annotation, then back-convert to VCF. (The exact command line is usually printed when you run table_annovar, you can test on a tiny file first to see what's the command, and then follow the command)

4. Yes, that's what I recommended to many people especially when a file contains thousands of samples. VCF is tab-delimited (except the # lines) so this approach is very appropriate.

^ | v • Reply • Share ›



Seongmin Choi • 9 months ago

Dear Dr. Wang,

Would it be allowed for my open-source script to download hg19_dbnsfp33a.txt directly from the database link given in ANNOVAR? Thanks

^ | v • Reply • Share ›



Kai Wang Mod → Seongmin Choi • 9 months ago

no problem.

^ | v • Reply • Share ›

v: latest ▼



Emma Athan • 9 months ago

Hi Dr. Wang,

can I use Annovar scripts to split a vcf to chr-vcf, for example during conversion from vcf to vcf4?

and at the end of the annotation to merge them?

Thanks

^ | v • Reply • Share ›



Kai Wang Mod ➔ Emma Athan • 9 months ago

I do not quite understand the question, what is chr-vcf? Do you mean a VCF that contains only one chromosome, so your original VCF is split into 24 VCFs? This is certainly okay.

^ | v • Reply • Share ›



Leandro Lima • 9 months ago

Hi, Kai.

I'm getting problems when trying to download some database files for hg38.

The command I'm running is:

```
"perl annotate\_variation.pl -buildver hg38 -downdb -webfrom annovar esp6500si_all humandb/"
```

And the output is:

```
"NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
```

2010-2018 ANNOVAR

Documentation built with MkDocs (<http://www.mkdocs.org/>).