ANNOVAR Tutorial

BYOB Presentation 04-14-15

Thomas A. Peterson

Outline

- Background
- Installing & Configuring ANNOVAR to work with different reference genomes
- Gene-Based Annotation
- Region-Based Annotation
- Filter-Based Annotation

ANNOVAR: What is this thing?

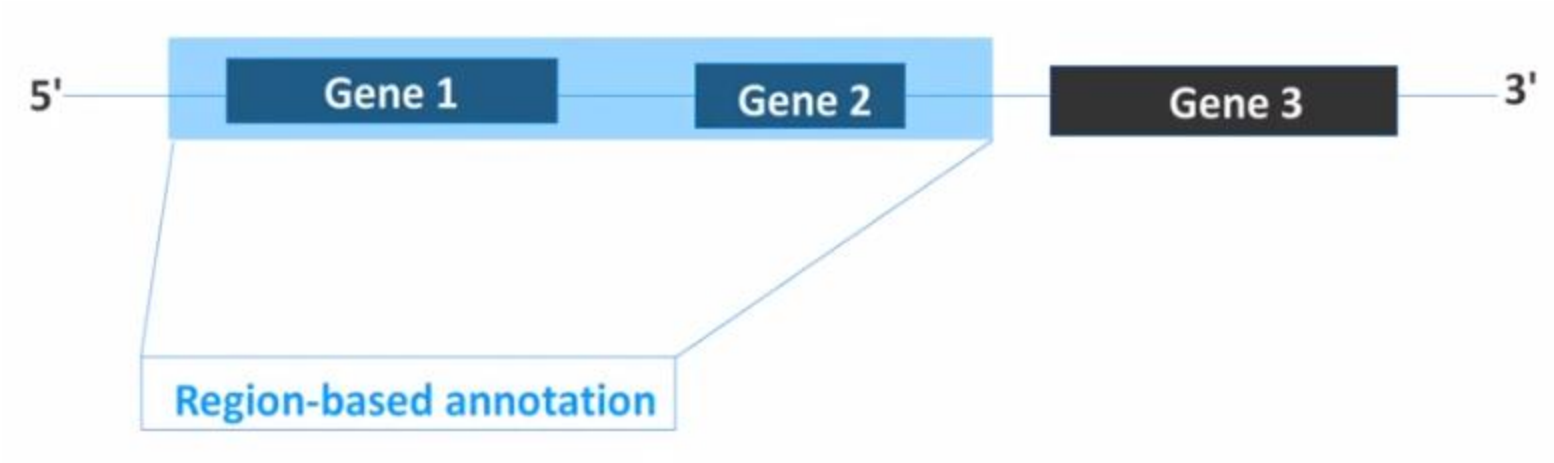
- ANNOVAR™ is a command line program that annotates (DNA-Level) genetic variants from high-throughput sequencing data
- Works with most reference genome versions (hg18, hg19, etc.) and several organisms (human, mouse, fly, yeast, etc.)
- The tool is free with premium options.

BIOBASE
BIOLOGICAL DATABASES
A QIAGEN® Company

Gene-Based Annotations

- On what gene is the variant located?
  - RefSeq
  - UCSC genes
  - ENSEMBL
  - GENCODE
  - UniProt
  - Etc.
- Distance to the nearest gene (intergenic)
- Distance to nearest exon/intron boundary

Region-Based Annotations



- ENCODE regions
  - FAIRE/Dnase/Methylation Peaks
  - Conserved regions
  - Etc.
- Transcription Factor Binding Sites (TFBS)
  - The Transfac® database can be purchased (another BioBase resource), which contains putative TFBS.
- Segmental duplication regions
- User-specified regions can also be used.

Variant-Based Annotations pt.1

- Known variants from (dbSNP, OMIM, etc) and regions with known function (e.g., TFBS, ENCODE FAIRE/DNase Peaks, conserved regions, etc.)
  - The Genome Trax™ database can be purchased (another BioBase tool), which contains variants from the Human Gene Mutation Database (HGMD), and the Pharmacogenomic Gene Mutation Database (PGMD®)
  - Additional variants specified by the user can also be used.

Variant-Based Annotations pt.2

- Any tool that has annotated dbNSFP (the database for non-synonymous SNPs functional prediction)
  - SIFT scores, PolyPhen2 HDIV scores, PolyPhen2 HVAR scores, LRT scores, MutationTaster scores, MutationAssessor score, FATHMM scores, GERP++ scores, PhyloP scores and SiPhy scores

Outline

- Background
- Installing & Configuring ANNOVAR to work with different reference genomes
- Gene-Based Annotation
- Region-Based Annotation
- Filter-Based Annotation

ANNOVAR Documentation

Search docs

User Guide

Download ANNOVAR

ANNOVAR main package

Additional databases

- For gene-based annotation

- For filter-based annotation

Quick Start-Up Guide

Prepare Input Files

Gene-based Annotation

Region-based Annotation

Filter-based Annotation

Misc

Accessoary Programs

FAQ

What is New

Version History

Credit

GitHub « Previous Next »

Docs » Download ANNOVAR

Edit on GitHub

ANNOVAR main package

Please join the ANNOVAR mailing list at google groups here to receive announcements on software updates.

The latest version of ANNOVAR (2015Mar22) can be downloaded here (registration required).

ANNOVAR is written in Perl and can be run as a standalone application on diverse hardware systems where standard Perl modules are installed.

Additional databases

Many of the databases that ANNOVAR uses can be directly retrieved from UCSC Genome Browser Annotation Database by `-downdb` argument.

Several very commonly used annotation databases for human genomes are additionally provided below. In general, users can use `-downdb -webfrom annovar` in ANNOVAR directly to download these databases.

- For gene-based annotation

Build Table Name	Explanation	Date
hg18 refGene	FASTA sequences for all annotated transcripts in RefSeq Gene	20150322

Download the latest version from the “Download ANNOVAR” section and extract the annovar.latest.tar.gz to the install directory

ANNOVAR Setup

```
tpeterson@dionysus:~/annovar$ ls -lh
total 448K
-rwxr-xr-x 1 tpeterson tpeterson 197K Mar 25 02:33 annotate_variation.pl
-rwxr-xr-x 1 tpeterson tpeterson  12K Mar 25 02:33 coding_change.pl
-rwxr-xr-x 1 tpeterson tpeterson 151K Mar 25 02:33 convert2annovar.pl
drwxr-xr-x 2 tpeterson tpeterson 4.0K Mar 25 02:33 example
drwxr-xr-x 3 tpeterson tpeterson 4.0K Mar 25 02:33 humandb
-rwxr-xr-x 1 tpeterson tpeterson  19K Mar 25 02:33 retrieve_seq_from_fasta.pl
-rwxr-xr-x 1 tpeterson tpeterson  32K Mar 25 02:33 table_annovar.pl
-rwxr-xr-x 1 tpeterson tpeterson  21K Mar 25 02:33 variants_reduction.pl
```

Several Perl scripts are in the install directory, we'll go over what most of these do.

Don't worry, you won't need to read or write any Perl!

ANNOVAR Setup

```
tpeterson@dionysus:~/annovar$ ls -lh
total 448K
-rwxr-xr-x 1 tpeterson tpeterson 197K Mar 25 02:33 annotate_variation.pl
-rwxr-xr-x 1 tpeterson tpeterson  12K Mar 25 02:33 coding_change.pl
-rwxr-xr-x 1 tpeterson tpeterson 151K Mar 25 02:33 convert2annovar.pl
drwxr-xr-x 2 tpeterson tpeterson 4.0K Mar 25 02:33 example
drwxr-xr-x 3 tpeterson tpeterson 4.0K Mar 25 02:33 humandb
-rwxr-xr-x 1 tpeterson tpeterson  19K Mar 25 02:33 retrieve_seq_from_fasta.pl
-rwxr-xr-x 1 tpeterson tpeterson  32K Mar 25 02:33 table_annovar.pl
-rwxr-xr-x 1 tpeterson tpeterson  21K Mar 25 02:33 variants_reduction.pl
tpeterson@dionysus:~/annovar$ ls -lh ./humandb/
total 171M
drwxr-xr-x 2 tpeterson tpeterson 4.0K Mar 25 02:33 genometrax-sample-files-gff
-rw-r--r-- 1 tpeterson tpeterson  20K Mar 25 02:33 GRCh37_MT_ensGeneMrna.fa
-rw-r--r-- 1 tpeterson tpeterson 3.1K Mar 25 02:33 GRCh37_MT_ensGene.txt
-rw-r--r-- 1 tpeterson tpeterson 5.9K Mar 25 02:33 hg19_example_db_generic.txt
-rw-r--r-- 1 tpeterson tpeterson 2.0M Mar 25 02:33 hg19_example_db_gff3.txt
-rw-r--r-- 1 tpeterson tpeterson  23K Mar 25 02:33 hg19_MT_ensGeneMrna.fa
-rw-r--r-- 1 tpeterson tpeterson 3.2K Mar 25 02:33 hg19_MT_ensGene.txt
-rw-r--r-- 1 tpeterson tpeterson 155M Mar 25 02:33 hg19_refGeneMrna.fa
-rw-r--r-- 1 tpeterson tpeterson  15M Mar 25 02:33 hg19_refGene.txt
tpeterson@dionysus:~/annovar$
```

- The default install comes with tables from the UCSC Genome Browser Annotation Database for human GRCh37 and hg19.
- We'll need to configure ANNOVAR to work with the reference genomes we're interested in.

Configuring ANNOVAR: Downloading Reference Genomes

- The –downdb command will retrieve reference genomes from a specified source, in this case we're using ANNOVAR's web resource.
- First, we'll download the hg18 human reference genome using the command:

Version

perl annotate_variation.pl -buildver hg18 -downdb -webfrom annovar refGene humandb/

Source Database Output Directory

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -buildver hg18 -downdb -webfrom annovar refGene humandb/
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg18_refGene.txt.gz ... OK
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg18_refGeneMrna.fa.gz ... OK
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg18_refGeneVersion.txt.gz ... OK
NOTICE: Uncompressing downloaded files
NOTICE: Finished downloading annotation files for hg18 build version, with files saved at the 'humandb' directory
tpeterson@dionysus:~/annovar$
```

Configuring ANNOVAR: Downloading Gene Definitions

- ANNOVAR comes pre-loaded with gene definitions from RefSeq. If you want genes from UCSC KnownGenes, Ensembl, or UniProt, you can download them:

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl --downdb knownGene humandb
NOTICE: The --buildver is set as 'hg18' by default
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/knownGene.txt.gz ... OK
NOTICE: Downloading annotation database http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/kgXref.txt.gz ... OK
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg18_knownGeneMrna.fa.gz ... OK
NOTICE: Uncompressing downloaded files
NOTICE: Finished downloading annotation files for hg18 build version, with files saved at the 'humandb' directory
tpeterson@dionysus:~/annovar$
```

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -buildver hg19 --downdb knownGene humandb
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/knownGene.txt.gz ... OK
NOTICE: Downloading annotation database http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/kgXref.txt.gz ... OK
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg19_knownGeneMrna.fa.gz ... OK
NOTICE: Uncompressing downloaded files
NOTICE: Finished downloading annotation files for hg19 build version, with files saved at the 'humandb' directory
```

- By default, RefGene hg18 is used. (Always use the –buildver hg19 flag)

Configuring ANNOVAR: Downloading dbSNP, 1000 Genomes, etc.

- Downloading known variants is as simple as downloading reference genomes and gene definitions:

dbSNP build 138:

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -downdb -buildver hg19 -webfrom annovar snp138 humandb
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg19_snp138.txt.gz ...
```

1000 Genomes:

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -downdb 1000g2012apr humandb -buildver hg19
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg19_1000g2012apr.zip ...
```

Configuring ANNOVAR: Downloading dbSNP, 1000 Genomes, etc.

- Downloading known variants is as simple as downloading reference genomes and gene definitions:

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -downdb -buildver hg19 -webfrom annovar snp138 humandb
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg19_snp138.txt.gz ... ^Z
[5]+  Stopped                 perl annotate_variation.pl -downdb -buildver hg19 -webfrom annovar snp138 humandb
tpeterson@dionysus:~/annovar$
tpeterson@dionysus:~/annovar$
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -downdb 1000g2012apr humandb -buildver hg19
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg19_1000g2012apr.zip ... ^Z
[6]+  Stopped                 perl annotate_variation.pl -downdb 1000g2012apr humandb -buildver hg19
tpeterson@dionysus:~/annovar$
```

- I stopped these downloads because they will take hours!

Configuring ANNOVAR: Downloading SIFT, PolyPhen Scores, etc.

- Similar to previous downloads:

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -downdb ljb23_sift humandb -buildver hg19
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/ljb23_sift.txt.gz .
```

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -downdb ljb23_pp2hdiv humandb -buildver hg19
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/ljb23_pp2hdiv.txt.gz ...
```

Outline

- Background
- Installing & Configuring ANNOVAR to work with different reference genomes
- Input Files
- Gene-Based Annotation
- Region-Based Annotation
- Filter-Based Annotation

Variant Call Format (VCF) Files

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER INFO                 FORMAT     SAMPLE1   SAMPLE2
1        1   .     ACG  A,AT    .   PASS   .                    GT:DP      1/2:13    0/0:29
1        2   rs1   C    T,CT    .   PASS   H2;AA=T              GT:GQ      0|1:100   2/2:70
1        5   .     A    G       .   PASS   .                    GT:GQ      1|0:77    1/1:95
1      100   .     T    <DEL>   .   PASS   SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3 0/0:20
```


Image source: http://bioinf.comav.upv.es/courses/sequence_analysis/snp_calling.html

Our Example VCF File

- We'll use the first 20,000 lines of dbSNP build 141

```
##fileformat=VCFv4.0
##fileDate=20140813
##source=dbSNP
##dbSNP_BUILD_ID=141
##reference=GRCh37.p13
##phasing=partial
##variationPropertyDocumentationUrl=ftp://ftp.ncbi.nlm.nih.gov/snp/specs/dbSNP_BitField
##INFO=<ID=RS,Number=1,Type=Integer,Description="dbSNP ID (i.e. rs number)">
##INFO=<ID=RSPOS,Number=1,Type=Integer,Description="Chr position reported in dbSNP">
##INFO=<ID=RV,Number=0,Type=Flag,Description="RS orientation is reversed">
```

……

```
#CHROM  POS     ID        REF     ALT     QUAL    FILTER  INFO
1       10019   rs376643643     TA      T       .       .       RS=376643643;RSPOS=10020
1       10055   rs373328635     T       TA      .       .       RS=373328635;RSPOS=10056
1       10108   rs62651026      C       T       .       .       RS=62651026;RSPOS=10108;
1       10109   rs376007522     A       T       .       .       RS=376007522;RSPOS=10109
1       10139   rs368469931     A       T       .       .       RS=368469931;RSPOS=10139
1       10144   rs144773400     TA      T       .       .       RS=144773400;RSPOS=10145
1       10146   rs375931351     AC      A       .       .       RS=375931351;RSPOS=10147
```

First we need to convert VCF into ANNOVAR's input format

```
tpeterson@dionysus:~/annovar$ perl convert2annovar.pl -format vcf4 ./example/Exam
ple_SNVs.vcf -outfile ./example/Example_SNVs.annovar
NOTICE: Finished reading 200000 lines from VCF file
```

1	10020	10020	A	-	.	.	.
1	10055	10055	-	A	.	.	.
1	10108	10108	C	T	.	.	.
1	10109	10109	A	T	.	.	.
1	10139	10139	A	T	.	.	.
1	10145	10145	A	-	.	.	.
1	10147	10147	C	-	.	.	.
1	10150	10150	C	T	.	.	.
1	10177	10177	A	C	.	.	.
1	10177	10177	-	C	.	.	.
1	10180	10180	T	C	.	.	.
1	10229	10229	A	-	.	.	.
1	10229	10255	AACCCCTAACCCTAACCCTAAACCCTA	-	.	.	.
1	10231	10231	C	-	.	.	.
1	10231	10231	C	A	.	.	.
1	10234	10234	C	T	.	.	.
1	10248	10248	A	T	.	.	.
1	10250	10251	AC	-	.	.	.
1	10250	10250	A	C	.	.	.
1	10255	10255	A	-	.	.	.
1	10257	10257	A	C	.	.	.
1	10259	10259	C	A	.	.	.
1	10291	10291	C	T	.	.	.
1	10327	10327	T	C	.	.	.
1	10329	10352	ACCCCTAACCCTAACCCTAACCCT	-	.	.	.
1	10330	10330	C	-	.	.	.
1	10352	10352	-	A	.	.	.
1	10383	10383	-	C			

Outline

- Background
- Installing & Configuring ANNOVAR to work with different reference genomes
- Input Files
- Gene-Based Annotation
- Region-Based Annotation
- Filter-Based Annotation

Gene-Based Annotation

- To annotate for genes, we run the annotate_variation.pl command on our VCF file:

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -buildver hg19 ./example
/Example_SNVs.annovar ./humandb/
NOTICE: The --geneanno operation is set to ON by default
NOTICE: Reading gene annotation from humandb/hg19_refGene.txt ... Done with 50914
 transcripts (including 11516 without coding sequence annotation) for 26271 uniqu
e genes
NOTICE: Reading FASTA sequences from humandb/hg19_refGeneMrna.fa ... Done with 22
5 sequences
WARNING: A total of 345 sequences will be ignored due to lack of correct ORF anno
tation
NOTICE: Finished gene-based annotation on 202514 genetic variants in ./example/Ex
ample_SNVs.annovar (including 49 with invalid format written to ./example/Example
_SNVs.annovar.invalid_input)
NOTICE: Output files were written to ./example/Example_SNVs.annovar.variant_funct
ion, ./example/Example_SNVs.annovar.exonic_variant_function
tpeterson@dionysus:~/annovar$
```

- Creates the “.variant_function” file

Variant Function File:

downstream	MIR1302-10,MIR1302-11,MIR1302-2,MIR1302-9	1	31029	31029	G	A		
downstream	MIR1302-10,MIR1302-11,MIR1302-2,MIR1302-9	1	31166	31166	C	T		
intergenic	MIR1302-2(dist=2647),FAM138F(dist=1461)	1	33150	33150	A	T	.	
intergenic	MIR1302-2(dist=2984),FAM138F(dist=1124)	1	33487	33487	C	T	.	
intergenic	MIR1302-2(dist=2992),FAM138F(dist=1116)	1	33495	33495	C	T	.	
intergenic	MIR1302-2(dist=3002),FAM138F(dist=1106)	1	33505	33505	T	C	.	
intergenic	MIR1302-2(dist=3005),FAM138F(dist=1103)	1	33508	33508	A	T	.	
intergenic	MIR1302-2(dist=3018),FAM138F(dist=1090)	1	33521	33521	T	A	.	
intergenic	MIR1302-2(dist=3090),FAM138F(dist=1018)	1	33593	33593	G	A	.	
downstream	FAM138A,FAM138F	1	33724	33724	T	C	.	.
downstream	FAM138A,FAM138F	1	33734	33734	T	C	.	.
downstream	FAM138A,FAM138F	1	33971	33971	C	G	.	.
ncRNA_exonic	FAM138A,FAM138F	1	34771	34771	G	C	.	.
ncRNA_exonic	FAM138A,FAM138F	1	34810	34810	A	T	.	.
upstream	FAM138A,FAM138F	1	37055	37055	C	T	.	.
intergenic	FAM138F(dist=2661),OR4F5(dist=30349)	1	38742	38742	C	T	.	
intergenic	FAM138F(dist=3099),OR4F5(dist=29911)	1	39180	39180	A	G	.	
intergenic	FAM138F(dist=3149),OR4F5(dist=29861)	1	39230	39230	G	A	.	
intergenic	FAM138F(dist=3166),OR4F5(dist=29844)	1	39247	39247	A	G	.	
intergenic	FAM138F(dist=3174),OR4F5(dist=29836)	1	39255	39255	A	C	.	
intergenic	FAM138F(dist=3174),OR4F5(dist=29836)	1	39255	39255	A	C	.	
intergenic	FAM138F(dist=3180),OR4F5(dist=29830)	1	39261	39261	T	C	.	
intergenic	FAM138F(dist=3550),OR4F5(dist=29460)	1	39631	39631	T	C	.	
intergenic	FAM138F(dist=3595),OR4F5(dist=29415)	1	39676	39676	C	T	.	
intergenic	FAM138F(dist=3811),OR4F5(dist=29199)	1	39892	39892	C	T	.	

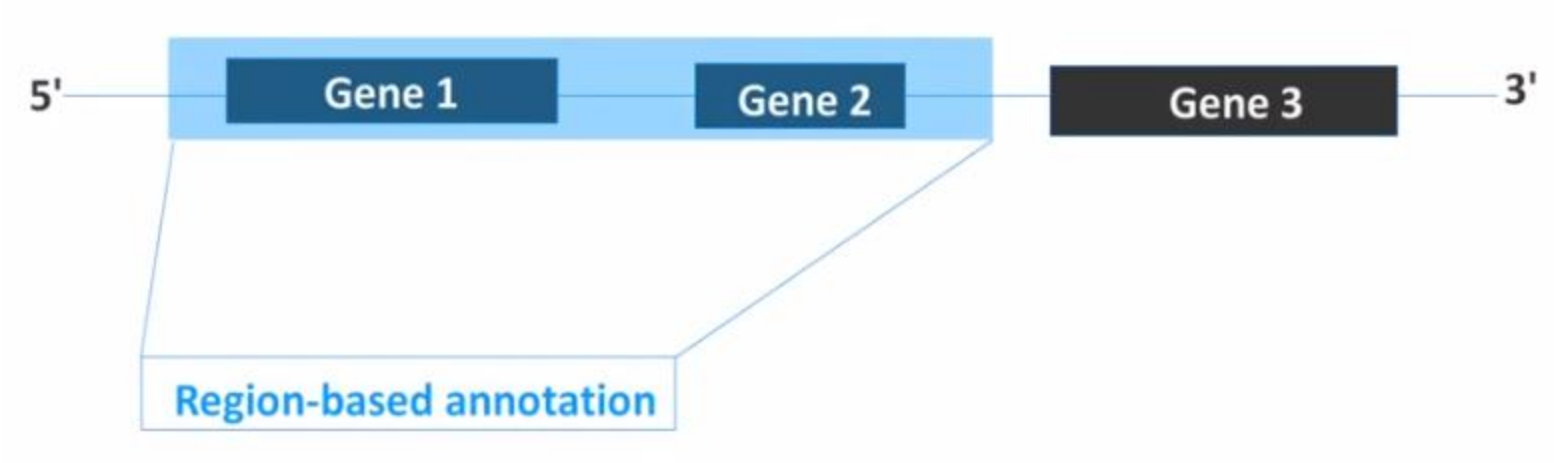
Exonic Variant Function File

line501	synonymous SNV	OR4F5:NM_001005484:exon1:c.G462C:p.A154A,	1	69552	69552	G	C	.
line502	synonymous SNV	OR4F5:NM_001005484:exon1:c.G462C:p.A154A,	1	69552	69552	G	C	.
line503	nonsynonymous SNV	OR4F5:NM_001005484:exon1:c.T479C:p.L160P,	1	69569	69569	T	C	
line504	nonsynonymous SNV	OR4F5:NM_001005484:exon1:c.T479C:p.L160P,	1	69569	69569	T	C	
line505	nonsynonymous SNV	OR4F5:NM_001005484:exon1:c.T500A:p.V167D,	1	69590	69590	T	A	
line506	synonymous SNV	OR4F5:NM_001005484:exon1:c.T504C:p.D168D,	1	69594	69594	T	C	.
line507	synonymous SNV	OR4F5:NM_001005484:exon1:c.T513C:p.Y171Y,	1	69603	69603	T	C	.
line508	nonsynonymous SNV	OR4F5:NM_001005484:exon1:c.C520T:p.L174F,	1	69610	69610	C	T	
line509	nonsynonymous SNV	OR4F5:NM_001005484:exon1:c.A671T:p.D224V,	1	69761	69761	A	T	
line510	synonymous SNV	OR4F5:NM_001005484:exon1:c.G678A:p.S226S,	1	69768	69768	G	A	.
line511	synonymous SNV	OR4F5:NM_001005484:exon1:c.T807C:p.S269S,	1	69897	69897	T	C	.
line8314	nonsynonymous SNV	SAMD11:NM_152486:exon2:c.A8G:p.K3R,	1	861329	861329	A	G	
line8315	nonsynonymous SNV	SAMD11:NM_152486:exon2:c.C28T:p.P10S,	1	861349	861349	C	T	
line8316	nonsynonymous SNV	SAMD11:NM_152486:exon2:c.C36G:p.I12M,	1	861357	861357	C	G	
line8458	synonymous SNV	SAMD11:NM_152486:exon3:c.G81A:p.G27G,	1	865543	865543	G	A	.
line8459	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.C82T:p.R28W,	1	865544	865544	C	T	
line8460	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.G83A:p.R28Q,	1	865545	865545	G	A	
line8461	synonymous SNV	SAMD11:NM_152486:exon3:c.C105T:p.V35V,	1	865567	865567	C	T	.
line8462	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.C113T:p.P38L,	1	865575	865575	C	T	
line8463	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.G118A:p.A40T,	1	865580	865580	G	A	
line8464	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.G122A:p.R41Q,	1	865584	865584	G	A	
line8465	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.G149C:p.S50T,	1	865611	865611	G	C	
line8466	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.G163A:p.D55N,	1	865625	865625	G	A	
line8467	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.G166A:p.G56S,	1	865628	865628	G	A	
line8468	synonymous SNV	SAMD11:NM_152486:exon3:c.C192T:p.T64T,	1	865654	865654	C	T	.
line8469	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.G200A:p.R67Q,	1	865662	865662	G	A	
line8470	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.C202T:p.R68W,	1	865664	865664	C	T	
line8471	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.G203A:p.R68Q,	1	865665	865665	G	A	
line8472	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.C232T:p.H78Y,	1	865694	865694	C	T	
line8473	nonsynonymous SNV	SAMD11:NM_152486:exon3:c.C238T:p.R80C,	1	865700	865700	C	T	
line8474	synonymous SNV	SAMD11:NM_152486:exon3:c.C243T:p.I81I,	1	865705	865705	C	T	.
line8492	nonsynonymous SNV	SAMD11:NM_152486:exon4:c.T257A:p.V86D,	1	866421	866421	T	A	

Outline

- Background
- Installing & Configuring ANNOVAR to work with different reference genomes
- Input Files
- Gene-Based Annotation
- Region-Based Annotation
- Filter-Based Annotation

Region-Based Annotations



- ENCODE regions
  - FAIRE/Dnase/Methylation Peaks
  - Conserved regions
  - Etc.
- Transcription Factor Binding Sites (TFBS)
  - The Transfac® database can be purchased (another BioBase resource), which contains putative TFBS.
- Segmental duplication regions
- User-specified regions can also be used.

Downloading Regions Annotated by Transfac

- This example uses the tfbsConsSites region annotation, which contains the location and score of transcription factor binding sites conserved in the human/mouse/rat alignment, where score and threshold are computed with the Transfac Matrix Database.

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -build hg19 -downdb tfbsConsSites humandb/
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/tfbsConsSites.txt.gz ...
NOTICE: Uncompressing downloaded files
NOTICE: Finished downloading annotation files for hg19 build version, with files saved at the 'humandb' directory
tpeterson@dionysus:~/annovar$
```

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -regionanno -build hg19
-dbtype tfbsConsSites ./example/Example_SNVs.annovar humandb/
NOTICE: Reading annotation database humandb/hg19_tfbsConsSites.txt ... Done with
5797266 regions
NOTICE: Finished region-based annotation on 202514 genetic variants in ./example/
Example_SNVs.annovar (including 49 with invalid format written to ./example/Examp
le_SNVs.annovar.invalid_input)
NOTICE: Output file is written to ./example/Example_SNVs.annovar.hg19_tfbsConsSit
es
tpeterson@dionysus:~/annovar$
```

Transfac Hits

tfbsConsSites	Score=898;Name=V$ELK1_01	1	894644	894644	C	T
tfbsConsSites	Score=971;Name=V$CETS1P54_01	1	894646	894646	A	T
tfbsConsSites	Score=871;Name=V$FAC1_01	1	896775	896775	C	T
tfbsConsSites	Score=871;Name=V$FAC1_01	1	896778	896778	C	T
tfbsConsSites	Score=757;Name=V$SEF1_C 1	896865	896865	C	G	.
tfbsConsSites	Score=757;Name=V$SEF1_C 1	896868	896868	G	A	.
tfbsConsSites	Score=829;Name=V$EVI1_03	1	897009	897009	A	G
tfbsConsSites	Score=861;Name=V$HAND1E47_01	1	897032	897032	A	G
tfbsConsSites	Score=828;Name=V$MYCMAX_03	1	897043	897043	G	A
tfbsConsSites	Score=788;Name=V$TCF11MAFG_01	1	897053	897053	A	G
tfbsConsSites	Score=865;Name=V$COUP_01	1	897118	897118	G	A
tfbsConsSites	Score=865;Name=V$COUP_01	1	897119	897119	G	A
tfbsConsSites	Score=865;Name=V$COUP_01	1	897120	897120	G	C
tfbsConsSites	Score=865;Name=V$COUP_01	1	897124	897124	T	C
tfbsConsSites	Score=741;Name=V$PPARG_01	1	897133	897133	G	A
tfbsConsSites	Score=788;Name=V$CMYB_01	1	897214	897214	C	T
tfbsConsSites	Score=788;Name=V$CMYB_01	1	897216	897216	C	T
tfbsConsSites	Score=788;Name=V$CMYB_01	1	897218	897218	G	A
tfbsConsSites	Score=824;Name=V$CMYB_01	1	897299	897299	C	T
tfbsConsSites	Score=765;Name=V$PAX5_02	1	897336	897336	G	T
tfbsConsSites	Score=765;Name=V$PAX5_02	1	897337	897337	C	T
tfbsConsSites	Score=765;Name=V$PAX5_02	1	897340	897340	C	T
tfbsConsSites	Score=765;Name=V$PAX5_02	1	897341	897341	C	T
tfbsConsSites	Score=765;Name=V$PAX5_02	1	897344	897344	C	T
tfbsConsSites	Score=765;Name=V$PAX5_02	1	897349	897349	G	A

Outline

- Background
- Installing & Configuring ANNOVAR to work with different reference genomes
- Input Files
- Gene-Based Annotation
- Region-Based Annotation
- Filter-Based Annotation





Configuring ANNOVAR: Downloading dbSNP, 1000 Genomes, etc.

- Downloading known variants is as simple as downloading reference genomes and gene definitions:

dbSNP build 138:

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -downdb -buildver hg19 -webfrom annovar snp138 humandb
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg19_snp138.txt.gz ...
```

1000 Genomes:

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -downdb 1000g2012apr humandb -buildver hg19
NOTICE: Web-based checking to see whether ANNOVAR new version is available ... Done
NOTICE: Downloading annotation database http://www.openbioinformatics.org/annovar/download/hg19_1000g2012apr.zip ...
```

Annotating with dbSNP or Other Variant Databases

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -filter -dbtype snp138
-buildver hg19 ./example/Example_SNVs.annovar humandb/
```

Annotating with dbSNP or Other Variant Databases

```
tpeterson@dionysus:~/annovar$ perl annotate_variation.pl -filter -dbtype snp138
-buildver hg19 ./example/Example_SNVs.annovar humandb/
NOTICE: Variants matching filtering criteria are written to ./example/Example_SN
Vs.annovar.hg19_snp138_dropped, other variants are written to ./example/Example_
SNVs.annovar.hg19_snp138_filtered
NOTICE: Processing next batch with 198801 unique variants in 202514 input lines
NOTICE: Database index loaded. Total number of bins is 2894320 and the number of
 bins to be scanned is 7976
NOTICE: Scanning filter database humandb/hg19_snp138.txt...Done
NOTICE: Variants with invalid input format were written to ./example/Example_SNV
s.annovar.invalid_input
tpeterson@dionysus:~/annovar$
```

Annotating with dbSNP or Other Variant Databases

```
5.5M  Example_SNVs.annovar
   0  Example_SNVs.annovar.hg19_snp138_dropped
 12M  Example_SNVs.annovar.hg19_snp138_filtered
   0  Example_SNVs.annovar.invalid_input
```

- All variants were “filtered”, meaning they were found in the dbSNP database
- This file now contains all dbSNP ids associated with the variants.

Questions?