Exome-Seq and Whole Genome Analysis: Overview and Best Practices

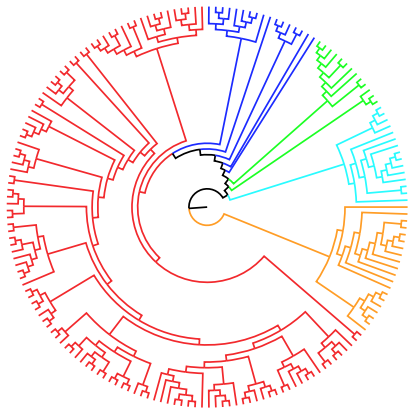
Justin Lack

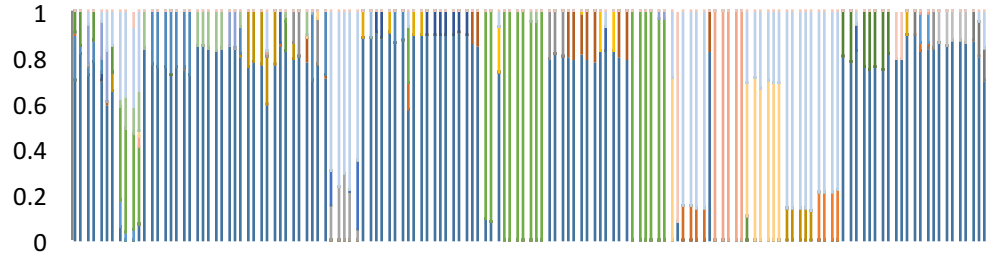
Variant Calling at CCBR

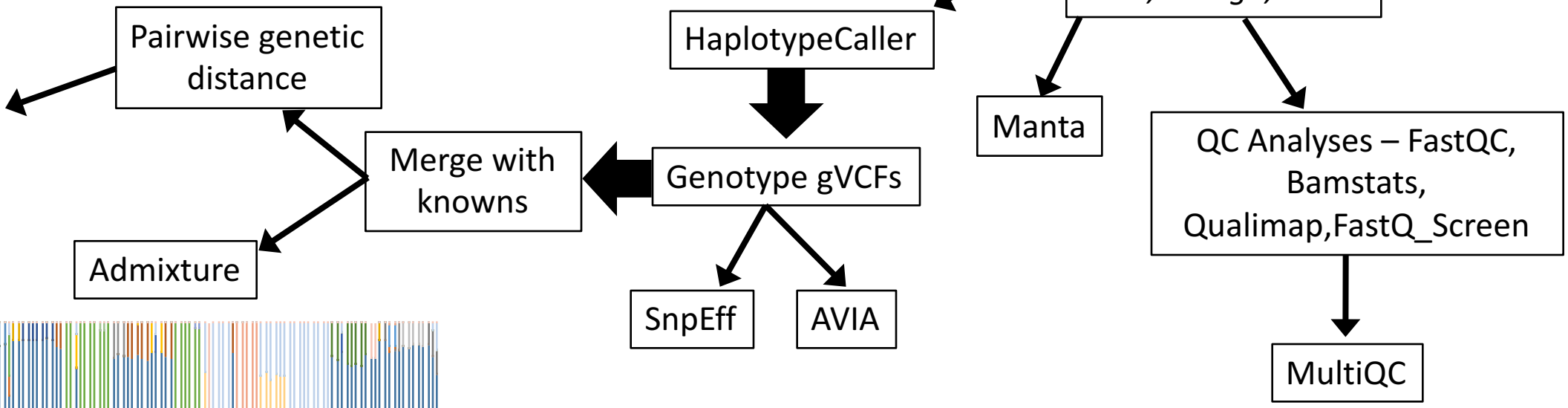
- Multiple Variant Calling CCBR Pipelines
  - Whole genome
  - Whole exome/targeted sequencing
  - RNAseq
- Generally follow GATK Best Practices, with modifications

Variant Calling at CCBR

- Multiple Variant Calling CCBR Pipelines
  - Whole genome
  - Whole exome/targeted sequencing



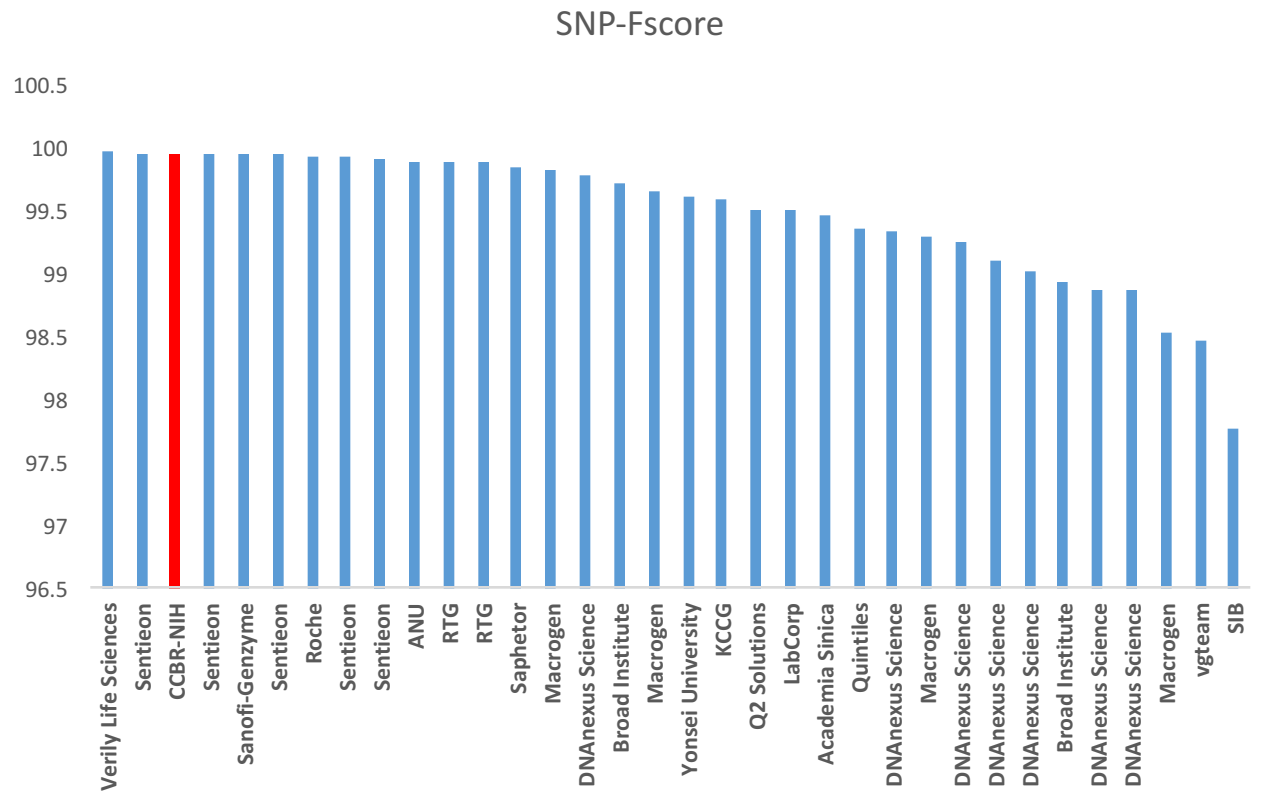




Variant Calling at CCBR

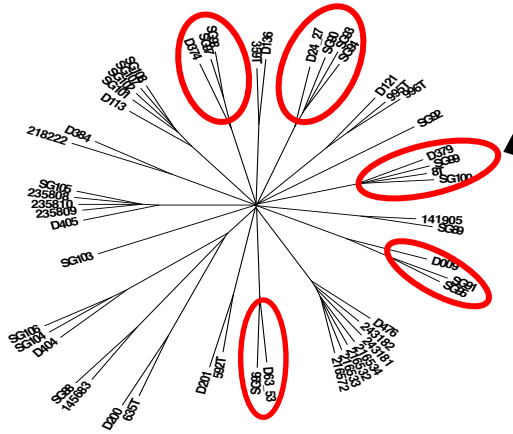
- Multiple Variant Calling CCBR Pipelines
  - Whole genome
  - Whole exome/targeted sequencing
  - Excellent performance in Precision FDA Challenge

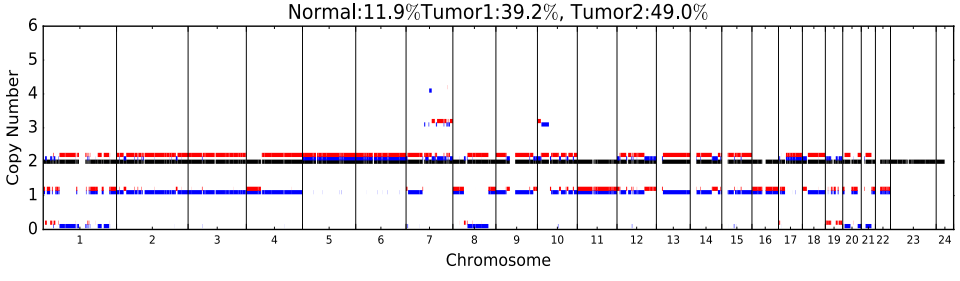
Germline Variant Calling

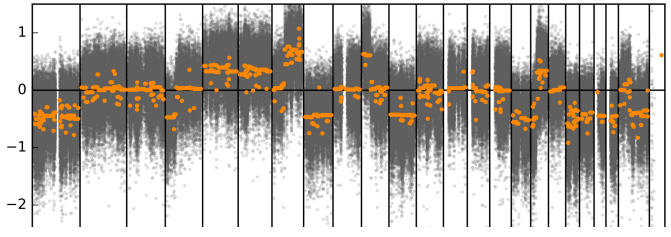


Variant Calling at CCBR

- Multiple Variant Calling CCBR Pipelines
  - Whole genome
  - Whole exome/targeted sequencing

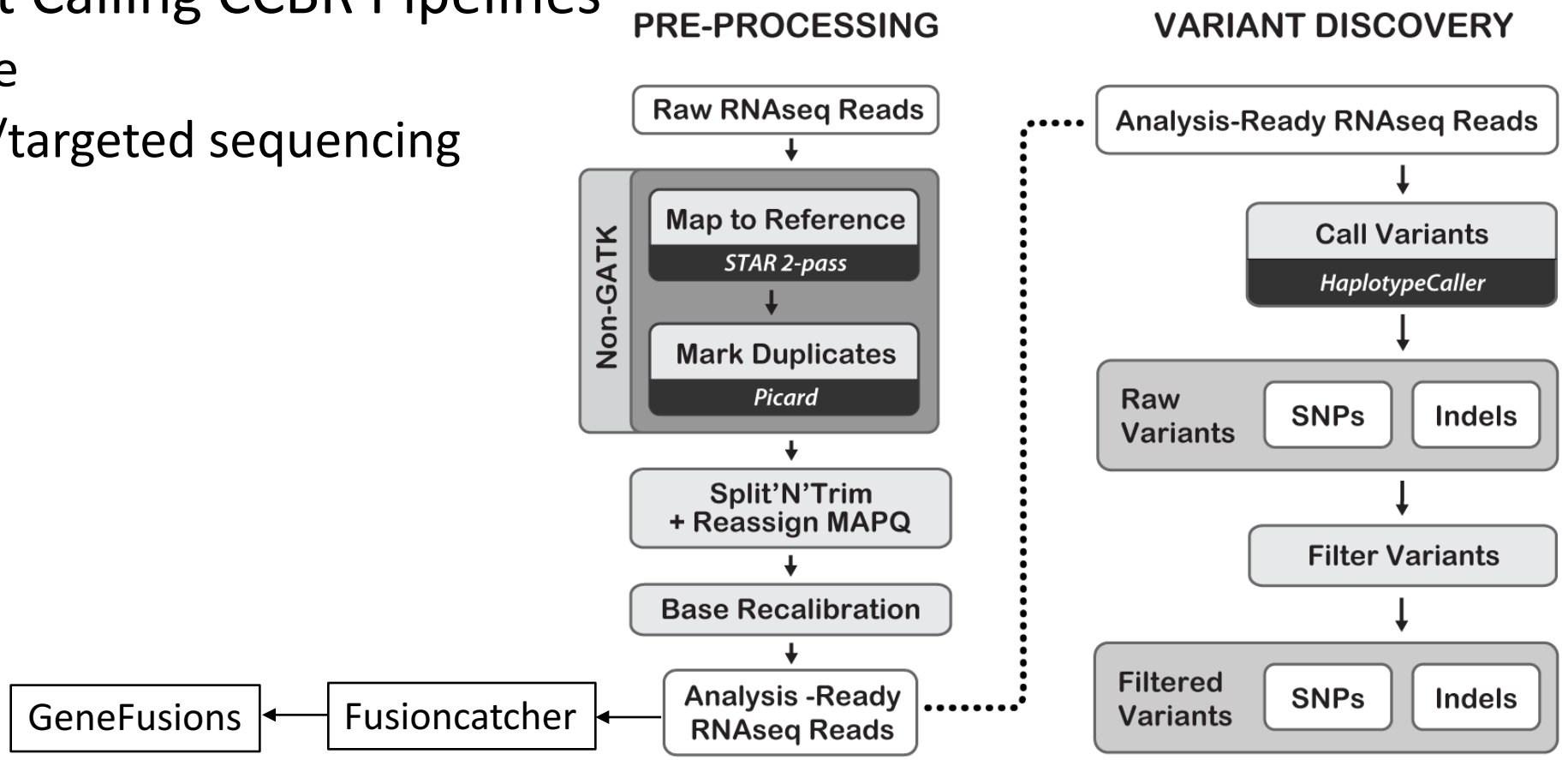






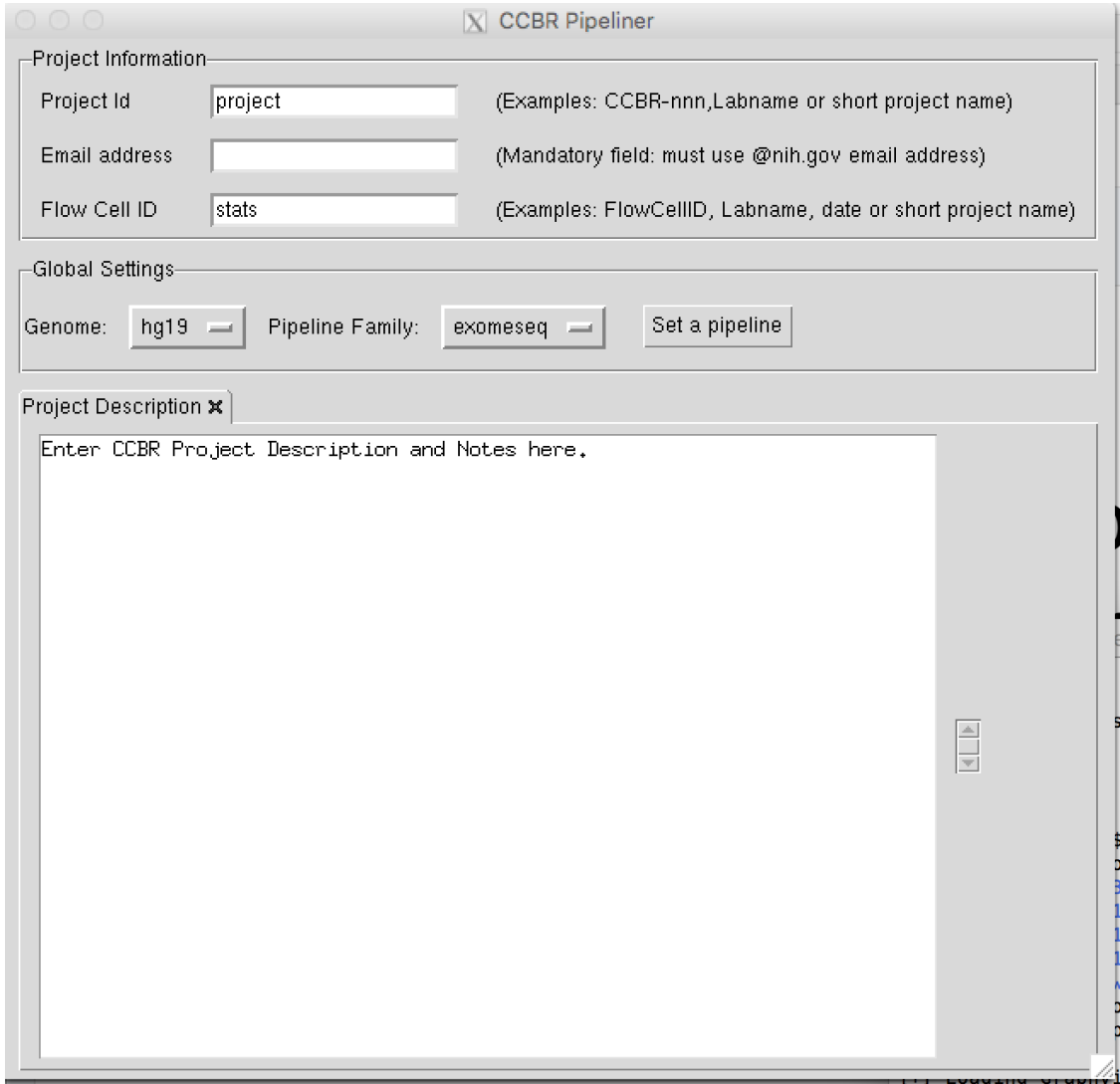
Variant Calling at CCBR

- Multiple Variant Calling CCBR Pipelines
  - Whole genome
  - Whole exome/targeted sequencing
  - RNAseq



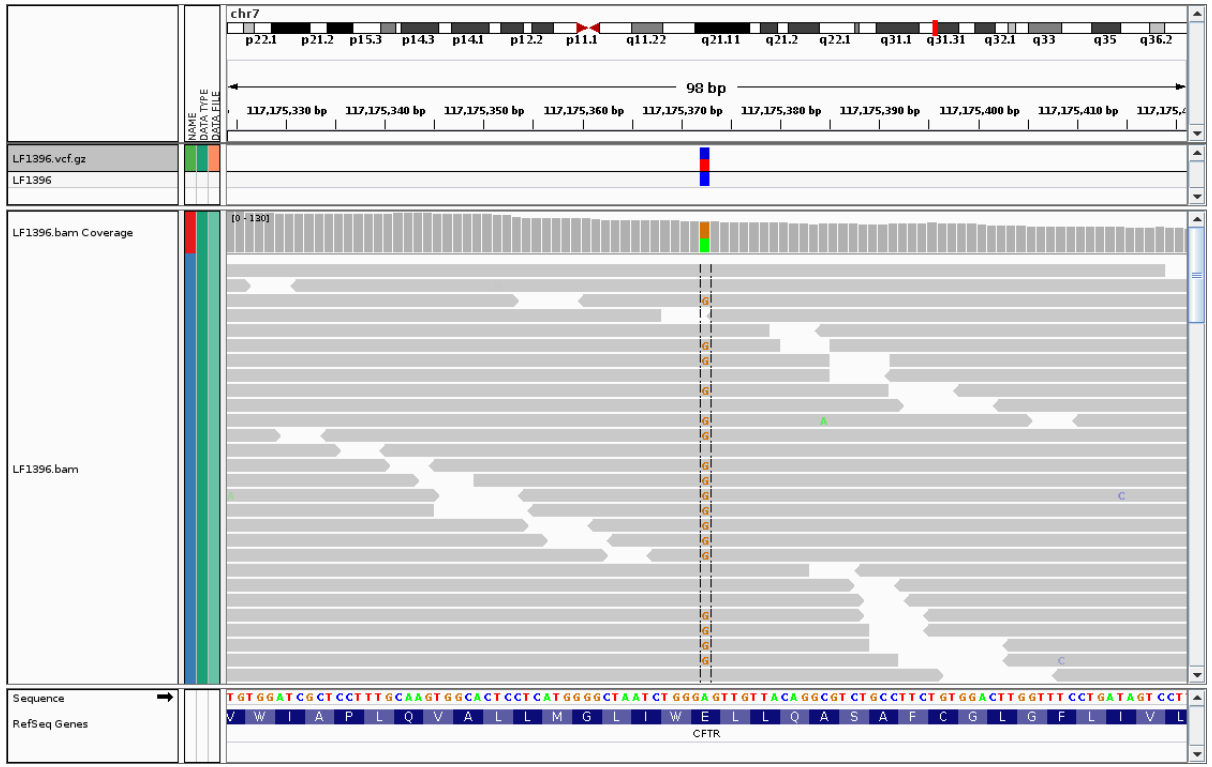
Variant Calling at CCBR

- Multiple Variant Calling CCBR Pipelines
  - Whole genome
  - Whole exome/targeted sequencing
  - RNAseq
- All pipelines (and several others) available through CCBR_Pipeliner app
  - Just need Biowulf account
  - https://github.com/CCBR/Pipeliner
  - module load ccbrpipeliner
  - BTEP Training Feb. 21/22

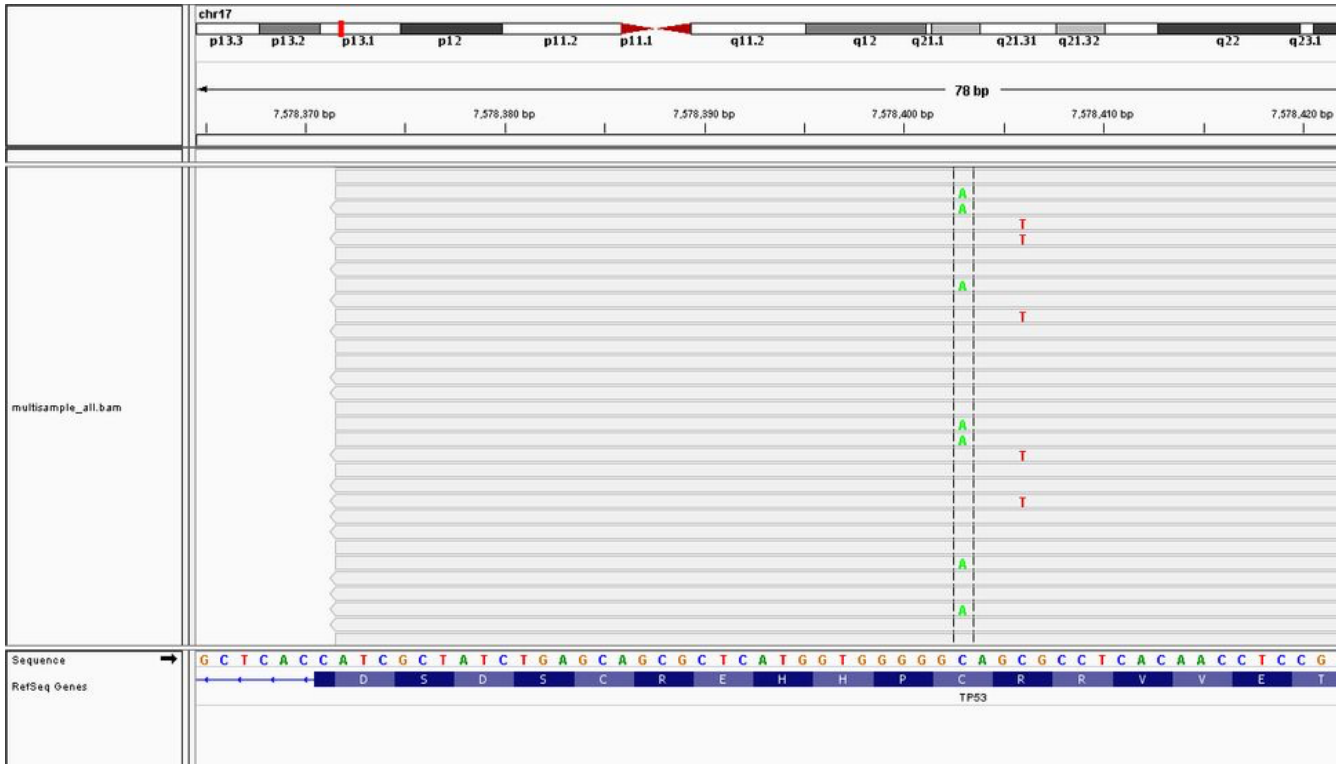


Germline vs Somatic Variant Calling

- Potentially very different allele frequency expectations



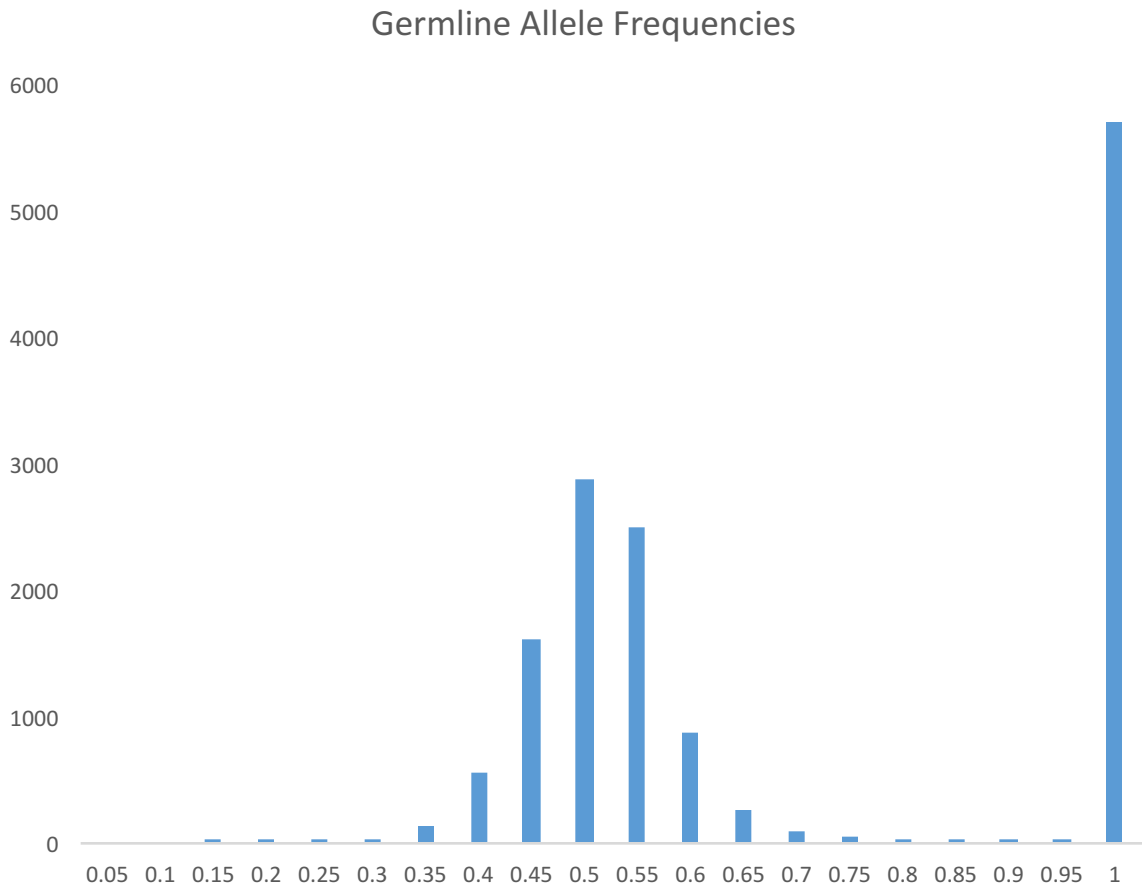
Germline - ~0.5 read proportions

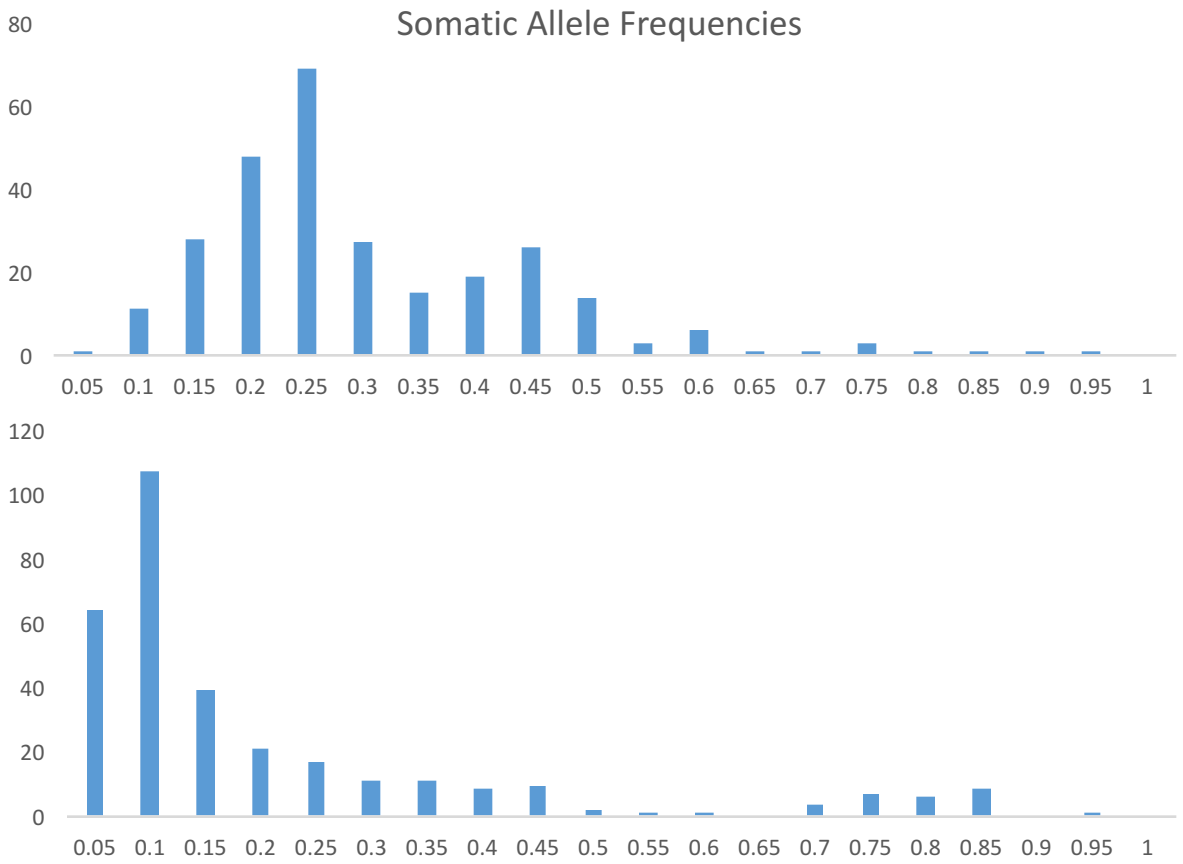


Somatic - ~0.3 read proportions

Germline vs Somatic Variant Calling

- Potentially very different allele frequency expectations



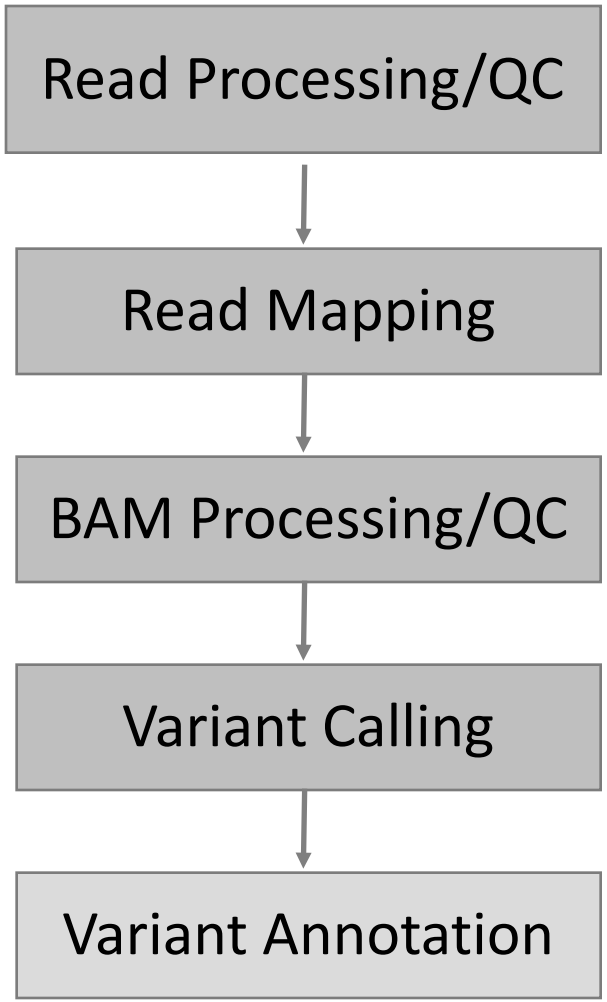


Exome vs Whole Genome Sequencing

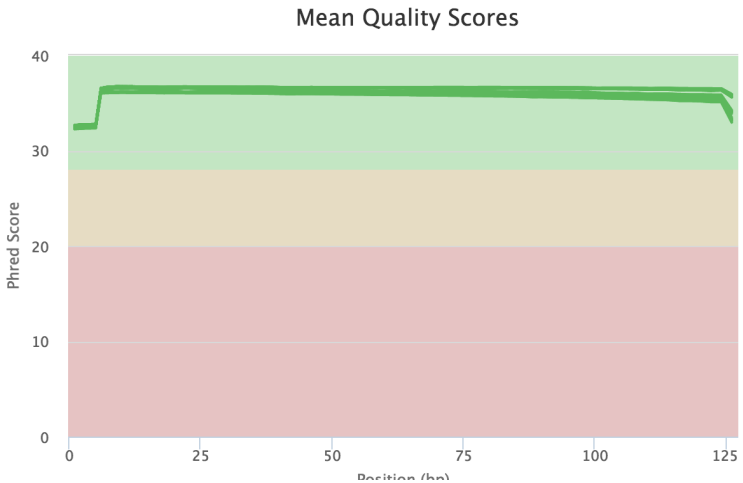
- Exome Sequencing
  - Covers ~5% of genome
  - Allows for high depth targeting
  - Most reasonable option for somatic variant analysis
  - Low-confidence copy number/structural variant calling
- Genome Sequencing
  - Confidently call >85% of reference genome (hg38)
  - Confidently call copy number/structural variant calling due to reduced depth variance
  - Significantly more accurate variant (SNP/INDEL) calling relative to exome
  - Price for WGS comparable to exome for germline-only projects

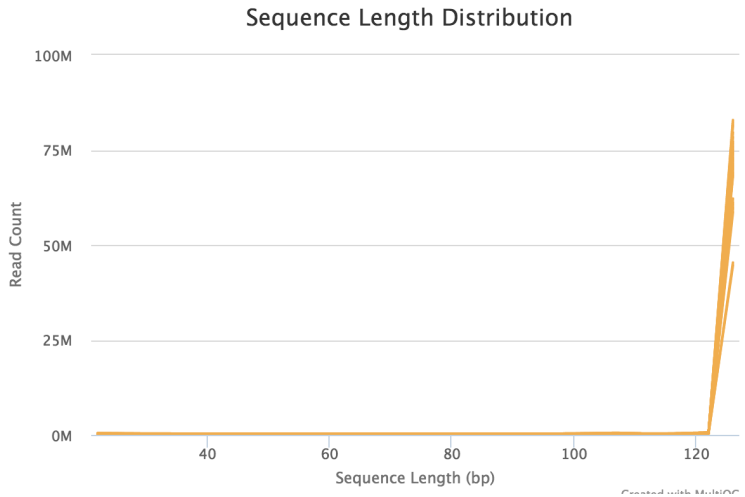
Variant Calling at CCBR

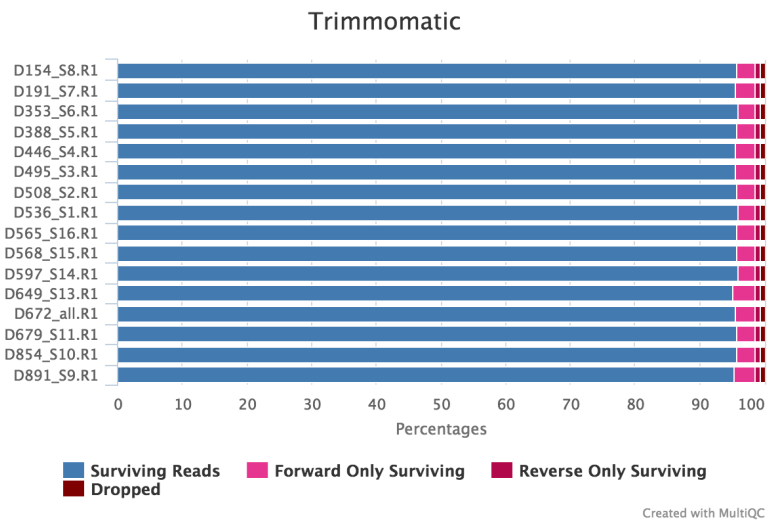
- All variant calling follows the same basic approach

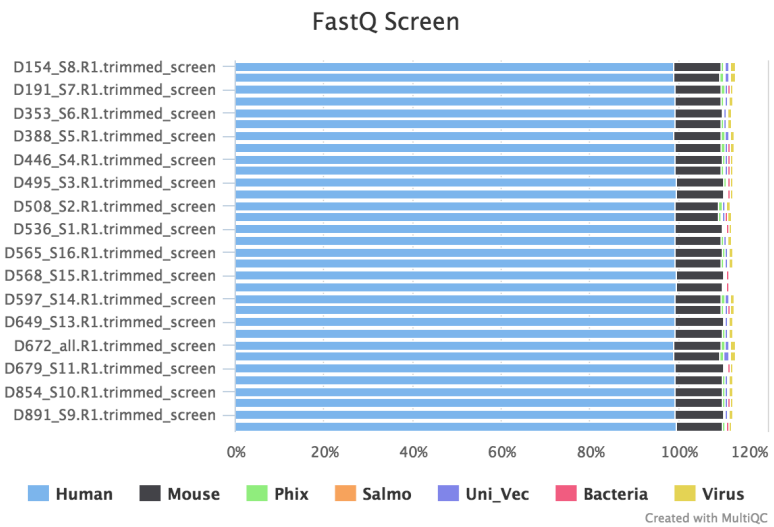


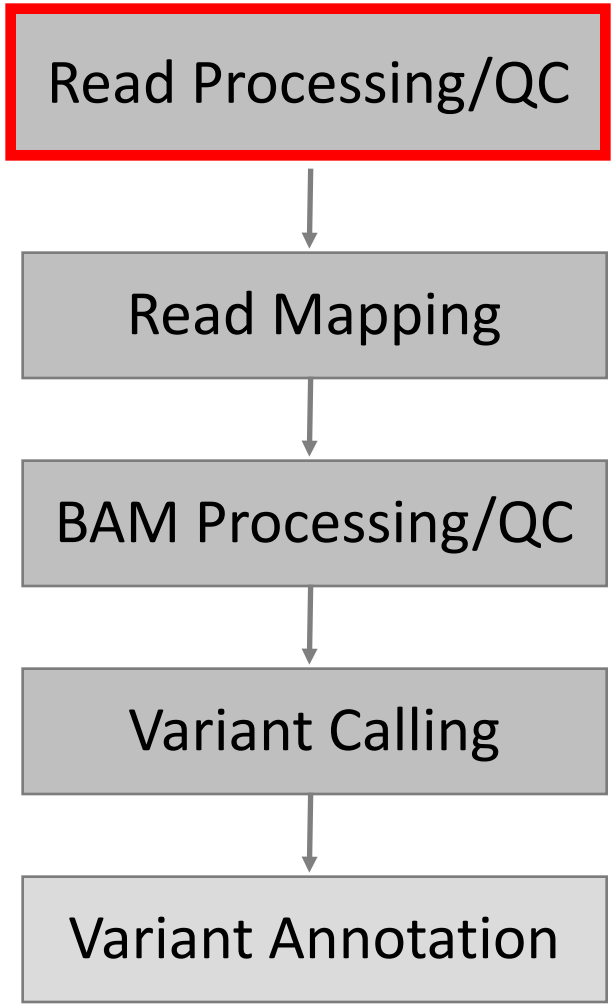
Variant Calling at CCBR



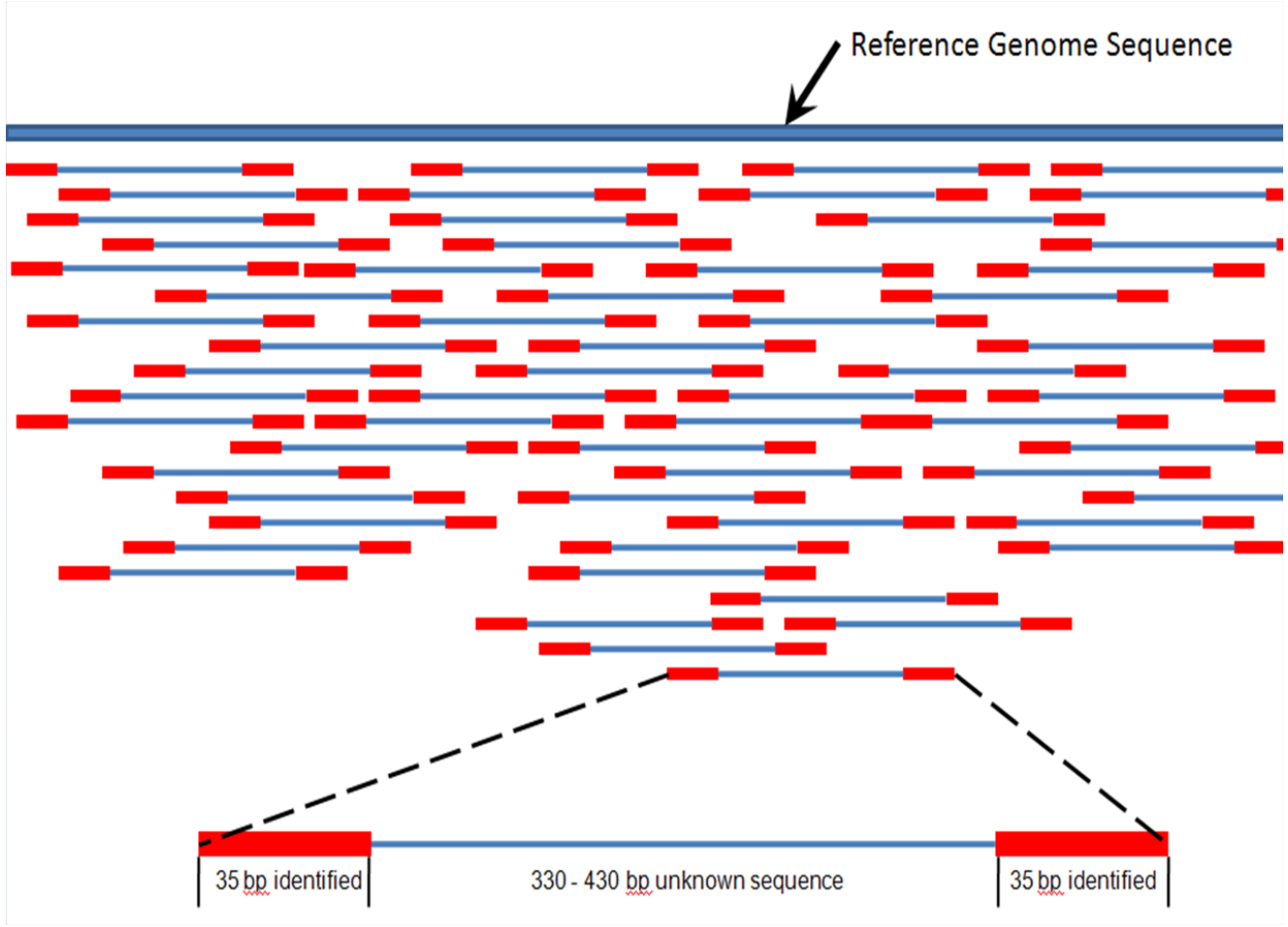


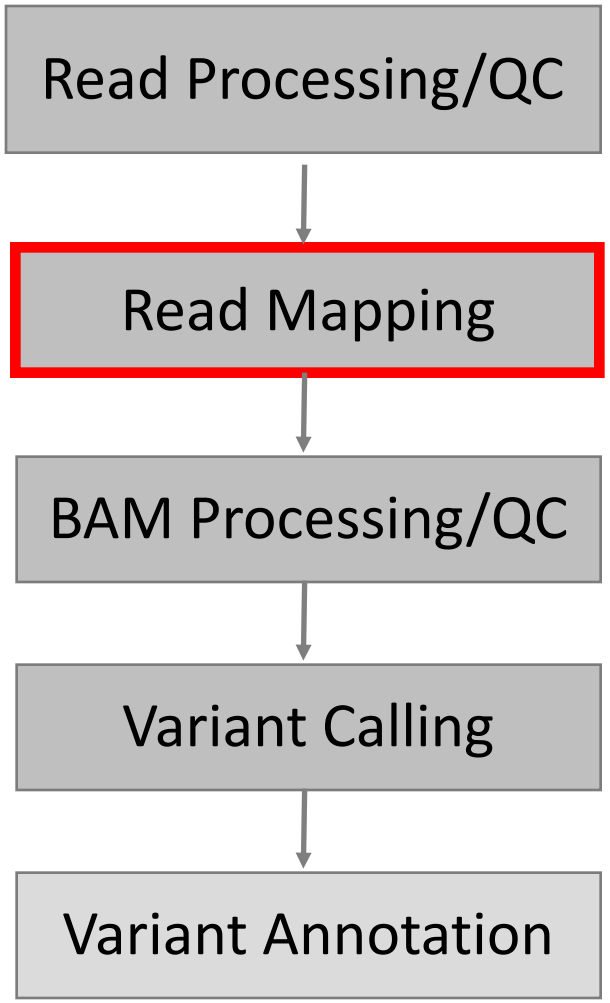






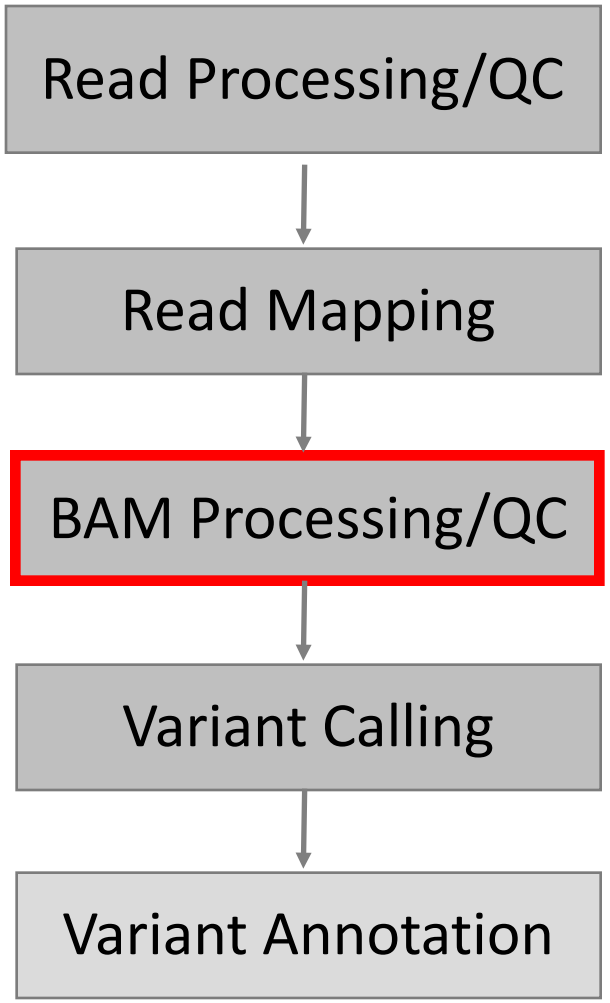
Variant Calling at CCBR





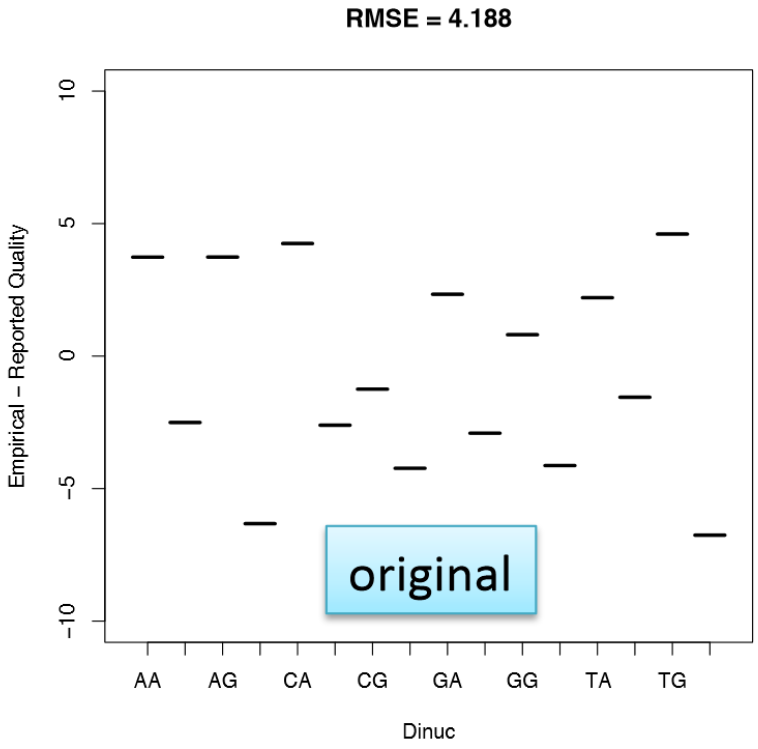
Variant Calling at CCBR

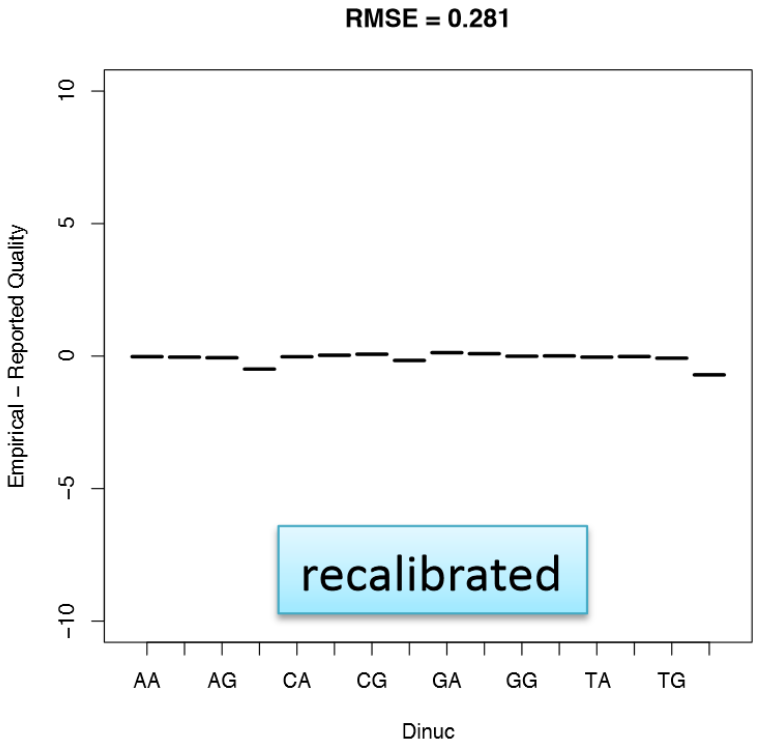
- Indel realignment

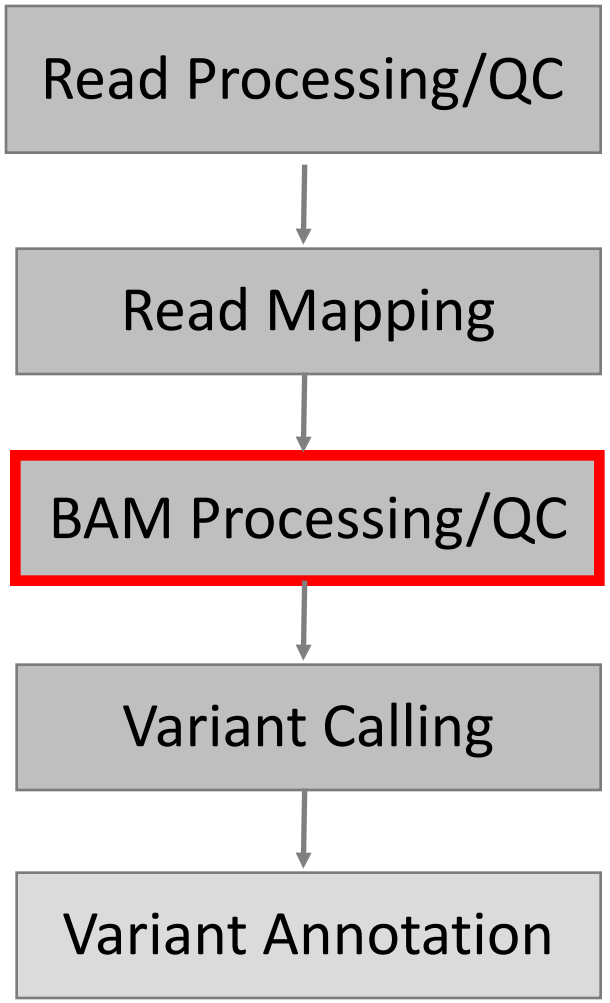


Variant Calling at CCBR

- Multiple sources of quality score bias

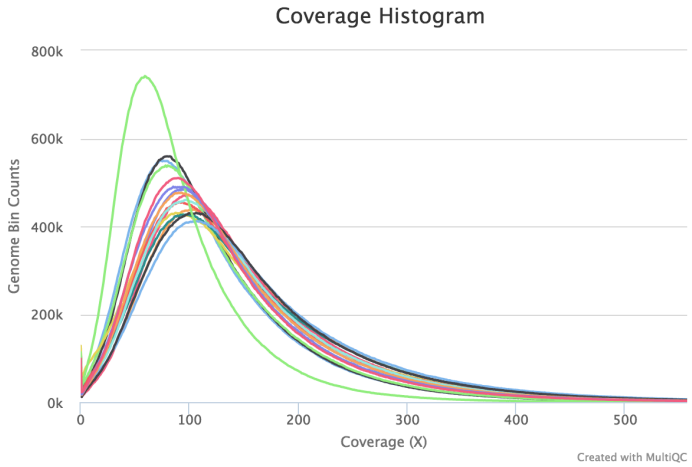


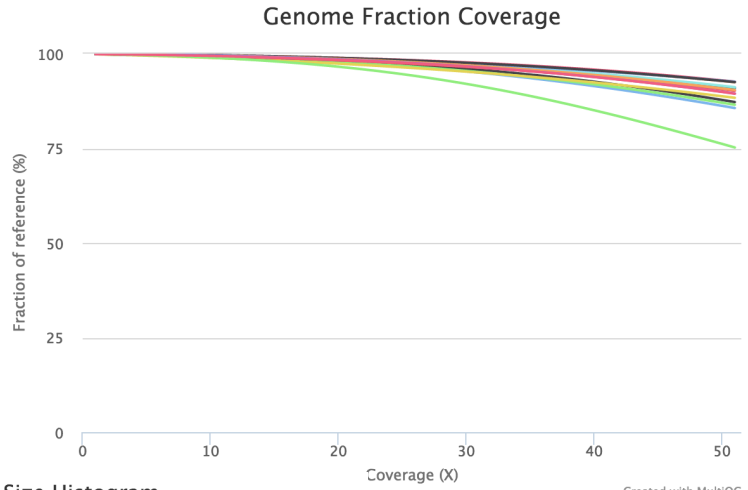


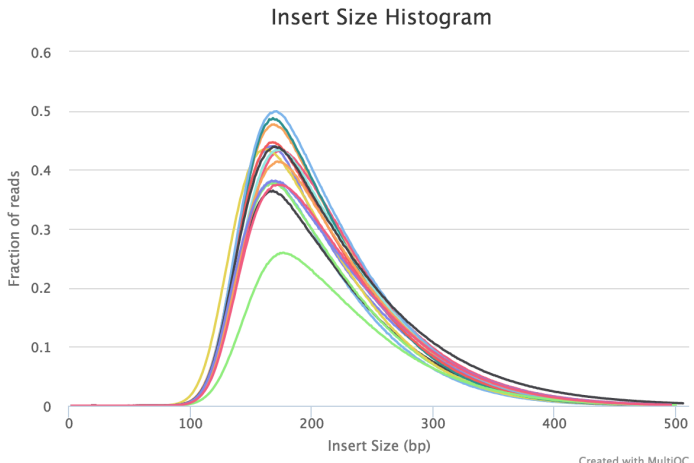


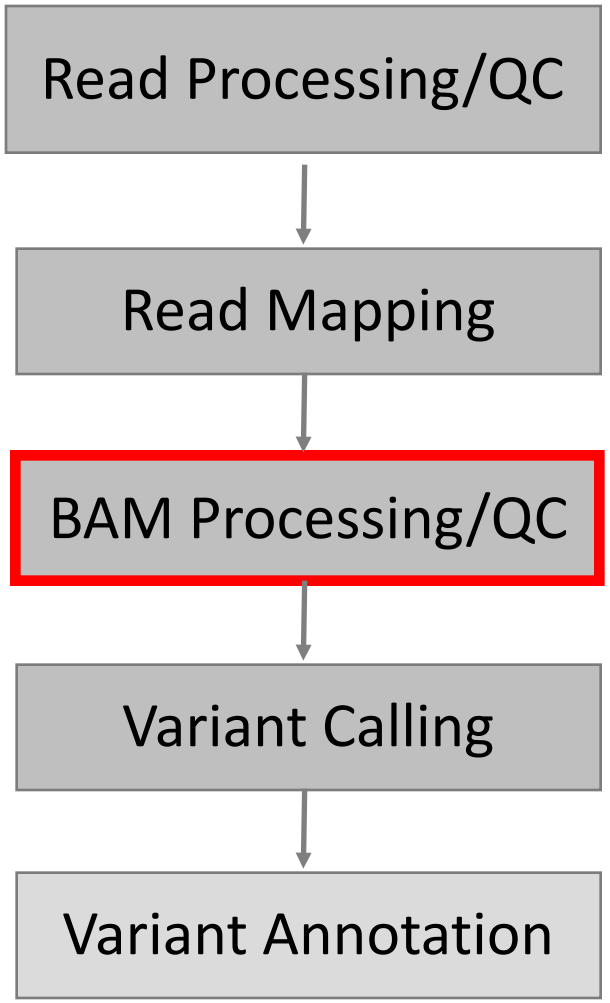
Variant Calling at CCBR

- Alignment QC



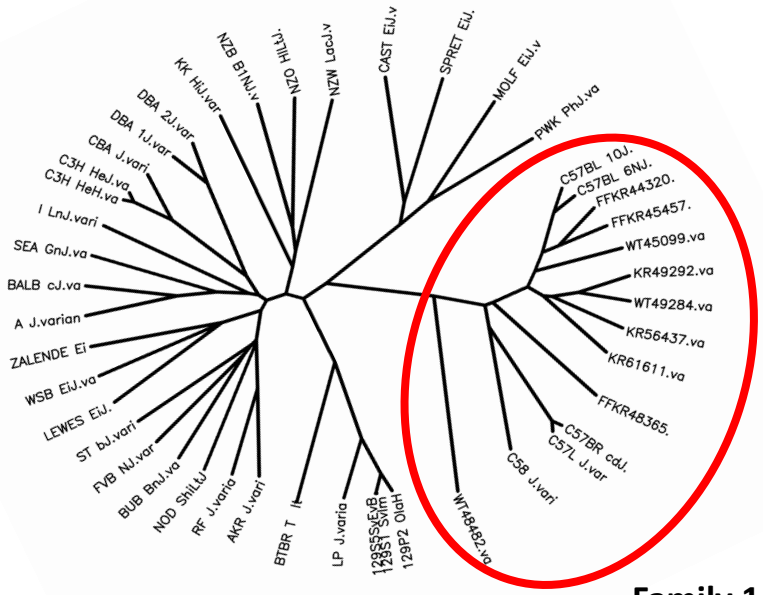


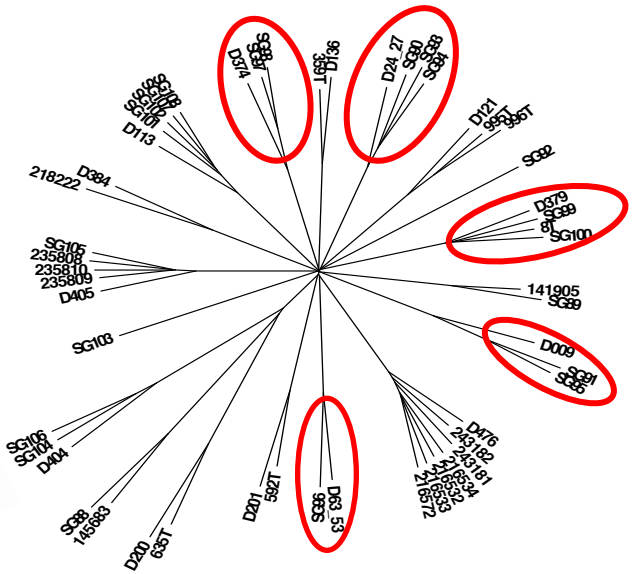


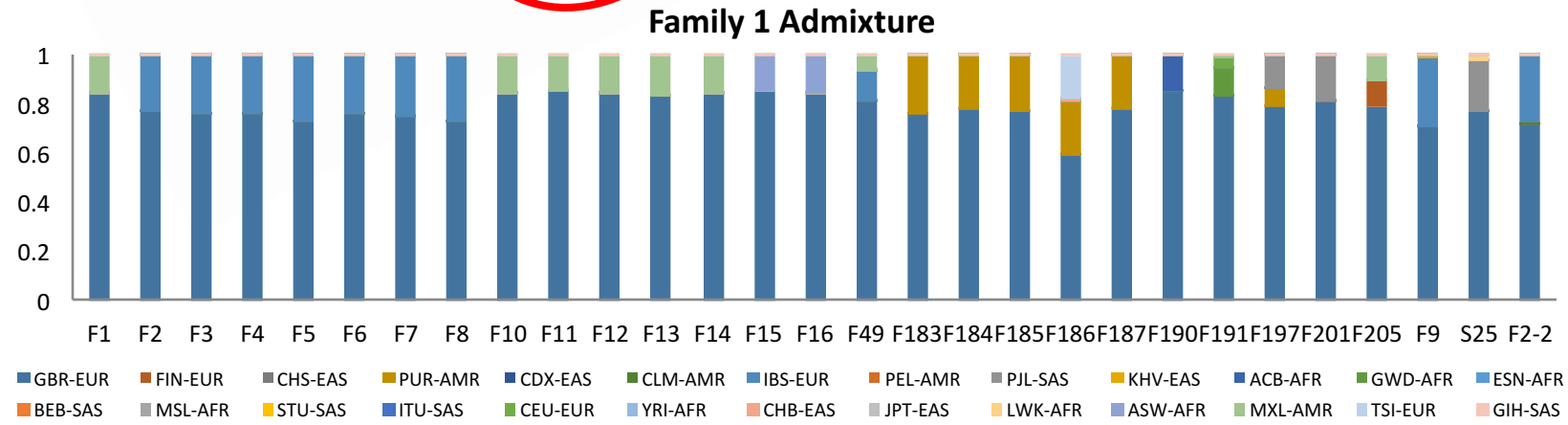


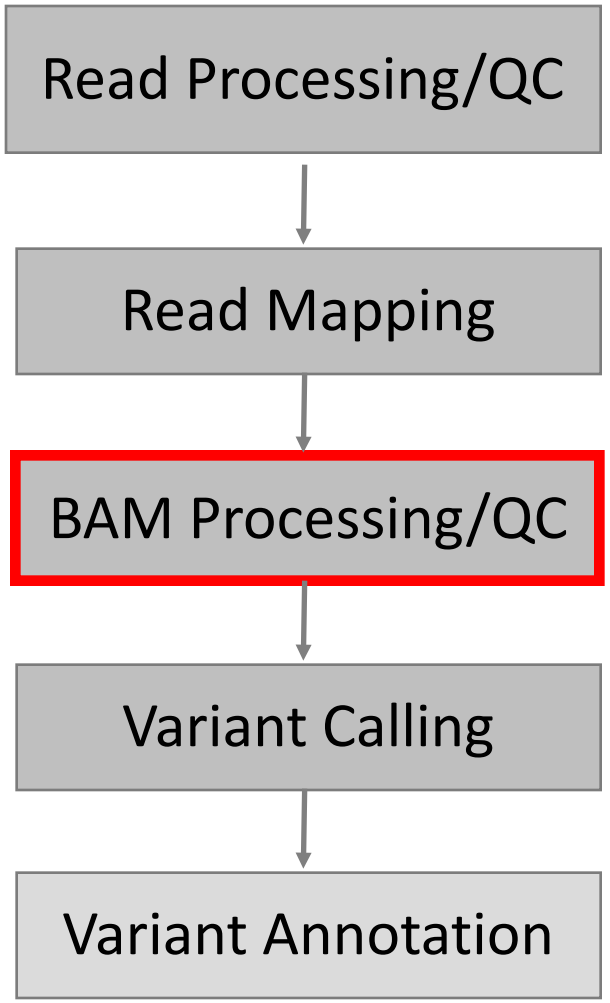
Variant Calling at CCBR

- Additional QC





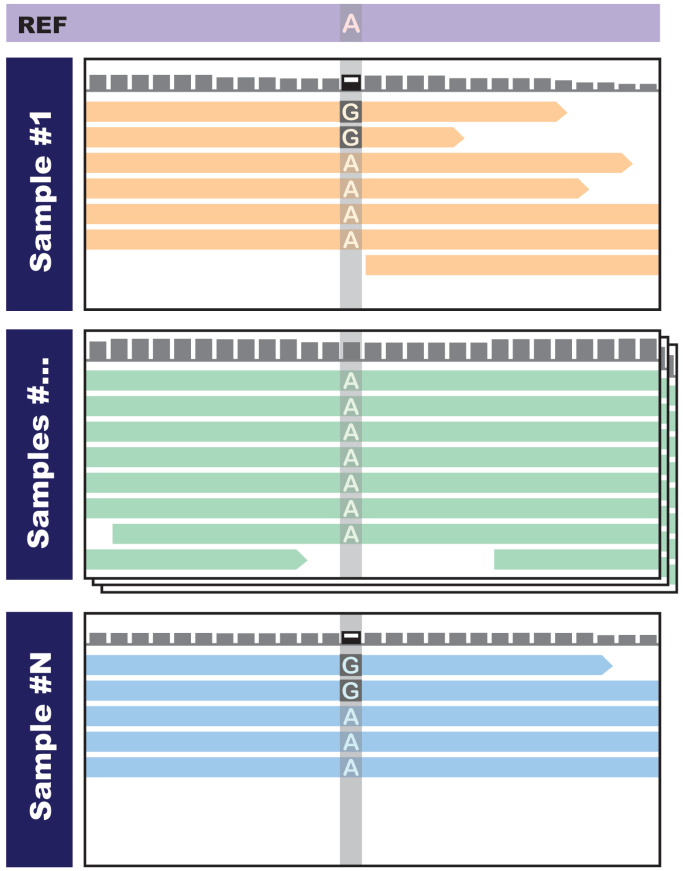


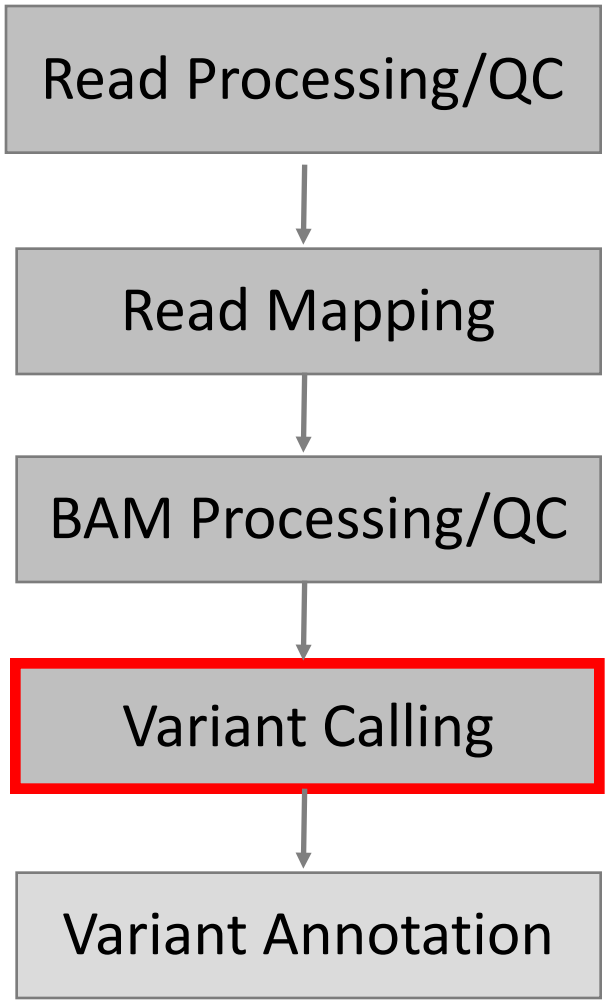


Variant Calling at CCBR

Germline

- Joint genotype with GATK HaplotypeCaller
  - SNPs/short INDELs
- MANTA
  - Large INDELs
  - Translocations
  - Inversions
  - Duplications

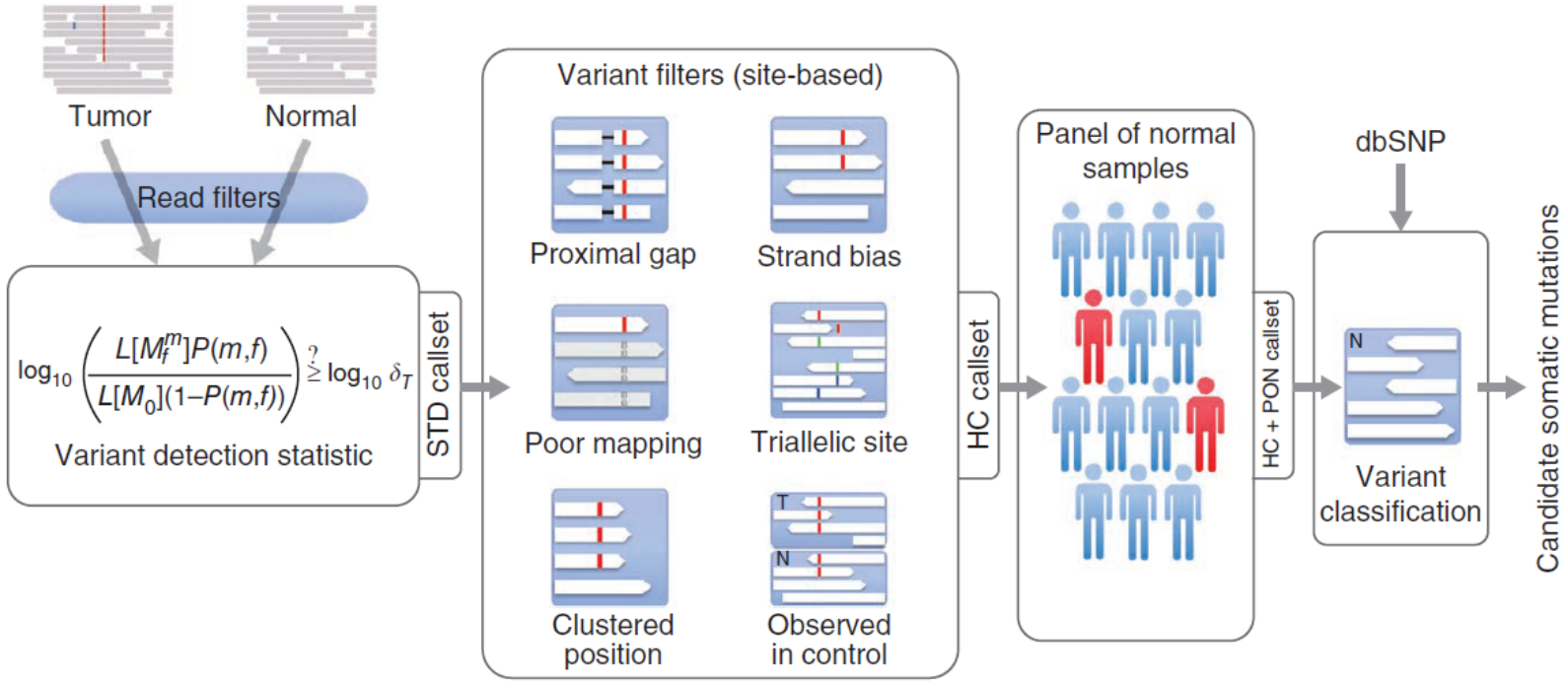


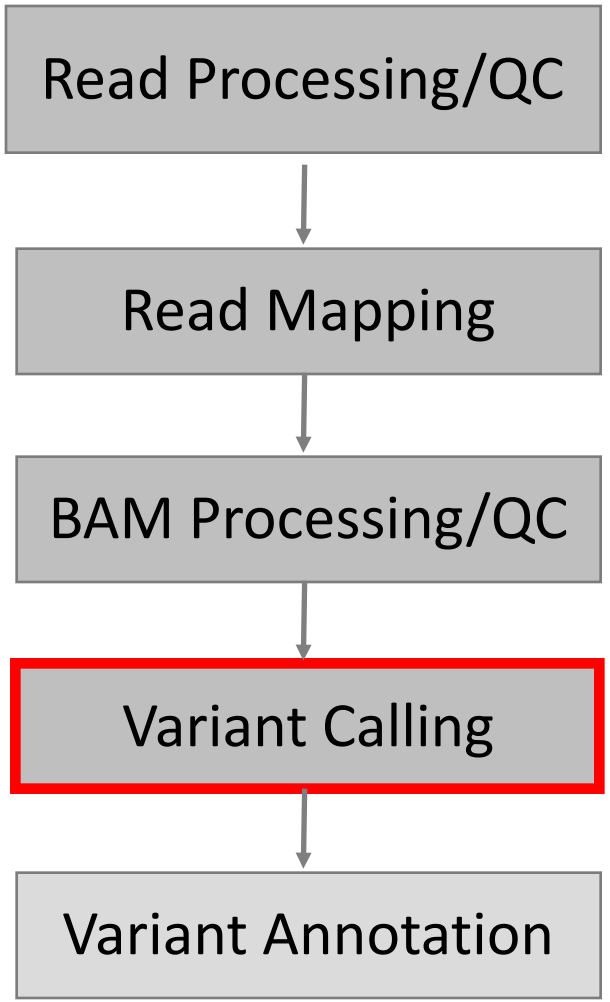


Variant Calling at CCBR

Somatic

- MuTect, MuTect2, Strelka

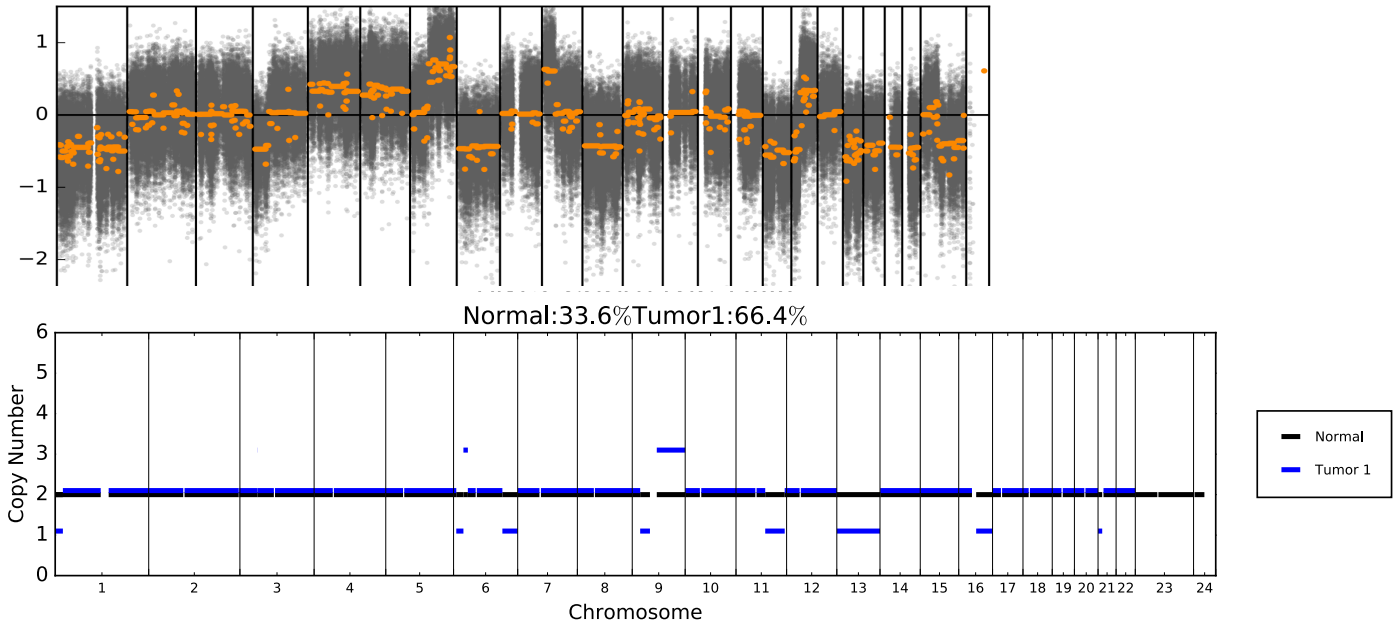


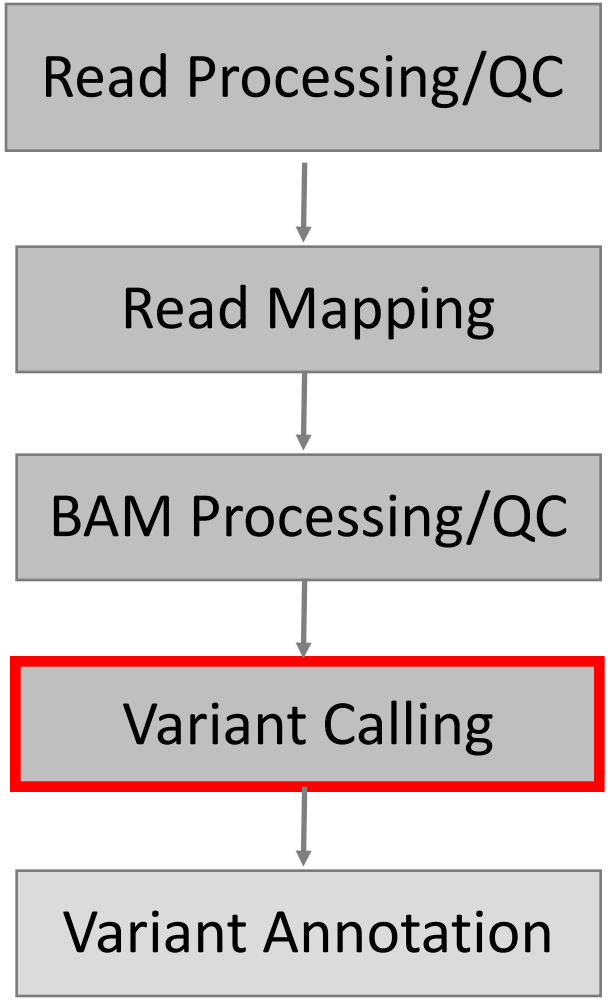


Variant Calling at CCBR

Somatic

- MuTect, MuTect2, Strelka
- Copy number – CNVkit, THetA2

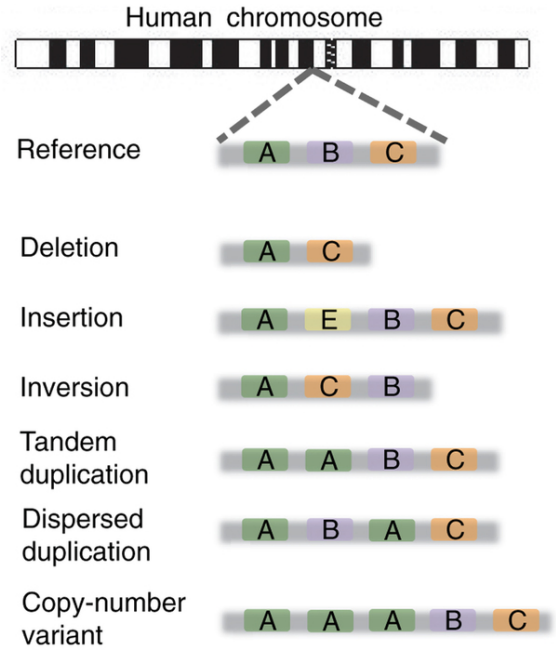


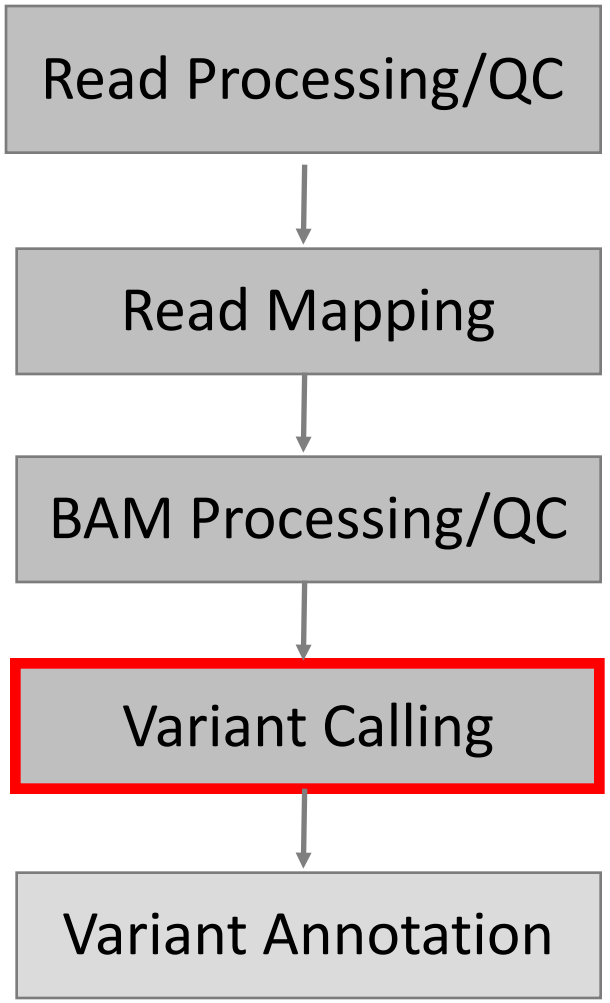


Variant Calling at CCBR

Somatic

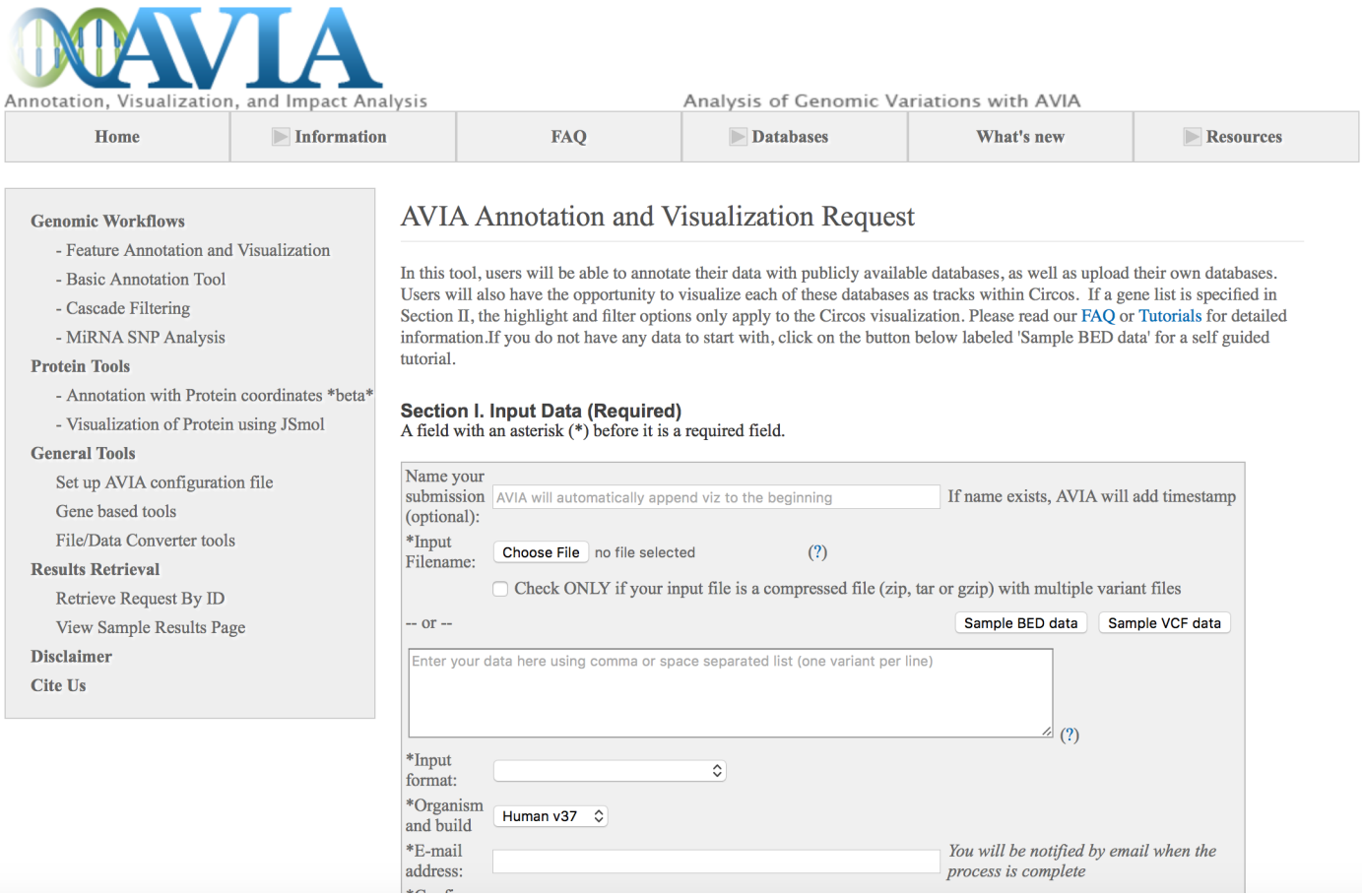
- MuTect, MuTect2, Strelka
- Copy number – CNVkit, THetA2
- Structural Variation
  - MANTA
  - DELLY

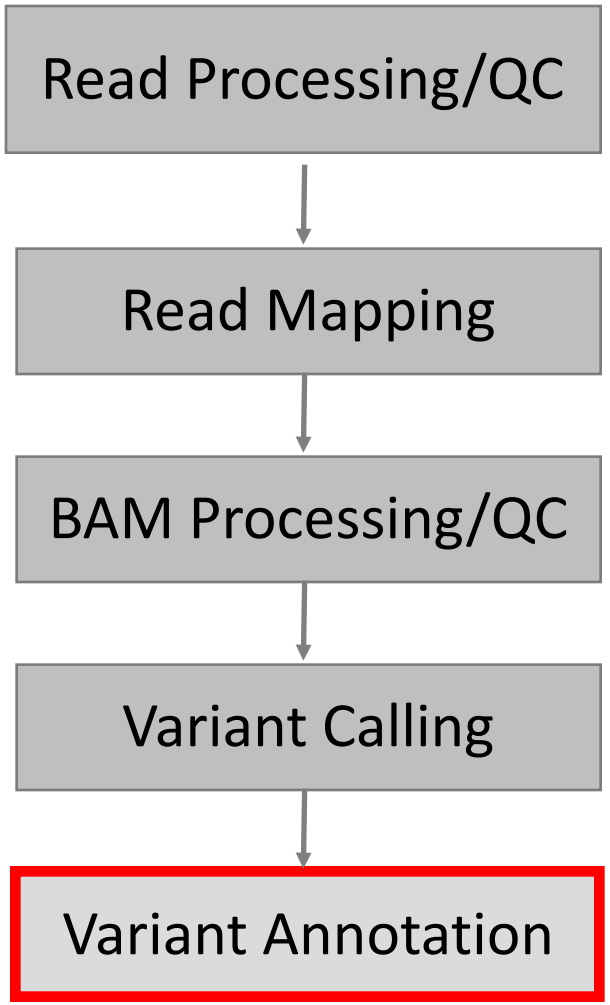




Variant Calling at CCBR

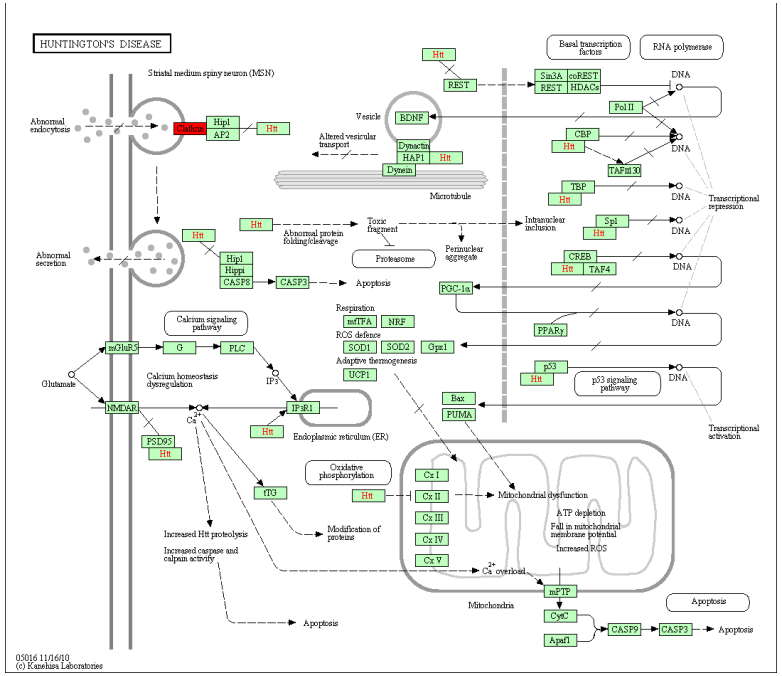
- AVIA! https://avia-abcc.ncifcrf.gov

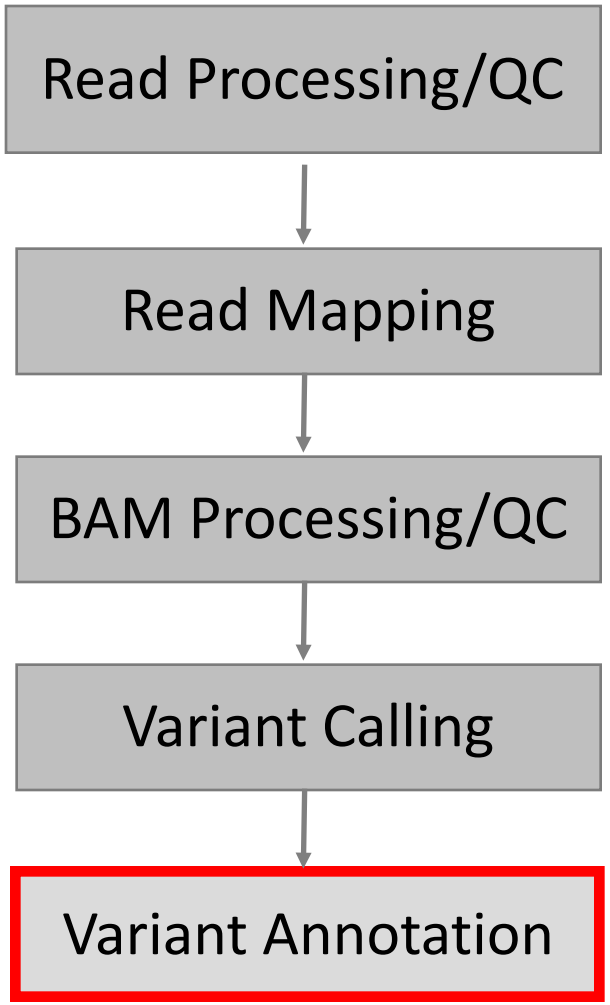




Variant Calling at CCBR

- AVIA! https://avia-abcc.ncifcrf.gov
- SnpEff
- Oncotator -> MutSigCV
- Pathway-level analysis

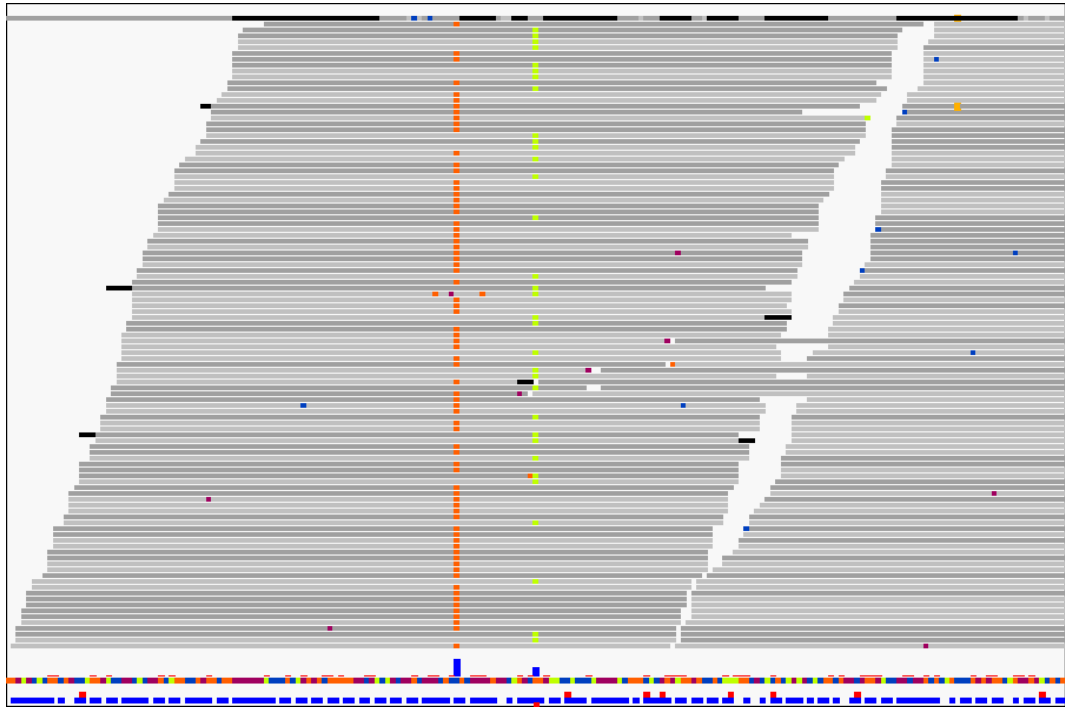




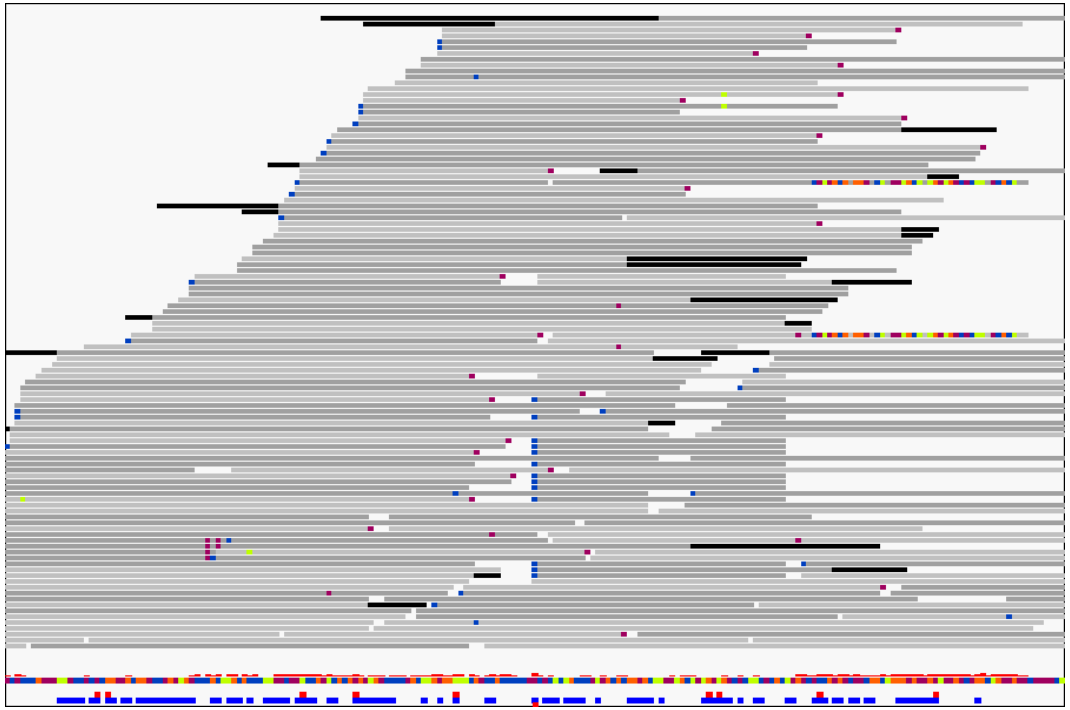
Variant Verification

- ABSOLUTELY CRUCIAL!!
- ALVIEW (https://github.com/NCIP/alview)
  - Internally-developed tool for BAM/SAM visualization (Richard Finney)

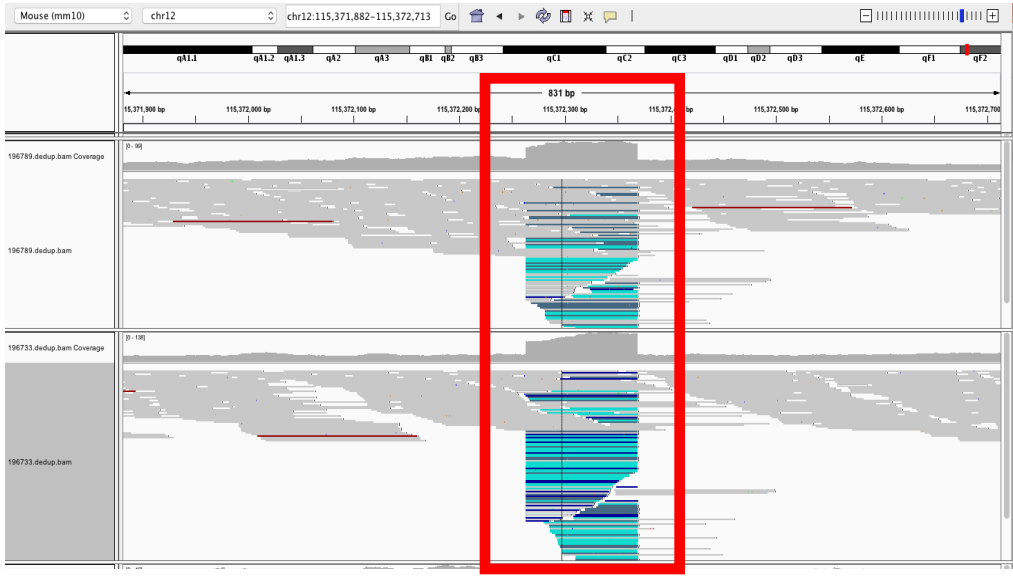
True Positive

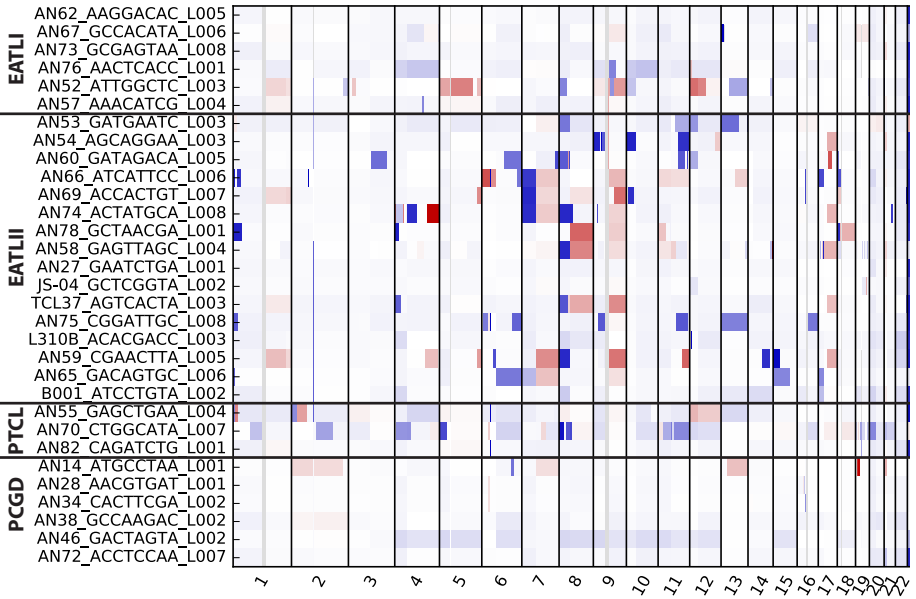


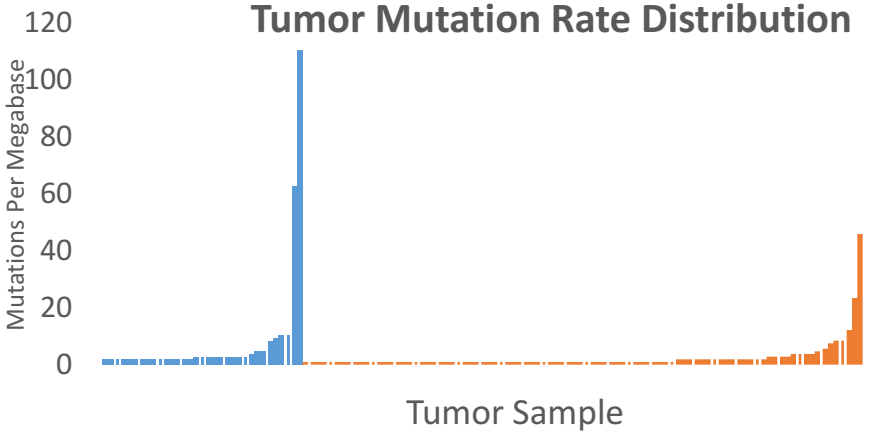
False Positive

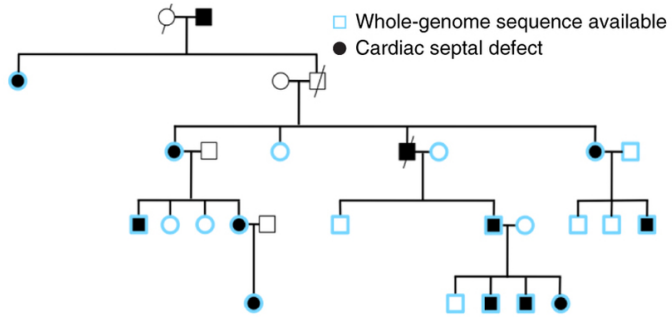


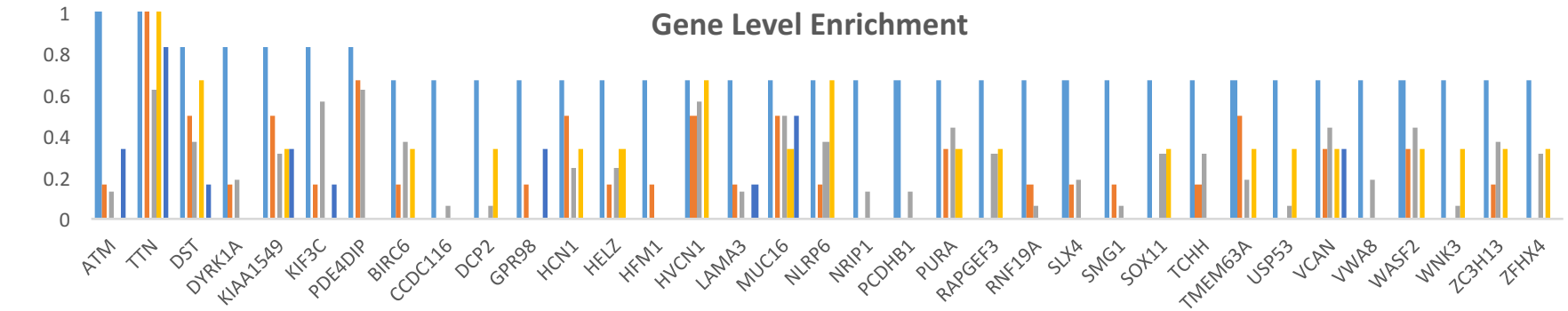
Downstream Analysis

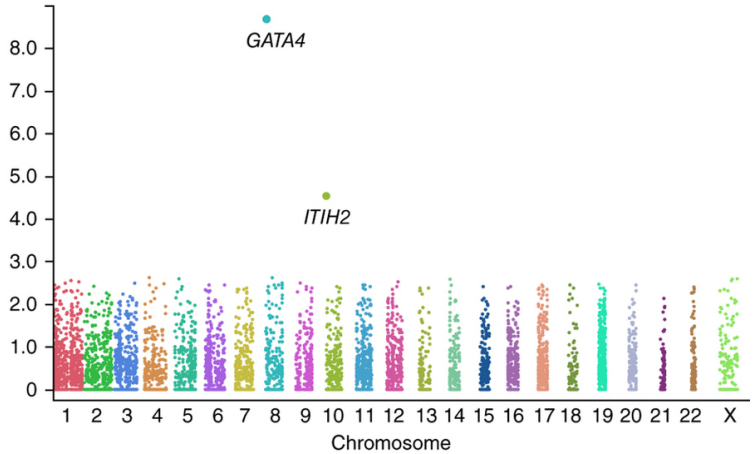








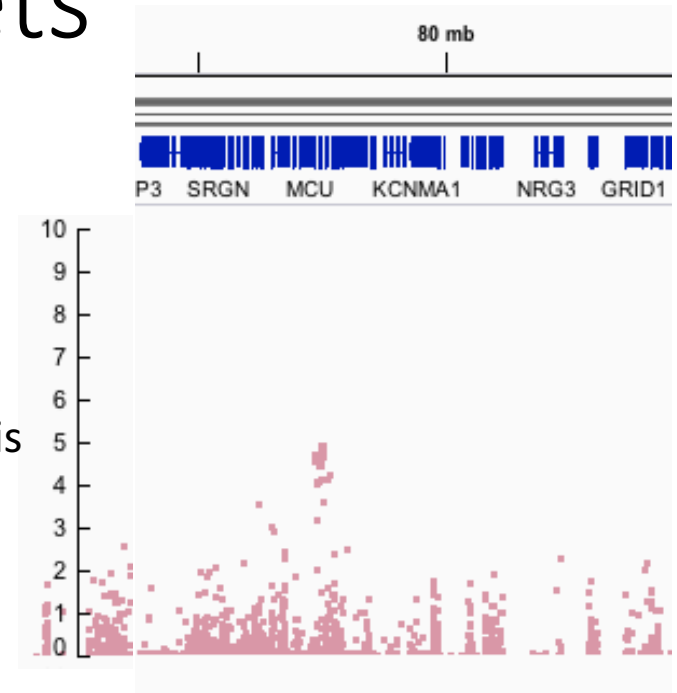


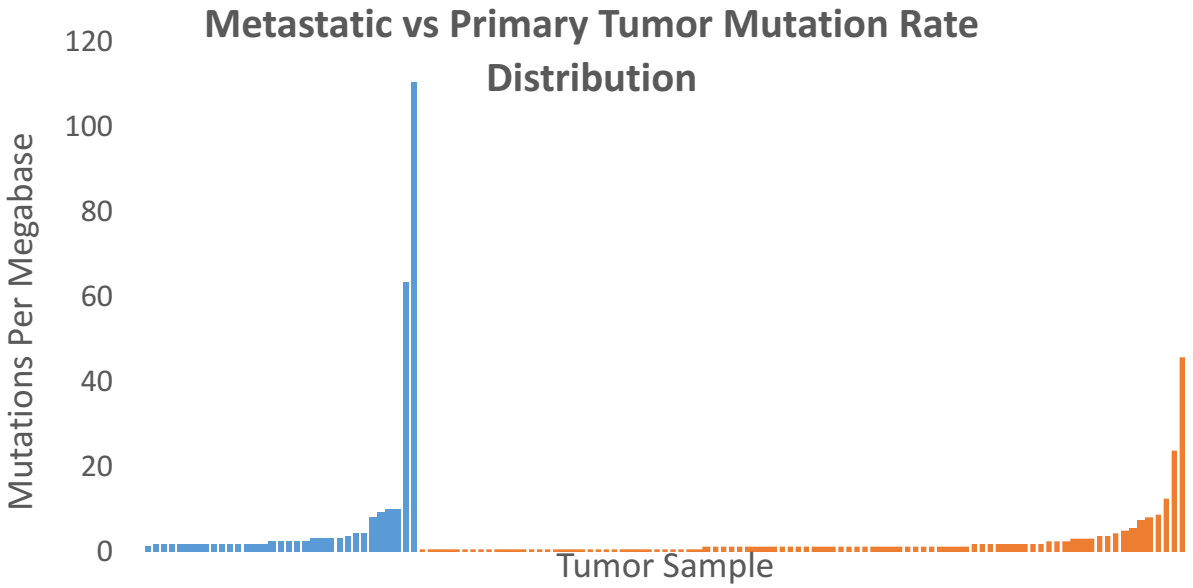


Analysis of Publicly Available Datasets

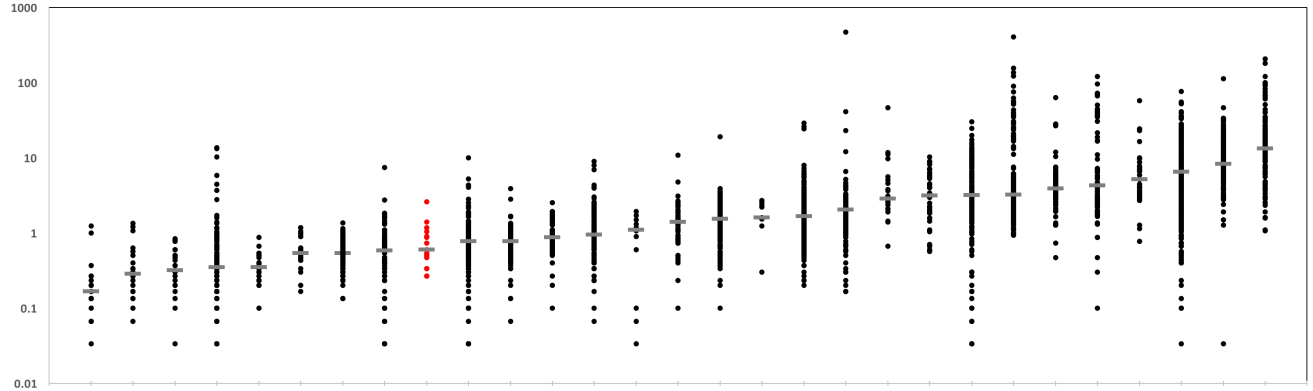
- In-depth analysis of large, public datasets
  - 1k Genomes, ExAC
  - TCGA

TCGA Germline Association Analysis





TCGA Mutational Load Analysis



Somatic Variant Calling – Best Practices

- STRONGLY favor paired tumor/normal design
  - Includes non-human samples
- For non-human samples
  - >=3 control/”germline” samples
- >=100X/50X mean depth for tumor/normal samples
- Significantly higher target depth for FFPE samples
- Tumor purity >50% (ideally, >60%)

Germline Variant Calling – Best Practices

- Whole genome strongly preferred
  - >=30X mean target depth
  - Superior to exome for structural variants, copy number analysis
- Germline exome
  - >=50X mean depth
- For familial/trio analyses, we strongly encourage early consultation
  - Selection of samples for sequencing can be CRUCIAL to maximizing power