Annovar Documentation

Introduction:

ANNOVAR (Annotate Variation) is a powerful tool used for annotating genetic variants detected from high-throughput sequencing data. This documentation provides step-by-step instructions for installing ANNOVAR and annotating a VCF (Variant Call Format) file with genomic annotations.

Step 1: Download ANNOVAR:

1. Visit the ANNOVAR website: http://annovar.openbioinformatics.org/en/latest/
2. Navigate to the Downloads section.
3. Download the latest version of ANNOVAR for your operating system (Linux, MacOS, or Windows).

Step 2: Extract ANNOVAR:

4. Once the download is complete, extract the ANNOVAR package to a directory of your choice.
5. You can use a file archiver utility or run the appropriate command for your operating system to extract the files.

Step 3: Install ANNOVAR:

6. Open a terminal or command prompt.
7. Navigate to the directory where ANNOVAR was extracted.
8. Run the following command to install ANNOVAR:

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

annotate_variation.pl -buildver hg19 -downdb cytoBand humandb/

annotate_variation.pl -buildver hg19 -downdb -webfrom annovar exac03 humandb/

annotate_variation.pl -buildver hg19 -downdb -webfrom annovar avsnp147 humandb/

annotate_variation.pl -buildver hg19 -downdb -webfrom annovar dbnsfp30a humandb/

Step 4: Prepare Input VCF File:

9. Obtain the VCF file containing genetic variants that you want to annotate.
10. Ensure that the VCF file is properly formatted and contains the necessary information, such as chromosome, position, reference allele, alternative allele, etc.

Step 5: Convert VCF to ANNOVAR Input Format:

11. Run the following command to convert the VCF file to ANNOVAR input format:

annovar/convert2annovar.pl -format vcf4 akj.vcf > akj_input.avinput

output file : “akj_input.avinput”

Step 6: Annotate Variants with ANNOVAR:

12. Run the following command to annotate the variants using ANNOVAR:

annovar/table_annovar.pl akj_input.avinput annovar/humandb/ -buildver hg19 -out akj_anno -remove -protocol refgene,cytoBand,exac03,avsnp147,dbnsfp30a -operation gx,r,f,f,f -nastring . -csvout -polish -xref annovar/example/gene_xref.txt

Here's a breakdown of the command and its options:

- **-buildver hg19:** Specifies the genome build version to be used for annotation, in this case, hg19.
- **-out akj_anno:** Defines the prefix for output files generated during the annotation process.
- **-remove:** Removes temporary files after the annotation process is complete.
- **-protocol refgene,cytoBand,exac03,avsnp147,dbnsfp30a:** Specifies the databases and annotation protocols to be used. In this command, annotations are requested from the RefSeq gene annotation (refgene), cytogenetic band (cytoBand), ExAC version 0.3 (exac03), dbSNP version 147 (avsnp147), and dbNSFP version 3.0a (dbnsfp30a).
- **-operation gx,r,f,f,f:** Specifies the type of operation to be performed for each database. "gx" stands for gene-based annotation, "r" for region-based annotation, and "f" for filter-based annotation. Multiple "f" operations indicate filtering based on multiple databases.
- **-nastring .:** Specifies the string to be used for missing values in the output files. In this case, "." is used.
- **-csvout:** Requests output files in CSV (comma-separated values) format.
- **-polish:** Polish the output format, making it more readable.
- **-xref annovar/example/gene_xref.txt:** Specifies a cross-reference file to be used for gene-based annotation. This file likely contains additional information about genes.

output file : "akj_anno.hg19_multianno.csv".

13. Execute the provided commands sequentially.
14. The initial commands will download the necessary databases into the directory named "humandb/".
15. The last command executes TABLE_ANNOVAR, utilizing the ExAC version 0.3 (abbreviated as exac03), dbNFSP version 3.0a (referred to as dbnsfp30a), and dbSNP version 147 with left-normalization (known as avsnp147) databases. It will also remove all temporary files and generate an output file named **"akj_anno.hg19_multianno.csv".**
16. Any fields lacking annotations will be denoted by a "." string.

A. For Gene-based annotation

The annotate_variation.pl program is the core program in ANNOVAR.

annovar/annotate_variation.pl -geneanno -dbtype refGene -buildver hg19 akj_input.avinput annovar/humandb/

output file :

1. **akj_input.avinput.exonic_variant_function**
2. **akj_input.avinput.variant_function**

√ In the variant_function file, the first and second column annotate variant effects on gene structure and the genes that are affected, yet the other columns are reproduced from input file.

√ In the exonic_variant_function file, the first, second and third column annotate variant line number in input file, the variant effects on coding sequences and the gene/transcript being affected, yet the other columns are reproduced from input file.

B. For region-based annotation

The annotate_variation.pl program is the core program in ANNOVAR.

annovar/annotate_variation.pl -regionanno -dbtype cytoBand -buildver hg19 akj_input.avinput annovar/humandb/

output file : akj_input.avinput.hg19_cytoBand

√ The first column shows cytoBand, the second column shows the annotation results, and the other columns are reproduced from input file.

C. For filter-based annotation

The annotate_variation.pl program is the core program in ANNOVAR.

annovar/annotate_variation.pl -filter -dbtype exac03 -buildver hg19 akj_input.avinput annovar/humandb/

output file :

1. **akj_input.avinput.hg19_exac03_dropped**
2. **akj_input.avinput.hg19_exac03_filtered**

√ The software distinguishes a subset of variants within the **akj_input.avinput** file that do not appear in the exac03 database (preserved in **akj_input.avinput.hg19_exac03_filtered**), and those that are present with allele frequencies (retained in **akj_input.avinput.hg19_exac03_dropped** file).