Sanity check of FASTQ files

```
fastqc seqfile1 seqfile2
```

Source: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Align to reference genome

Create BWA index of reference fasta:

```
bwa index GRCh37.fasta
```

BWA alignment to human reference:

```
bwa mem -t 6 bwa_index/GRCh37.fasta NA24385_R1.fastq.gz
NA24385_R2.fastq.gz | sambamba view --format=bam -S -p -l 0 -o
NA24385_bwamem.bam -t 6 /dev/stdin
```

Inspecting BAM file

# lets take a look at some alignments from the command line

```
Samtools tview NA24385_bwamem.bam GRCh37.fasta
```

You can toggle a mini help screen by pressing the ? key, and from here determine the main keys for navigating across your reads. For example, you can use the h,j,k,l keys (or the cursor keys) to make small movements. We will use this to spot check some variant calls made downstream in our pipeline.

# of mapped reads

```
samtools view -F 0x04 -c NA24385_bwamem.bam
```

# of unmapped reads

```
samtools view -f4 -c NA24385_bwamem.bam
```

Get all basic alignment metrics

```
/gpfs0/software/sambamba/latest flagstat -t 6 -p NA24385_bwamem.bam
```

```
less NA24385_samtools_flagstats.out
```

Sort BAM file

```
/gpfs0/software/sambamba/latest sort -m 12GB -o
NA24385_bwamem_sorted.bam -l 9 -u -p -t 7 --tmpdir=./
NA24385_bwamem.bam
```

Mark PCR duplicates

```
/gpfs0/software/sambamba/latest markdup -t 4 -l 9
NA24385_bwamem_sorted.bam NA12879_bwamem_sorted_dedup.bam
```

Run Variant Caller

```
python Platypus_0.8.1/Platypus.py callVariants \
--regions="1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,M
T,X,Y" \
--trimOverlapping=0 \
--refFile=
/gpfs0/genomics_tmp/datasources/human_g1k_v37_decoyPlus.fasta \
--maxReads=99999999 \
--bamFiles=$BAM \
--nCPU=8 \
--output=${OUT}.vcf
```

Extract PASS calls

```
awk -F '\t' '{if($0 ~ /\#/) print; else if($7 == "PASS") print}'
NA24385_bwamem_10pct_sorted_fixed.vcf >
NA24385_bwamem_10pct_sorted_fixed_PASS.vcf
```

Apply coverage and variant quality filters

```
vcffilter -f "TC > 9 & MQ > 30 & QD > 20"
NA24385_bwamem_10pct_sorted_fixed_PASS.vcf > NA24385_bwamem_10p
ct_sorted_fixed_PASS_QCfiltered.vcf
```

How many variants did we filter OUT so far?

```
grep -v “^#” NA24385_bwamem_10pct_sorted_fixed_PASS.vcf | wc -l

grep -v “^#” NA24385_bwamem_10pct_sorted_fixed_PASS_QCfiltered.vcf |
wc -l
```

Annotate VCF file

Make sure VCF file is sorted first:

```
vcf-sort -c NA24385_bwamem_10pct_sorted_fixed_PASS_QCfiltered.vcf >
NA24385_bwamem_10pct_sorted_fixed_PASS_QCfiltered.sorted.vcf

time perl vep --offline --numbers --domains --symbol --ccds
--check_existing --variant_class --sift b --polyphen b
--gene_phenotype --regulatory --total
_length --terms SO --fork 4 -i
NA24385_bwamem_10pct_sorted_fixed_PASS_QCfiltered.sorted.vcf -o
NA24385_bwamem_10pct_sorted_fixed_PASS_QCfiltered.vep_anno.out
--force --dont_skip --failed 1 --buffer_size 5000
```

Or, an easier path to annotate:
http://grch37.ensembl.org/Homo_sapiens/Tools/VEP/Ticket?tl=QBLMFgjomGam6BMY

For more information on the options:
http://www.ensembl.org/info/docs/tools/vep/script/vep_options.html

FIlters based on:

- Allele frequencies:
- NHLBI Exome Sequencing Project 5400
- 1000 Genomes Phase3
- GNOMAD
- ExAC

Predictions:

- dbSNFP

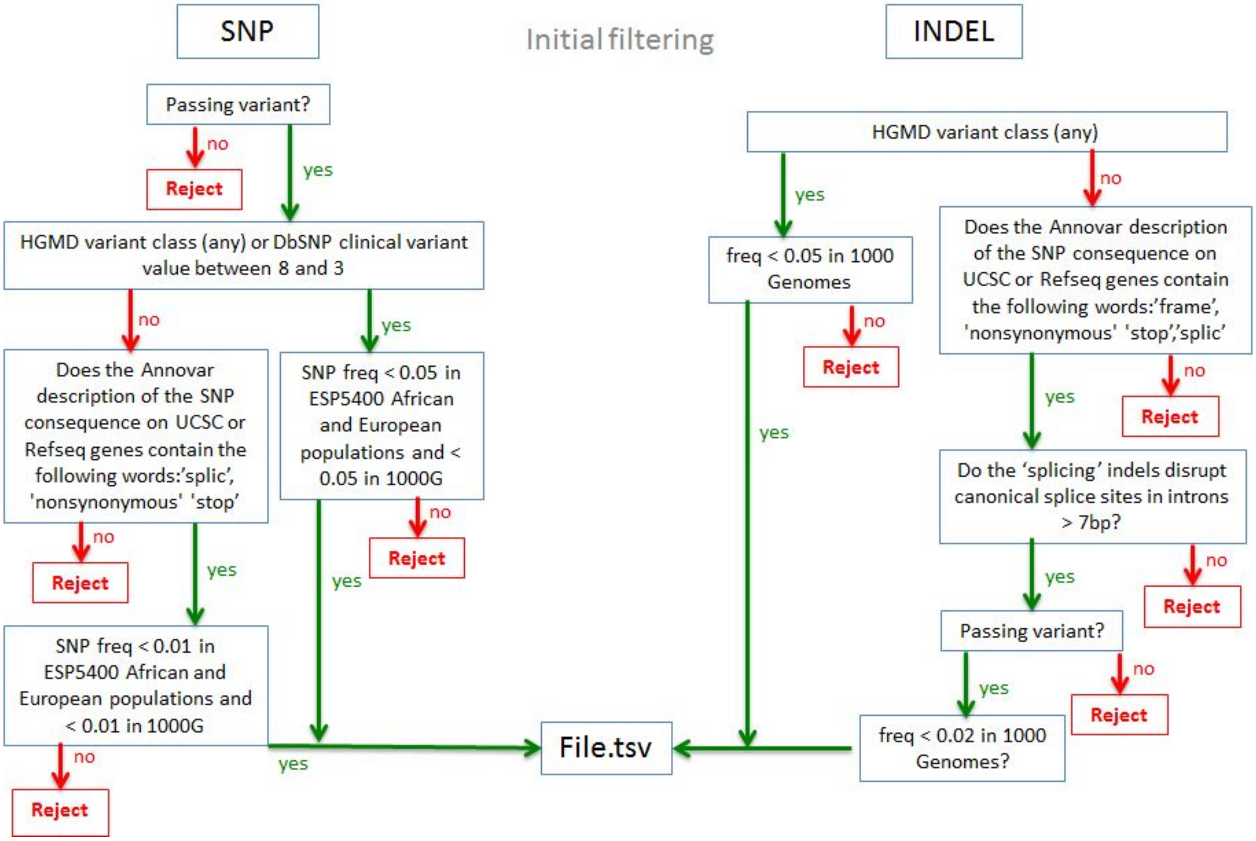
Public & private databases:

- dbSNP - common variants
- HGMD

Known genomic characteristics:

- Mappability Score: https://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=606555319_G7UCOGIsfPSkkiesRADUVNjqjs3E&c=chr21&g=wgEncodeMapability

Prioritize variants


SNP
Initial filtering
INDEL
Passing variant?
no
Reject
yes
HGMD variant class (any) or DbSNP clinical variant value between 8 and 3
no
yes
Does the Annovar description of the SNP consequence on UCSC or Refseq genes contain the following words:'splic', 'nonsynonymous' 'stop'
SNP freq < 0.05 in ESP5400 African and European populations and < 0.05 in 1000G
no
Reject
no
Reject
yes
yes
SNP freq < 0.01 in ESP5400 African and European populations and < 0.01 in 1000G
no
Reject
yes
File.tsv
HGMD variant class (any)
yes
no
freq < 0.05 in 1000 Genomes
no
Reject
yes
Does the Annovar description of the SNP consequence on UCSC or Refseq genes contain the following words:'frame', 'nonsynonymous' 'stop','splic'
yes
no
Reject
Do the 'splicing' indels disrupt canonical splice sites in introns > 7bp?
yes
no
Reject
Passing variant?
yes
no
Reject
freq < 0.02 in 1000 Genomes?

Assignment: Write code in your favorite language to accomplish the above workflow