Introduction to Exome Sequencing and Variant Analysis

Bioinformatics and Computational Bioscience Branch (BCBB)

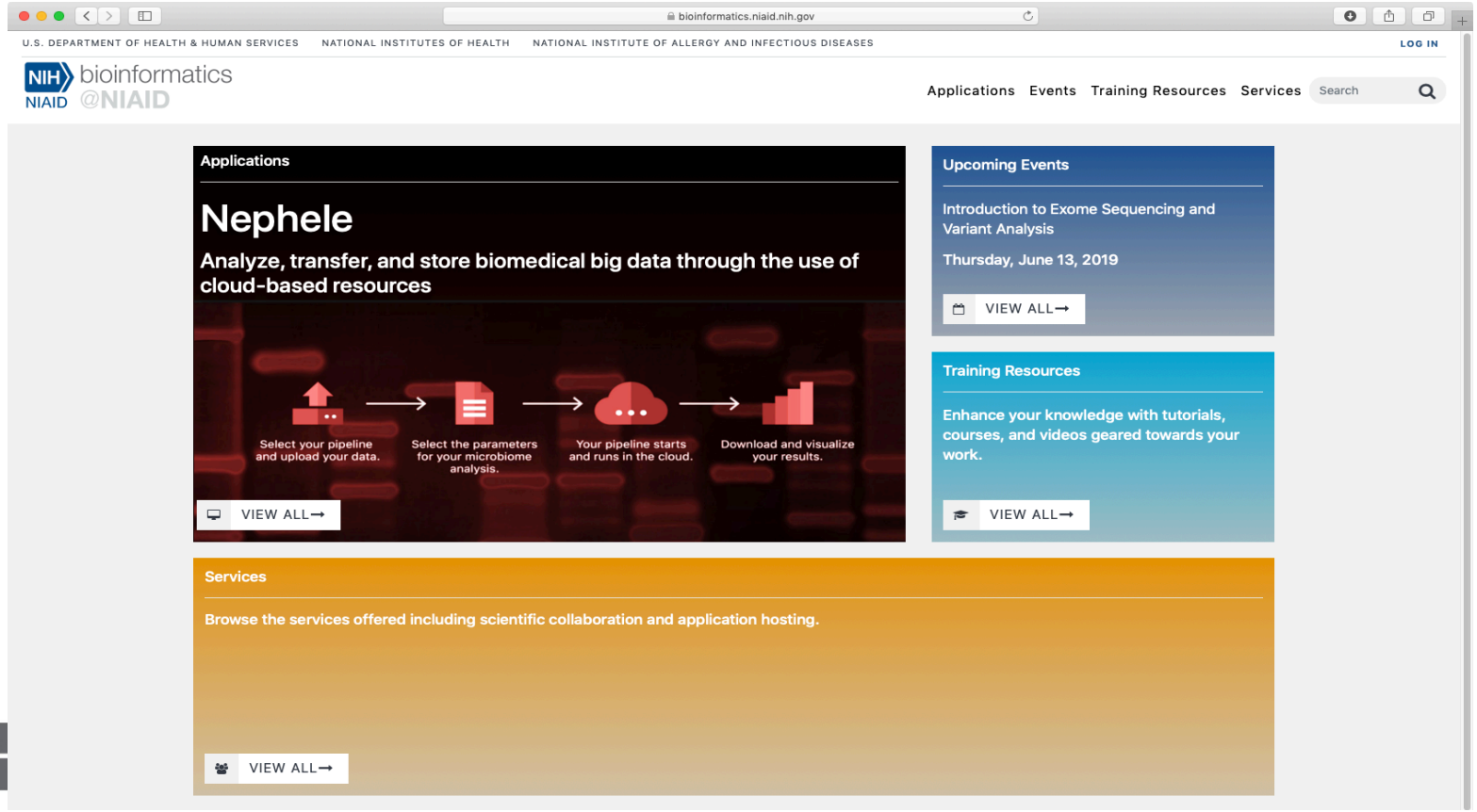
Ke Huang, Ph.D.
Daniel Veltri, Ph.D.
Sandhya Xirasagar, Ph.D.

2019-6-13

NIH National Institute of Allergy and Infectious Diseases

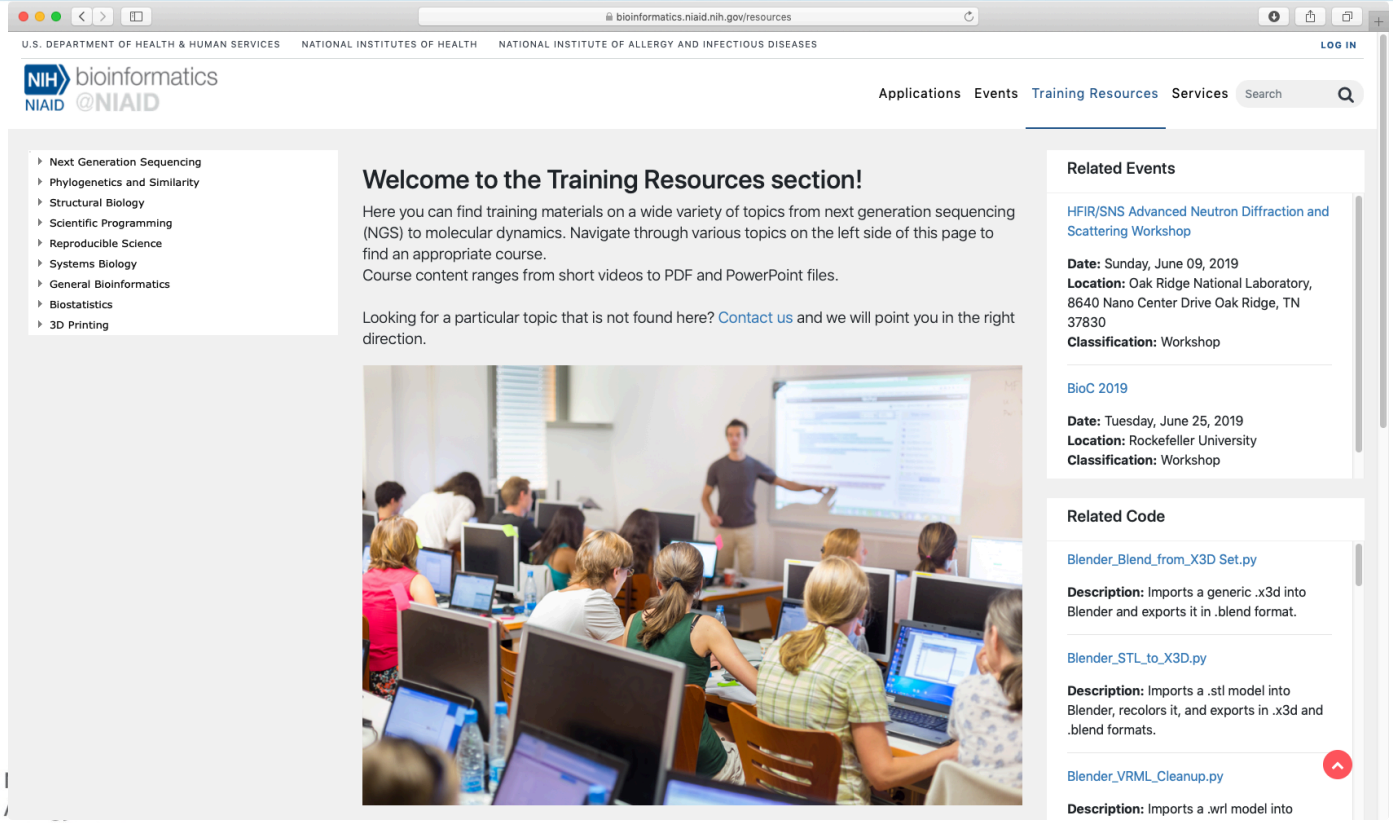
NIAID

https://bioinformatics.niaid.nih.gov



NIAID

Additional Training Resources



NIH Infectious Diseases

NIAID

Find more information on BCBB

- Bioinformatics @ NIAID portal
  - https://bioinformatics.niaid.nih.gov/
- For bioinformatics questions, training, or analysis help:
  - bioinformatics@niaid.nih.gov
- Inside NIAID – Bioinformatics Resources
  - http://inside.niaid.nih.gov/topic/IT/bioinformatics

Connect to NIH-Wireless

https://github.com/niaid/ACE/tree/master/WES_Data_Analysis

Overview

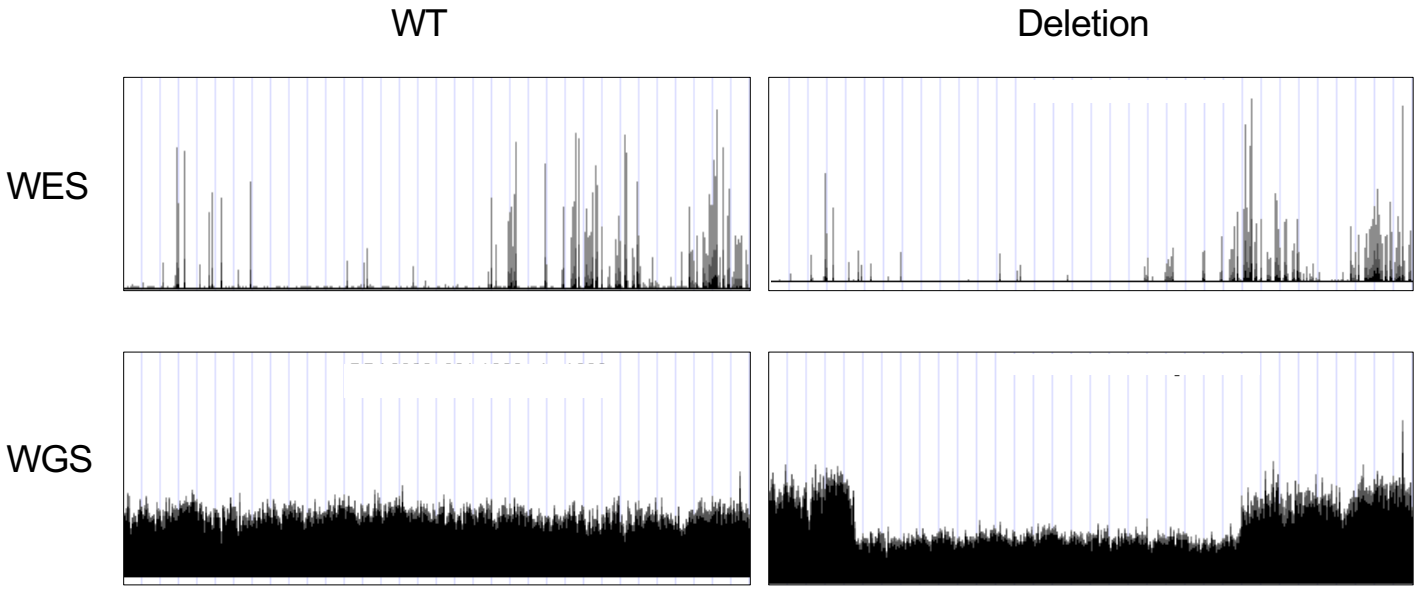
- Part 1: *Seminar* - Introduction to Whole Exome Sequencing

- Part 2: *Hands on with Command Line*
  - Data Preprocessing and QC
  - Variant Discovery
  - Annotation and Downstream Analysis

- Part 3: *Seminar* – Introduction to GWAS and CNV analysis

Difference between WES and WGS

- Whole genome sequencing (WGS) has even coverage
- Even a deletion is observed by eye



All about exomes

- What is an exome?
  - Sequencing targeted to exonic regions
  - ~2% of human genome
- Important to know
  - You will NOT get a whole exome!
  - Not all exons in all genes are captured!
  - Must recognize *negative results* vs. *no data*
- Coverage will vary with targets

What to know about your data

- What is sequence coverage/depth?
  - The depth of your sequence
  - 10x, 30x, 50x, 100x
- Read length
  - How long are your reads?
- What software was used to align?
- What variant caller was used to call variants?
- Which reference sequence/genome was used?
- Which capture kit was used? What does it cover?

Sequencers

Vendors

Illumina NovaSeq 6000

PacBio

Ion Torrent GeneStudio S5

Applications determine platform choice

Applications

- Large Whole-Genome Sequencing (human, plant, animal)
- Small Whole-Genome Sequencing
- Exome Sequencing
- Targeted Gene Sequencing
- Whole-Transcriptome Sequencing
- mRNA-Seq for gene expression
- Targeted Gene Expression Profiling

- Long-Range Amplicon Sequencing
- miRNA & Small RNA Analysis
- DNA-Protein Interaction Analysis
- Methylation
- 16S Metagenome

Sequencers

Machine	Cost of Machine	Methods/Sample Throughput	Reads per run	Read length (bp)	Error rate
Ion Torrent GeneStudio S5	$80K	Small genomes, targeted gene	2-130 Million	200-600	1.71%
Illumina NextSeq 550	$250K	3 Exomes 13 Transcriptomes 30 Small Genomes	130-400 Million	2x150	0.8%
Illumina HiSeq 4000	$654K	12 Genomes 96 Exomes 100 Transcriptomes	2.5-5 Billion	2X150	0.76%
Illumina NovaSeq 6000	$900 K	48 Genomes 500 Exomes 400 Transcriptomes	16-20 Billion	2X250	0.8 %
PacBio Sequel II	$695K	Small Genomes, targeted gene, methylation	100 MB	15 kb	13-15%

Error rates vary between different platforms as well as the types of error (indel versus mismatches etc) and in biases depending on the sequence content (for example, GC skewing) requiring different error handling algorithms

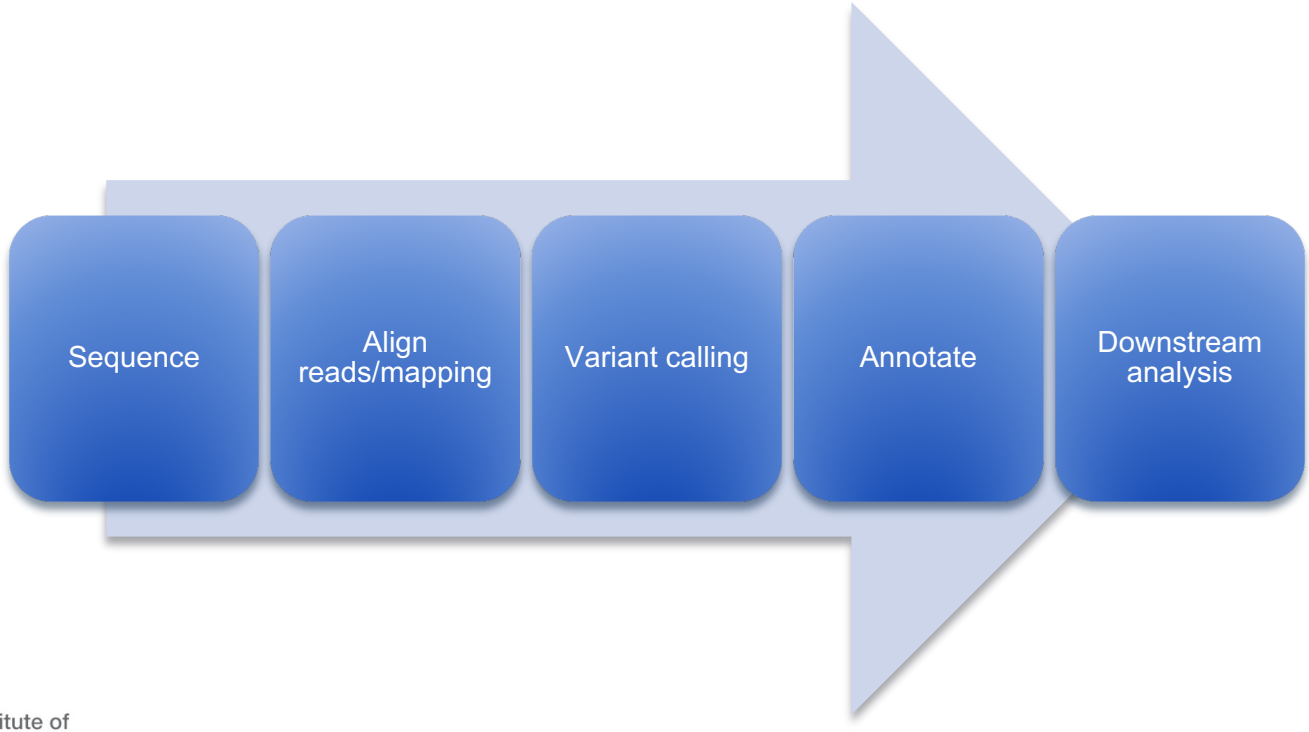
Comparison between WES and WGS

	Exomes (ES/WES)	Genomes (GS/WGS)
DNA	Targeted and captured	Sheared DNA
What you can get	Most coding regions (+UTR)	coding and non-coding
Variants that can be examined	SNVs, indels (CNVs)	SNVs, indels, CNVs, structural variations
Size of bam files	~10 Gb	~200 Gb
Cost (research vs clinical, platform, coverage etc.)	$300 - $1000	$500 - $6500

Partial list of capture enrichment kits

Manufacturer	Kits	Regions targeted	Bases covered
IDT	xGen Exome Research Panel	Exons	39Mb
Illumina	Nextera Rapid capture	Exons + UTRs + miRNA	62 Mb
Nimblegen	SeqCap EZ Exome	Exons + UTR	96 Mb
	SeqCap EZ MedExome	Disease-associated regions	47 Mb
Agilent	SureSelect Human All Exon V6	Exons + UTRs	60Mb
	Clinical Research Exome	Disease-relevant targets	51Mb

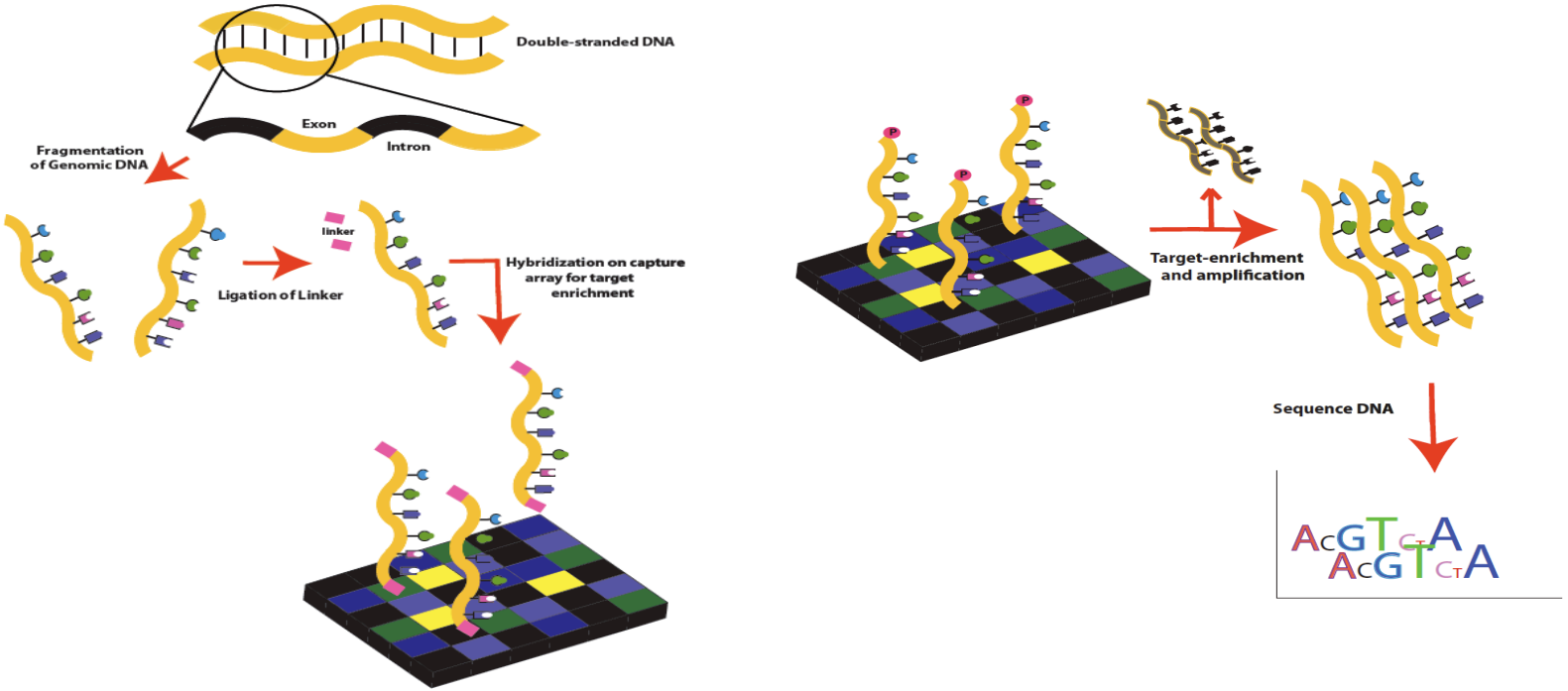
Overview of next-generation sequence processing



National Institute of Allergy and Infectious Diseases

NIAID

Exome sequencing overview



National Institute of Allergy and Infectious Diseases

Source: https://en.wikipedia.org/wiki/Exome_sequencing#/media/File:Exome_Sequencing_workflow_1b.png. 2012 Jan; 71(1):5-14

NIAID

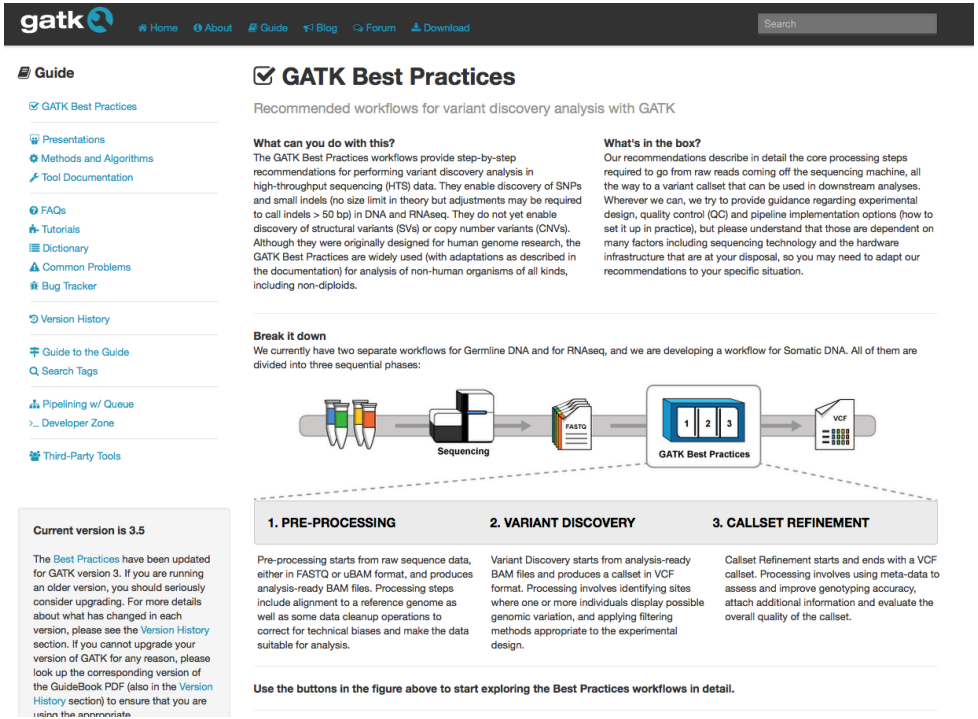
Popular tools

Task	Popular tools
Align reads/mapping	**BWA-mem** Bowtie NovoAlign BLASR (for PacBio reads)
Variant calling	**GATK4 (Broad)** DeepVariant (Google) SpeedSeq (Wash. Univ. St. Louis) Platypus (Wellcome Trust) Starling (Illumina)
Annotating variants	**VEP (EMBL-EBI)** AnnoVar snpEff Funcotator (Broad)

Popular resources

Task	Popular sources
Control population frequency	gnomAD TGAC ExAC 1000GP ESP
Annotation	RefSeq Ensembl UCSC genes GENCODE
Visualization	UCSC genome browser **IGV**
Clinical relevance	HGMD OMIM CGD ClinVar

GATK Best Practices Website



National Institute of Allergy and Infectious Diseases

https://www.broadinstitute.org/gatk/guide/best-practices

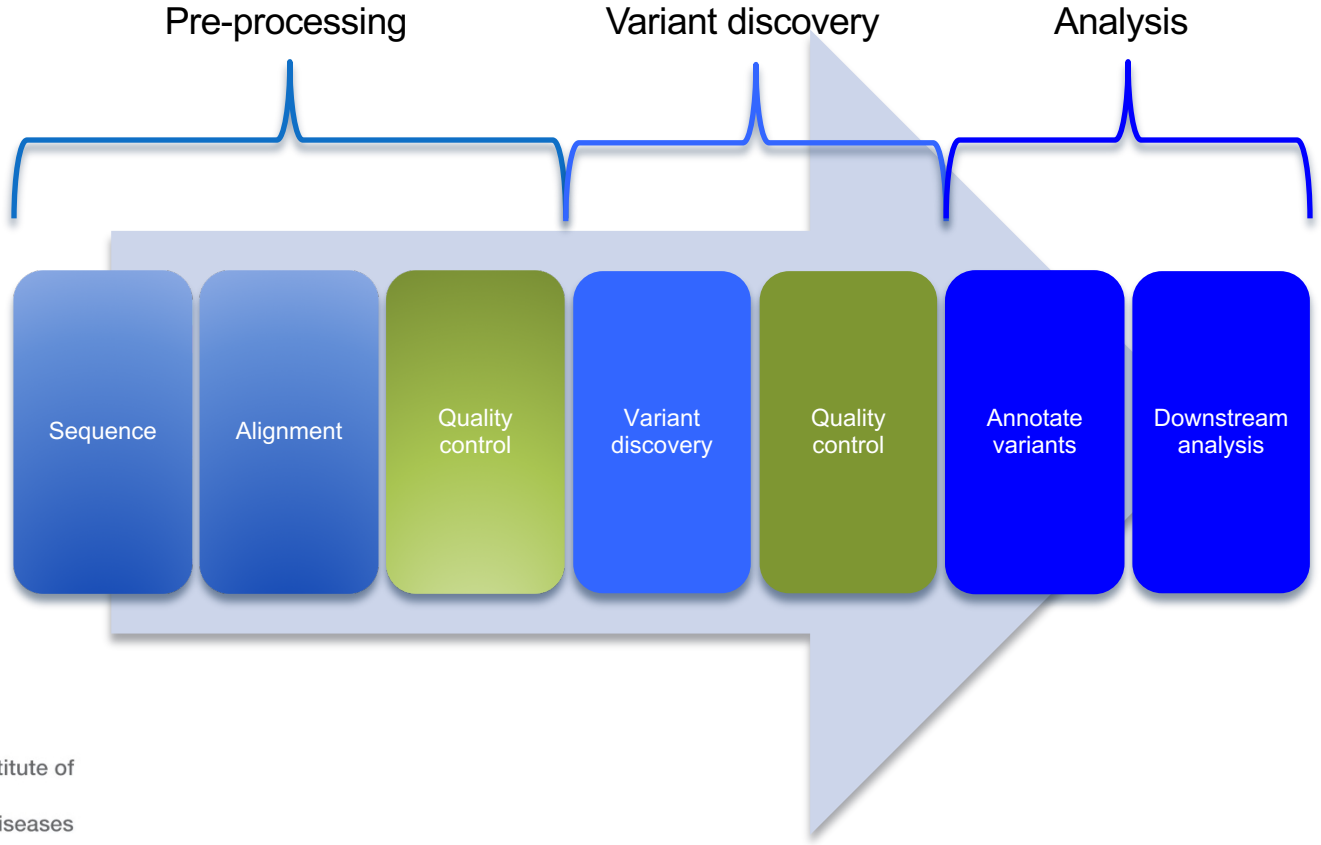
NIAID

Introduction to WES Analysis Workflow

National Institute of Allergy and Infectious Diseases

NIAID

WES pipeline overview



Files formats

File type	Origin
FASTA	Text-based format for reference
FASTQ	Raw reads from sequencer
SAM BAM CRAM	Sequence Alignment/Map Binary version of SAM Compressed BAM
gVCF/VCF	Variant call format

Index files

- Index files are needed for files in next-generation analysis, as file sizes are big!
- Enables program to efficiently access the data, rather than having to read the whole file

File type	Index
FASTA	*fai
BAM	*bai
CRAM	*crai
VCF	*vcf.idx

Tools

Tool	Purpose
BWA mem	Read alignment to reference
Picard	Mark duplicates
FastQC	Short reads QC
samtools	Sort sam/bam, convert sam<->bam
bcftools	VCF query, normalization
GATK4 (Haplotype caller)	QC/variant calling
verifyBamId	Contamination detection
PEDDY	Family relationship & gender check
VEP	Variant annotation
Gemini	Mendelian inheritance analysis

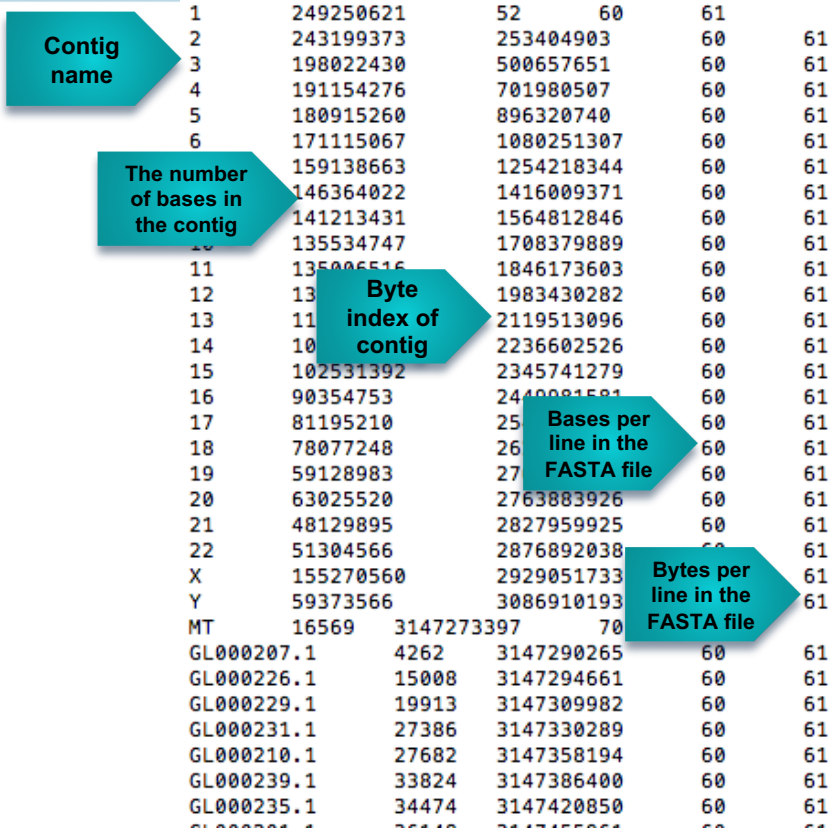
Reference genome

- There are different versions of human reference genome

Reference Name	Chromosome Notation	Mitochondrial Sequence	Additional Sequences Included
GRCh37 (Genome Reference Consortium)	1, 2, … , X, Y, MT	Yes	• Unlocalized • Unplaced • Alternate loci
hg19 (UCSC Genome Browser)	chr1, chr2, … , chrX, chrY, chrM	Copied from previous release	• (same as above)
b37/b37+decoy/hs37d5 (1000 Genomes Project)	1, 2, … , X, Y, MT	Yes	• Unlocalized • Unplaced • “decoy” sequence • Human herpesvirus 4 type 1
GRCh38 (Genome Reference Consortium)	chr1, chr2, … , chrX, chrY, chrM	Yes	• Unlocalized • Unplaced • Alternate loci • Model centromeres • GRCh37 Assembly corrections
hg38 (UCSC Genome Browser)	chr1, chr2, … , chrX, chrY, chrM	Yes	• (same as above)

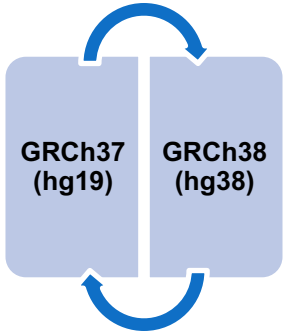
* Unlocalized: chromosome known, exact location unknown
* Unplaced: known to originate from human genome, chromosome unknown
* Alternate loci: alternate representation of specific human regions

Reference FASTA index file



Converting Coordinates from One Genome Version to Another

- UCSC provides tools (*liftOver*) to convert BED file from one genome assembly to another



Lift over can fail for different reasons:

1) **Genome position cannot be lifted --** a SNV resides in a contig that only exists in older/newer reference build, liftOver cannot give it new genome.
2) **Refence allele and alternative allele flipped in different reference build -- G**(ref)/**C**(alt) in GRCh37 and **C**(ref)/**G**(alt) in GRCh38

Why BWA+ GATK Haplotype caller?

- Widely accepted as the “conventional” way of processing NGS data
- Well assessed (many citations)
- Well documented
- Software is supported and still being updated
- Community support for troubleshooting or information

QC Checks for Data Quality

- Sequencer generated short reads – FastQC
- Coverage at different levels of partitioning and aggregation – GATK3 DepthOfCoverage
- Sample contamination – verifyBamId
- Familial-relationships and gender check – peddy

QC on FASTQ/BAMs by fastQC

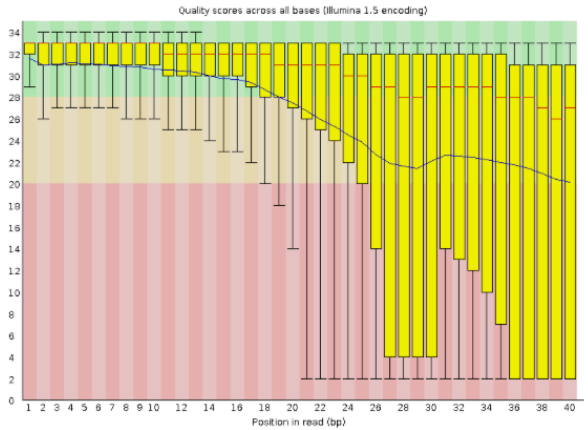
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

Per base sequence quality



Bad quality

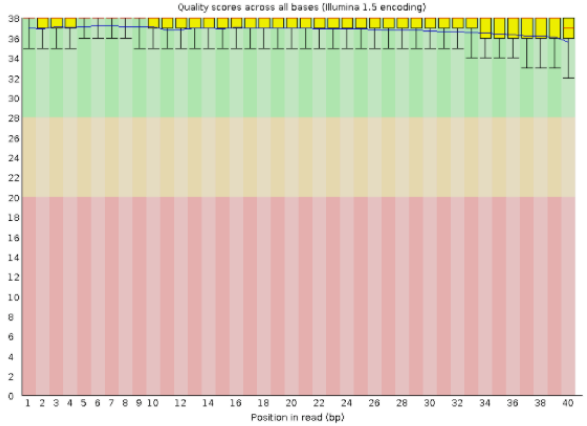
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Per base sequence quality

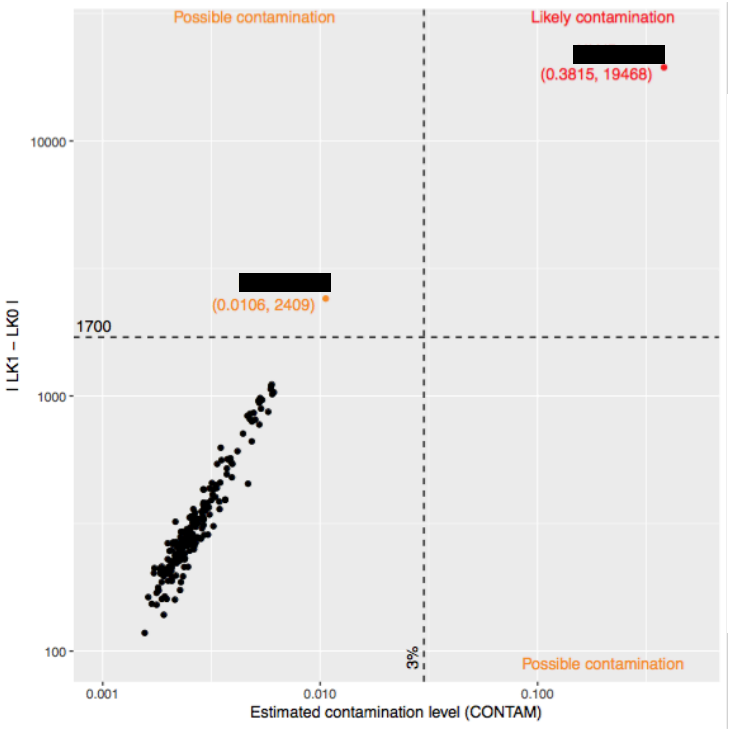


Good quality

National Institute of Allergy and Infectious Diseases

NIAID

QC on sample contamination by verifyBamId



Likely contaminated

ID	CONTAM	| LK1−LK0 |	NumSNPs	NumReads	AvgDP
[redacted]	0.38148	19468.34	15627	441693	28.26

CONTAM > 0.03 AND | LK1−LK0 | > 1700.
CONTAM is the estimated fraction of contamination based on verifyBamID.

Possibly contaminated

ID	CONTAM	| LK1−LK0 |	NumSNPs	NumReads	AvgDP
[redacted]	0.01062	2409.17	15627	634552	40.61

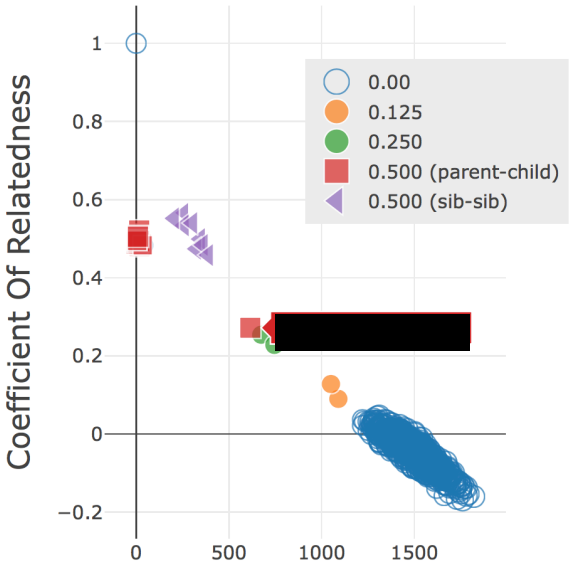
CONTAM > 0.03 OR | LK1−LK0 | > 1700 (not both)
CONTAM is the estimated fraction of contamination based on verifyBamID.

NIH National Institute of Allergy and Infectious Diseases

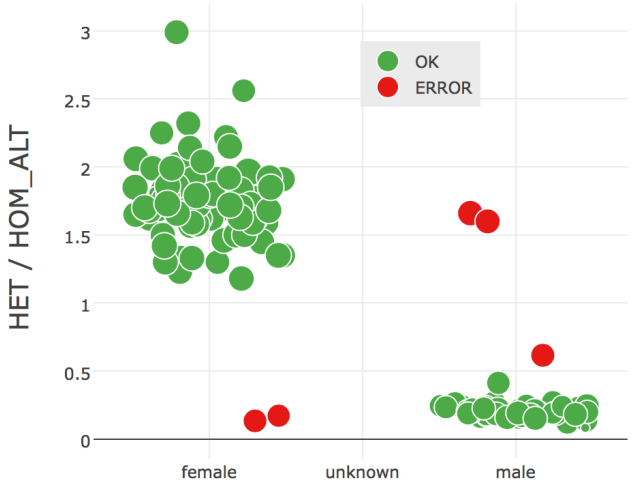
NIAID

QC on family-relationship and gender by peddy

Example of pedigree conflict



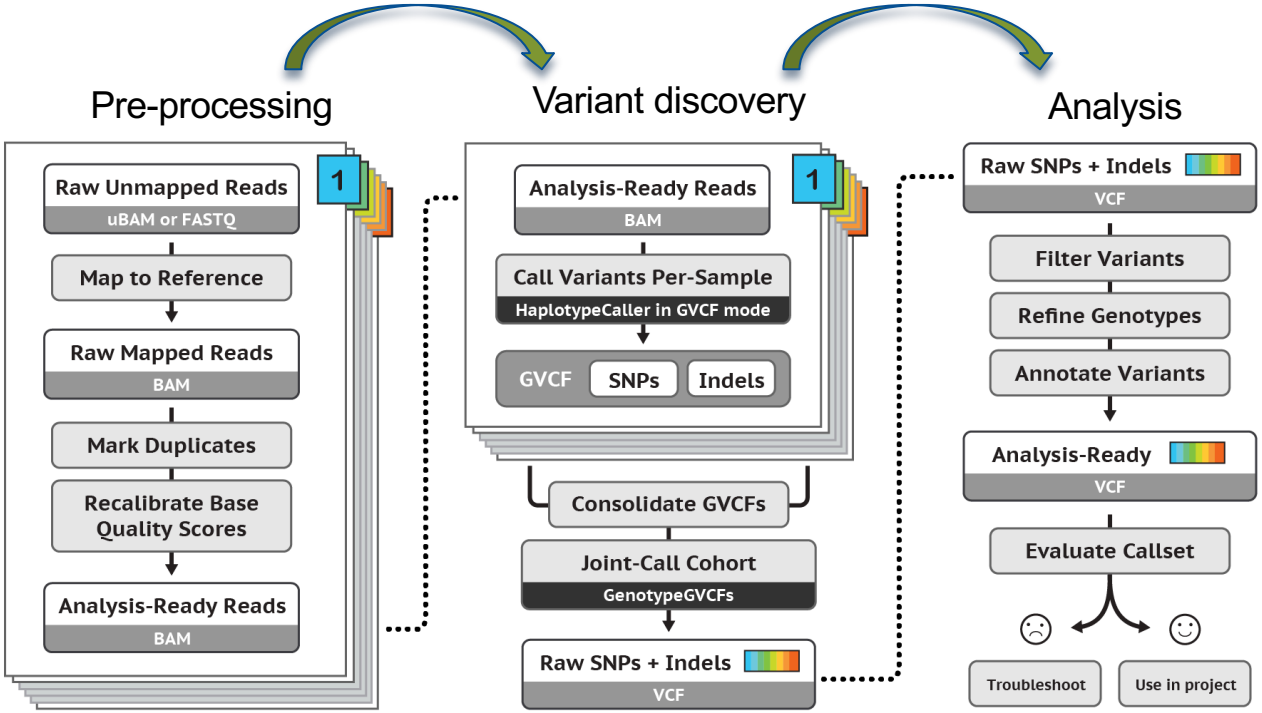
Example of gender conflict



National Institute of Allergy and Infectious Diseases

NIAID

GATK Best Practices Workflow



Sequence ID Sequence Sequence quality

FASTQ

```
@C7JHKANXX:7:2109:01323:47728
TCTTTAATTTAGTATGCAAGAAAACCAATTTGGAAATATCAACTGTTTTGGAAACCTTAGACCTAGGTCATCCTTAGTAAGATCTTCCCATTTATATAAATACTTGCAAGTAGTAGTGCCATAATT
+
CCBCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGCGGGGGGGFGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGDGGGGGGEGGGGGGGGGGGGGGG>FDG>FGGGGGGGGGGGGG
@C7HEMANXX:2:1104:01510:87765
TTAGAGGATTTATTTCCCATTTTTCTTAAAAAGGAAGAACAAACTGTGCCCTAGGGTTTACTGTGTCAGAACAGAGTGTGCCGATTGTGGTCAGGACTCCATAGCATTTCACCATTGAGTTATTTC
+
BBBB=FGGGGGGGGGGGCGGGGGGGGGCGG>GGGGGGGGGGEGGGGGGGFGGGGGGGGGGEGGGGGGGGGGGG>FEGGGGGGGGGGGGGG0@FGGGGGGGGGGGGGGCEGGGGGGGGGGGGGCGD
@C7JHKANXX:7:2109:01323:47728
ATTAATTAAAAGAAGGCTCATCCAGGAAACACATATAAAAGAGCTGTTTACTCAGTGTTTTATGCCCTCTTGGAGATAATAGACCGCTGAAGAGAGACACGTAAGGGGGCGGAAATAACTCAATGG
+
=BBBBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGCFGFCEGGGGGGGGGGGGDGGGE>GGGGGFG@CDGGGGGGGGGGGGGGGGGGGGGGD?GDGGGGGGGBGGGDD
@C7HEMANXX:2:1104:01510:87765
GTCTGGGGGTGACTCCCCATGACGTGCAGTGGCCCTAGAGGAAATCCCTTAACAAAATTAATTAAAAGAAGGCTCATCCAGGAAACACATATAAAAGAGCTGTTTACTCAGTGTTTTATGCCCTCT
+
BBBCCGGGGGGGGDGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG>FGGGGGGFGEGGFGGGGGGGGGGCGGGG>FGGGGGGGGGEGGGGGGGEGGGGGGGGGGGGGFFDFGGGGGGGGGGD
@C7HP0ANXX:7:1314:08094:49231
GATTTCCTCTAGGGCCACTGCACGTCATGGGGAGTCACCCCCAGACACTCCCAATTGGCCCCTTGTCACCCAGGGGCACATTTCAGCTATTTGTAAAACCTGAAATCACTAGAAAGGAATGTCTAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@C7HP0ANXX:7:1314:08094:49231
ATATATTTTTTGACATATTTTGAAATGGCTGTCACAGGGCCAACACATTAAAATGGCCCTGCAAGGATTTCTTTTGTGGAGGAAATTTTGCACATGGAGAGAATCTCCATTAATGCAGTCAGGCCT
+
```

Align

BAM/SAM

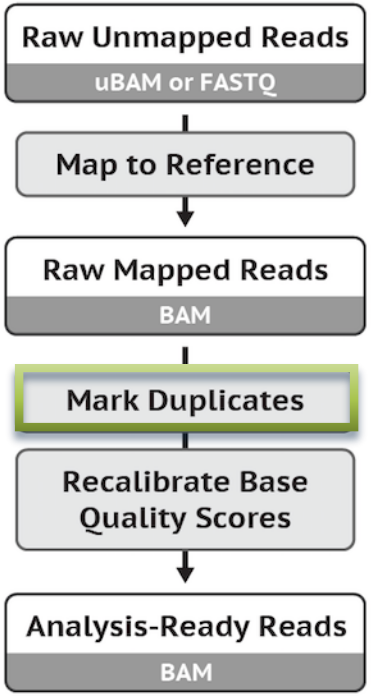
Sequence ID	Bitwise flags	chrosome	position	MappingQ	CIGAR	Paired-end chr	Paired-end position	Observed template length	
C7JHKANXX:7:2109:01323:47728	163	chr22	16050365	70	126M	=	16050663	424	TCTTTAATTTAGTATGCAAGAA
C7HEMANXX:2:1104:01510:87765	99	chr22	16050552	30	126M	=	16050719	293	TTAGAGGATTTATTTCCCATTT
C7JHKANXX:7:2109:01323:47728	83	chr22	16050663	70	126M	=	16050365	-424	CCATTGAGTTATTTCCGCCCCC
C7HEMANXX:2:1104:01510:87765	147	chr22	16050719	30	126M	=	16050552	-293	AGAGGGCATAAAACACTGAGTA

National Institute of Allergy and Infectious Diseases

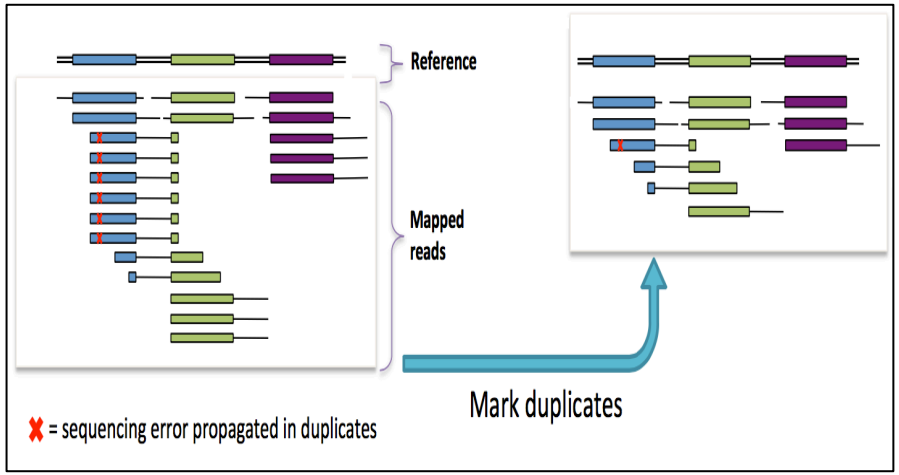
NIAID

http://samtools.github.io/hts-specs/SAMv1.pdf

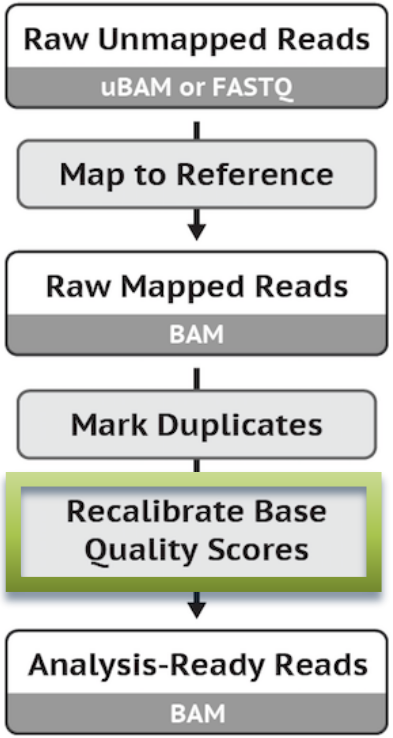
Marking duplicates and why is it necessary?



- It is assumed that each read corresponds to an independent DNA fragment from *randomly* sheared DNA
- However, PCR amplification and optic sensor artifact can cause duplicates
  - Identify based on start + stop of reads
  - Choose the best and ignore the rest

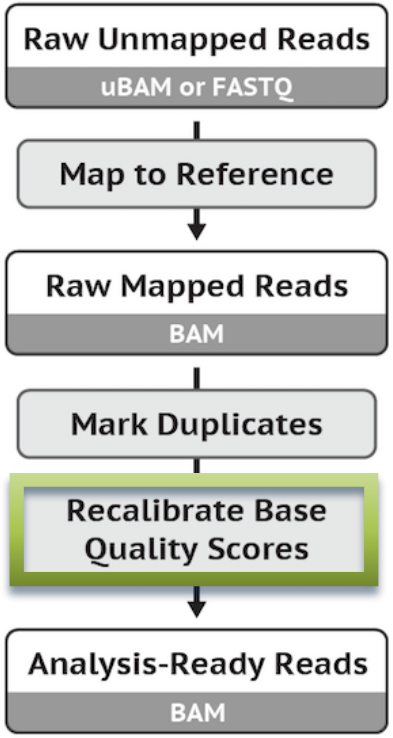


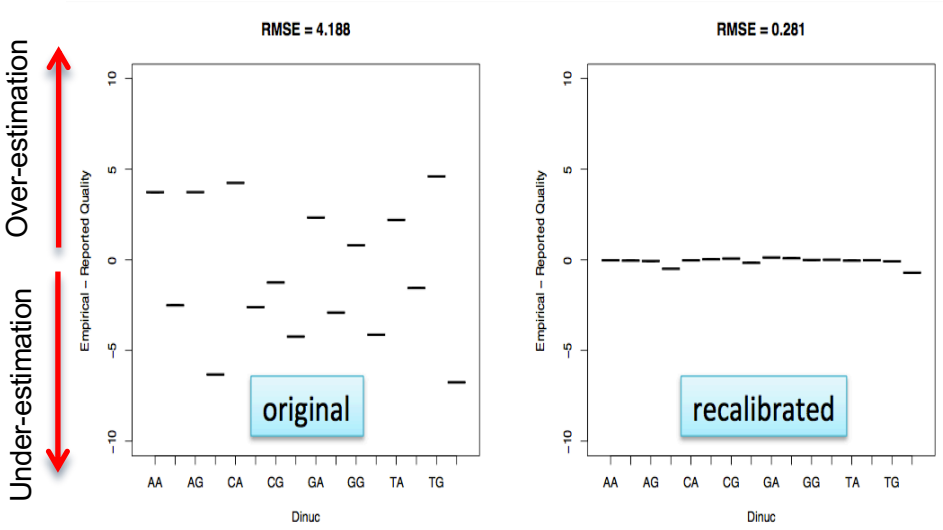
Recalibrate Base Quality Scores



- Base scores are produced by sequencers
- Quality scores are **inaccurate** and **biased**
  - Prone to various technical errors
  - QS are often over- or under estimated
- To identify and correct non-random technical error
  - Physics or the chemistry of sequencing reactions
  - Manufacturing flaws in the equipment
- Error covariates e.g.
  - Reported quality score
  - Position within the read (machine cycle)
  - Preceding and current nucleotide (sequencing chemistry)

Recalibrate Base Quality Scores





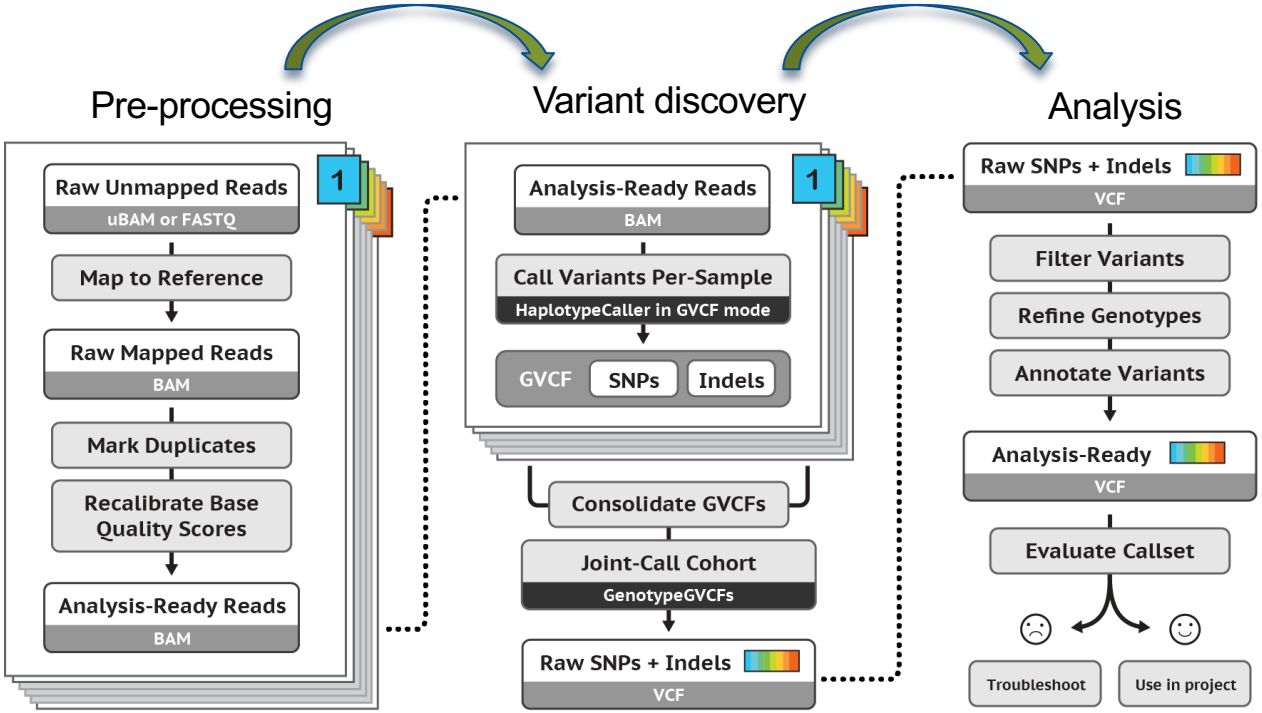
- GATK BSQR builds model based on the known variants set
- Adjusts the base quality scores in the data based on the model

National Institute of Allergy and Infectious Diseases

Broad Institute

NIAID

GATK Best Practices Workflow

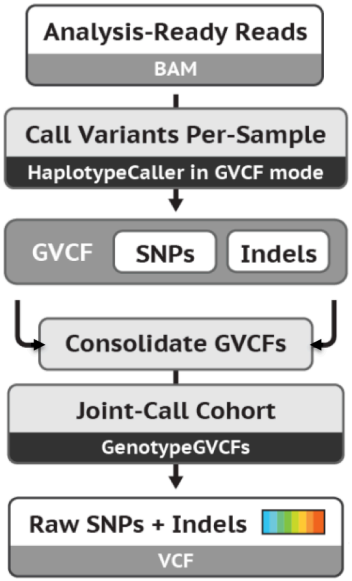


National Institute of Allergy and Infectious Diseases

Broad Institute

NIAID

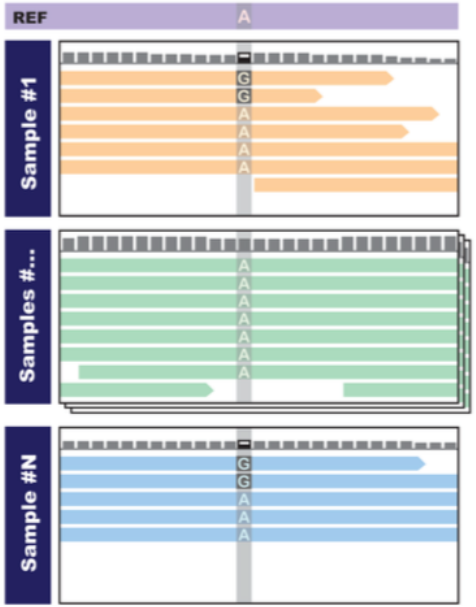
Generate genotype likelihoods in each sample (gVCF)



- For a single sample, calculates normalized Phred-scaled likelihoods (PL) for genotypes:
  - "likelihood of the genotype"= "the probability that the genotype is ***not*** correct"
  - Normalized so that the most likely genotype's PL is 0

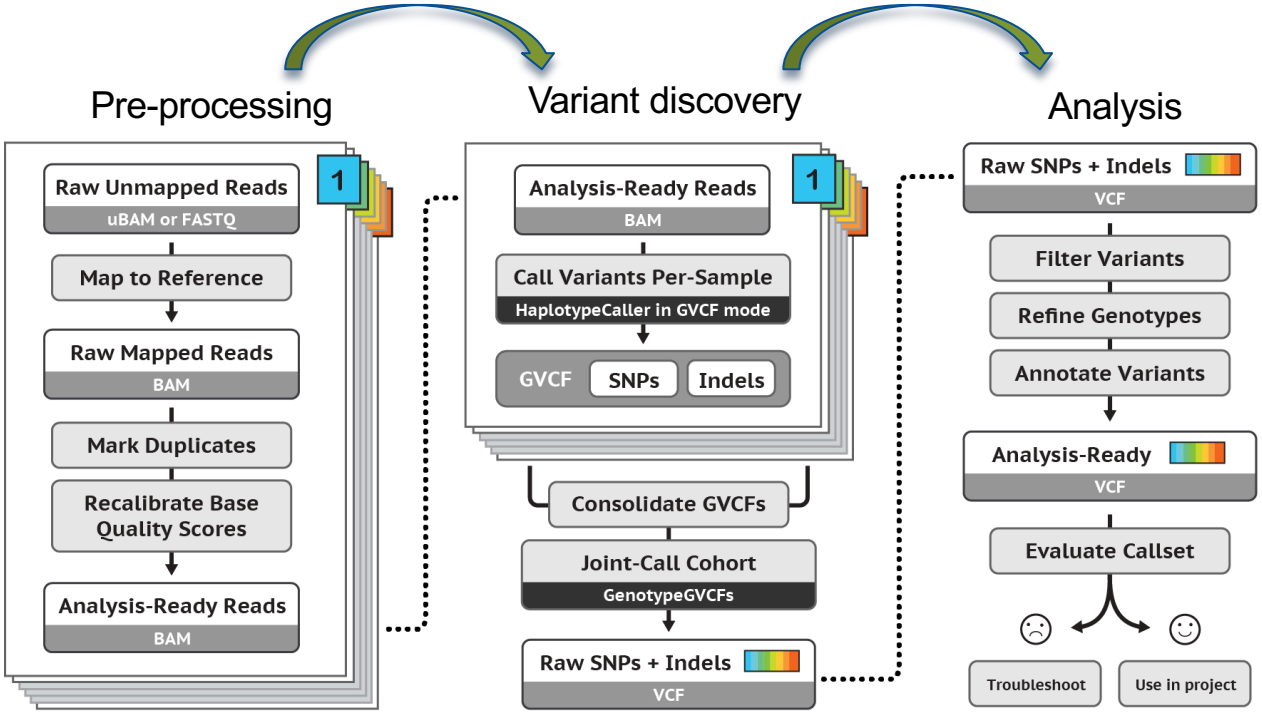
16	105459	.	G	<NON_REF>	.	.	.	GT:AD:DP:GQ:PL	0/0:60,0:60:9:0,120,1800
16	105460	.	T	<NON_REF>	.	.	.	GT:AD:DP:GQ:PL	0/0:60,0:60:9:0,120,1800
16	105461	.	T	<NON_REF>	.	.	.	GT:AD:DP:GQ:PL	0/0:61,0:61:9:0,120,1800
16	105462	.	A	<NON_REF>	.	.	.	GT:AD:DP:GQ:PL	0/0:61,1:62:9:0,120,1800
16	105463	.	C	<NON_REF>	.	.	.	GT:AD:DP:GQ:PL	0/0:64,0:64:9:0,120,1800
16	105464	.	T	<NON_REF>	.	.	.	GT:AD:DP:GQ:PL	0/0:64,0:64:9:0,120,1800
16	105465	.	C	<NON_REF>	.	.	.	GT:AD:DP:GQ:PL	0/0:66,0:66:9:0,120,1800

What is joint genotyping?

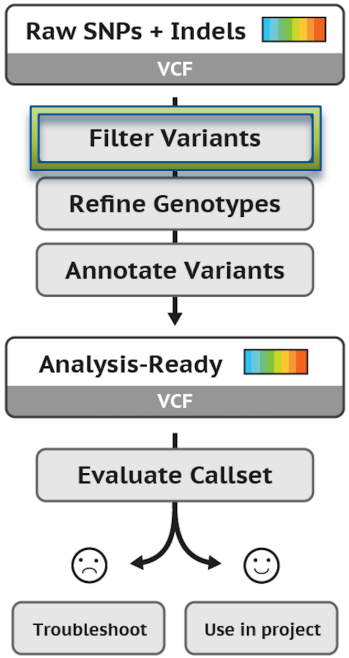


- If we analyze Sample 1 or Sample N alone, we are not confident that the variant is real
- If we see the variant in both samples, we are more confident that there is real variation at this site in this cohort

GATK Best Practices Workflow

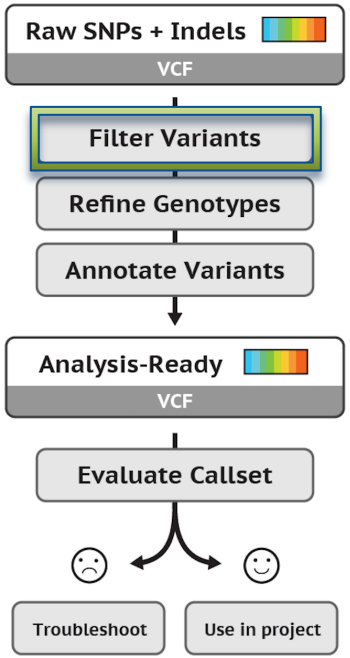


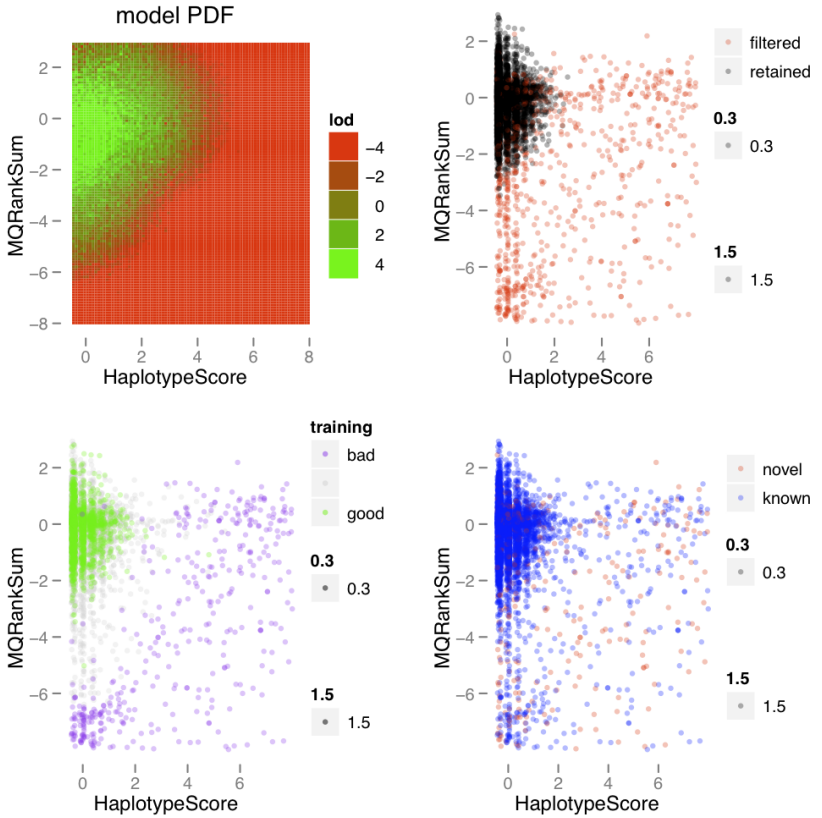
Generate new variant quality score using VQSR



- What is Variant Quality Score Recalibration
  - NOT adjusting scores!!
  - Generate new score VQSLOD (variant quality score log-odds)
- Approach
  - Machine learning to profile good variants vs bad variants
  - Using multiple dimensions (5-8, typically)
  - Uses INFO annotations for each variant (eg. Allele count, allele frequency, etc)

Generate new variant quality score using VQSR



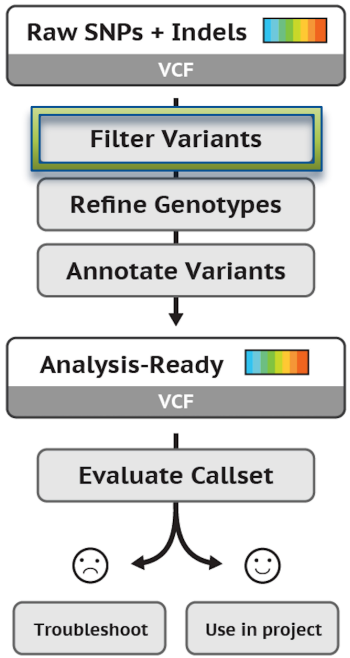


http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr

National Institute of Allergy and Infectious Diseases

NIAID

Alternatively! If you have few samples (< 30 exomes)



- Apply hard filter!
- Define how to filter your variant or use default filter parameters
  - QualByDepth (QD) 2.0
  - FisherStrand (FS) 60.0
  - RMSMappingQuality (MQ) 40.0
  - MappingQualityRankSumTest (MQRankSum) -12.5 (only het calls)
  - ReadPosRankSumTest (ReadPosRankSum) -8.0 (only het calls)

Final VCF

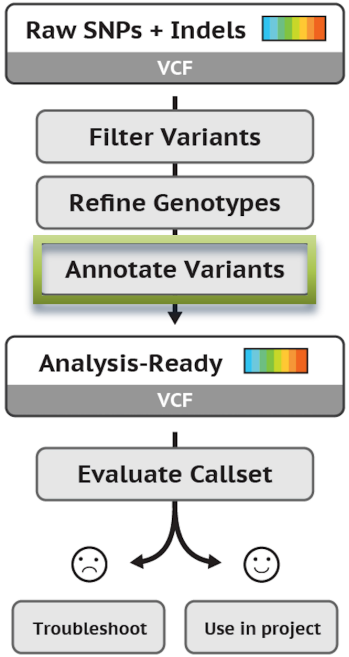
Header

```
##fileformat=VCFv4.1
##FILTER=<ID=LowQual,Description="Low quality">
##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="Approximate read depth (reads with MQ=255 or with bad mates are fi
ltered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in t
he VCF specification">
##GATKCommandLine.HaplotypeCaller=<ID=HaplotypeCaller,Version=3.4-3-gd1ac142,Date="Mon May 18 17:36:4
.
.
.
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as
listed">
##INFO=<ID=AF,Number=A,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">
##contig=<ID=chr1,length=249250621,assembly=b37>
##reference=file:human_genome_b37.fasta
```

Body

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

Annotating Variants with VEP



VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

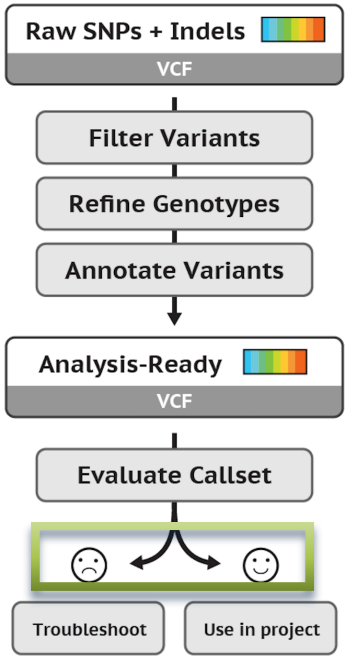
Paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4893825/

The specific annotations retrieved are:

- Genes and transcripts affected by the variants
- Location of the variants:
  - upstream of a transcript
  - in coding sequences
  - in non-coding RNAs
  - in regulatory regions
- Impact of variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- Known variants that match the variant, and associated minor allele frequencies from the 1000 Genomes Project
- SIFT and PolyPhen scores for changes to protein sequence

Family Based Study

Prioritize candidate variants with GEMINI



GEMINI (GEnome MINIng), a flexible software package for exploring all forms of human genetic variation.

It integrates genetic variation using an adaptable set of genome annotation sources (e.g., dbSNP, ENCODE, UCSC, ClinVar, KEGG)

Researchers can compose complex queries based on sample genotypes, inheritance patterns, and both pre-installed and custom genome annotations.

Both command line and graphical tools for common analyses are supported

Paper:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3715403/

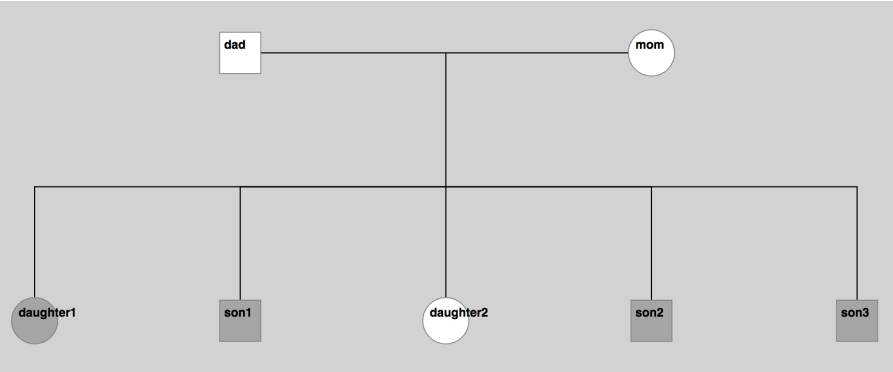
Similar tool to consider: pVAAST http://www.hufflab.org/software/pvaast/

Next: Part 2 - Hands-on exercises!

Please visit this URL for walkthrough:
https://github.com/niaid/ACE/blob/master/WES_Data_Analysis/wes_data_analysis.md

Hands on settings

- We have FASTQ files from a family with 7 members



- We'll only be looking at subset regions on chr6, 11, X, and Y
- ***Are there any deleterious variants in this family?***

Steps to get started on LOCUS…

1. **Log into the Locus Cluster:** *ssh username@ai-submit1.niaid.nih.gov*
   -Note: if you want to use and X-windows application see the walkthrough about adding the –X flag to your ssh command and using Xquartz or Xceed!

2. **Start an Interactive Session:** `qrsh -pe threaded 8`

3. **Make a Tutorial Folder:** `mkdir WES_Tutorial`

4. **Move into Folder:** `cd WES_Tutorial`

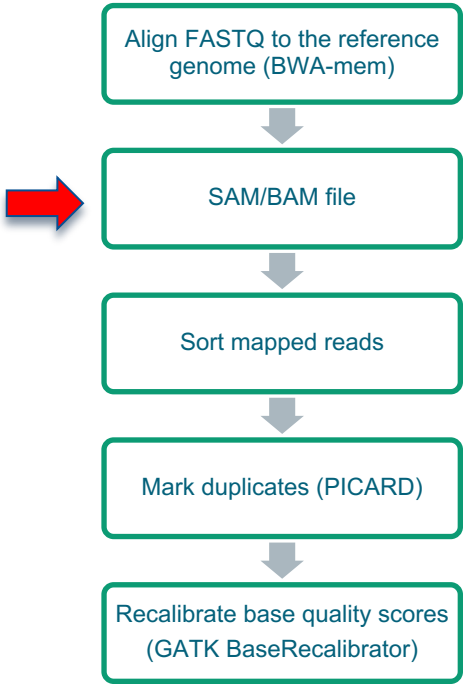
5. **Copy Example Files:**
   `cp -r /hpcdata/bcbb/wes_training/data/* .`

6. **Verify Files Copied:** `ls -l`
   -You should see files like: dad, daughter1, daughter2, etc.

Modules we will use…
(load only as needed in walkthrough!)

- gatk/3.8.1-Java-1.8.0_92
- gatk/4.0-8-Java-1.8.0_92
- bwa
- samtools
- picard
- FastQC
- IGV
- verifyBamID
- peddy
- VEP/89-goolf-1.7.20
- GEMINI

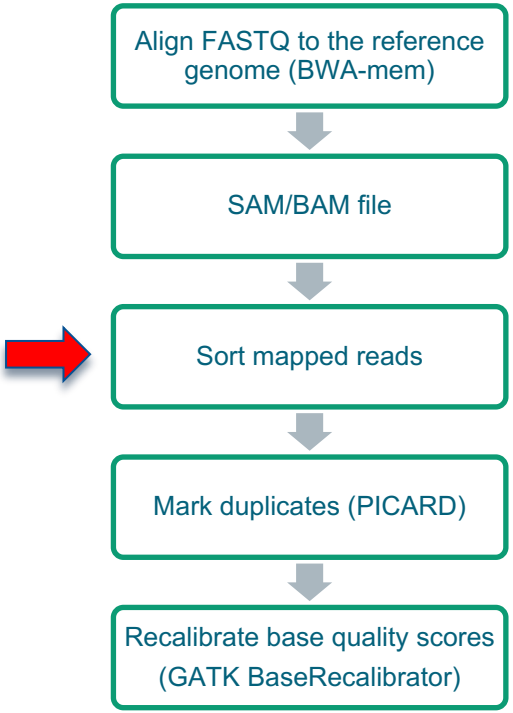
Step 1: align fastq files



- Align with BWA MEM using **-M** to mark secondary alignments and **-R** to annotate Read Groups (e.g., different samples)
  - **ID, LB, SM, and PL tags are required**

```
module load bwa
sample='son1'

bwa mem -t 8 -M -R
"@RG\tID:${sample}\tLB:${sample}\tSM:${sample}\tPL:ILLUMINA" \
/hpcdata/bcbb/wes_training/reference/human_g1k_v37_decoy.fasta \
./${sample}/${sample}.R1.fq ./${sample}/${sample}.R2.fq \
-o ./${sample}/${sample}.sam

ls son1
```

Step 2: sort sam, convert to bam

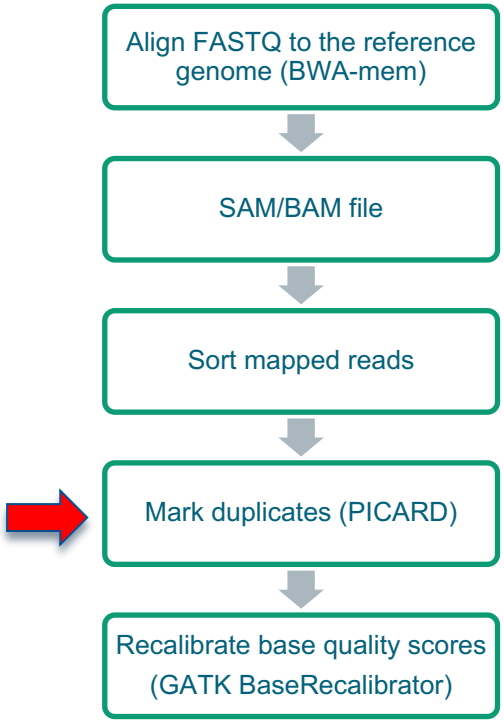


- Aligned reads need to be sorted

```
module load picard

java -jar ${EBROOTPICARD}/picard.jar SortSam \
I=./${sample}/${sample}.sam \
O=./${sample}/${sample}.sorted.bam\
MAX_RECORDS_IN_RAM=1000000 CREATE_INDEX=true \
VALIDATION_STRINGENCY=LENIENT \
SORT_ORDER=coordinate

ls son1
```

*Now we have **mapped** and **sorted** reads*

Step 3: Mark duplicates



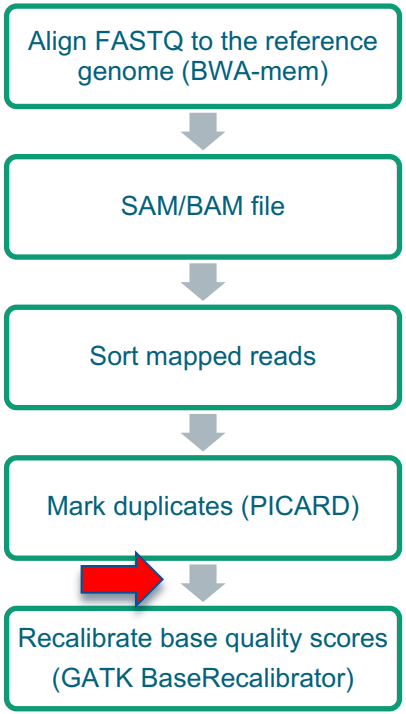
- Using picard to mark duplicates

```
java -jar ${EBROOTPICARD}/picard.jar MarkDuplicates \
I=./${sample}/${sample}.sorted.bam \
O=./${sample}/${sample}.markdup.bam \
VALIDATION_STRINGENCY=LENIENT AS=true CREATE_INDEX=true \
M=./${sample}/${sample}.bwa.markdup.log

ls son1
```

How many were marked as duplicates??

91,193 duplicates & 24 optical duplicate clusters

Step 4: QC - FastQC



1) run FastQC on fastq files
   - generates .html file with qc statistics
   - Generates compressed folders with images

```
module load FastQC

fastqc -f bam -o ./${sample} ./${sample}/${sample}.markdup.bam

ls son1
```

2) Check FastQC html output

Need this dot!

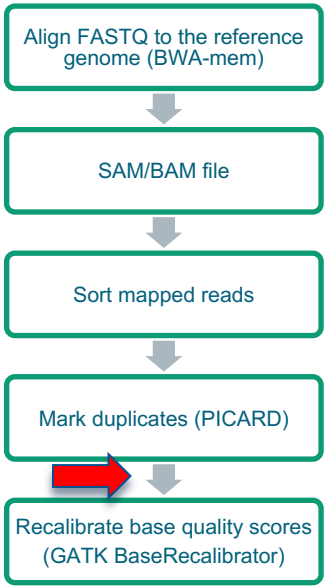
- Copy the html file to your local machine

scp username@ai-submit1.niaid.nih.gov:~/WES_Tutorial/son1/son1.markdup_fastqc.html .

- Open the html file in your local browser

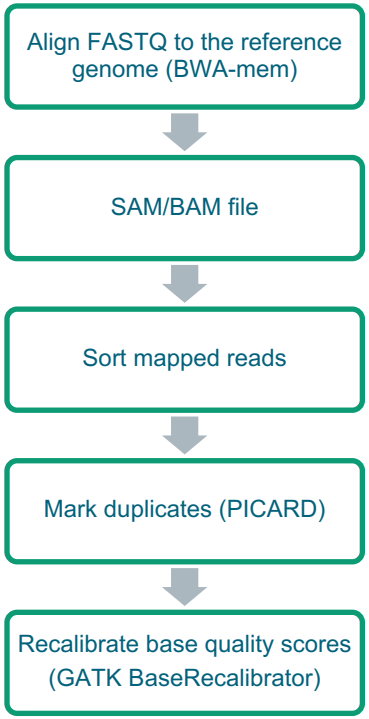
Step 5: QC - DepthOfCoverage

NOTE: GenomeAnalysisTK is currently only in GATK3 – Broad suggests it will be released for GATK4 later this year!




```
module load gatk/3.8.1-Java-1.8.0_92

find . -name "*markdup.bam" > fam001_bams.list

cat fam001_bams.list

java -jar $EBROOTGATK/GenomeAnalysisTK.jar -T DepthOfCoverage \
 -R /hpcdata/bcbb/wes_training/reference/human_g1k_v37_decoy.fasta \
 -o ./fam001_dov \
 -I fam001_bams.list \
 -geneList:REFSEQ /hpcdata/bcbb/wes_training/reference/refSeq.sorted.txt \
 -ct 4 -ct 10 -ct 15 -ct 20 \
 -L /hpcdata/bcbb/wes_training/reference/exome.interval_list

module unload gatk/3.8.1-Java-1.8.0_92
```

Step 6: Recalibrate base QS



Will need references data sets in GATK resource bundles

- Build a model

```
module load gatk/4.0-8-Java-1.8.0_92

#Part1
${EBROOTGATK}/gatk BaseRecalibrator \
-R /hpcdata/bcbb/wes_training/reference/human_g1k_v37_decoy.fasta \
-I ./${sample}/${sample}.markdup.bam \
-O ./${sample}/${sample}.recal_data.table \
--known-sites /hpcdata/bcbb/wes_training/reference/dbsnp_138.b37.vcf.gz \
--known-sites/hpcdata/bcbb/wes_training/reference/Mills_and_1000G_gold_standard.indels.b37.vcf.gz
```

- Recalibrate scores

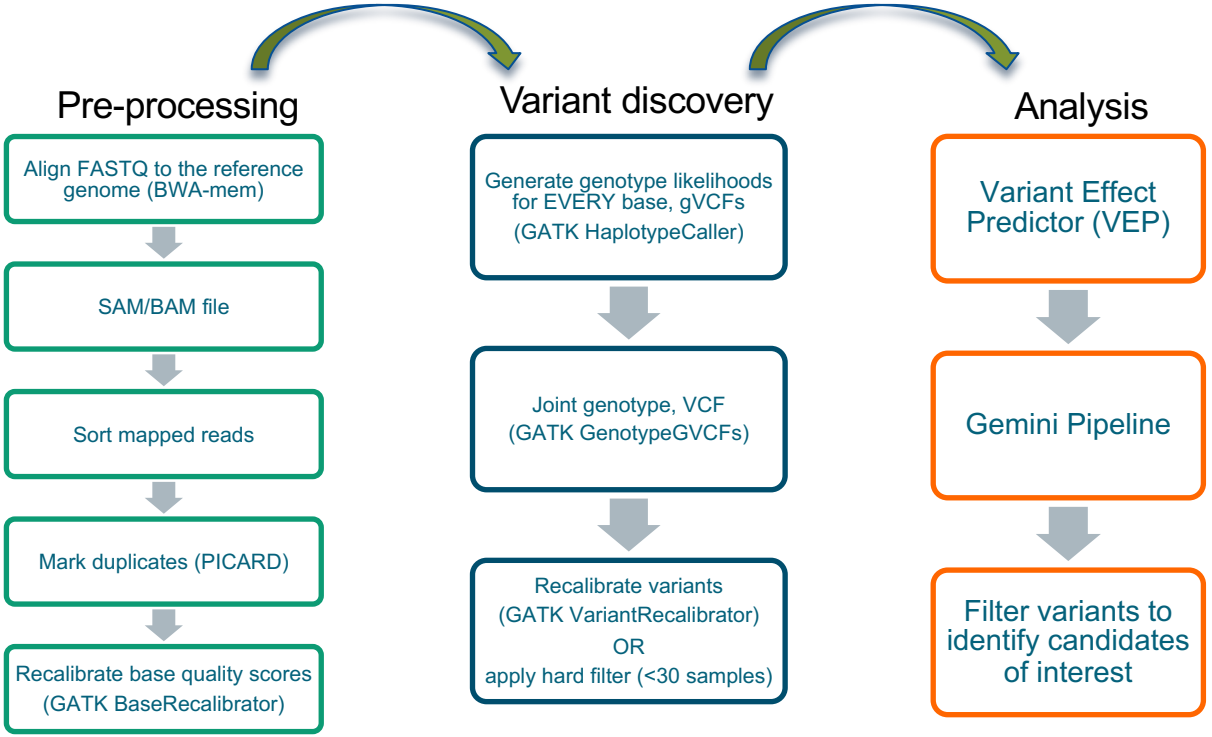
```
#Part2
$EBROOTGATK/gatk ApplyBQSR \
 -R /hpcdata/bcbb/wes_training/reference/human_g1k_v37_decoy.fasta \
 -I ./${sample}/${sample}.markdup.bam \
 -bqsr ./${sample}/${sample}.recal_data.table \
 -O ./${sample}/${sample}.bqsr.bam
```

QC Bonus: Sample contamination estimates!

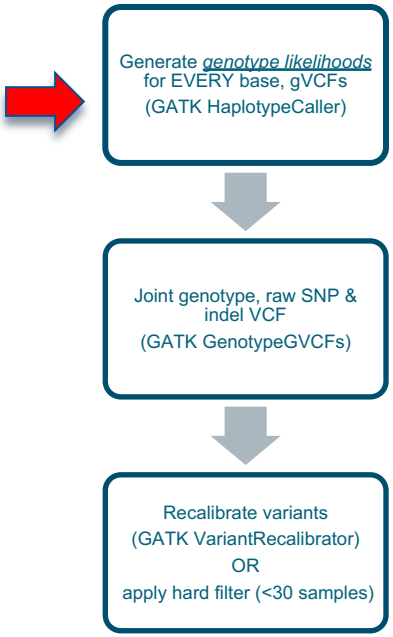
```
module load verifyBamID

verifyBamID --ignoreRG --noPhoneHome \
 --vcf /hpcdata/bcbb/wes_training/reference/1kg.sites.coding.ExAC.vcf.gz \
 --bam ${sample}/${sample}.bqsr.bam \
 --out ${sample}/${sample}.selfSM \
 --maxDepth 1000 –precise

cat ./son1/son1.selfSM.selfSM

module unload verifyBamID
```

Contamination: When genotype data is not available but allele-frequency-based estimates of [FREEMIX] >= 0.03 or [FREELK1] - [FREELK0] is large, then it is possible that the sample is contaminated with other samples!

Data processing



Step 1: Generate gVCF




```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  son1
6   82204749    .   A   <NON_REF>   .   .   END=82204869    GT:DP:GQ:MIN_DP:PL  0/0:116:99:79:0,120,1800
6   82459379    .   A   <NON_REF>   .   .   END=82460199    GT:DP:GQ:MIN_DP:PL  0/0:75:99:42:0,108,1620
6   82460200    .   A   <NON_REF>   .   .   END=82460205    GT:DP:GQ:MIN_DP:PL  0/0:41:90:39:0,90,1350
6   82460206    .   A   <NON_REF>   .   .   END=82460219    GT:DP:GQ:MIN_DP:PL  0/0:34:84:31:0,84,1260
6   82460790    .   C   <NON_REF>   .   .   END=82460910    GT:DP:GQ:MIN_DP:PL  0/0:62:99:40:0,102,1482
6   82461270    .   C   <NON_REF>   .   .   END=82461272    GT:DP:GQ:MIN_DP:PL  0/0:33:96:33:0,96,1440
6   82461273    .   G   <NON_REF>   .   .   END=82461458    GT:DP:GQ:MIN_DP:PL  0/0:52:99:33:0,99,1211
6   82461459    .   C   <NON_REF>   .   .   END=82461463    GT:DP:GQ:MIN_DP:PL  0/0:37:93:36:0,93,1395
6   82461464    .   T   <NON_REF>   .   .   END=82461464    GT:DP:GQ:MIN_DP:PL  0/0:37:99:37:0,99,1485
```

```
${EBROOTGATK}/gatk HaplotypeCaller \
 -R /hpcdata/bcbb/wes_training/reference/human_g1k_v37_decoy.fasta \
 -I ./${sample}/${sample}.bqsr.bam \
 -O ./${sample}/${sample}.g.vcf.gz \
 -stand-call-conf 20.0 \
 --dbsnp /hpcdata/bcbb/wes_training/reference/dbsnp_138.b37.vcf.gz \
 -L /hpcdata/bcbb/wes_training/reference/exome.interval_list \
 -ERC GVCF
```

Step 2: Make Joint Calling VCF

Generate *genotype likelihoods* for EVERY base, gVCFs (GATK HaplotypeCaller)

Joint genotype, raw SNP & indel VCF (GATK GenotypeGVCFs)

Recalibrate variants (GATK VariantRecalibrator) OR apply hard filter (<30 samples)

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  dad     daughter1       daughter2       mom
     son1    son2    son3
6       82930437        .       G       GA      4774.86 .       AC=7;AF=0.500;AN=14;BaseQRankSum=-2.850e-01;DP=582;Ex
cessHet=17.2937;FS=0.534;MLEAC=7;MLEAF=0.500;MQ=59.97;MQRankSum=0.00;QD=13.08;ReadPosRankSum=-8.880e-01;SOR=0.651  GT
:AD:DP:GQ:PL  0/1:12,25:40:99:730,0,172       0/1:28,14:42:99:279,0,583       0/1:8,39:51:20:1145,0,20        0/1:44,
20:68:99:468,0,895      0/1:41,38:88:99:869,0,846       0/1:13,25:41:99:508,0,246       0/1:24,34:66:99:824,0,457
6       82933309        .       C       T       7796.17 .       AC=9;AF=0.643;AN=14;BaseQRankSum=-5.360e-01;DP=386;Ex
cessHet=7.7512;FS=28.470;MLEAC=9;MLEAF=0.643;MQ=60.00;MQRankSum=0.00;QD=20.20;ReadPosRankSum=0.836;SOR=0.177       GT
:AD:DP:GQ:PL  1/1:0,39:39:99:1316,117,0       0/1:21,15:36:99:426,0,601       1/1:0,46:46:99:1621,138,0       0/1:39,
43:82:99:1325,0,1159    0/1:32,38:70:99:1236,0,1007     0/1:25,32:57:99:965,0,739       0/1:24,32:56:99:950,0,649
6       83075914        .       G       A       12212.17        .       AC=9;AF=0.643;AN=14;BaseQRankSum=1.38;DP=630;
```

```
#Set up our samples for processing
sample_list=("dad" "mom" "daughter1" "son1" "daughter2" "son2" "son3")
gvcf_list=""
for i in ${sample_list[@]}; do gvcf_list=${gvcf_list}" -V ./${i}/${i}.g.vcf.gz"; done

echo "$gvcf_list"

#Using GenomicsDBImport to consolidate GVCFs for input to GenotypeGVCFs
${EBROOTGATK}/gatk GenomicsDBImport ${gvcf_list} \
 --genomicsdb-workspace-path demo_database \
 -L /hpcdata/bcbb/wes_training/reference/exome.interval_list

#Perform joint genotyping on one or more samples pre-called with HaplotypeCaller
${EBROOTGATK}/gatk GenotypeGVCFs -R /hpcdata/bcbb/wes_training/reference/human_g1k_v37_decoy.fasta \
 -V gendb://demo_database -O fam001.joint.vcf.gz

#Sort the VCF
${EBROOTGATK}/gatk SortVcf -I fam001.joint.vcf.gz \
 -O fam001.joint.sorted.vcf.gz
```

QC Bonus: Family and Gender Check!

- **Check that the PED file looks normal…**

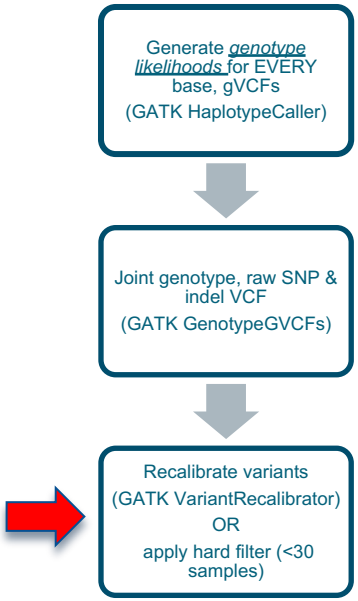
```
cat fam001.ped
# fam001 dad 0 0 1 1
# fam001 mom 0 0 2 1
# fam001 daughter1 dad mom 2 2
# fam001 daughter2 dad mom 2 1
# fam001 son1 dad mom 1 2
# fam001 son2 dad mom 1 2
# fam001 son3 dad mom 1 2
```

- **Run PEDDY to check status (https://github.com/brentp/peddy)**

```
module load peddy
python -m peddy -p 8 --prefix fam001 --plot fam001.joint.vcf.gz fam001.ped
```

PEDDY will print any errors to STDOUT (e.g. 'Individual 123 is marked as dad but has female sex' etc.)

Step 3: Recalibrate variants with VQSR or Deep Learning



At this stage we need a way to evaluate the quality of our variant calls. However, how we do this depends heavily on how many samples we have to consider. The GATK team offers a number of options…

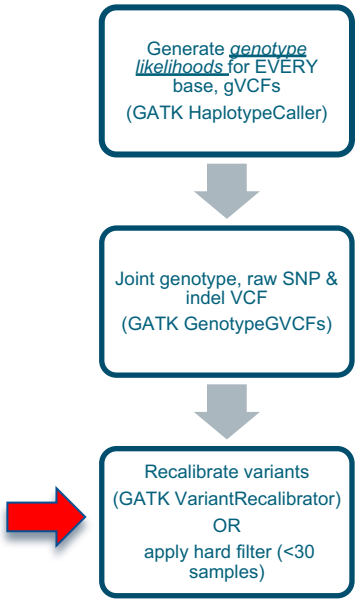
- Working on individual (non-joint called) samples: You can use VQSR or get better results using the new "GATK4 CNN" (convolutional neural network) tool (more details on the tool here)
- Working on **>= *30*** joint-called samples: You can use VQSR (tool details here) to train a model on known SNPs and INDELs and apply this model to your samples to score and rank variant calls
- For **< *30*** samples (*our case today!)* use HARD FILTERS – set thresholds and throw away any variants that fall above/below your cutoffs

Variant Call Performance Comparison [GATK4 CNN team]

Area Under Receiver Operating Characteristic Curve Table (high number = better performance)

–	INDEL AUC	INDEL AUC	INDEL AUC	SNP AUC	SNP AUC	SNP AUC	Metric
-	NA12878	NA24385	CHM WGS1	NA12878	NA24385	CHM WGS1	**Sample**
Architecture	NIST GiaB	NIST GiaB	SynDip	NIST GiaB	NIST GiaB	SynDip	**Truth**
VQSR Single Sample	.808	.557	.531	.954	.917	.808	
VQSR gnomAD	.869	.766	.571	.979	.955	.940	
Deep Variant	.874	.739	.779	**.987**	**.985**	.967	
GATK4 CNN	**.949**	**.836**	**.817**	**.987**	.979	**.973**	

Step 3 Continued: Apply hard filters (<30 samples)



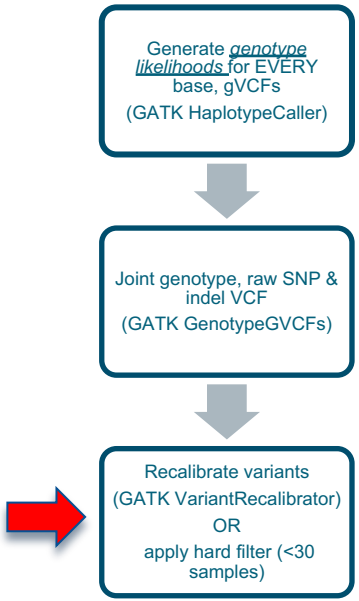
- Extract SNPs

```
${EBROOTGATK}/gatk SelectVariants -V fam001.joint.sorted.vcf.gz \
-select-type SNP -O fam001.snps.vcf.gz
```

- Extract non-SNPs

```
${EBROOTGATK}/gatk SelectVariants -V fam001.joint.sorted.vcf.gz \
-xl-select-type SNP -O fam001.nonsnps.vcf.gz
```

Step 3 Continued: Apply hard filters (<30 samples)



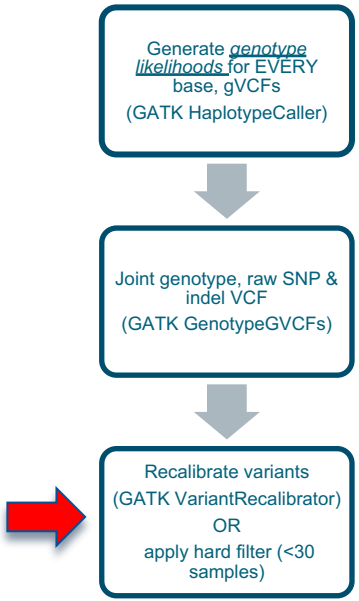
- Filter SNPs

```
${EBROOTGATK}/gatk VariantFiltration \
 -V fam001.snps.vcf.gz \
 -filter "QD < 2.0" --filter-name "QD2" \
 -filter "QUAL < 30.0" --filter-name "QUAL30" \
 -filter "SOR > 3.0" --filter-name "SOR3" \
 -filter "FS > 60.0" --filter-name "FS60" \
 -filter "MQ < 40.0" --filter-name "MQ40" \
 -filter "MQRankSum < -12.5" --filter-name "MQRankSum12.5" \
 -filter "ReadPosRankSum < -8.0" --filter-name "ReadPosRankSum8" \
 -O fam001.snps_filtered.vcf.gz
```

- Filter non-SNPs

```
${EBROOTGATK}/gatk VariantFiltration \
 -V fam001.nonsnps.vcf.gz \
 -filter "QD < 2.0" --filter-name "QD2" \
 -filter "QUAL < 30.0" --filter-name "QUAL30" \
 -filter "FS > 200.0" --filter-name "FS200" \
 -filter "ReadPosRankSum < -20.0" --filter-name "ReadPosRankSum20" \
 -O fam001.nonsnps_filtered.vcf.gz
```

Step 3 Continued: Apply hard filters (<30 samples)



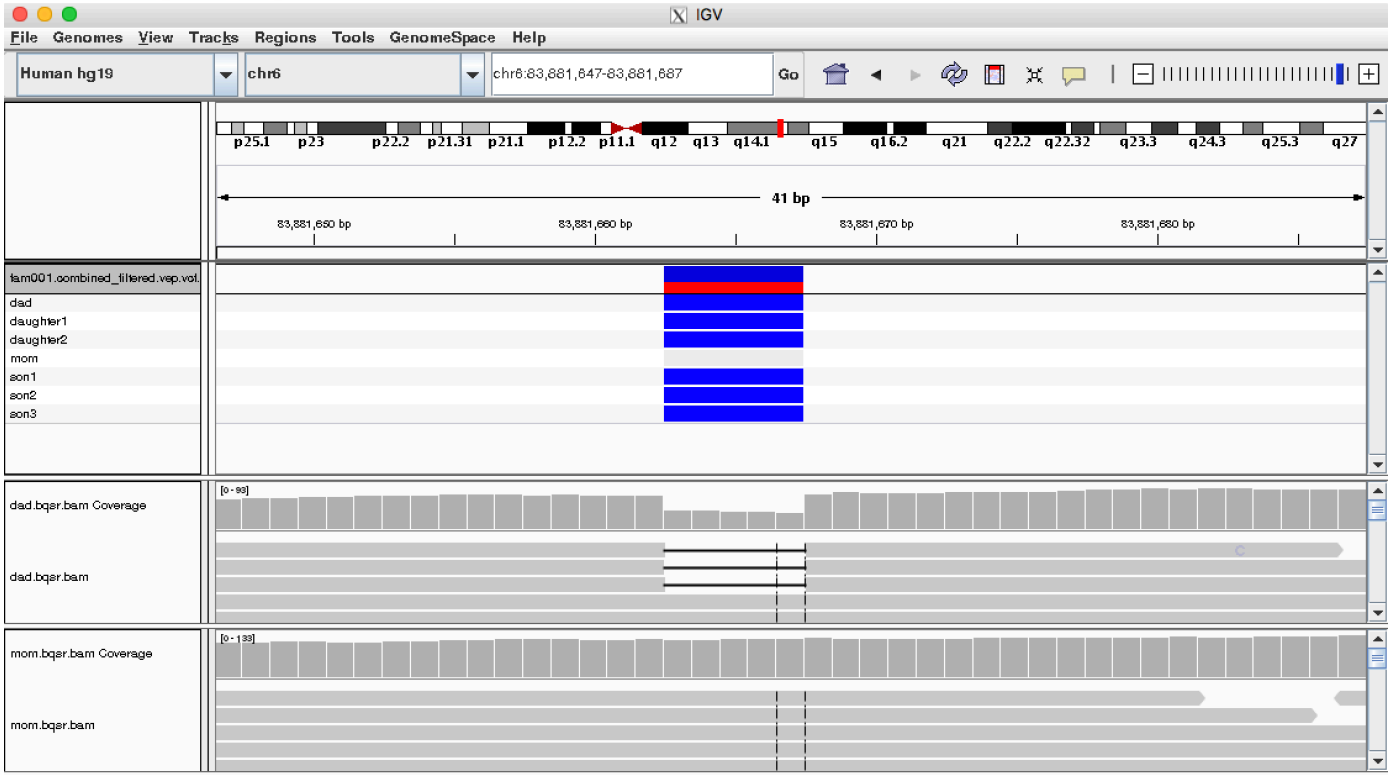
- Now combine filtered variants into a single VCF …

```
${EBROOTGATK}/gatk MergeVcfs \
 -I fam001.snps_filtered.vcf.gz \
 -I fam001.nonsnps_filtered.vcf.gz \
 -O fam001.combined_filtered.vcf.gz
```

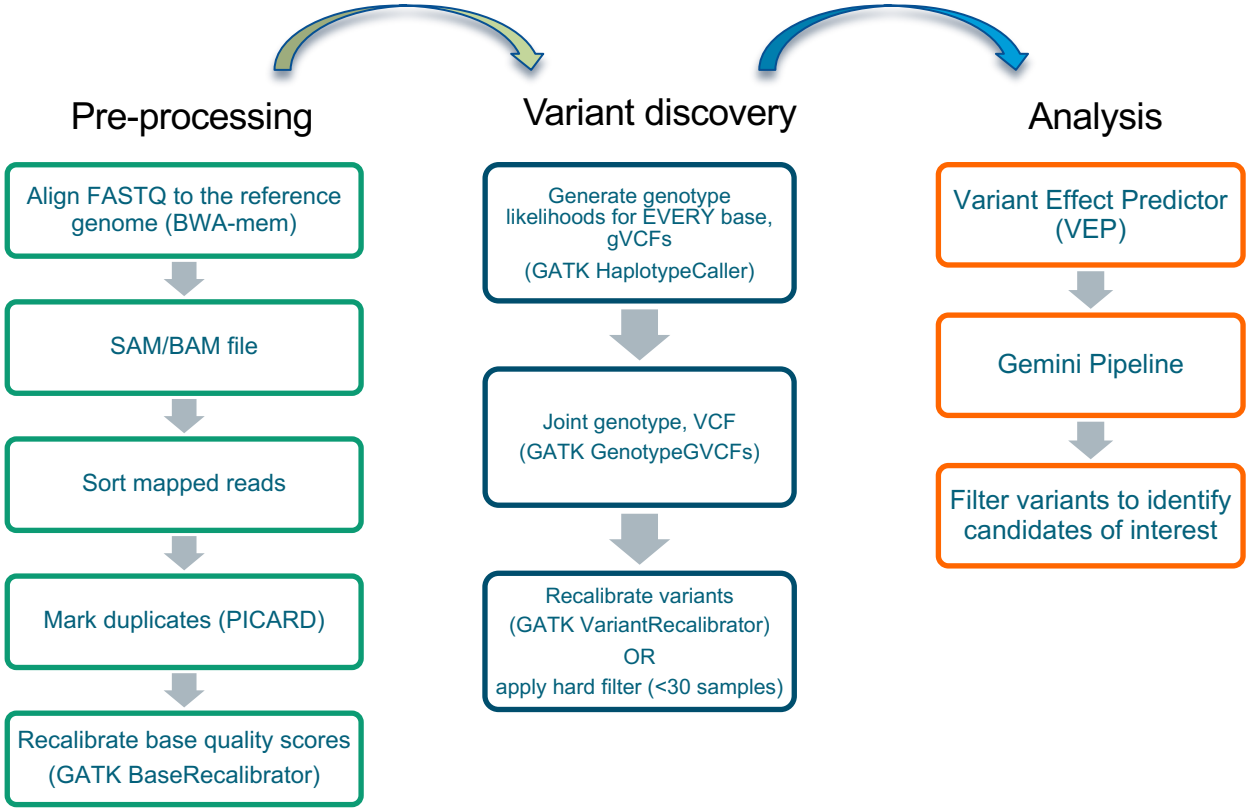
Example of filtered output

```
Y       15027529        .       T       G       9330.14 SOR3    AC=8;AF=1.00;AN=8;DP=268;ExcessHet=3.0103;FS=0.000;ML
EAC=8;MLEAF=1.00;MQ=59.98;QD=34.81;SOR=3.158        GT:AD:DP:GQ:PL  1/1:0,65:65:99:2184,195,0        ./.:0,0:0:.:0,0,0
        ./.:0,0:0:.:0,0,0       ./.:0,0:0:.:0,0,0       1/1:0,74:74:99:2758,223,0       1/1:0,67:67:99:2314,202,0
        1/1:0,62:62:99:2099,186,0
Y       15467824        .       A       G       2203.14 PASS    AC=8;AF=1.00;AN=8;DP=64;ExcessHet=3.0103;FS=0.000;MLE
AC=8;MLEAF=1.00;MQ=60.00;QD=34.42;SOR=1.721         GT:AD:DP:GQ:PL  1/1:0,10:10:30:319,30,0 ./.:0,0:0:.:0,0,0        ./
.:0,0:0:.:0,0,0       ./.:0,0:0:.:0,0,0       1/1:0,18:18:54:670,54,0 1/1:0,21:21:63:707,63,0 1/1:0,15:15:45:532,45,0
Y       15591537        .       G       C       1237.16 SOR3    AC=8;AF=1.00;AN=8;DP=41;ExcessHet=3.0103;FS=0.000;MLE
AC=8;MLEAF=1.00;MQ=60.00;QD=30.17;SOR=3.312         GT:AD:DP:GQ:PL  1/1:0,10:10:30:308,30,0 ./.:0,0:0:.:0,0,0        ./
.:0,0:0:.:0,0,0       ./.:0,0:0:.:0,0,0       1/1:0,17:17:51:536,51,0 1/1:0,6:6:18:179,18,0   1/1:0,8:8:24:239,24,0
```

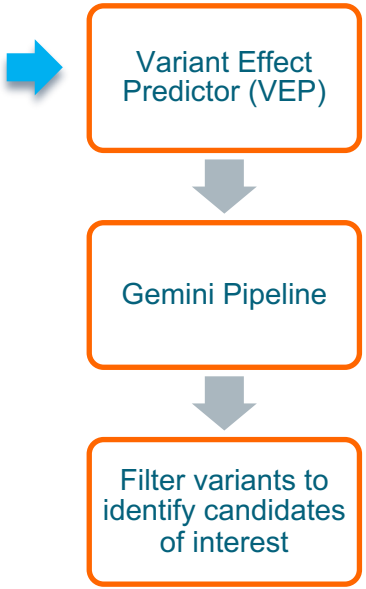
Check Output in IGV



Data processing



VEP




```
module load bcftools

bcftools norm -m -any --threads 8 -Oz \
-o fam001.combined_filtered.norm.vcf.gz fam001.combined_filtered.vcf.gz

tabix -p vcf fam001.combined_filtered.norm.vcf.gz

module load VEP/89-goolf-1.7.20

dbNSFP_dir="/sysapps/cluster/software/VEP/84-goolf-1.7.20/.vep/dbNSFP/2.9.3"

${EBROOTVEP}/vep --fork 8 --dir ${EBROOTVEP}/.vep --assembly GRCh37 --merged \
--everything --vcf --allele_number --no_stats --cache --offline --force_overwrite \
--compress_output bgzip --plugin \
dbNSFP,${dbNSFP_dir}/dbNSFP.gz,Polyphen2_HVAR_pred,CADD_phred,SIFT_pred,FATHMM_pred,MutationTaster_pr
ed,MetaSVM_score,MetaSVM_pred,MetaLR_score,MetaLR_pred,Reliability_index –plugin \
LoF,human_ancestor_fa:$EBROOTVEP/.vep/Plugins/loftee/human_ancestor.fa.gz,filter_position
:0.05,min_intron_size:15 \
-i fam001.combined_filtered.norm.vcf.gz -o fam001.combined_filtered.vep.vcf.gz

tabix -p vcf fam001.combined_filtered.vep.vcf.gz # create index for the normalized VCF
```

Now we have annotated VCF file!!

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  dad     daughter1       daughter2       mom
     son1    son2    son3
6       82930437        .       G       GA      4774.86 PASS    AC=7;AF=0.5;AN=14;BaseQRankSum=-0.285;DP=582;ExcessHe
t=17.2937;FS=0.534;MLEAC=7;MLEAF=0.5;MQ=59.97;MQRankSum=0;QD=13.08;ReadPosRankSum=-0.888;SOR=0.651;CSQ=A|splice_regio
n_variant&intron_variant|LOW|IBTK|ENSG00000005700|Transcript|ENST00000306270|protein_coding||8/28|ENST00000306270.7:c
.1125-8dupT|||||||rs34032511|1||-1||insertion|HGNC|17853|YES|||CCDS34490.1|ENSP00000305721|Q9P2D0||UPI000041929F||Ens
embl|||||||||||||||||||||||||||||||||||||||||||||||||,A|splice_region_variant&intron_variant&non_coding_transcript_varia
nt|LOW|IBTK|ENSG00000005700|Transcript|ENST00000369751|retained_intron||8/23|ENST00000369751.2:n.1675-8dupT|||||||rs3
4032511|1||-1||insertion|HGNC|17853|||||||||Ensembl|||||||||||||||||||||||||||||||||||||||||||||||||,A|splice_region_va
riant&intron_variant&NMD_transcript_variant|LOW|IBTK|ENSG00000005700|Transcript|ENST00000503400|nonsense_mediated_dec
ay||8/28|ENST00000503400.1:c.*436-8dupT|||||||rs34032511|1||-1||insertion|HGNC|17853|||||ENSP00000422136||D6R9F4|UPI0
001D3BC1A||Ensembl|||||||||||||||||||||||||||||||||||||||||||||||||,A|splice_region_variant&intron_variant|LOW|IBTK|ENSG
00000005700|Transcript|ENST00000503631|protein_coding||7/26|ENST00000503631.1:c.1125-8dupT|||||||rs34032511|1||-1||in
sertion|HGNC|17853|||||ENSP00000422762||E9PDR5|UPI0001D3BC1B||Ensembl|||||||||||||||||||||||||||||||||||||||||||||||||,A
|splice_region_variant&intron_variant&non_coding_transcript_variant|LOW|IBTK|ENSG00000005700|Transcript|ENST000005052
22|retained_intron||7/25|ENST00000505222.1:n.1204-8dupT|||||||rs34032511|1||-1||insertion|HGNC|17853|||||||||Ensembl
|||||||||||||||||||||||||||||||||||||||||||||||||,A|splice_region_variant&intron_variant|LOW|IBTK|ENSG00000005700|Transc
ript|ENST00000510291|protein_coding||7/27|ENST00000510291.1:c.1125-8dupT|||||||rs34032511|1||-1||insertion|HGNC|17853
|||||ENSP00000426405||E7EPI0|UPI0001D3BBE0||Ensembl|||||||||||||||||||||||||||||||||||||||||||||||||,A|splice_region_var
iant&intron_variant|LOW|IBTK|25998|Transcript|NM_015525.2|protein_coding||8/28|NM_015525.2:c.1125-8dupT|||||||rs34032
511|1||-1||insertion|||YES||||NP_056340.2|||||RefSeq|||||||||||||||||||||||||||||||||||||||||||||||||    GT:AD:DP:GQ:PL
  0/1:12,25:40:99:730,0,172       0/1:28,14:42:99:279,0,583       0/1:8,39:51:20:1145,0,20        0/1:44,20:68:99:468
,0,895       0/1:41,38:88:99:869,0,846       0/1:13,25:41:99:508,0,246       0/1:24,34:66:99:824,0,457
```

Now what?

What do you do with all these variants?

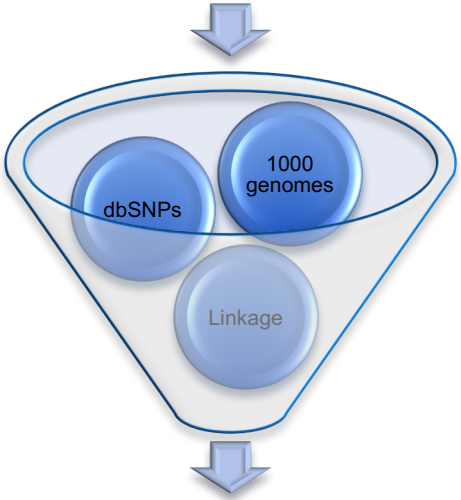
Variant Analysis…
like finding a needle in a ‘deep’ haystack

Look for evidence of variants of interest

Further filters needed

- High number of variants
- The goal is to narrow down your list of variants
- Eliminate variants that are not interesting

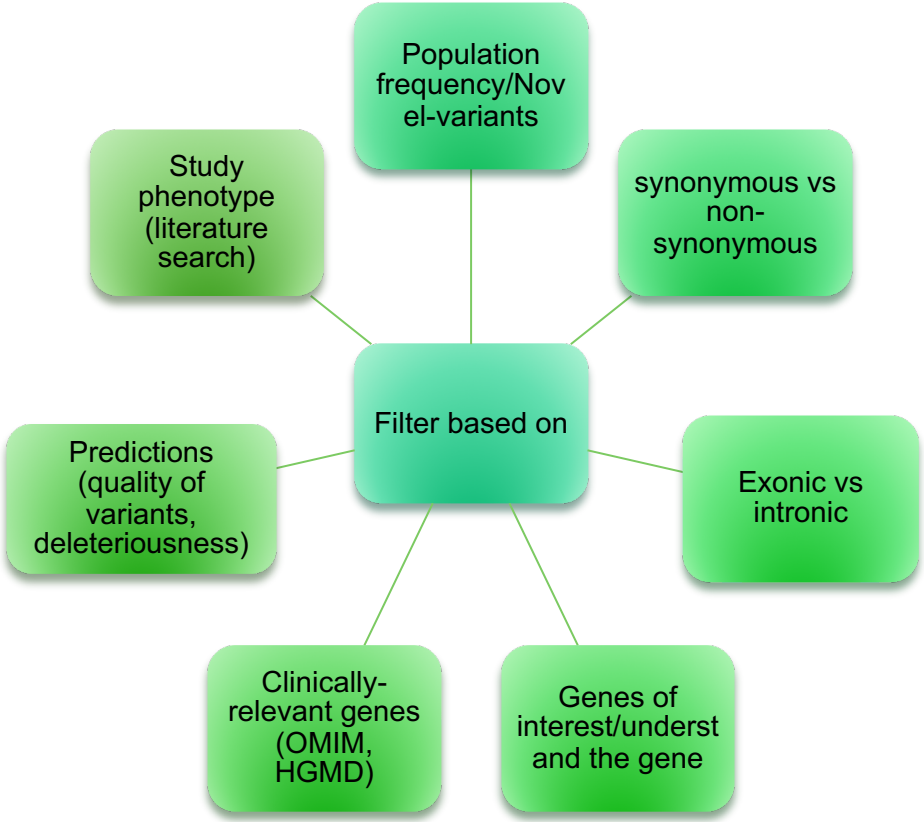
17,687 coding SNPs



2 novel in chr10

PLoS One. 2012;7(1):e29708

Things to consider



GEMINI

GEMINI can be used to consider inheritance models to find candidate variants to explore further!

- Step1 - Create a reference DB:

```
module load GEMINI
gemini load -v fam001.combined_filtered.vep.vcf.gz \
 -t VEP -p fam001.ped --cores 8 --tempdir /hpcdata/scratch/ fam001.db
```

- Step2 - Try different inheritance models:

```
gemini <INSERT_MODEL> --columns \
 "gene, chrom, start, end, ref, alt, impact_severity,vep_cadd_phred, impact, in_1kg" fam001.db
```

*Where **<INSERT_MODEL>** above can be "**autosomal_dominant**","**autosomal_recessive**", "**comp_hets**", "**de_novo**", "**x_linked_de_novo**", etc. (type "`gemini -h`" on the command line for the full list of options)*

Did any of your results show any promising candidates???

GEMINI Output

```
gemini autosomal_recessive --columns  "gene, chrom, start, end, ref, alt, impact_severity,vep_cadd_phred, impact, in_1kg" fam001.db
```

OUTPUT:

```
gene chrom start end ref alt impact_severity vep_cadd_phred impact in_1kg variant_id family_id family_members
family_genotypes samples family_count
OR5AN1 chr11 59132797 59132798 G C MED 24.2 missense_variant 1 27 fam001
dad(dad;unaffected;male),daughter1(daughter1;affected;female),daughter2(daughter2;unaffected;female),mom(mom;unaffected;fe
male),son1(son1;affected;male),son2(son2;affected;male),son3(son3;affected;male) G/C,C/C,G/C,G/C,C/C,C/C,C/C
daughter1,son1,son2,son3 1
OR5A1 chr11 59211187 59211188 G A MED 27.2 missense_variant 1 33 fam001
dad(dad;unaffected;male),daughter1(daughter1;affected;female),daughter2(daughter2;unaffected;female),mom(mom;unaffected;fe
male),son1(son1;affected;male),son2(son2;affected;male),son3(son3;affected;male) G/A,A/A,G/A,G/A,A/A,A/A,A/A
daughter1,son1,son2,son3 1
OR5A1 chr11 59211264 59211265 T C LOW synonymous_variant 1 34 fam001
dad(dad;unaffected;male),daughter1(daughter1;affected;female),daughter2(daughter2;unaffected;female),mom(mom;unaffected;fe
male),son1(son1;affected;male),son2(son2;affected;male),son3(son3;affected;male) T/C,C/C,T/C,T/C,C/C,C/C,C/C
daughter1,son1,son2,son3 1
```

GEMINI Output

```
gemini comp_hets --columns  "gene, chrom, start, end, ref, alt, impact_severity,vep_cadd_phred, impact, in_1kg" fam001.db
```

OUTPUT:

```
gene chrom start end ref alt impact_severity vep_cadd_phred impact in_1kg variant_id family_id family_members
family_genotypes samples family_count comp_het_id priority
PGM3 chr6 83880061 83880062 C G MED 28.4 missense_variant 0 13 fam001
dad(dad;unaffected;male),daughter1(daughter1;affected;female),daughter2(daughter2;unaffected;female),mom(mom;unaffected;
female),son1(son1;affected;male),son2(son2;affected;male),son3(son3;affected;male) C/C,G|C,C/C,C/G,G|C,G|C,G|C
daughter1,son1,son2,son3 1 1_13_15 1
PGM3 chr6 83881661 83881667 TTTAAG T HIGH frameshift_variant 0 15 fam001
dad(dad;unaffected;male),daughter1(daughter1;affected;female),daughter2(daughter2;unaffected;female),mom(mom;unaffected;
female),son1(son1;affected;male),son2(son2;affected;male),son3(son3;affected;male)
TTTAAG/T,TTTAAG|T,TTTAAG|T,TTTAAG/TTTAAG,TTTAAG|T,TTTAAG|T,TTTAAG|T daughter1,son1,son2,son3 1 1_13_15 1
```

GEMINI

- Step3 – Get basic stats for your DB:

```
gemini stats --tstv fam001.db
gemini stats --tstv-coding fam001.db
gemini stats --tstv-noncoding fam001.db
```

- Step4 – Get number of variants for each sample:

```
gemini stat –vars-by-sample fam001.db
```

Follow-up

- A literature search *IS A MUST*! It is time consuming, but required to interpret for prioritization and further follow up of variants.

  - Valuable resources for exploring variants further:
    - gnomAD Browser: https://gnomad.broadinstitute.org
    - Human Gene Mutation Database (*NIH employees get a free subscription through the library!*): http://www.hgmd.cf.ac.uk/ac/index.php
    - dbSNP: https://www.ncbi.nlm.nih.gov/snp/

Next: Part 3 - Additional analyses

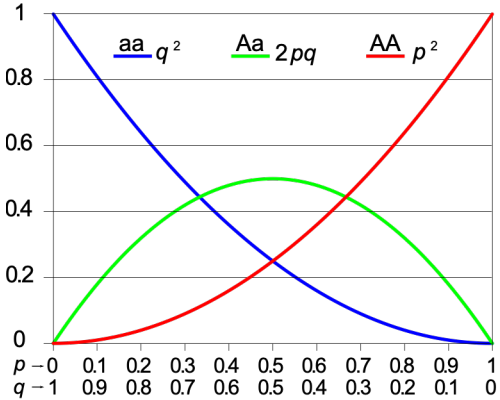
Additional Analysis

- Association analysis
  - Single variant based case-control study (common variants)
  - Group-wise burden test (rare variants)
- Copy number variation analysis
- Somatic Variation
  - MuTect2
  - VarDict

Population-based Cohort Study

GWAS Association studies

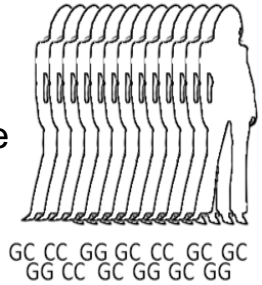
- A typical analysis:
  - Identify SNPs where one allele is significantly more common in cases than controls
  - Hardy-Weinberg equilibrium (HWE) testing

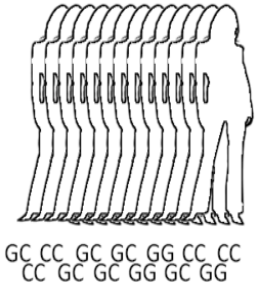


χ^2 test for HWE deviation

$p + q = 1$
$Exp(\text{aa}) = p^2 n$
$Exp(\text{Aa}) = 2pqn$
$Exp(\text{AA}) = q^2 n$

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$





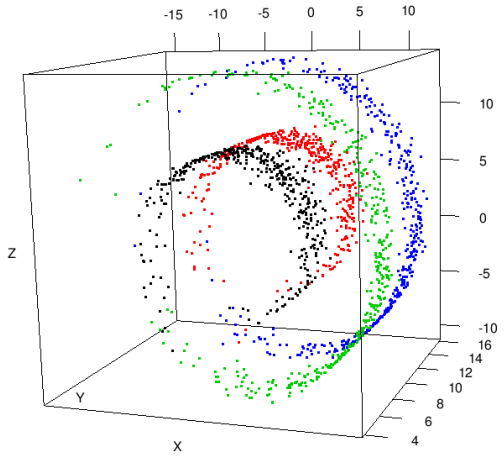
SNP1	*SNP2*	*SNP…*
Cases	**Cases**	*Repeat for all SNPs*
Count of G: 2104 of 4000	Count of G: 1648 of 4000	
Frequency of G: 52.6%	Frequency of G: 41.2%	
Controls	**Controls**	
Count of G: 2676 of 6000	Count of G: 2532 of 6000	
Frequency of G: 44.6%	Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

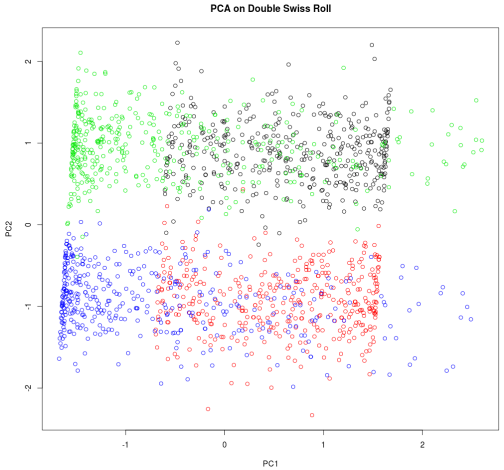
http://en.wikipedia.org/wiki/Genome-wide_association_study

http://www.ebi.ac.uk/gwas/

Population-based Cohort Study

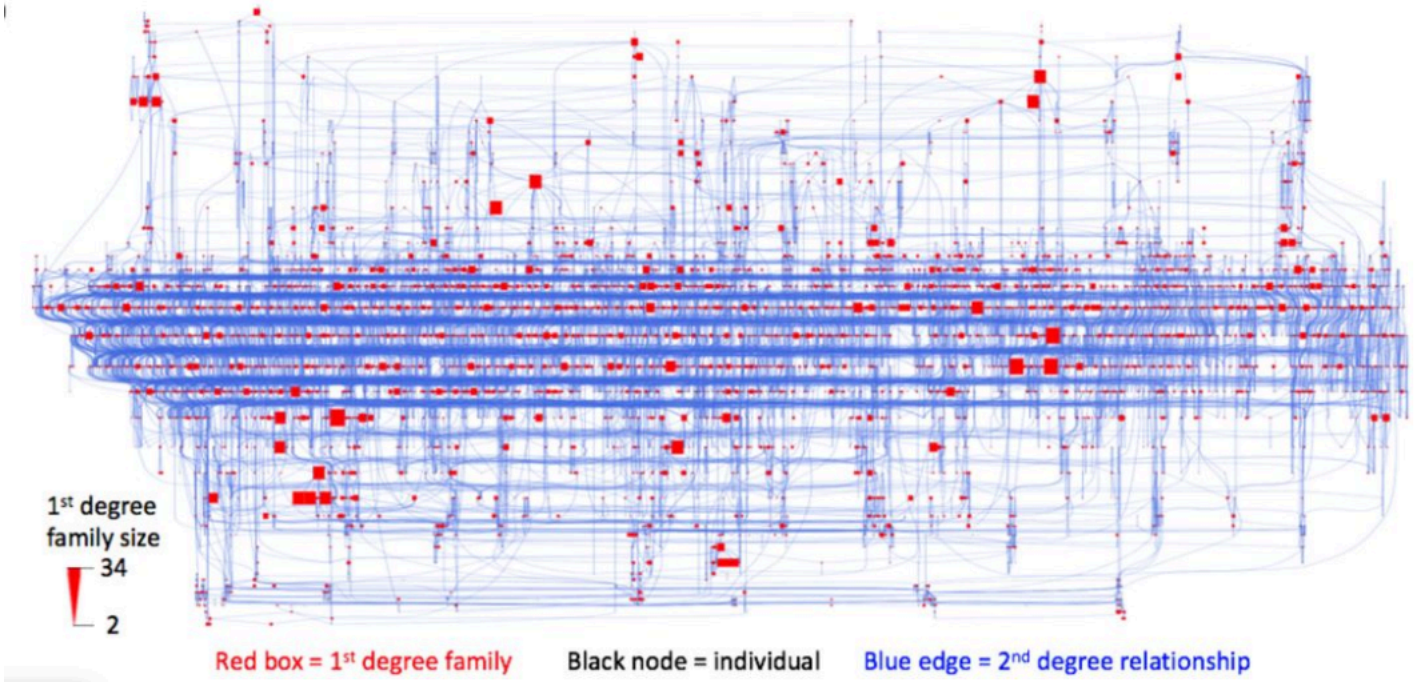
- Population stratification
  - Genomic Control λ (λ_{GC})
  - Principal components analysis (PCA)
  - Mixed-Model





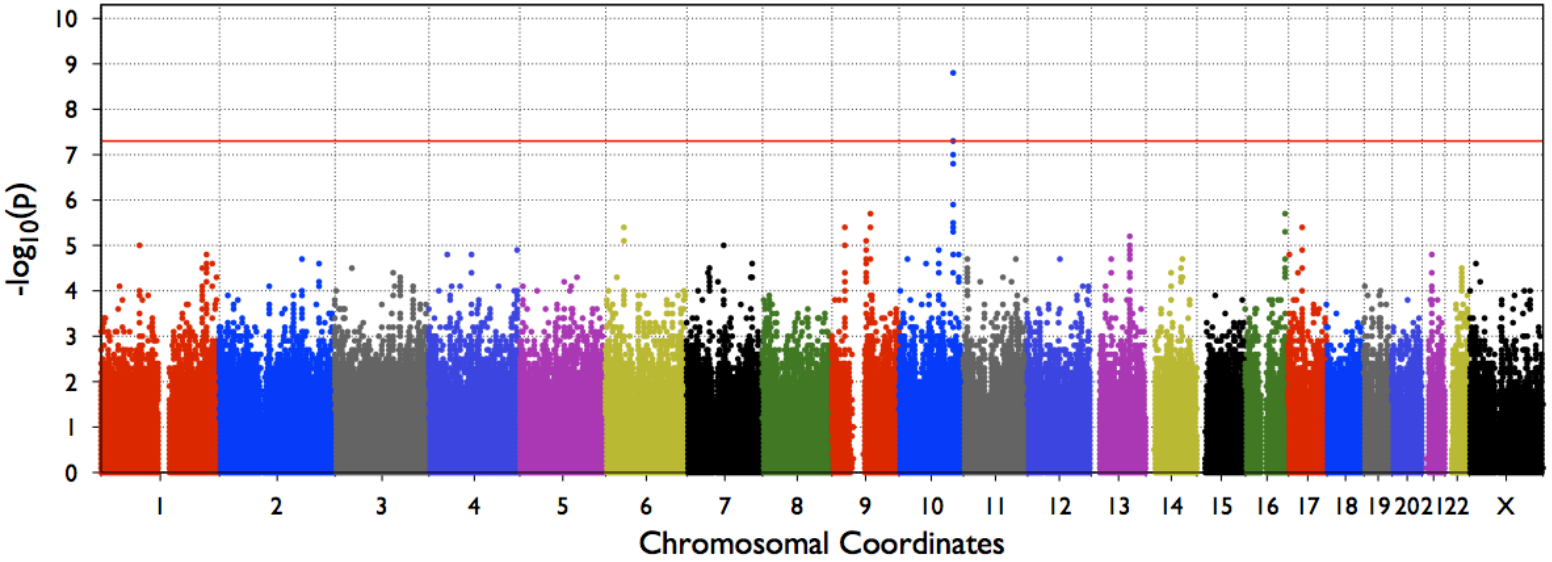
- Multiple-testing correction
  - Bonferroni ($P_{adjusted} = P_{raw} * n$)
  - Benjamini & Hochberg FDR ($P_{adjusted} = P_{raw} * n/k$)

4,062 first-degree family networks (red boxes) and 5,584 additional individuals (black nodes) connected by 11,430 second-degree relationships (blue edges)



Staples et al., Am J Hum Genet. 2018 May 3;102(5):874-889.

GWAS Manhattan plot



- Only applicable for large cohort

Analysis Tools for genotype-phenotype association from sequencing data

- PLINK/SEQ - https://atgu.mgh.harvard.edu/plinkseq/
- EPACTS - https://genome.sph.umich.edu/wiki/EPACTS
- VAAST 2 - http://www.yandell-lab.org/software/vaast.html
- SNPTEST – https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html

Copy Number Variations (CNVs) from exome

- CNV is a stretch of DNA, which is duplicated in some people, and sometimes even triplicated or quadruplicated.
- High variability of read-depth in exomes
- CNV prediction is very challenging problem
  - High false positive rate
  - Ambiguous break-points due to targets
  - Long range CNVs have higher positive predictive value
- Lots of prediction tools!! (some examples below)

Population Caller	Somatic Caller
CNVnator CONTRA CoNIFER ExomeDepth GATK XHMM	ADTEx CONTRA Control-FREEC ExomeCNV EXCAVATOR GATK Varscan2

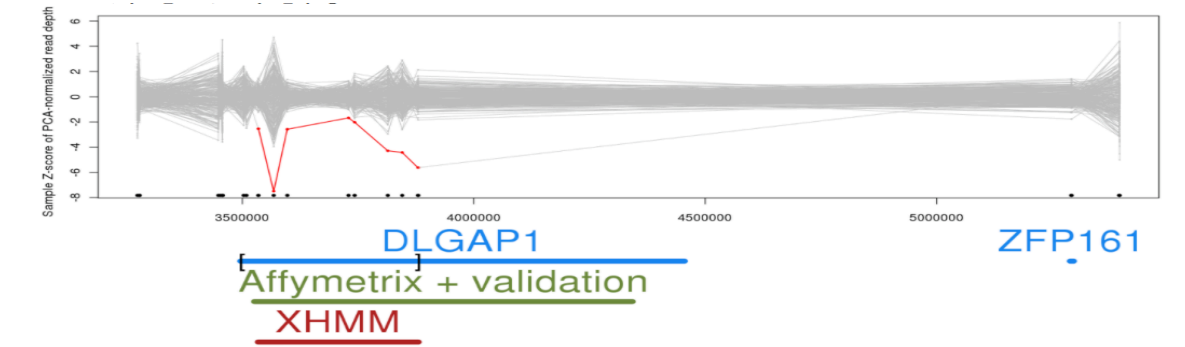
Evaluation of structural variation detection algorithms for ***WGS***

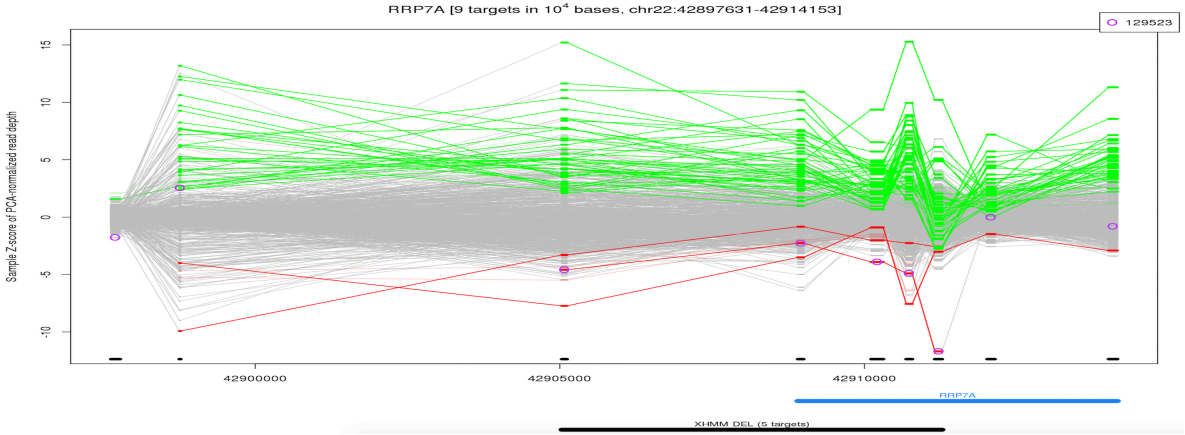
Table 1 List of tools providing good SV calling results for both the simulated and NA12878 real datasets

SV type	Tools	Simulated data		Real data		nF[*1]
		Precision	Recall	Precision	Recall	
DEL	GRIDSS	98.9 (5)	86.6 (2)	87.6 (7)	28.9 (2)	3.57 (1)
	Lumpy	99.1 (4)	81.4 (6)	87.1 (8)	26.1 (4)	3.41 (2)
	SVseq2	96.2 (11)	86.1 (3)	75.7 (17)	24.9 (5)	3.28 (3)
	SoftSV	96.8 (10)	83.6 (4)	80.2 (13)	23.2 (8)	3.25 (7)
	Manta	95.9 (12)	83.1 (5)	74.2 (20)	24.3 (6)	3.21 (5)
	MATCHCLIP	99.4 (2)	71.7 (10)	91.6 (4)	20.9 (11)	3.12 (6)
	inGAP-sv	91.1 (18)	78.6 (7)	78.3 (14)	22.5 (8)	3.10 (7)
DUP	Wham	96.9 (4)	81.7 (4)	57.1 (4)	10.2 (5)	3.92 (1)
	SoftSV	84.2 (14)	67.8 (13)	47.3 (6)	14.3 (3)	3.91 (2)
	MATCHCLIP	87.6 (11)	77.5 (8)	58.0 (3)	9.9 (6)	3.79 (3)
	GRIDSS	91.1 (9)	77.9 (7)	58.4 (2)	9.6 (7)	3.78 (4)
	Manta	99.0 (1)	83.2 (1)	40.4 (9)	6.5 (11)	3.35 (5)
	SvABA	82.6 (15)	69.6 (11)	42.7 (8)	7.2 (9)	3.02 (6)
INS [Unspecified]	pbsv	89.7 (3)	38.2 (5)	72.7 (8)	27.5 (2)	6.68 (1)
	inGAP-sv	99.7 (1)	58.5 (2)	85.5 (2)	11.8 (3)	6.27 (2)
	Sniffles	74.8 (5)	52.5 (3)	65.9 (10)	9.0 (5)	5.08 (3)
	SVseq2	70.4 (8)	64.2 (1)	38.5 (19)	7.1 (9)	4.87 (4)
INS [MEI]	MELT	99.7 (3)	68.9 (3)	88.9 (1)	85.6 [*2] (1)	3.21 (1)
	Mobster	100 (1)	67.1 (4)	88.3 (2)	71.9 [*2] (2)	3.04 (2)
INV	DELLY	94.7 (8)	81.8 (4)	38.9 (4)	15.6 (2)	3.07 (1)
	TIDDIT	89.2 (14)	77.9 (8)	49.1 (1)	11.7 (5)	2.89 (2)
	1–2-3-SV	70.7 (19)	81.2 (5)	31.8 (9)	14.8 (3)	2.67 (3)
	GRIDSS	96.6 (6)	84.7 (3)	34.2 (8)	10.4 (7)	2.67 (4)

[*1]Sum of normalized *F*-measures of the simulated and the real data. Normalized *F*-measure = *F*-measure/the mean *F*-measure for the corresponding category
[*2]Provisional recall value: the number of true positives was calculated by dividing by the provisional number of reference MEIs (1350), which was estimated using the data from the 1000 Genome project
Ranks of tools for each result (precision, recall, or *F*-measure) are indicated within parentheses

Popular utility of exome sequencing
eXome Hidden Markov Model





National Institute of Allergy and Infectious Diseases

http://atgu.mgh.harvard.edu/xhmm/tutorial.shtml

NIAID

Suggestions

- Request for a LOCUS/Biowulf account (if you don't have one already, you will automatically receive notification when logging into the Locus Support Portal (https://locus.niaid.nih.gov) for the first time and will be given instructions to request an account.)
  - Programs are already installed
  - Easy to keep track of software version
  - You can request programs to be installed

    NIAIDHPCSUPPORT@niaid.nih.gov

Questions ?

National Institute of Allergy and Infectious Diseases

Credit: http://omogemura.com/thank-you/

NIAID