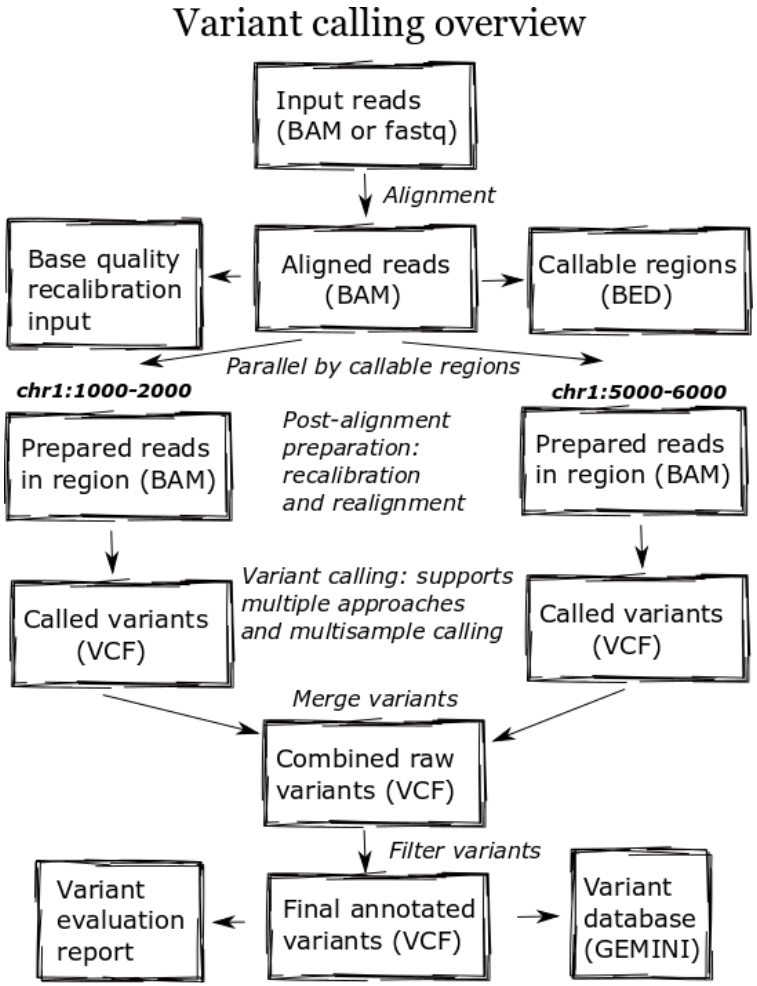
1000 Genomes Data Analysis Demo

Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2015

FastQ or BAM Files to Variants



http://bcb.io/category/variation/

From VCF file to SKAT analysis

- Example Dataset: 1000 Genome Exome Seq. Data (Chr 22)
- 16k variants
- Analysis Flow
  - Convert VCF to Plink File
  - Annotation using ANNOVAR software
  - Association test using the SKAT package

DATA format: VCF file

```
##fileformat=VCFv4.0
##FILTER=<ID=LowQual,Description="QUAL < 50.0">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth (only filtered reads used for calling)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=3,Type=Float,Description="Normalized, Phred-scaled likelihoods for AA,AB,BB genotypes where A=re
f and B=alt; not applicable if site is not biallelic">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as l
isted">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direc
tion">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with two (and only two) segregating
haplotypes">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=SB,Number=1,Type=Float,Description="Strand Bias">
##INFO=<ID=VQSLOD,Number=1,Type=Float,Description="log10-scaled probability of variant being true under the trained gau
ssian mixture model">
##UnifiedGenotyperV2="analysis_type=UnifiedGenotyperV2 input_file=[TEXT CLIPPED FOR CLARITY]"
#CHROM  POS ID      REF ALT QUAL    FILTER  INFO    FORMAT  NA12878
chr1    873762  .       T   G   5231.78 PASS    AC=1;AF=0.50;AN=2;DP=315;Dels=0.00;HRun=2;HaplotypeScore=15.11;MQ=91.05
;MQ0=15;QD=16.61;SB=-1533.02;VQSLOD=-1.5473 GT:AD:DP:GQ:PL   0/1:173,141:282:99:255,0,255
chr1    877664  rs3828047   A   G   3931.66 PASS    AC=2;AF=1.00;AN=2;DB;DP=105;Dels=0.00;HRun=1;HaplotypeScore=1.59;MQ
=92.52;MQ0=4;QD=37.44;SB=-1152.13;VQSLOD= 0.1185 GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
chr1    899282  rs28548431  C   T   71.77   PASS    AC=1;AF=0.50;AN=2;DB;DP=4;Dels=0.00;HRun=0;HaplotypeScore=0.00;MQ=9
9.00;MQ0=0;QD=17.94;SB=-46.55;VQSLOD=-1.9148 GT:AD:DP:GQ:PL  0/1:1,3:4:25.92:103,0,26
chr1    974165  rs9442391   T   C   29.84   LowQual AC=1;AF=0.50;AN=2;DB;DP=18;Dels=0.00;HRun=1;HaplotypeScore=0.16;MQ=
95.26;MQ0=0;QD=1.66;SB=-0.98 GT:AD:DP:GQ:PL  0/1:14,4:14:60.91:61,0,255
```

DATA format: VCF file

```
[HEADER LINES]
#CHROM  POS ID      REF ALT QUAL    FILTER  INFO          FORMAT          NA12878
chr1    873762  .       T   G   5231.78 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:173,141:282:99:255,0,255
chr1    877664  rs3828047   A   G   3931.66 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
chr1    899282  rs28548431  C   T   71.77   PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:1,3:4:25.92:103,0,26
chr1    974165  rs9442391   T   C   29.84   LowQual [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:14,4:14:60.91:61,0,255
```

gatkForums

- VCFtools can be used to handle a VC file and convert it to a Plink file

Format conversion: VCF to Plink

- Use vcftools to convert VCF to plink tped file
- http://vcftools.sourceforge.net/

```
# command
vcftools --gzvcf ExomeChr22.vcf.gz --plink-tped --out ExomeChr22

# LOG
VCFtools - v0.1.10
(C) Adam Auton 2009

Parameters as interpreted:
--gzvcf ExomeChr22.vcf.gz
--out ExomeChr22
--plink-tped

Using zlib version: 1.2.5
Reading Index file.
File contains 16885 entries and 1092 individuals.
....
```

Format conversion: Plink Tped to Binary

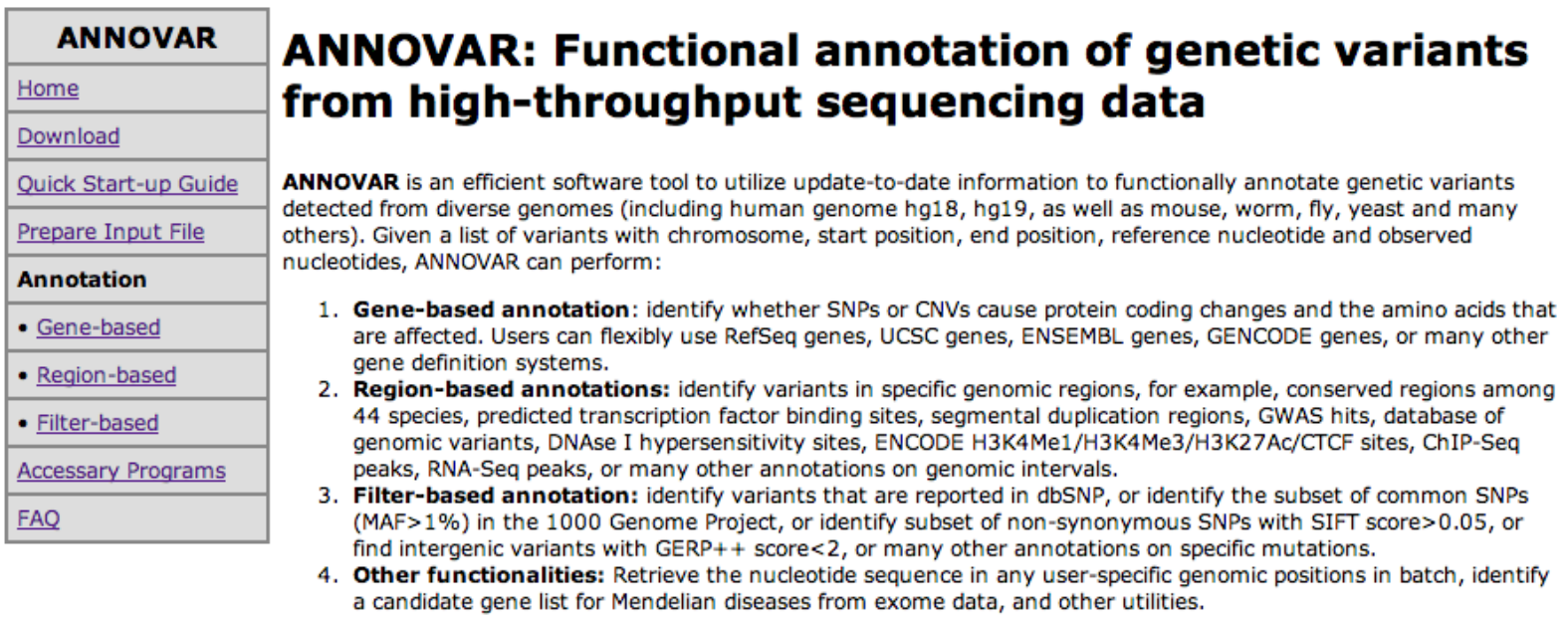
- Generate binary Plink files from TPED files.

```
plink --tfile ExomeChr22 --make-bed --out ExomeChr22
```

```
host10-125:1000Genome LEE7801$ plink --tfile ExomeChr22 --make-bed --out ExomeChr22

@----------------------------------------------------------@
|        PLINK!       |     v1.07      |   10/Aug/2009     |
|----------------------------------------------------------|
|  (C) 2009 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:  |
|        http://pngu.mgh.harvard.edu/purcell/plink/        |
@----------------------------------------------------------@

Web-based version check ( --noweb to skip )
Recent cached web-check found... OK, v1.07 is current

Writing this text to log file [ ExomeChr22.log ]
Analysis started: Thu May 15 16:22:50 2014

Options in effect:
        --tfile ExomeChr22
        --make-bed
        --out ExomeChr22
```

Annotation

- Annotate variants to genes
- Use Annovar (http://www.openbioinformatics.org/annovar/)



ANNOVAR

Home
Download
Quick Start-up Guide
Prepare Input File
Annotation
• Gene-based
• Region-based
• Filter-based
Accessary Programs
FAQ

ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data

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, 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:

1. **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, or many other gene definition systems.
2. **Region-based annotations:** 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.
3. **Filter-based annotation:** identify variants that are reported in dbSNP, or identify the subset of common SNPs (MAF>1%) in the 1000 Genome Project, or identify subset of non-synonymous SNPs with SIFT score>0.05, or find intergenic variants with GERP++ score<2, or many other annotations on specific mutations.
4. **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.

Annotation

- Command Line

```
# prepare input file
vcftools --gzvcf ExomeChr22.vcf.gz --recode  --out ExomeChr22Info

convert2annovar.pl -format vcf4 ExomeChr22Info.recode.vcf  > ExomeChr22.avinput


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

# Annovar
annotate_variation.pl -buildver hg19  ExomeChr22.avinput ./humandb/
```

- Two output files are generated
- ExomeChr22.avinput.variant_function
  - Annotation of all variants
- ExomeChr22.avinput.exonic_variant_function
  - Detailed annotation of exonic variants.

Annotation

- ExomeChr22.avinput.variant_function (annotation of all variants)

```
intronic POTEH 22 16287215 16287215 C T unknown 100
intronic POTEH 22 16287226 16287226 C T unknown 100
exonic POTEH 22 16287365 16287365 C T unknown 100
exonic POTEH 22 16287649 16287649 G A unknown 100
exonic POTEH 22 16287784 16287784 C T unknown 100
exonic POTEH 22 16287851 16287851 G A unknown 100
UTR5 POTEH 22 16287912 16287912 G A unknown 100
exonic OR11H1 22 16449075 16449075 G A unknown 100
```

Annotation

- ExomeChr22.avinput.exonic_variant_function (functional role of exonic variate)

```
line3 nonsynonymous SNV POTEH:NM_001136213:exon1:c.G521A:p.R174Q, 22 16287365 1
line4 synonymous SNV POTEH:NM_001136213:exon1:c.C237T:p.N79N, 22 16287649 16287
line5 stopgain SNV POTEH:NM_001136213:exon1:c.G102A:p.W34X, 22 16287784 1628778
line6 nonsynonymous SNV POTEH:NM_001136213:exon1:c.C35T:p.S12F, 22 16287851 162
line8 nonsynonymous SNV OR11H1:NM_001005239:exon1:c.C730T:p.P244S, 22 16449075
line9 nonsynonymous SNV CCT8L2:NM_014406:exon1:c.G1441C:p.G481R, 22 17072000 17
```

Advanced Usage of SKAT package

- SKAT package can read binary plink files
- The following files are needed:
  - Plink Bed, Bim and FAM files
  - Set ID file (2 columns): file to link variant (SNP) to variant sets.
    - Set ID
    - Variant ID (ex. rs number)
  - Covariate file : Plink covariate file.

SKAT File Preparation: Generate Set ID file

- Generate Set ID file using ANNOVA output
- Use GetSetID function

```
> File.Annovar<-"./ExomeChr22.avinput.variant_function"
> File.BIM<-"./ExomeChr22.bim"
> File.SetID<-"./ExomeChr22.SetID"
>
> GetSetID(File.Annovar, File.BIM, File.SetID, Include=c("exonic", "splicing"))
10029 variants in ANNOVAR file [./ExomeChr22.avinput.variant_function] belong t
BIM file [./ExomeChr22.bim] has 16885 variants.
399 genes (SNP sets) and total 10154 variants are saved in SetID file.
```