

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

Justin Lack

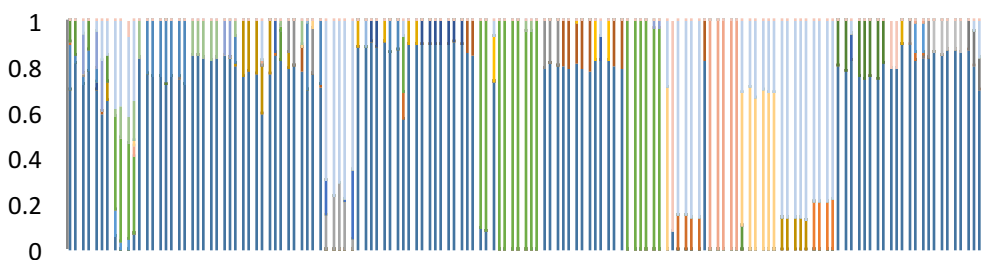
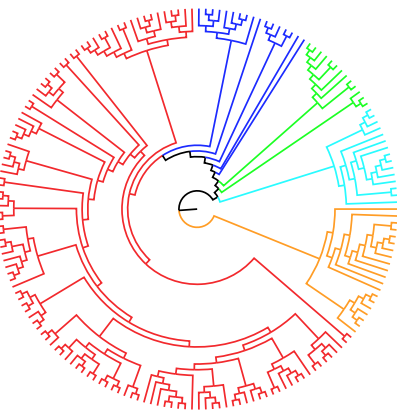
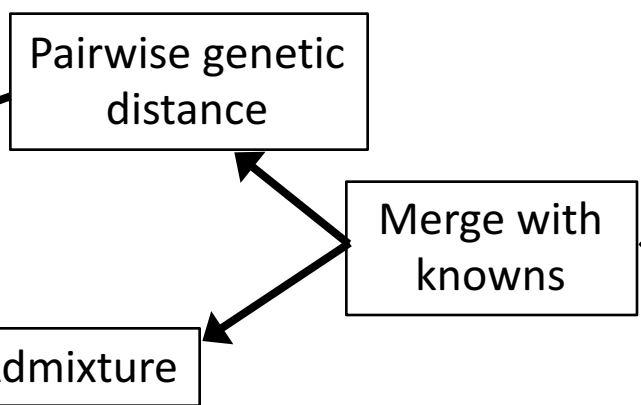
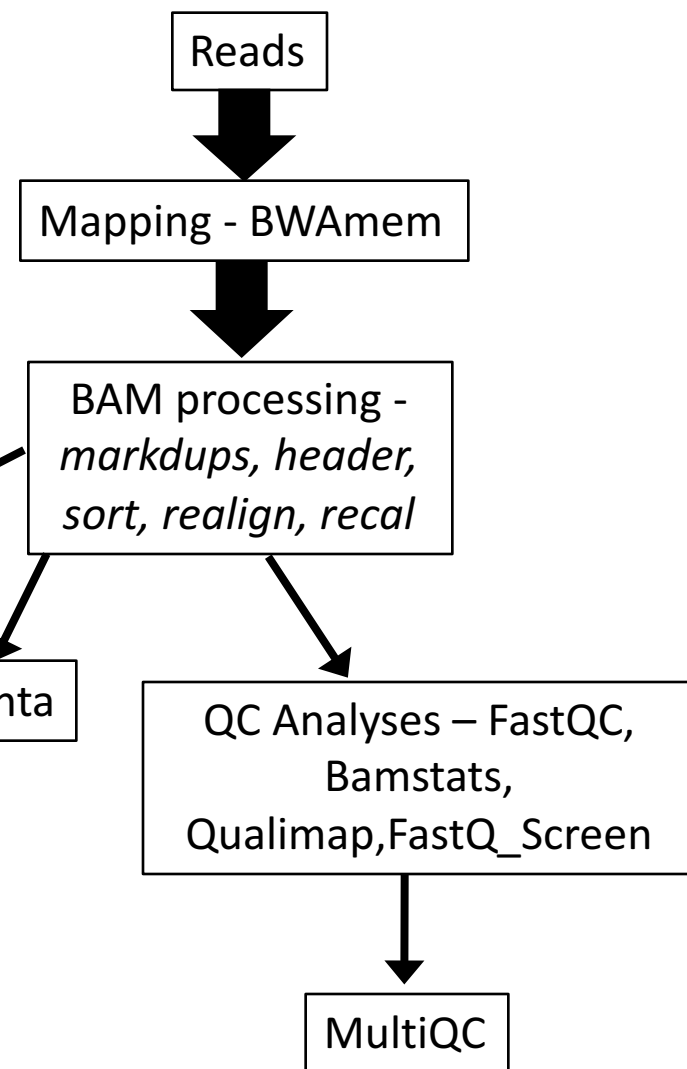
Variant Calling at CCBR

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

Variant Calling at CCBR

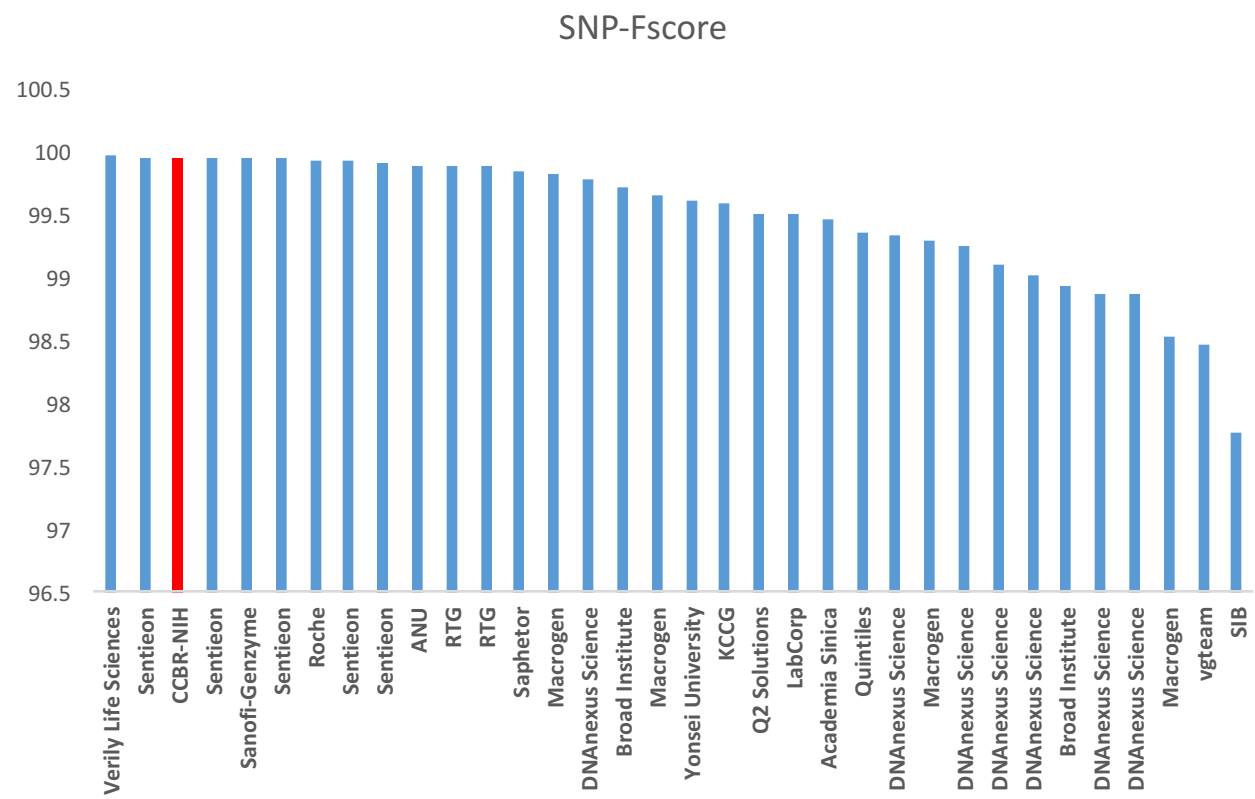
- Multiple Variant Calling CCBR Pipelines
 - Whole genome
 - Whole exome/targeted sequencing

Germline Variant Calling



Variant Calling at CCBR

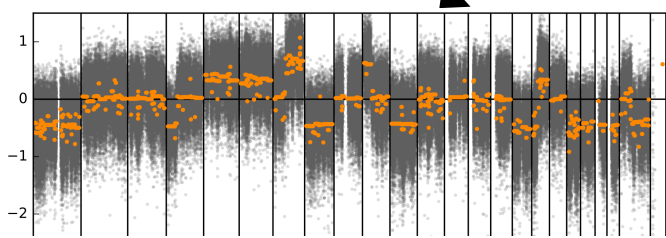
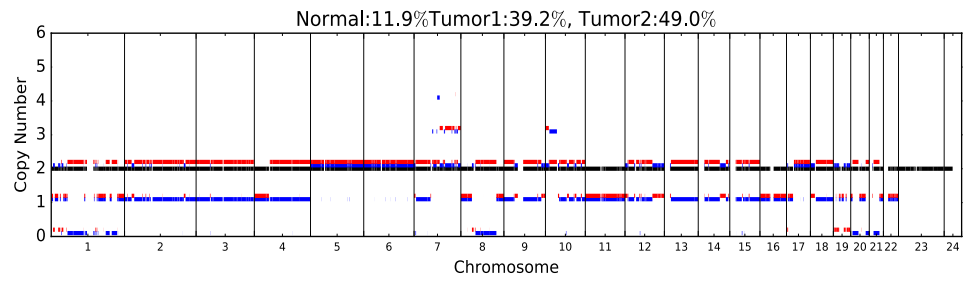
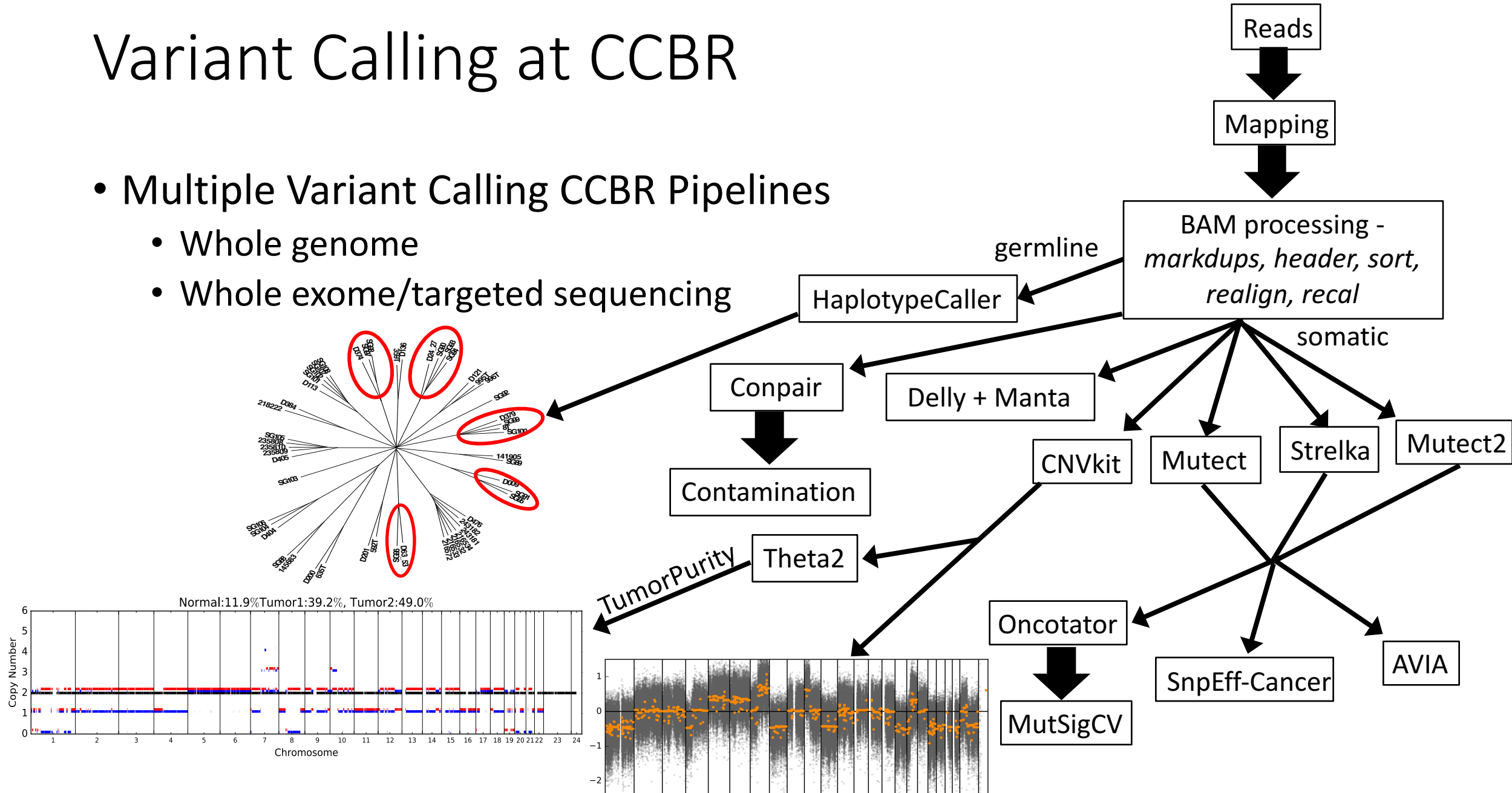
- Multiple Variant Calling CCBR Pipelines
 - Whole genome
 - Whole exome/targeted sequencing
 - Excellent performance in Precision FDA Challenge



Variant Calling at CCBR

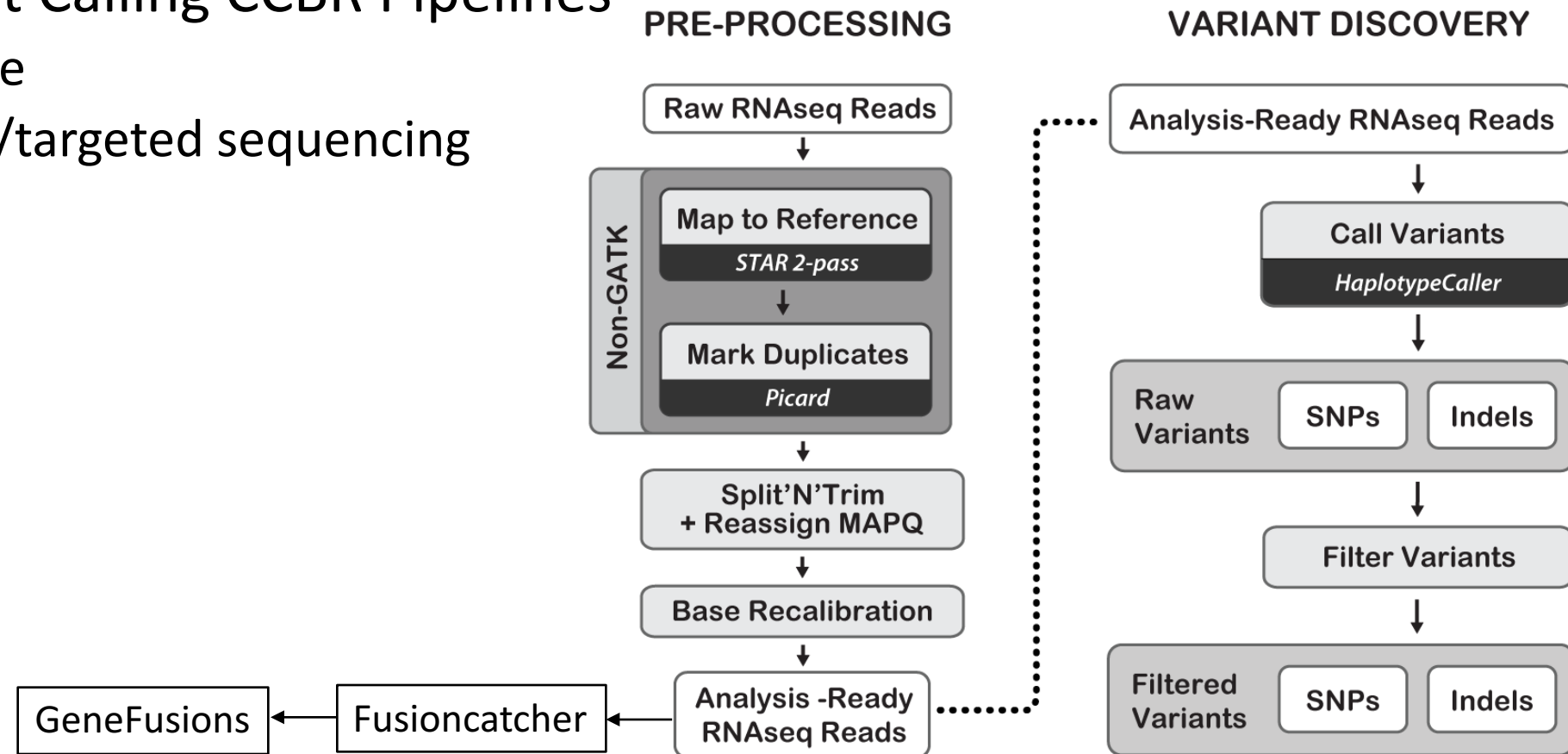
- Multiple Variant Calling CCBR Pipelines
 - Whole genome
 - Whole exome/targeted sequencing

Somatic Variant Calling



Variant Calling at CCBR

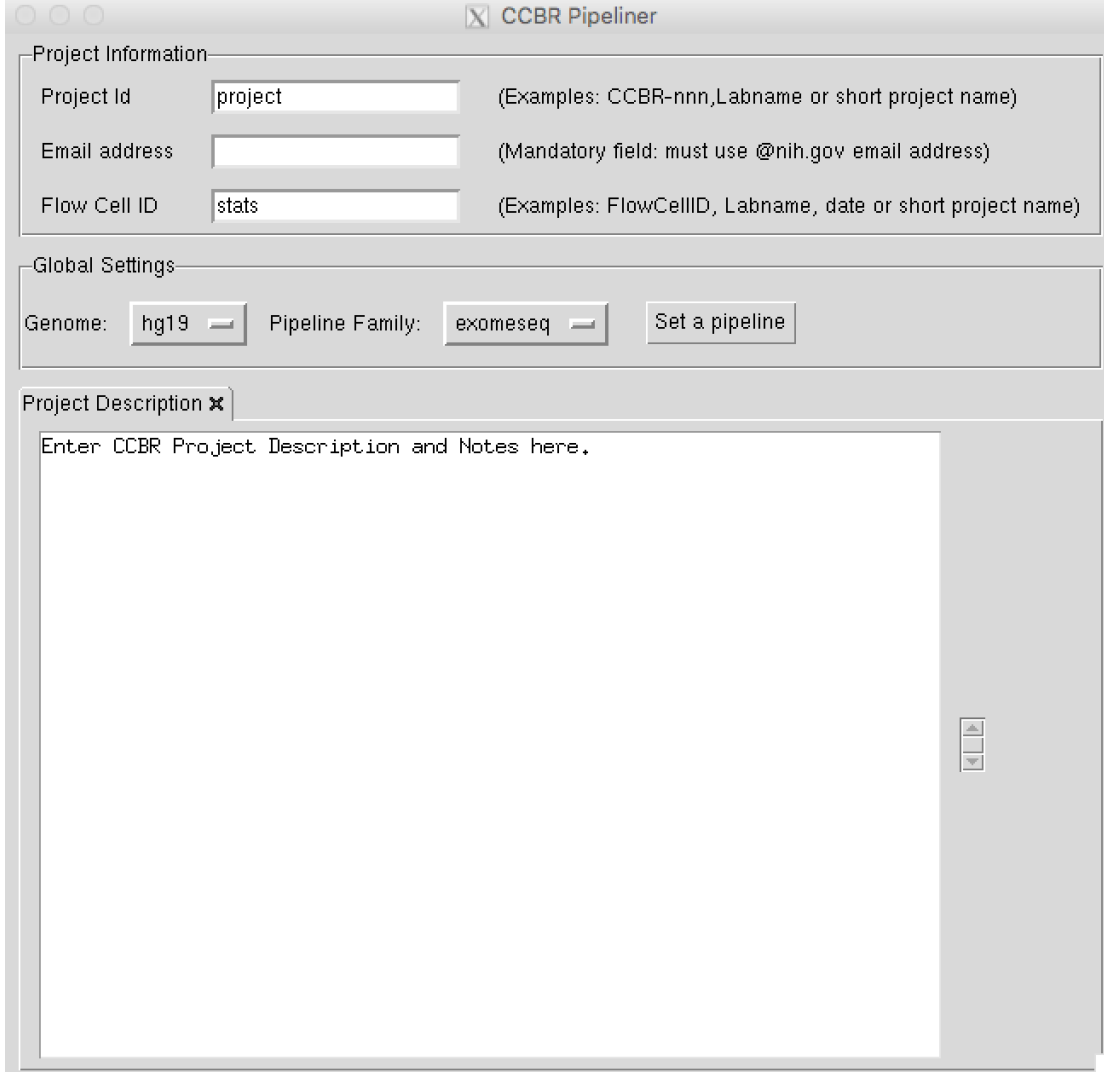
- Multiple Variant Calling CCBR Pipelines
 - Whole genome
 - Whole exome/targeted sequencing
 - RNAseq



Best Practices for Germline SNPs and Indels in 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_Pipelinier app
 - Just need Biowulf account
 - <https://github.com/CCBR/Pipelinier>
 - module load ccbripipelinier
 - BTEP Training Feb. 21/22

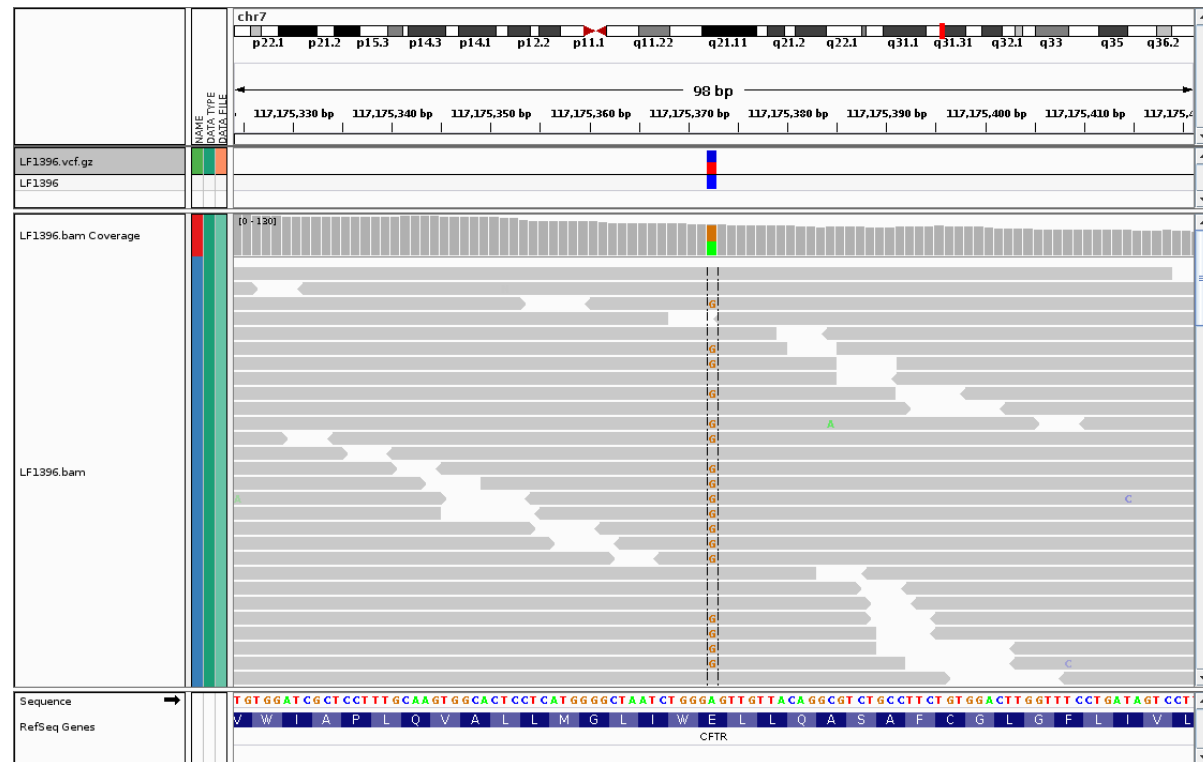


The screenshot shows the CCBR Pipelinier application window. It has a title bar with the text "CCBR Pipelinier". The interface is divided into three main sections:

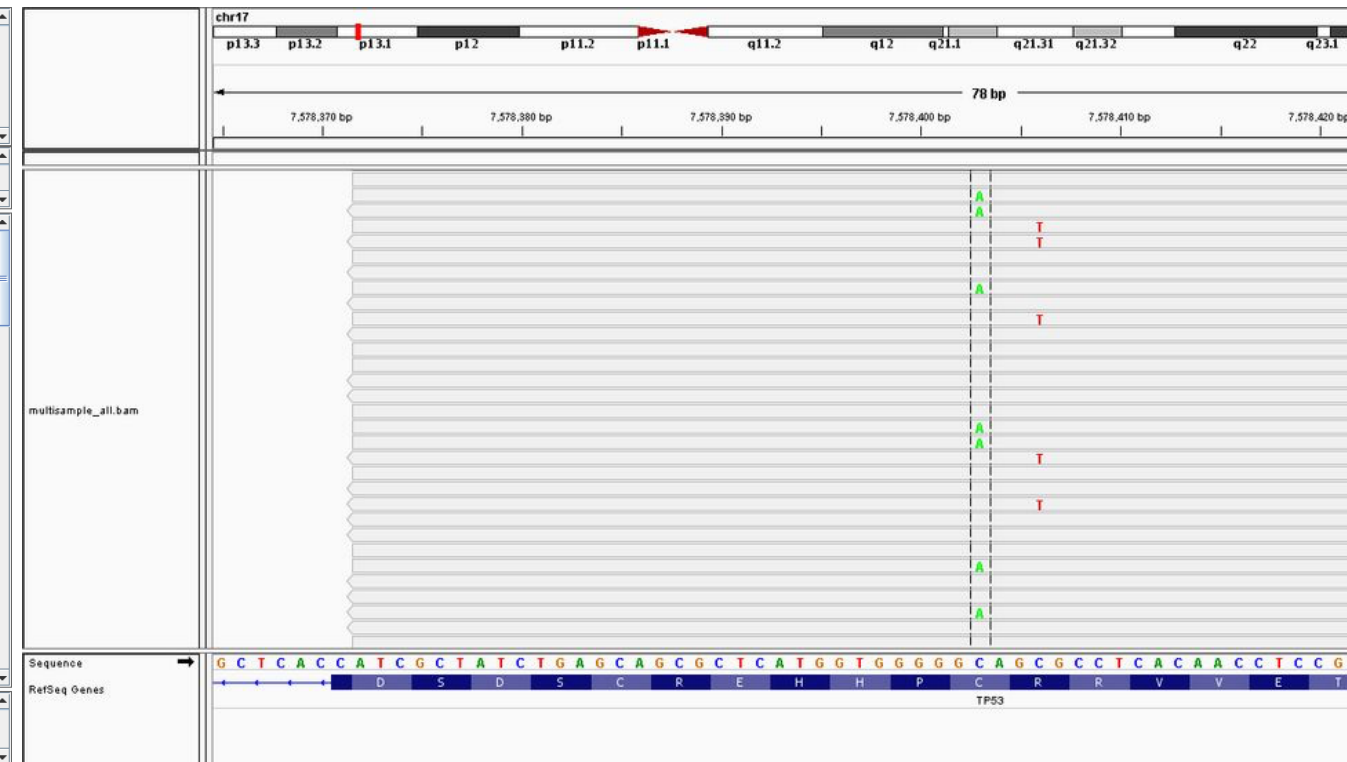
- Project Information:** Contains three input fields with labels and examples:
 - Project Id:** Input field with "project" entered. Example: "CCBR-*nnn*, Labname or short project name".
 - Email address:** Empty input field. Example: "(Mandatory field: must use @nih.gov email address)".
 - Flow Cell ID:** Input field with "stats" entered. Example: "FlowCellID, Labname, date or short project name".
- Global Settings:** Contains two dropdown menus and a button:
 - Genome:** Dropdown menu with "hg19" selected.
 - Pipeline Family:** Dropdown menu with "exomeseq" selected.
 - Set a pipeline:** A button to the right of the Pipeline Family dropdown.
- Project Description:** A large text area with a tab labeled "Project Description ✕". The text inside says "Enter CCBR Project Description and Notes here." and is currently empty.

Germline vs Somatic Variant Calling

- Potentially very different allele frequency expectations



Germline - ~0.5 read proportions

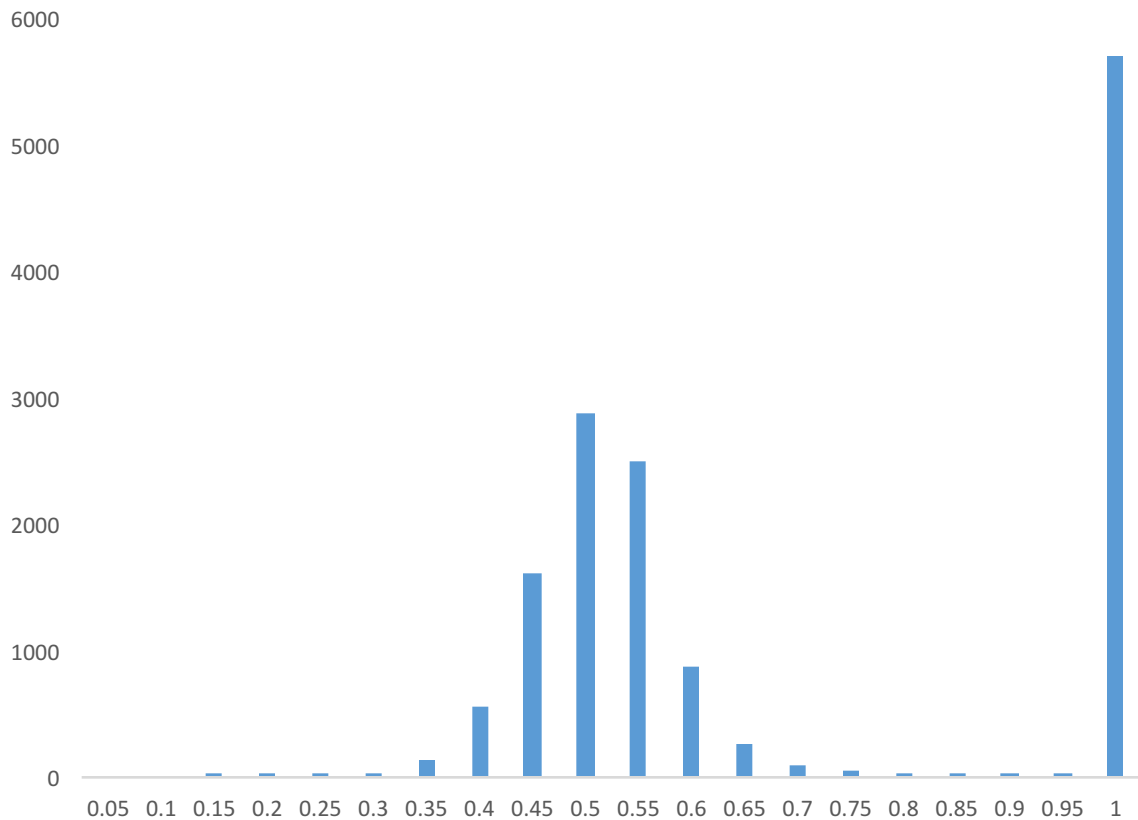


Somatic - ~0.3 read proportions

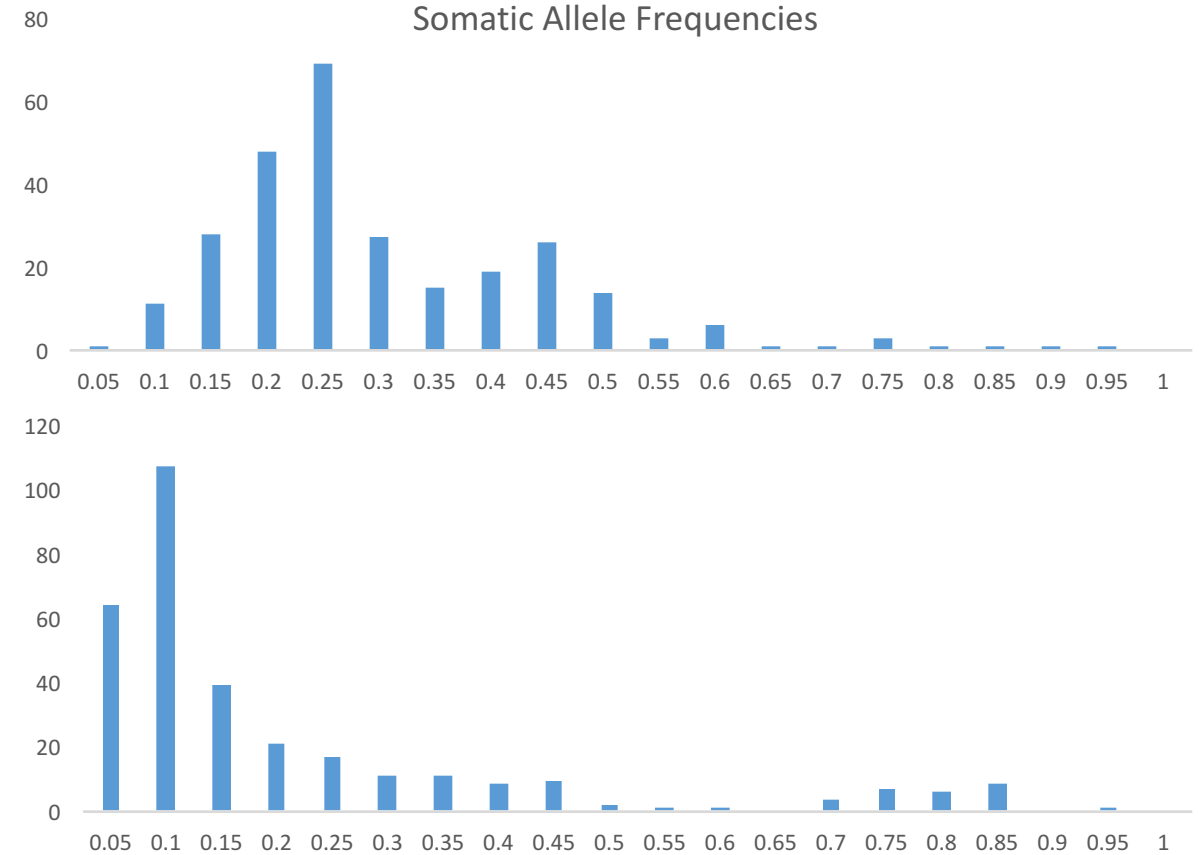
Germline vs Somatic Variant Calling

- Potentially very different allele frequency expectations

Germline Allele Frequencies



Somatic Allele Frequencies

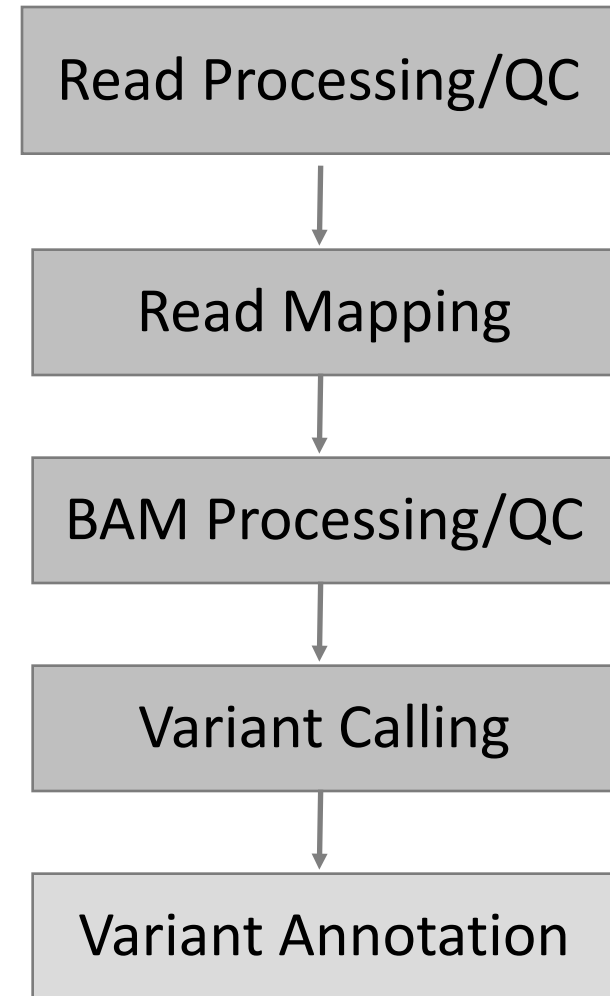


Exome vs Whole Genome Sequencing

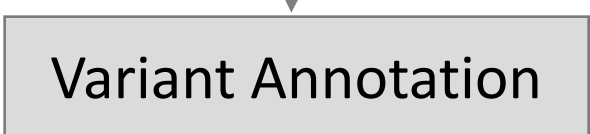
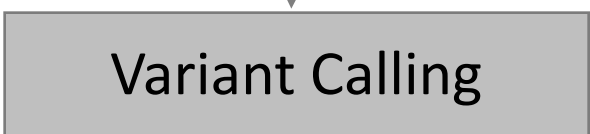
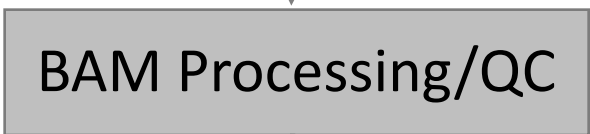
- Exome Sequencing
 - Covers ~5% of genome
 - Allows for high depth targeting
 - Most reasonable option for somatic variant analysis
 - Low-confidence copy number/structural variant calling
- Genome Sequencing
 - Confidently call >85% of reference genome (hg38)
 - Confidently call copy number/structural 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

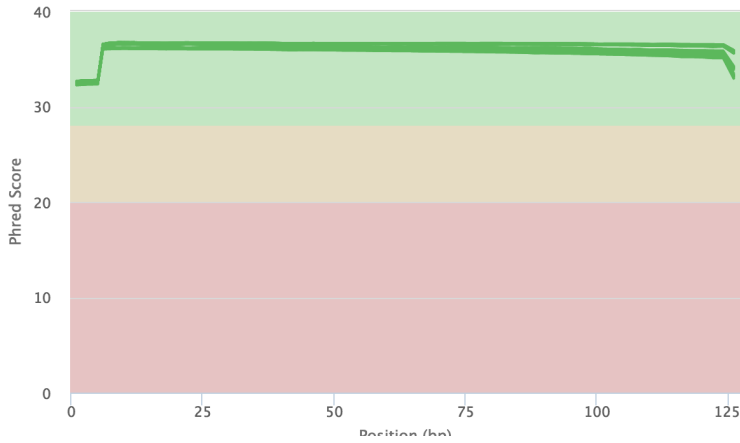
- All variant calling follows the same basic approach



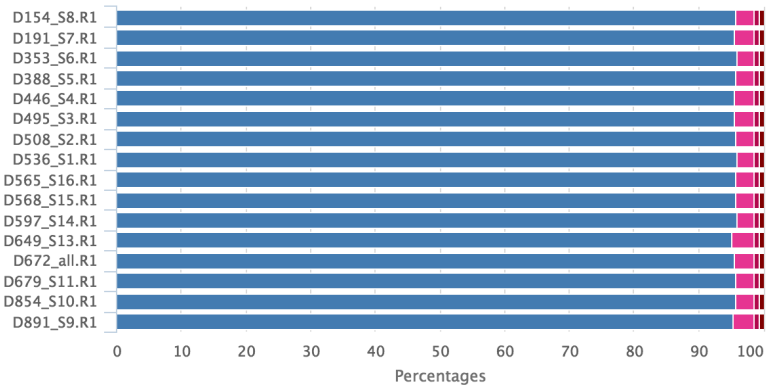
Variant Calling at CCBR



Mean Quality Scores

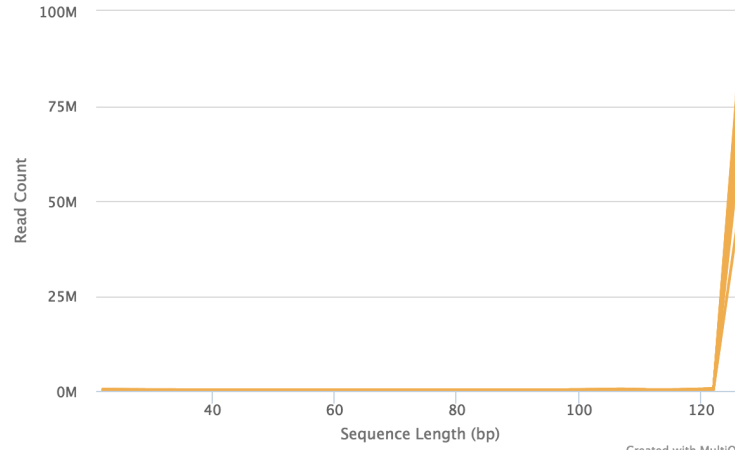


Trimmomatic

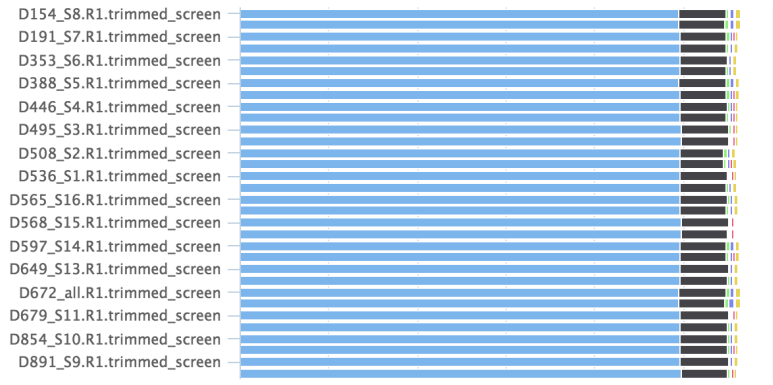


Created with MultiQC

Sequence Length Distribution

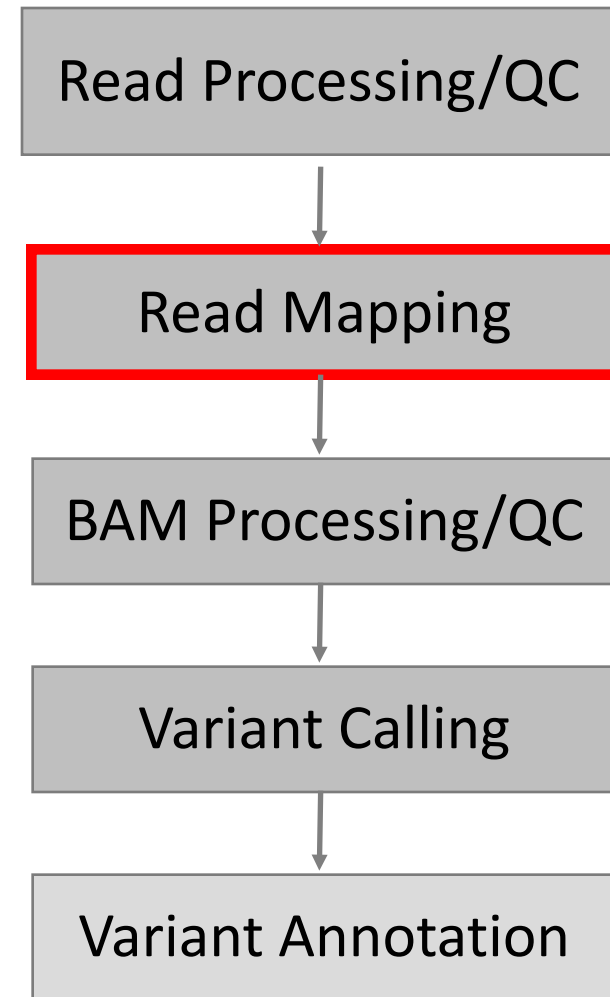
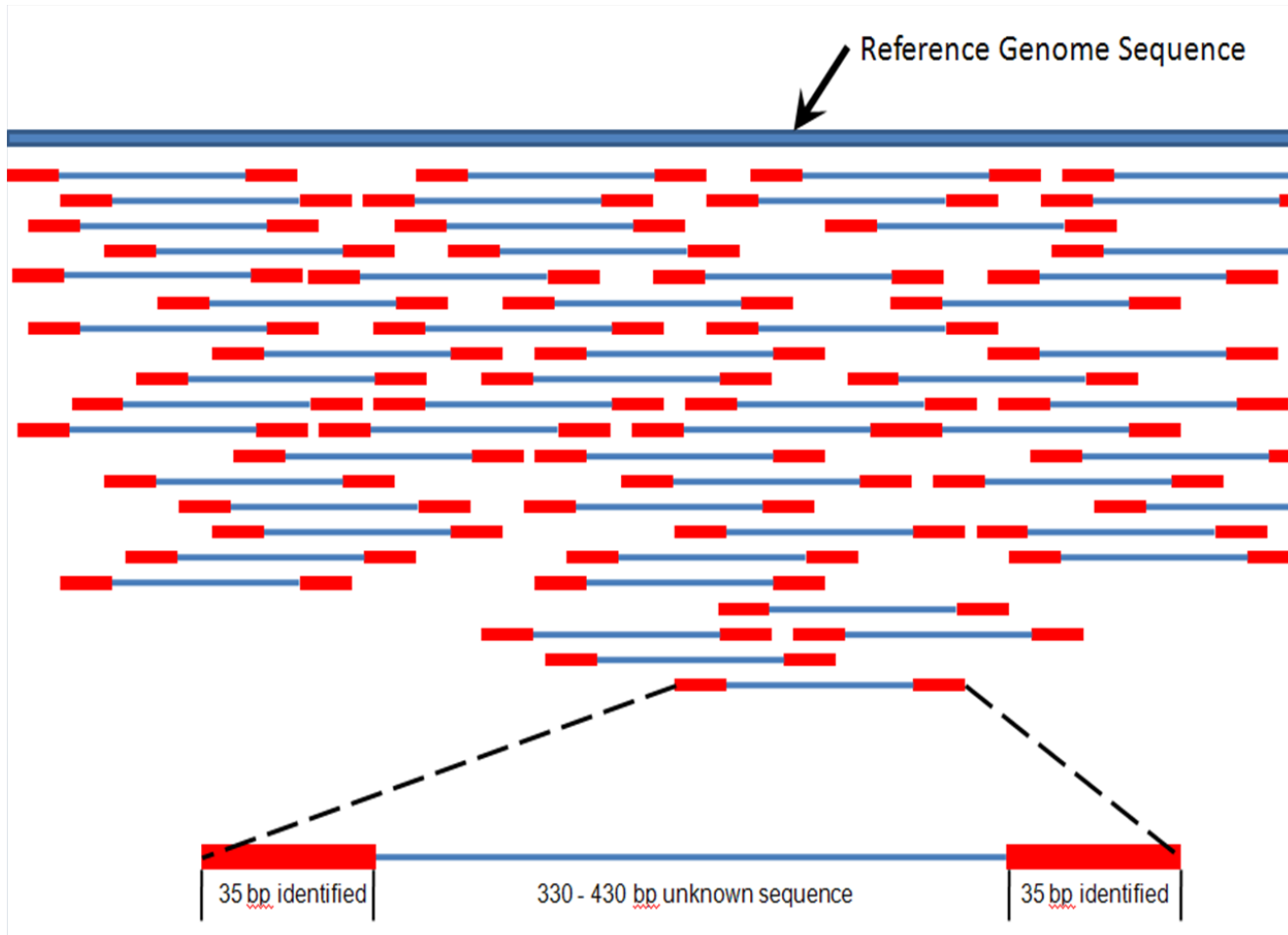


FastQ Screen



Created with MultiQC

Variant Calling at CCBR

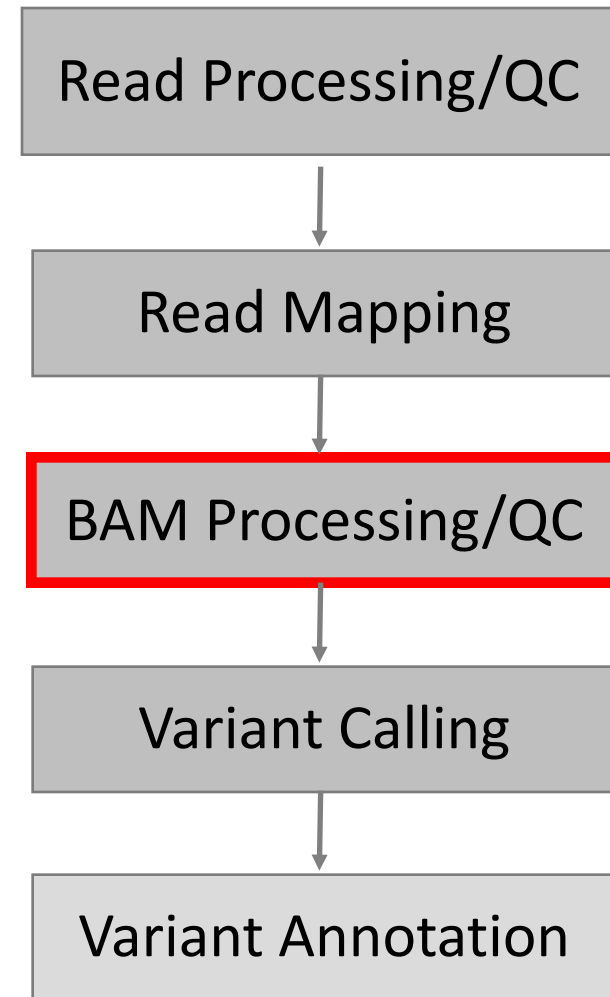
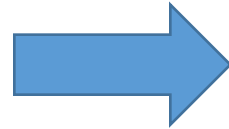


Variant Calling at CCBR

- Indel realignment



Local realignment

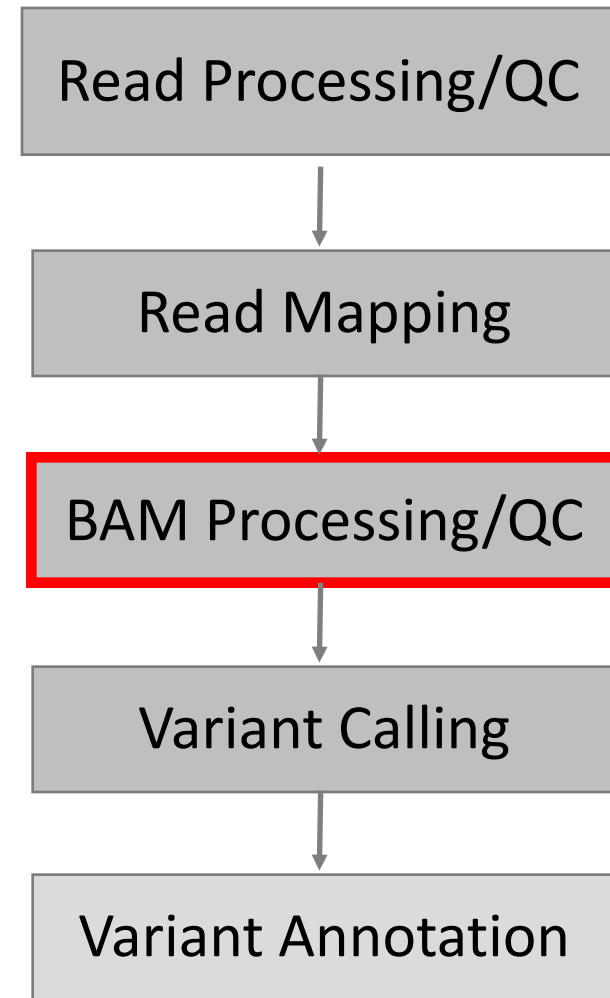
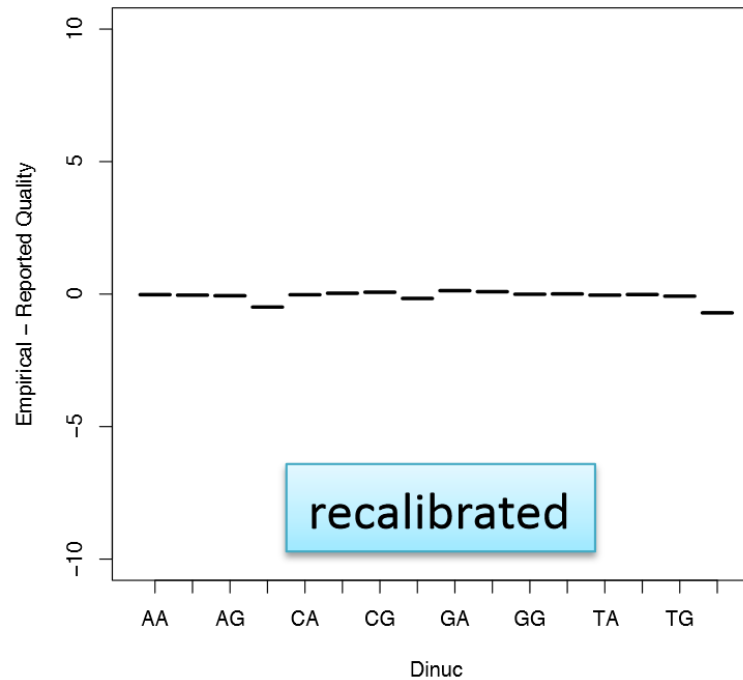
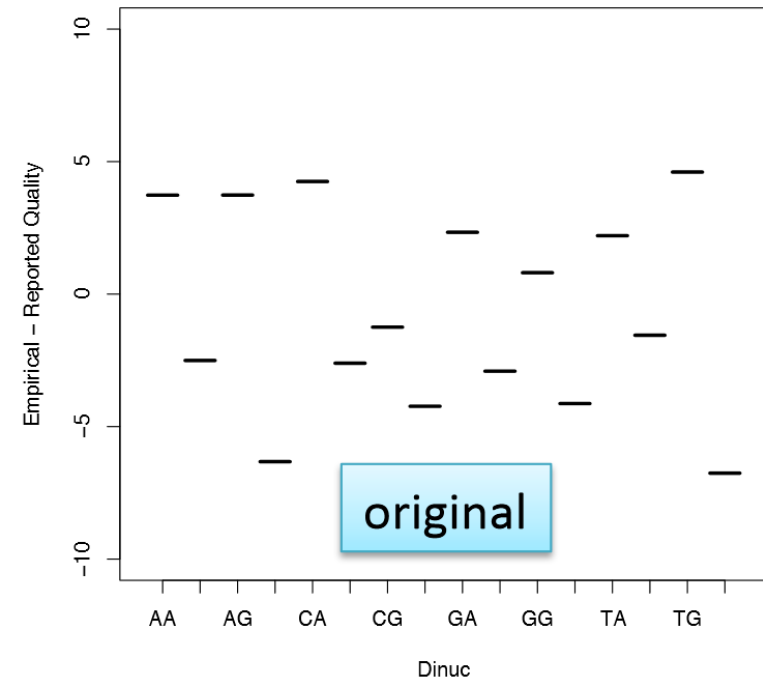


Variant Calling at CCBR

- Multiple sources of quality score bias

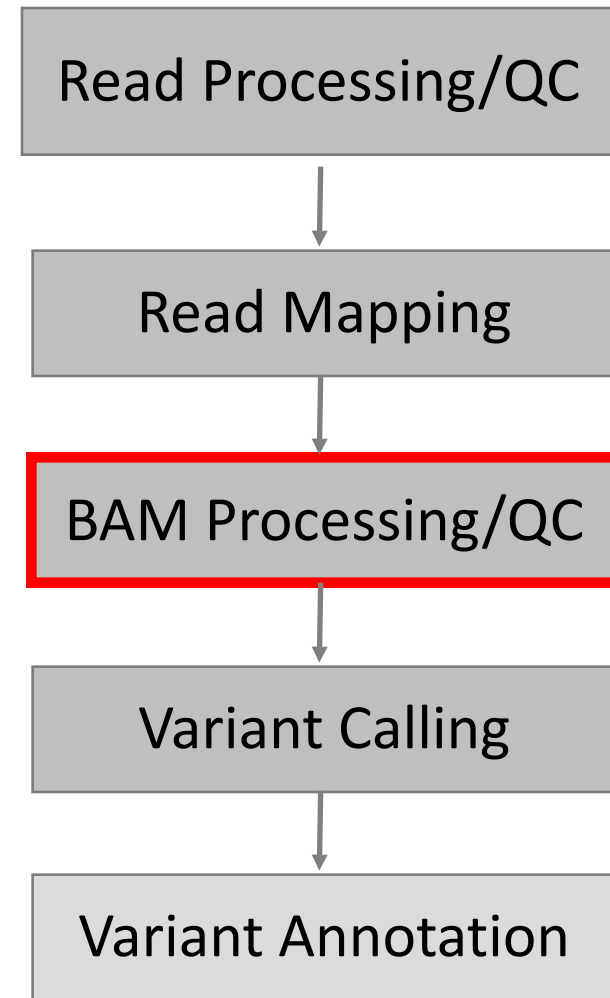
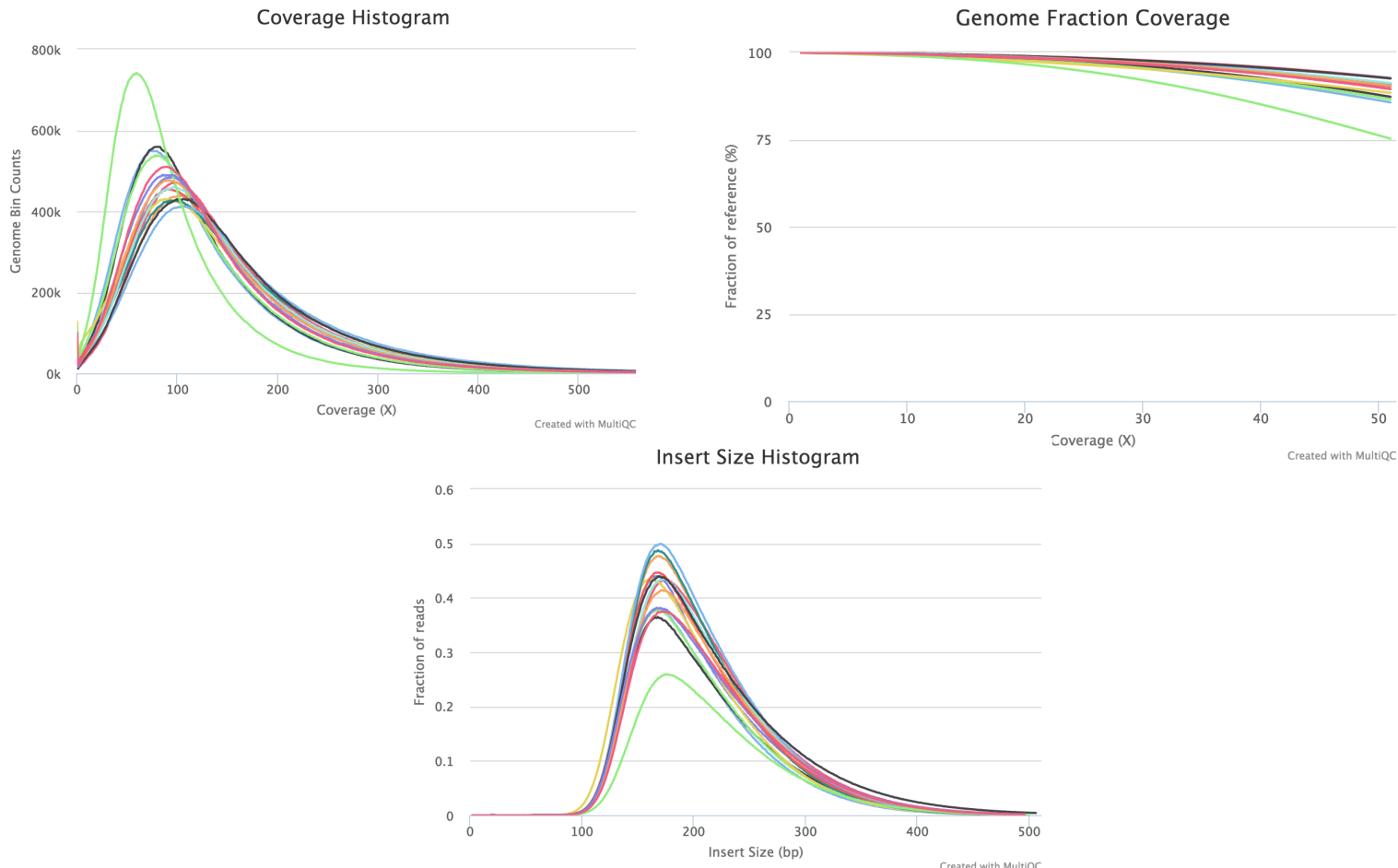
RMSE = 4.188

RMSE = 0.281



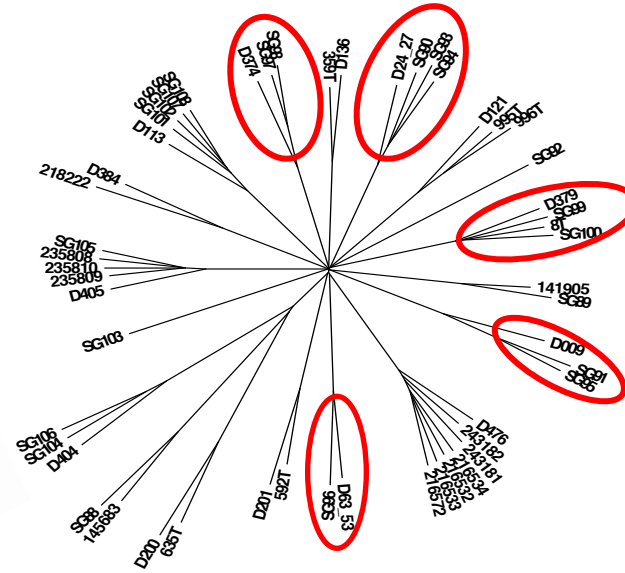
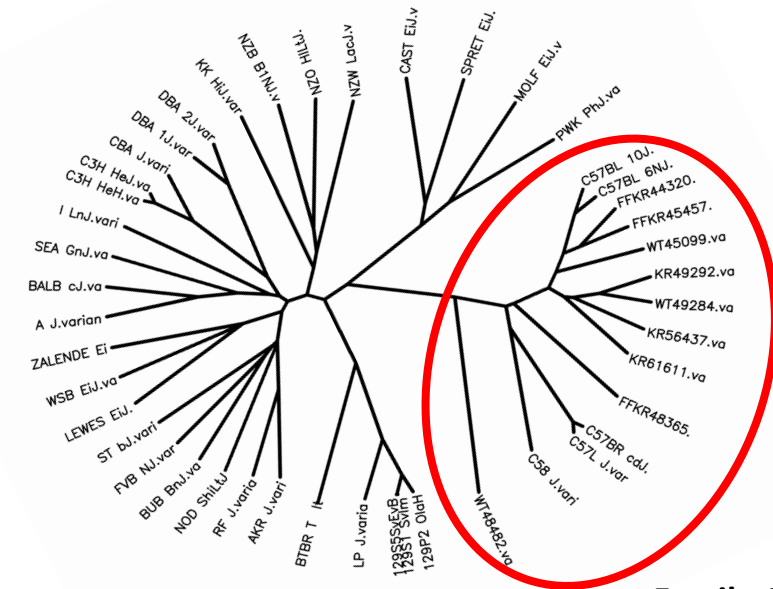
Variant Calling at CCBR

- Alignment QC

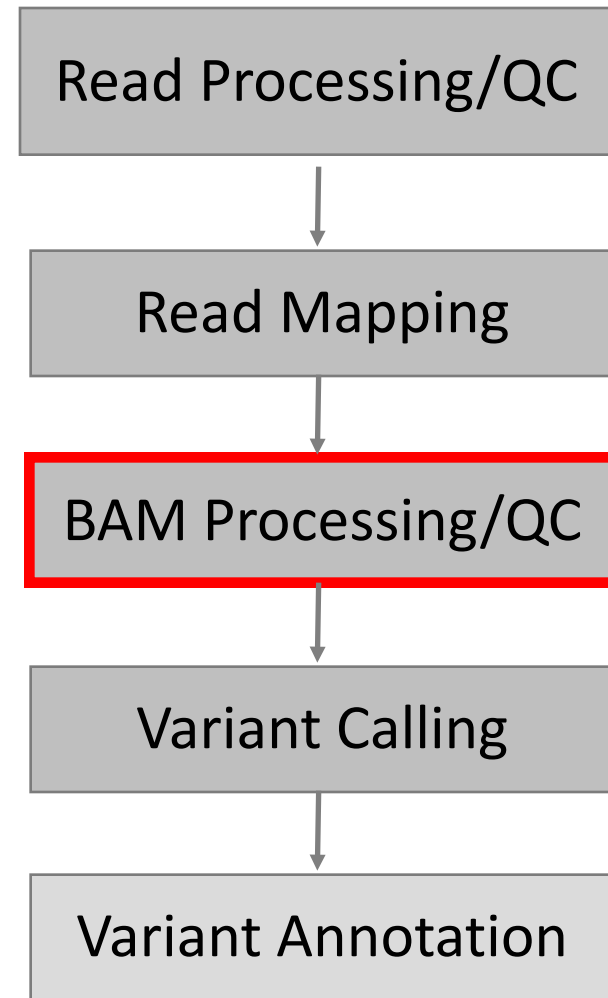
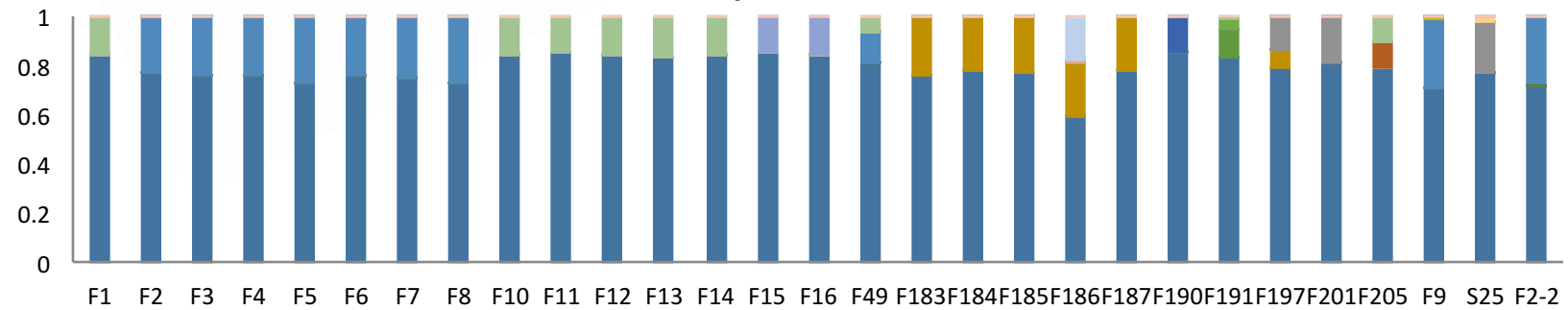


Variant Calling at CCBR

- Additional QC



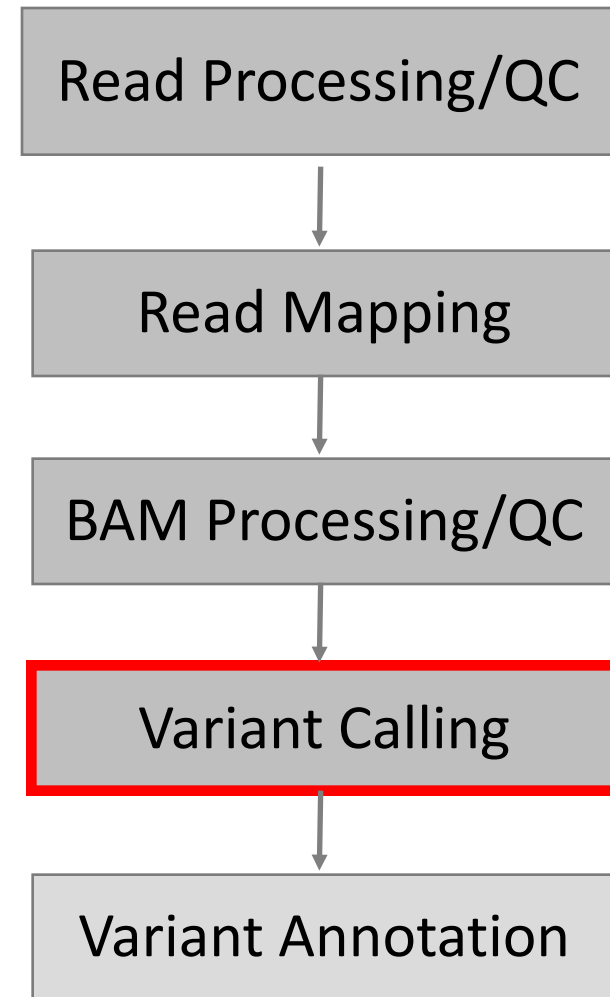
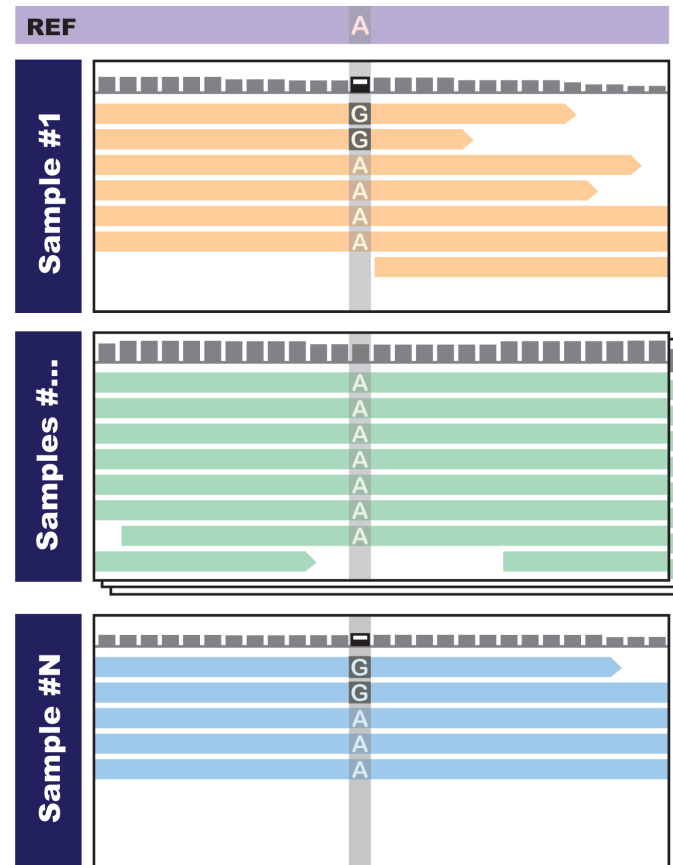
Family 1 Admixture



Variant Calling at CCBR

Germline

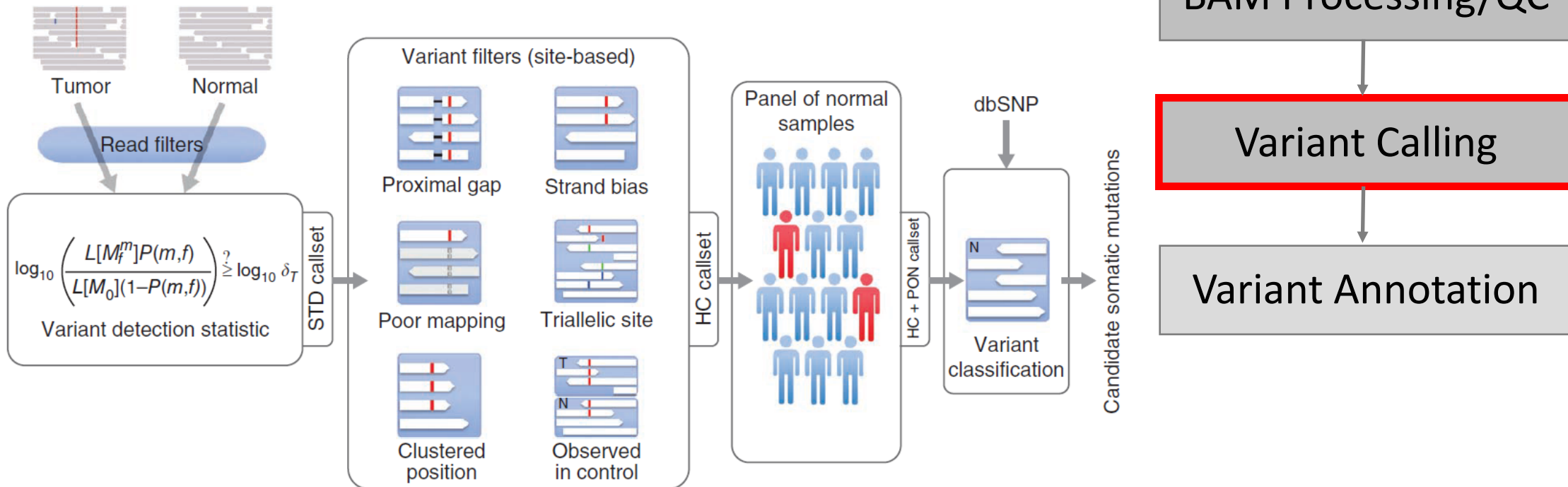
- Joint genotype with GATK HaplotypeCaller
 - SNPs/short INDELS
- MANTA
 - Large INDELS
 - Translocations
 - Inversions
 - Duplications



Variant Calling at CCBR

Somatic

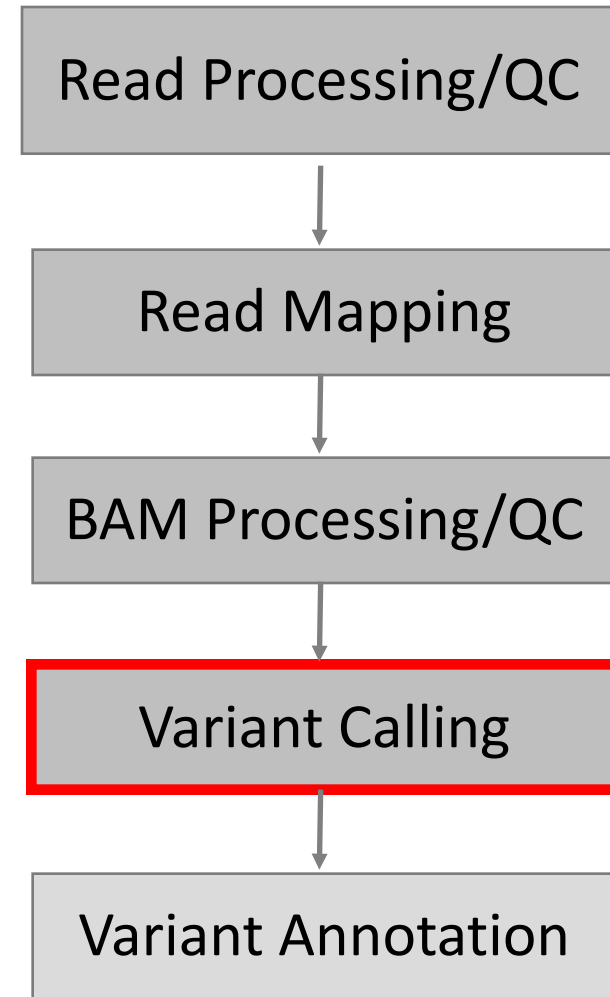
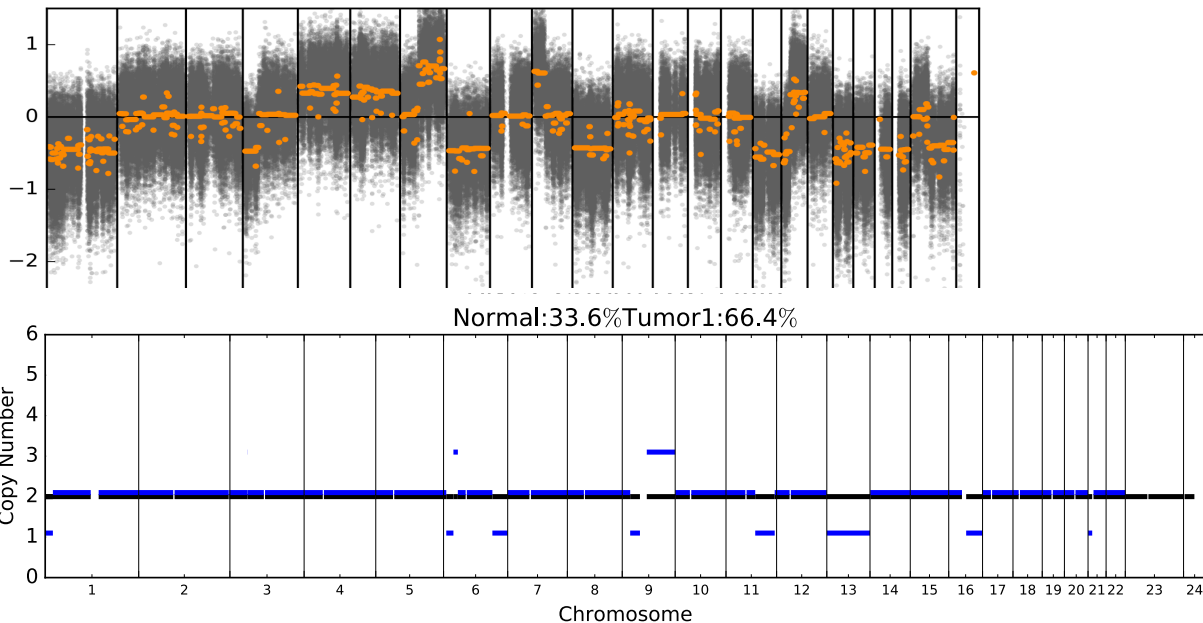
- MuTect, MuTect2, Strelka



Variant Calling at CCBR

Somatic

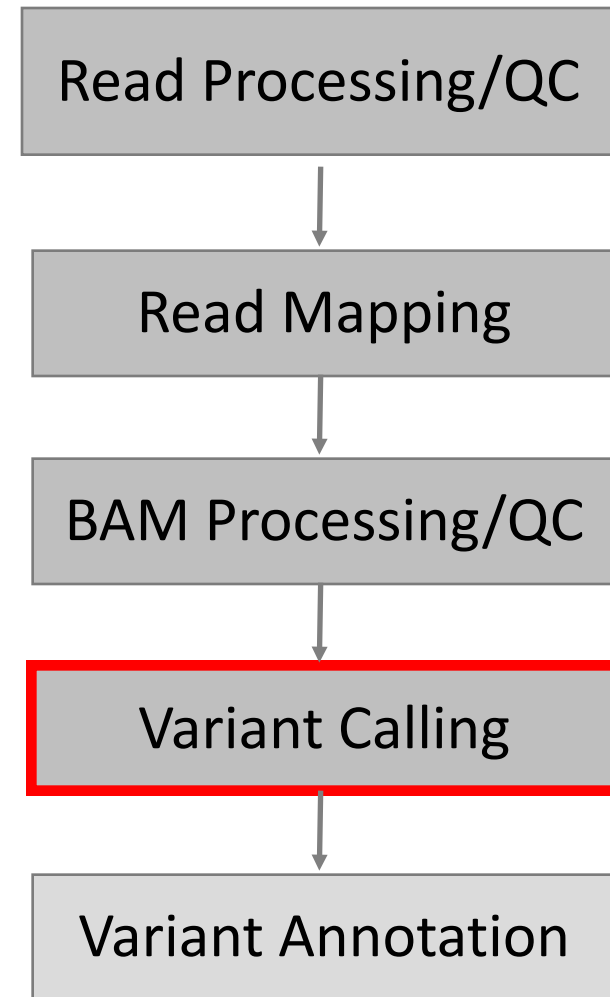
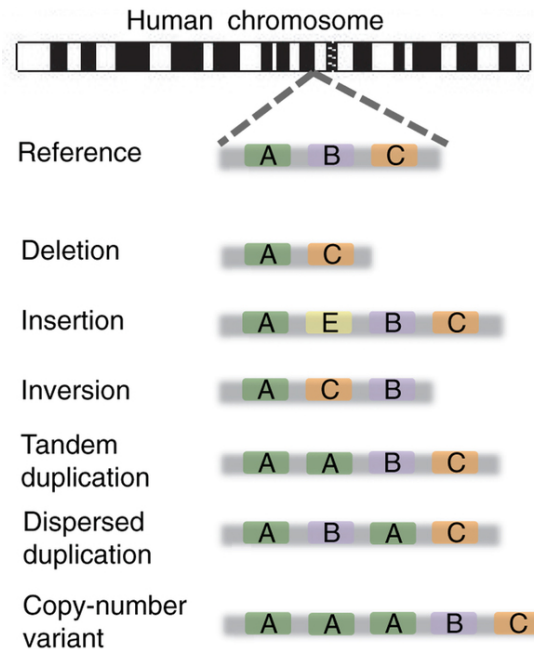
- MuTect, MuTect2, Strelka
- Copy number – CNVkit, THetA2



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>



Annotation, Visualization, and Impact Analysis

Analysis of Genomic Variations with AVIA

Home

Information

FAQ

Databases

What's new

Resources

Genomic Workflows

- Feature Annotation and Visualization
- Basic Annotation Tool
- Cascade Filtering
- MiRNA SNP Analysis

Protein Tools

- Annotation with Protein coordinates *beta*
- Visualization of Protein using JSmol

General Tools

- Set up AVIA configuration file
- Gene based tools
- File/Data Converter tools

Results Retrieval

- Retrieve Request By ID
- View Sample Results Page

Disclaimer

Cite Us

AVIA Annotation and Visualization Request

In this tool, users will be able to annotate their data with publicly available databases, as well as upload their own databases. Users will also have the opportunity to visualize each of these databases as tracks within Circos. If a gene list is specified in Section II, the highlight and filter options only apply to the Circos visualization. Please read our [FAQ](#) or [Tutorials](#) for detailed information. If you do not have any data to start with, click on the button below labeled 'Sample BED data' for a self guided tutorial.

Section I. Input Data (Required)

A field with an asterisk (*) before it is a required field.

Name your submission (optional): If name exists, AVIA will add timestamp

*Input Filename: no file selected (?)

Check ONLY if your input file is a compressed file (zip, tar or gzip) with multiple variant files

-- or --

Enter your data here using comma or space separated list (one variant per line)

*Input format:

*Organism and build:

*E-mail address: You will be notified by email when the process is complete

Read Processing/QC

Read Mapping

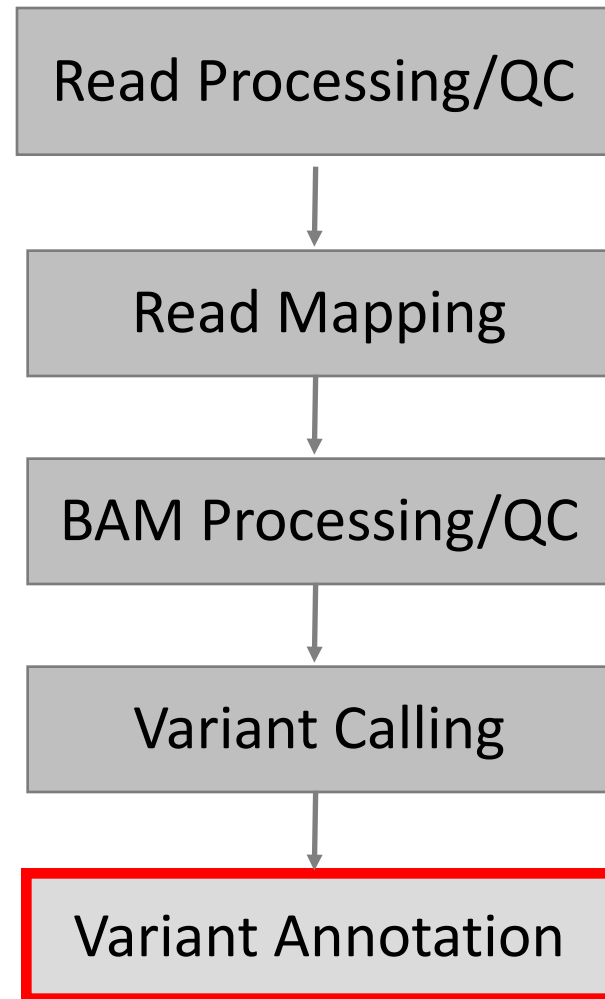
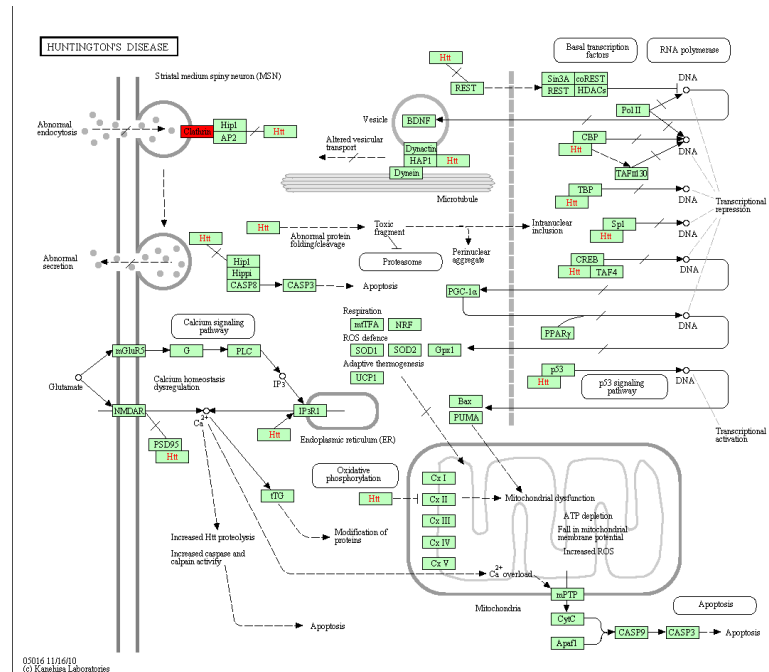
BAM Processing/QC

Variant Calling

Variant Annotation

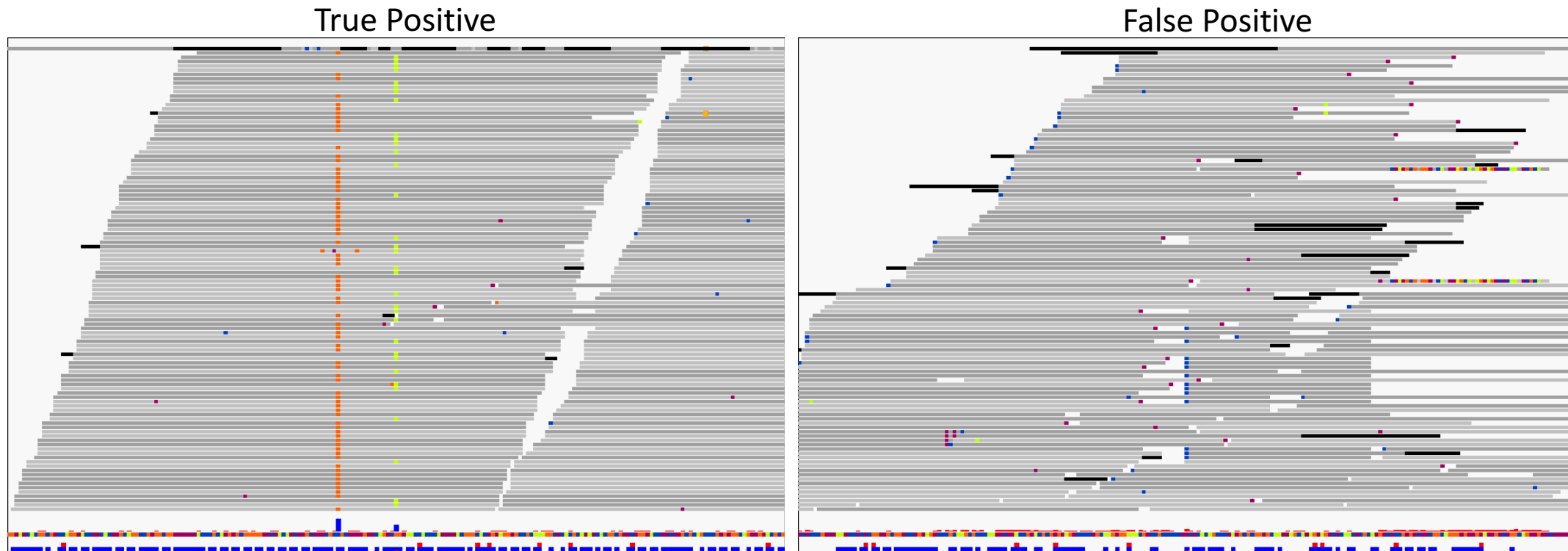
Variant Calling at CCBR

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

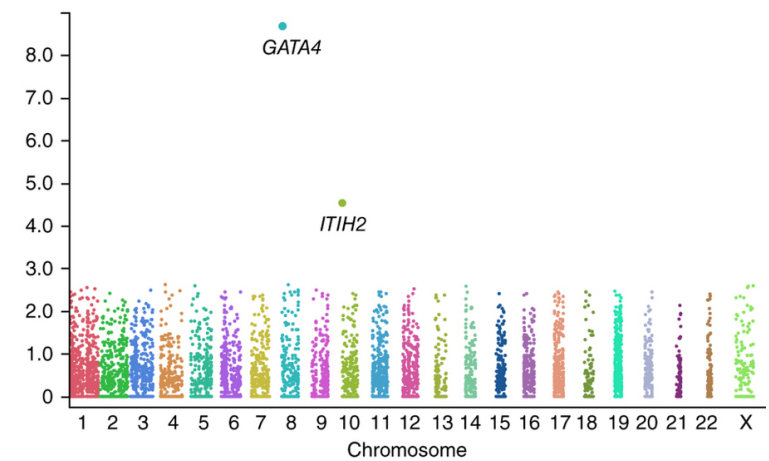
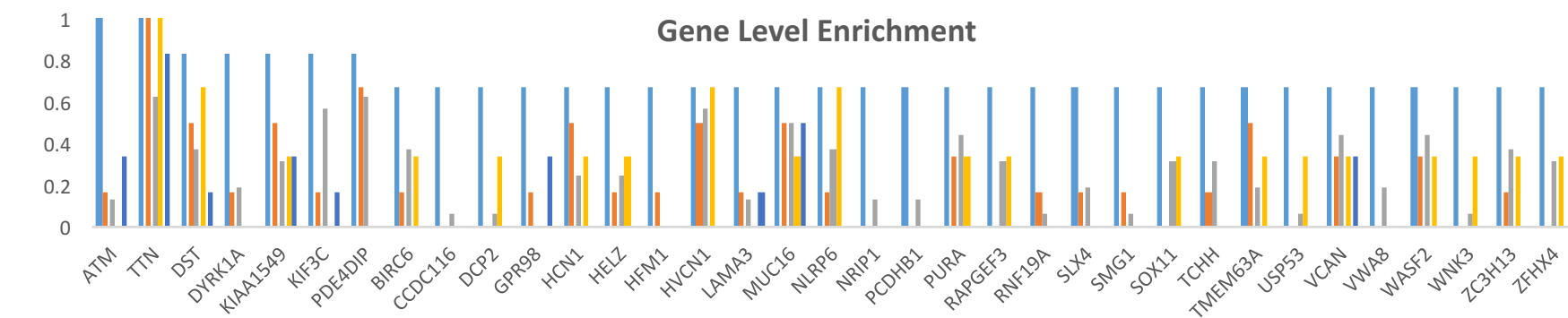
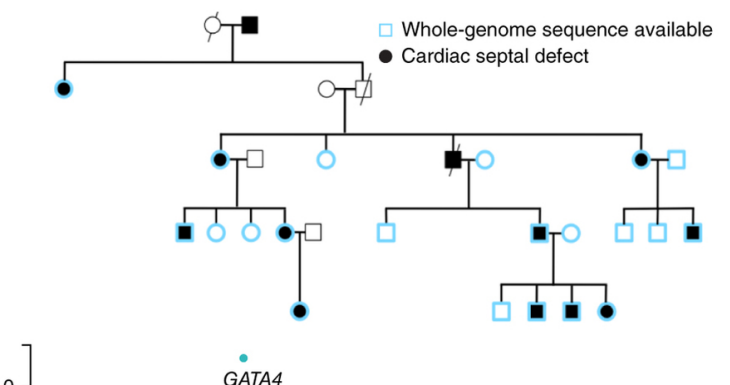
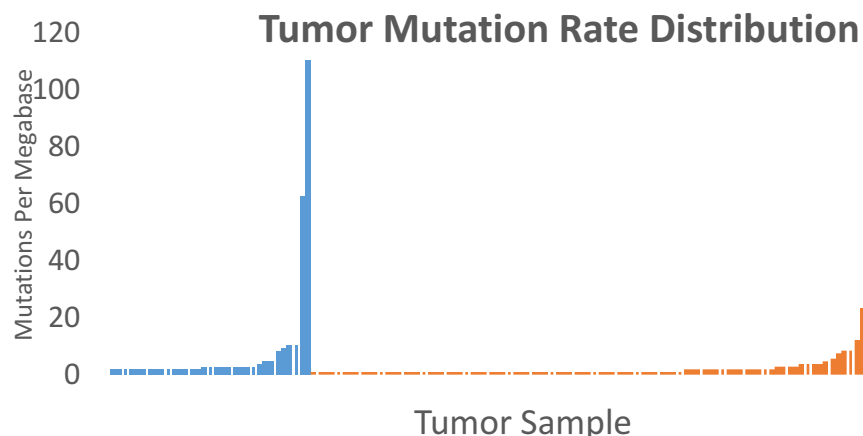
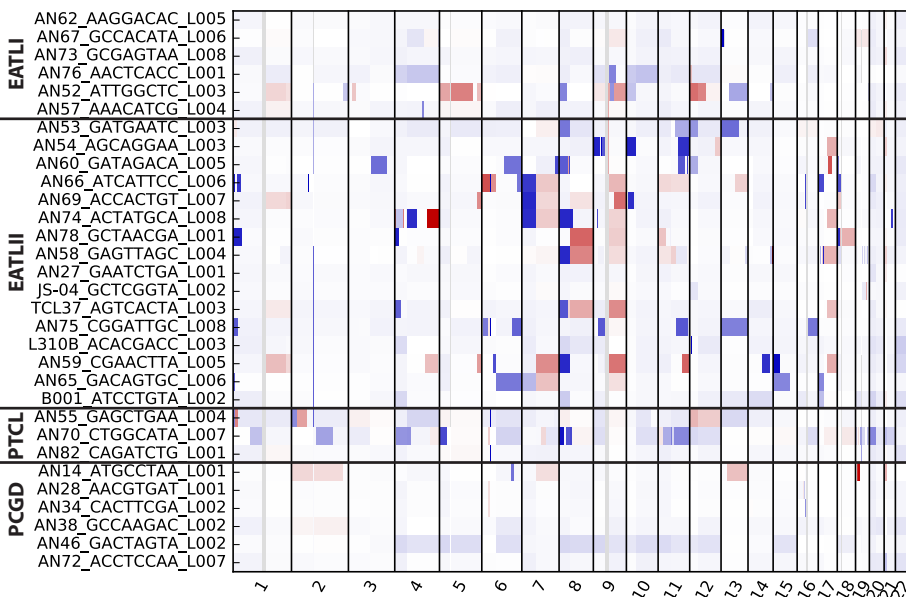
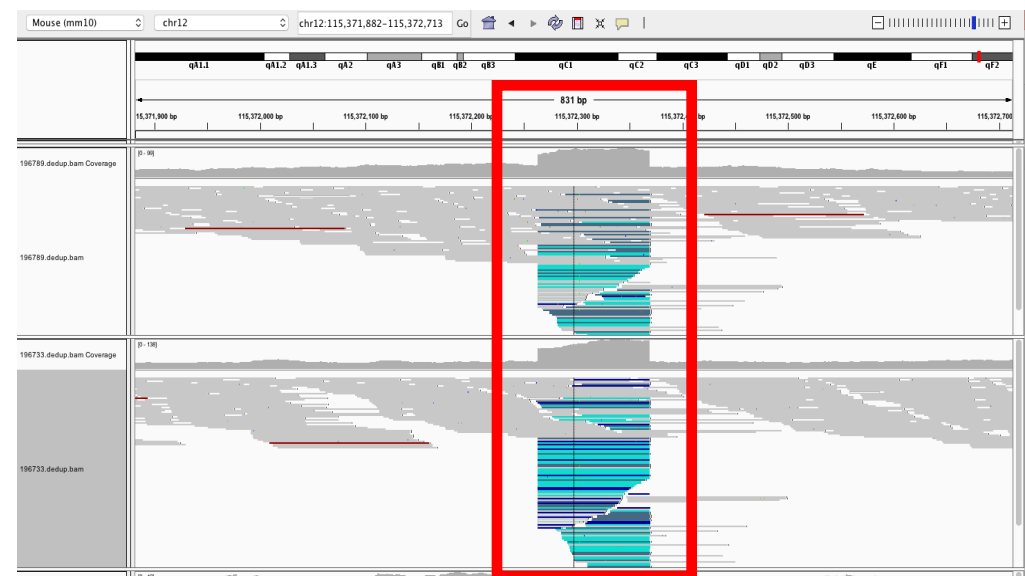


Variant Verification

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

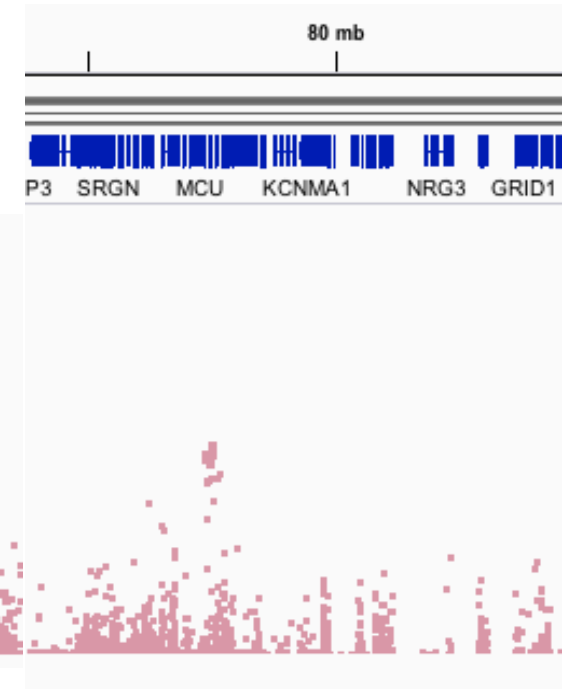


Downstream Analysis

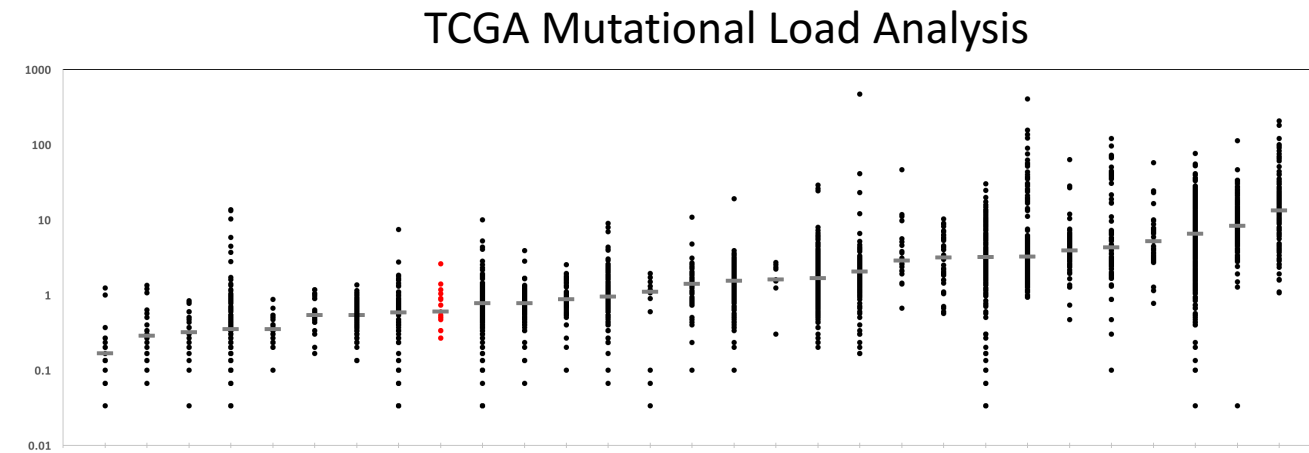
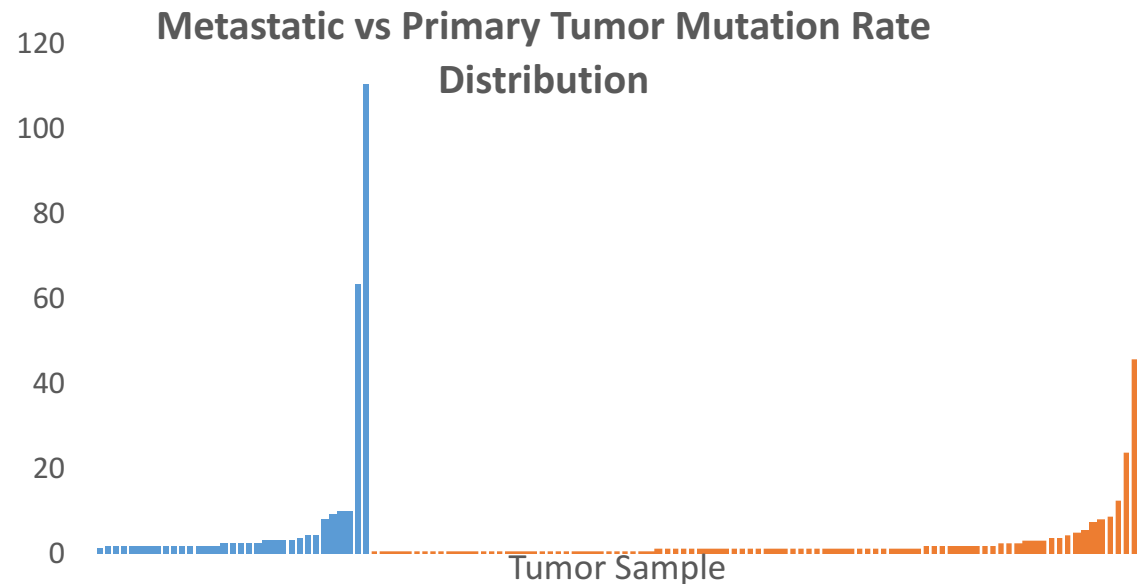


Analysis of Publicly Available Datasets

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



TCGA Germline
Association Analysis



Somatic Variant Calling – Best Practices

- STRONGLY favor paired tumor/normal design
 - Includes non-human samples
- For non-human samples
 - ≥ 3 control/"germline" samples
- $\geq 100X/50X$ 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
 - $\geq 30X$ mean target depth
 - Superior to exome for structural variants, copy number analysis
- Germline exome
 - $\geq 50X$ mean depth
- For familial/trio analyses, we strongly encourage early consultation
 - Selection of samples for sequencing can be CRUCIAL to maximizing power

