Exome-Seq Analysis: Overview and Best Practices

Justin Lack

CCBR Exome-seq Pipeline (and other pipelines, too!)

- Streamline and expedite delivery of actionable variants for a wide range of projects
  - Tumor/Normal, Tumor-only, and Germline variant discovery
  - Data sets ranging from one to thousands of samples
  - Both mouse and human (and potentially other model organisms, as well)
  - Easily used and interpreted by a wide range of expertise
  - Meet QC requirements of Sequencing Facilities for seamless delivery
  - Operate within framework for other pipelines…

Variant Calling at CCBR

- Multiple Variant Calling CCBR Pipelines
  - Whole genome
  - Whole exome/targeted sequencing
  - RNAseq-var (available soon)
- Other pipelines, too:
  - ChIP-seq
  - RNAseq
  - mirSeq
  - more coming…

Variant Calling at CCBR

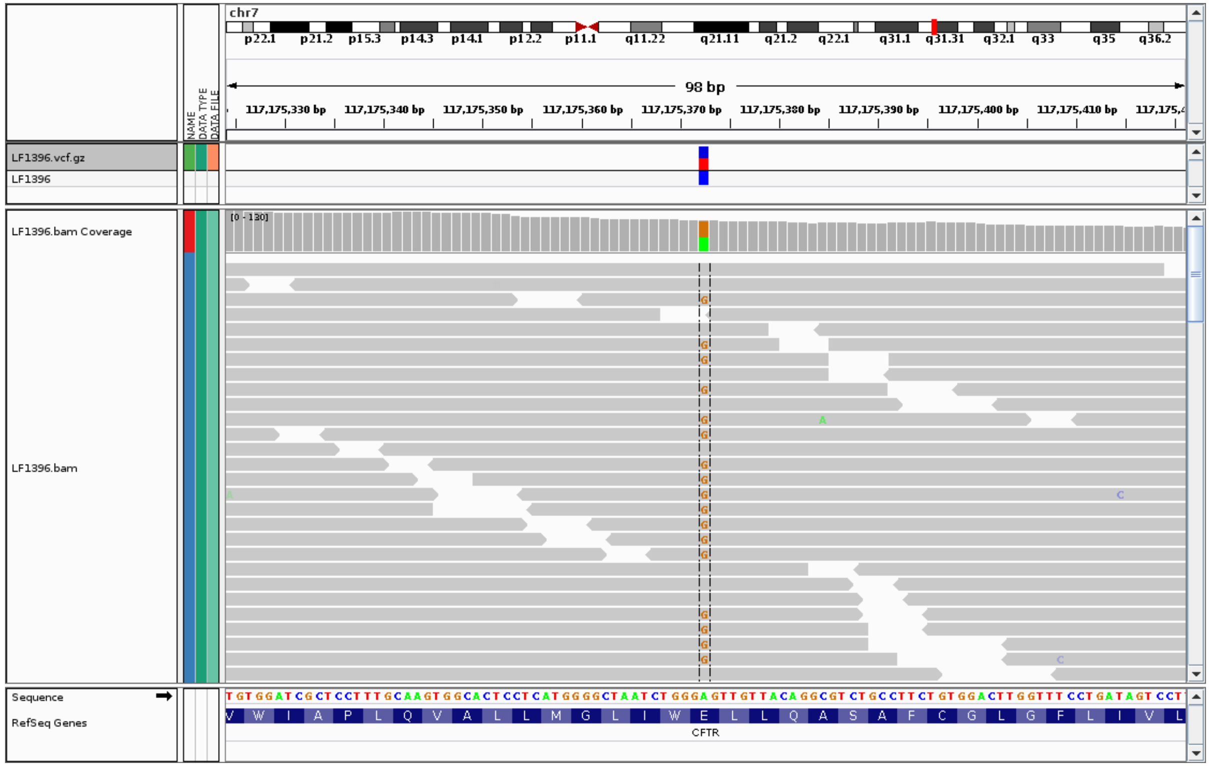
- Multiple Variant Calling CCBR Pipelines
  - **Whole genome**
  - **Whole exome/targeted sequencing**
  - RNAseq-var (available soon)

Variant Calling at CCBR

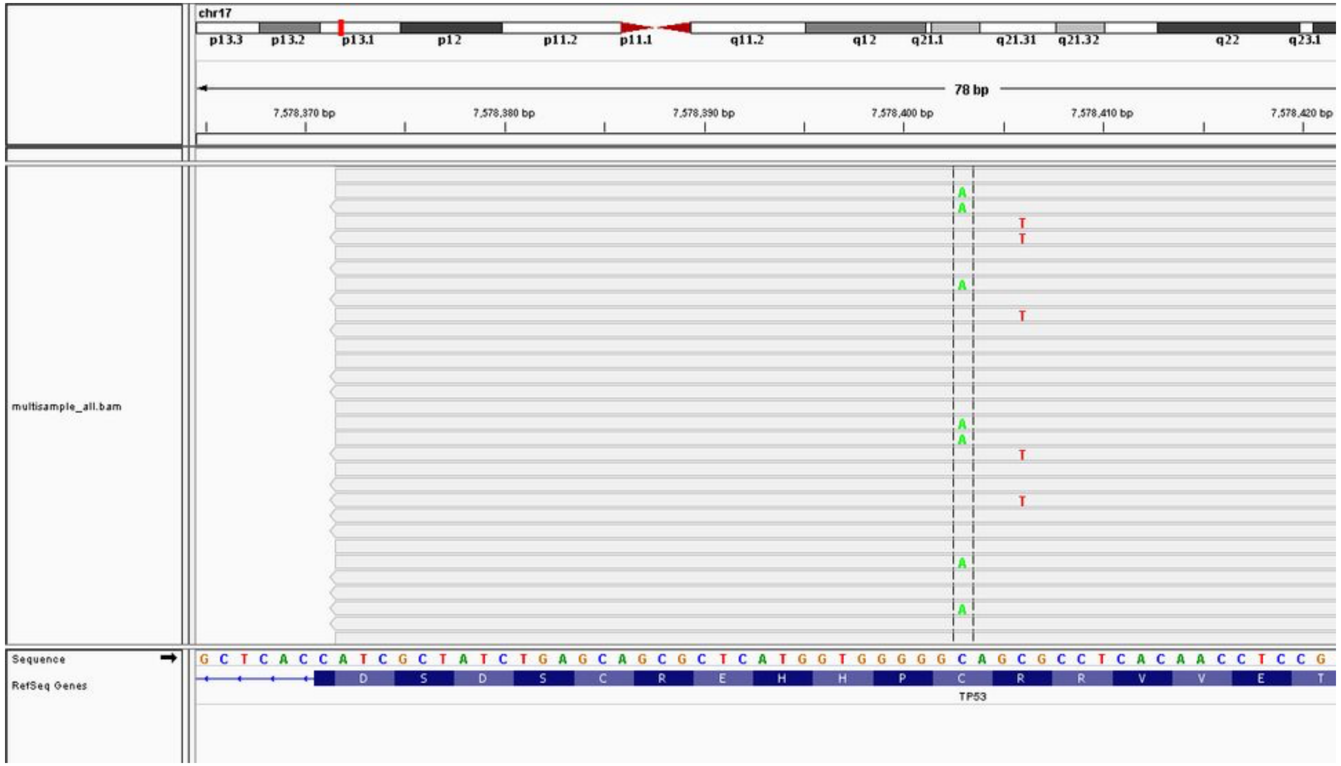
- Multiple Variant Calling CCBR Pipelines
  - **Whole genome**
  - **Whole exome/targeted sequencing**
- Two variant calling “flavors”
  - Germline
    - Heritable disease-causing variation (i.e., familial/trio design), population-level analyses (i.e., GWAS), cell lines, etc.
  - Somatic
    - Tumor/Normal or Tumor-only variants
- Very different expectations in terms of variant detection

Germline vs Somatic Variant Calling

- Potentially very different allele frequency expectations


chr7
p22.1 p21.2 p15.3 p14.3 p14.1 p12.2 p11.1 q11.22 q21.11 q21.2 q22.1 q31.1 q31.31 q32.1 q33 q35 q36.2
98 bp
117,175,330 bp 117,175,340 bp 117,175,350 bp 117,175,360 bp 117,175,370 bp 117,175,380 bp 117,175,390 bp 117,175,400 bp 117,175,410 bp
NAME
DATA TYPE
DATA FILE
LF1396.vcf.gz
LF1396
LF1396.bam Coverage
[0 - 130]
LF1396.bam
Sequence
RefSeq Genes
TGTGGATCGCTCCTTTGCAAGTGGCACTCCTCATGGGGCTAATCTGGGAGTTGTTACAGGCGTCTGCCTTCTGTGGACTTGGTTTCCTGATAGTCCT
V W I A P L Q V A L L M G L I W E L L Q A S A F C G L G F L I V L
CFTR

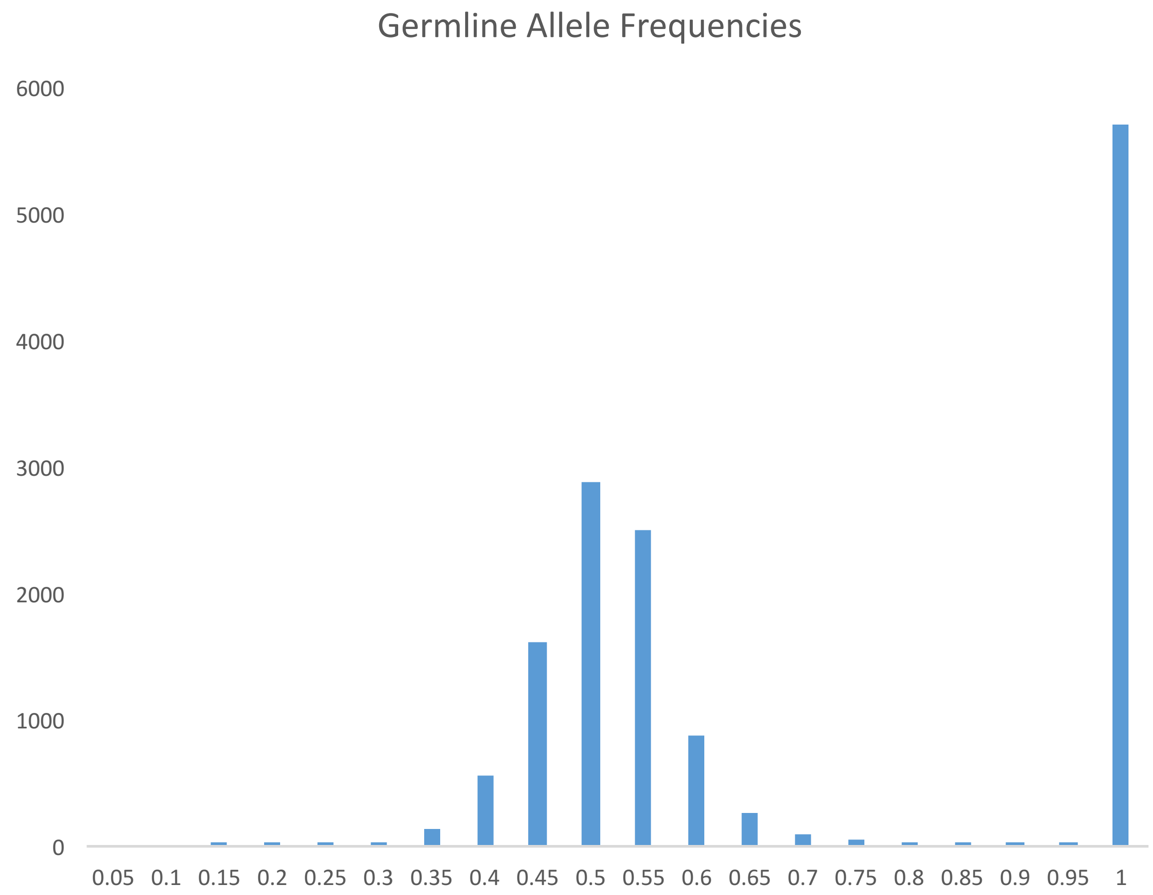
Germline - ~0.5 read proportions

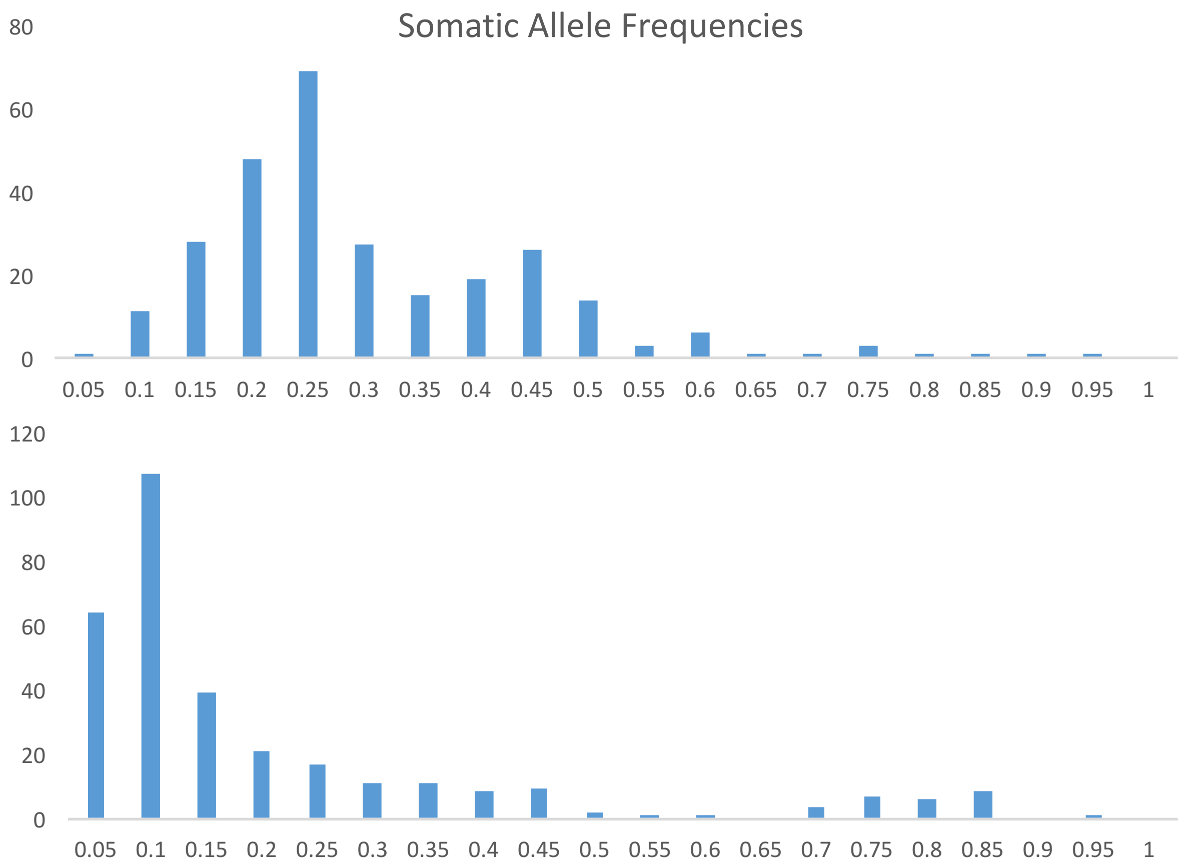

chr17
p13.3 p13.2 p13.1 p12 p11.2 p11.1 q11.2 q12 q21.1 q21.31 q21.32 q22 q23.1
78 bp
7,578,370 bp 7,578,380 bp 7,578,390 bp 7,578,400 bp 7,578,410 bp 7,578,420 bp
multisample_all.bam
Sequence
RefSeq Genes
GCTCACCATCGCTATCTGAGCAGCGCTCATGGTGGGGGCAGCGCCTCACAACCTCCG
D S D S C R E H H P C R R V V E T
TP53

Somatic - ~0.3 read proportions

Germline vs Somatic Variant Calling

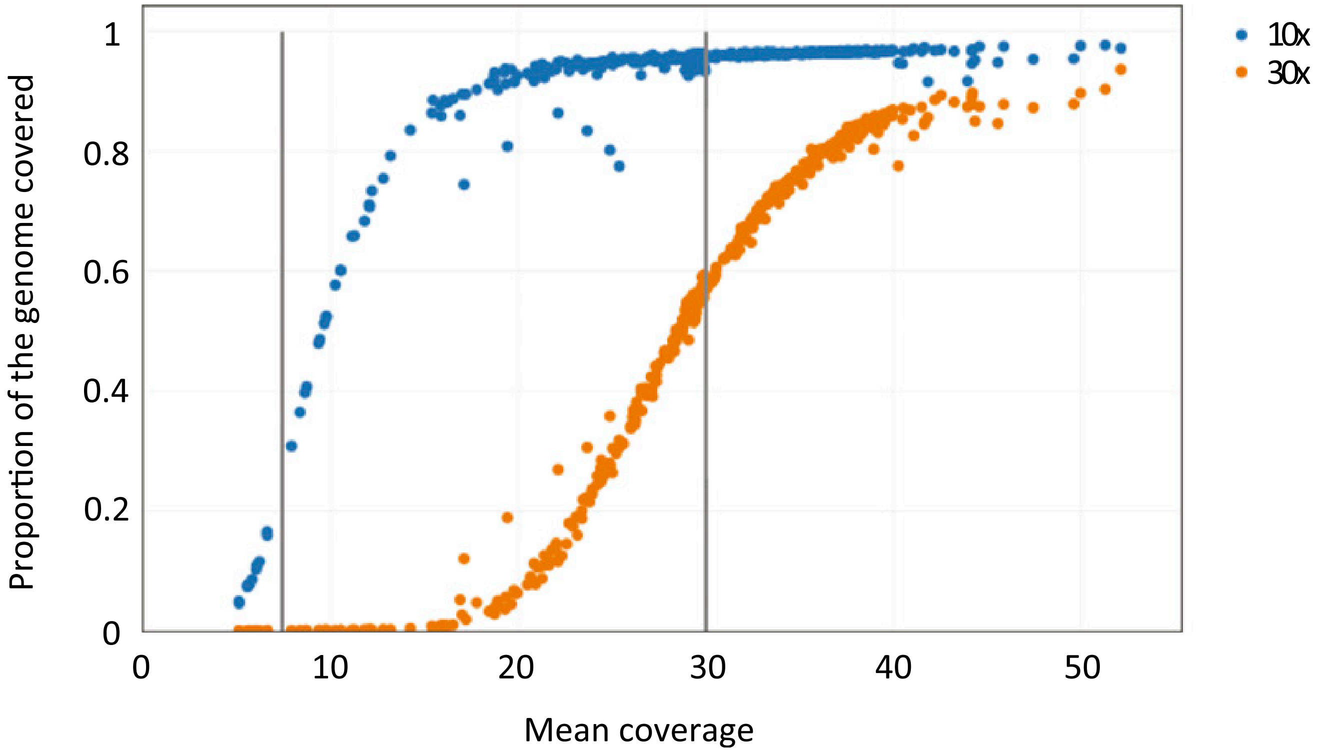
- Potentially very different allele frequency expectations

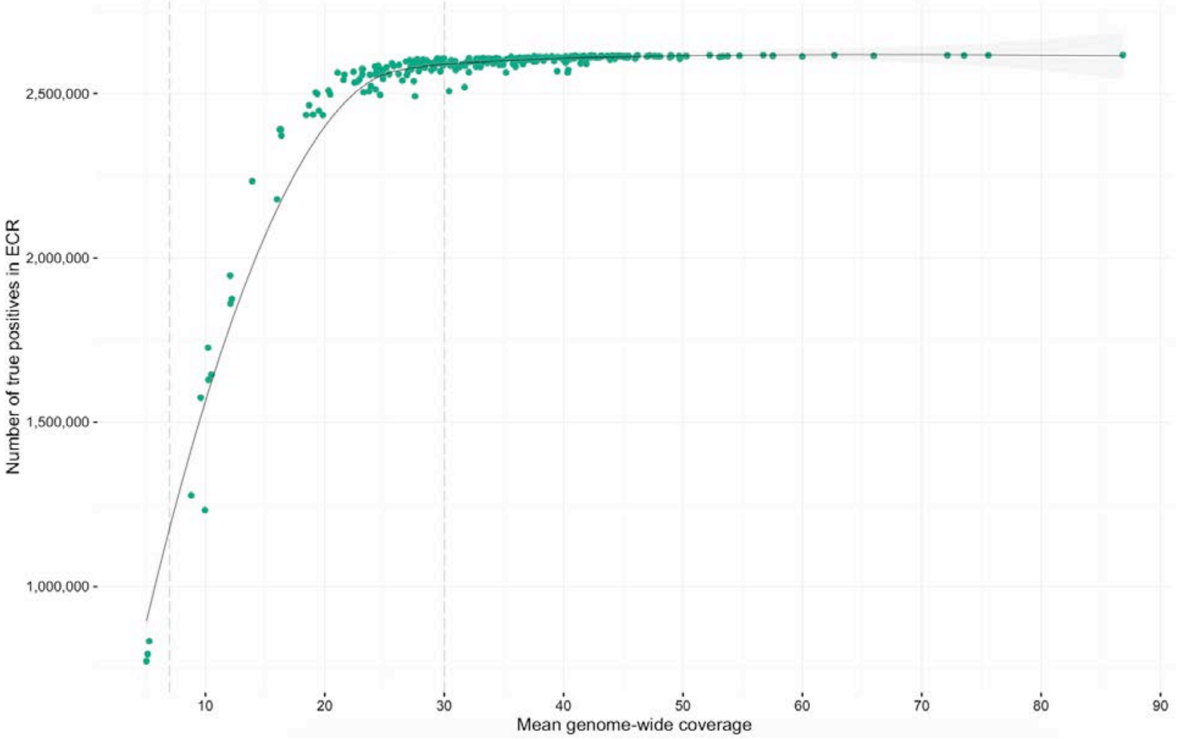




Depth Effects - Germline

- ~30X target for genome data (below)
- ~50X target for exome, due to increased depth variance

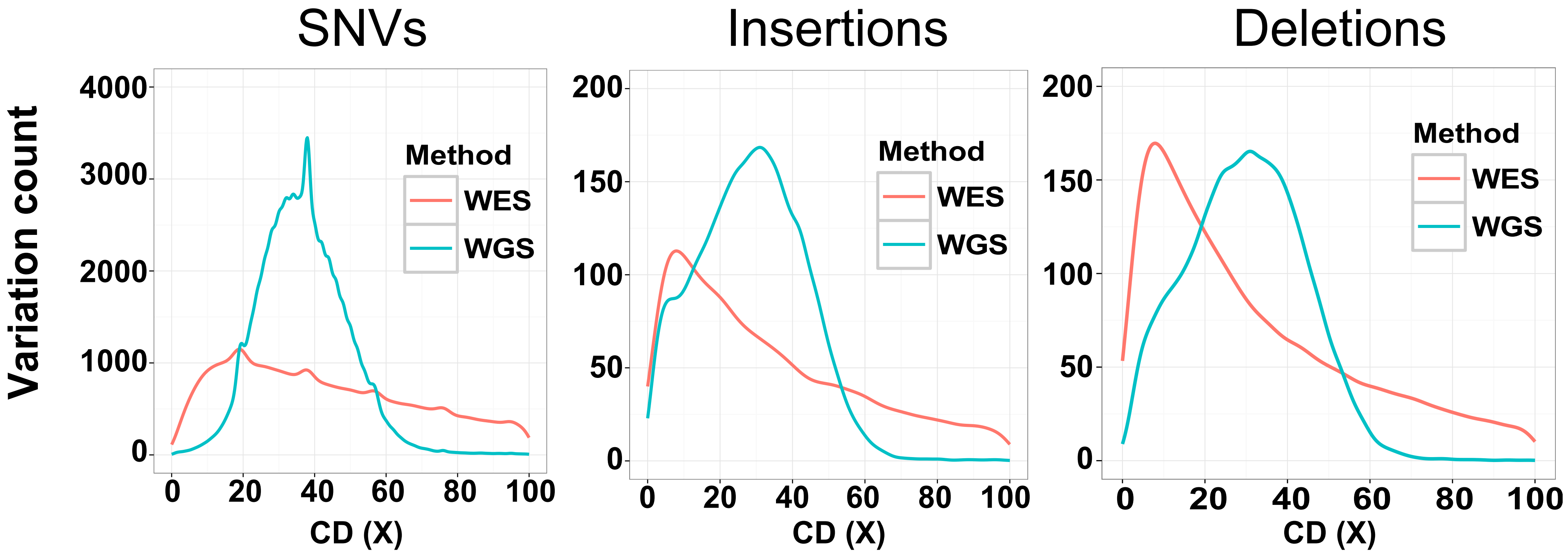




Telenti et al., 2016 *PNAS*

Depth Effects - Germline

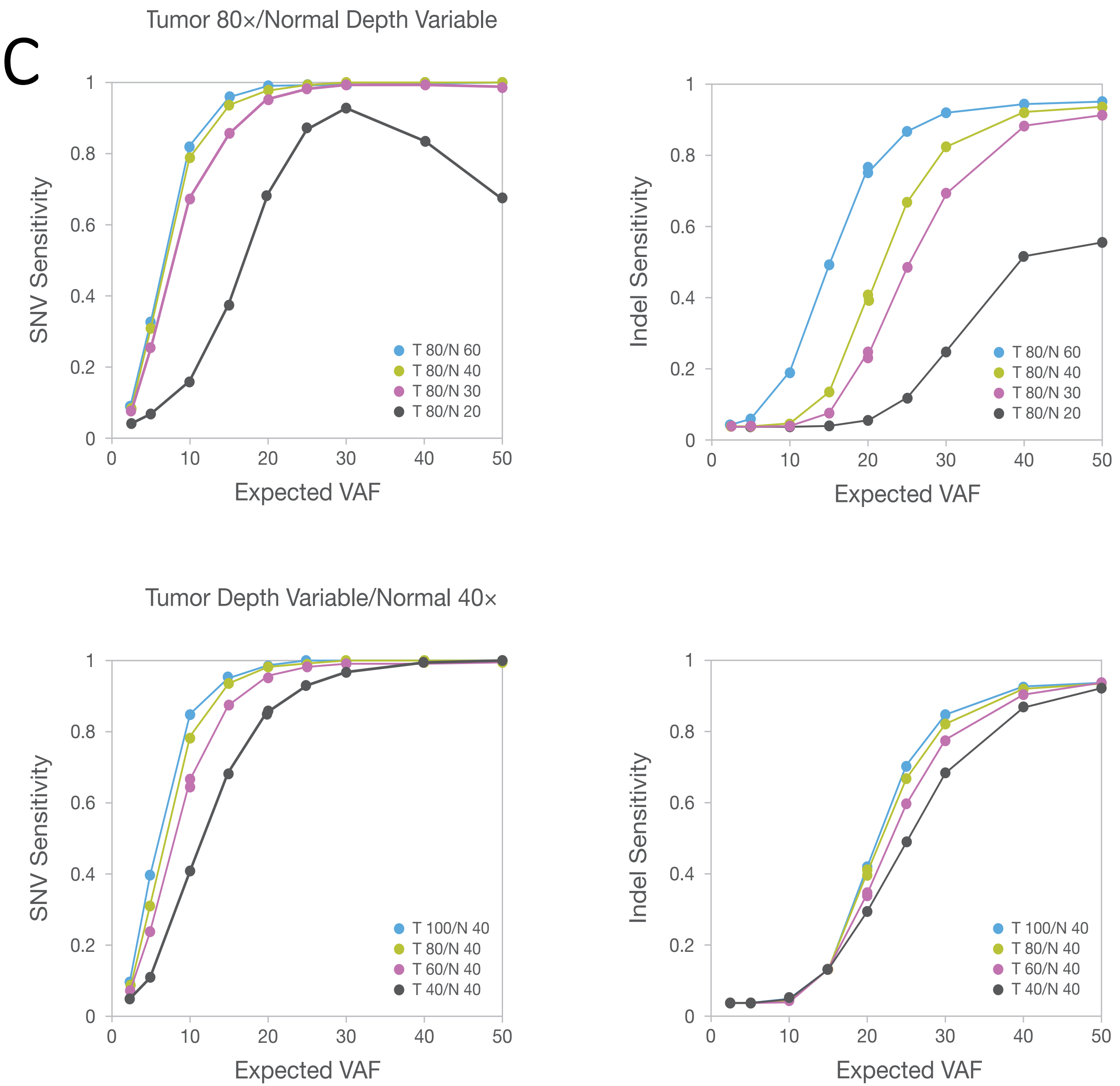
- ~30X target for genome data (below)
- ~50X target for exome, due to increased depth variance



Belkadi et al., 2015 *PNAS*

Depth Effects - Somatic

- >50X target for germline exome
- >100X target for somatic exome
- Tumor purity ≥50% (ideally ≥60% for copy number calling)



Tumor 80×/Normal Depth Variable

Tumor Depth Variable/Normal 40×

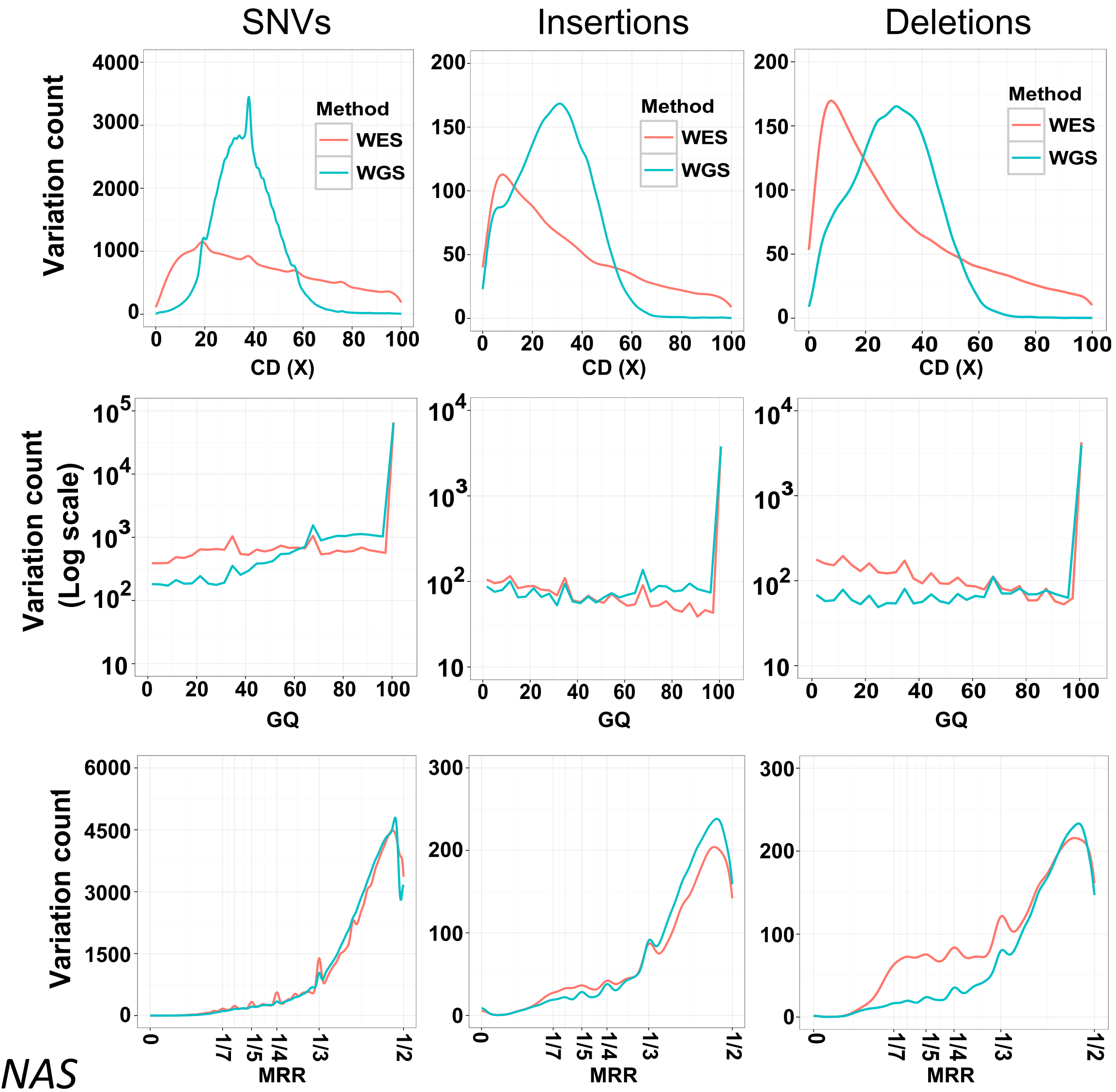
Exome vs Whole Genome Sequencing

Exome vs Whole Genome Sequencing

- Exome Sequencing
  - Covers ~2% 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
  - Confidently call copy number/structural variants
  - Significantly more accurate variant (SNP/INDEL) calling relative to exome
  - Price for WGS comparable to exome for germline-only projects

Exome vs Whole Genome Sequencing

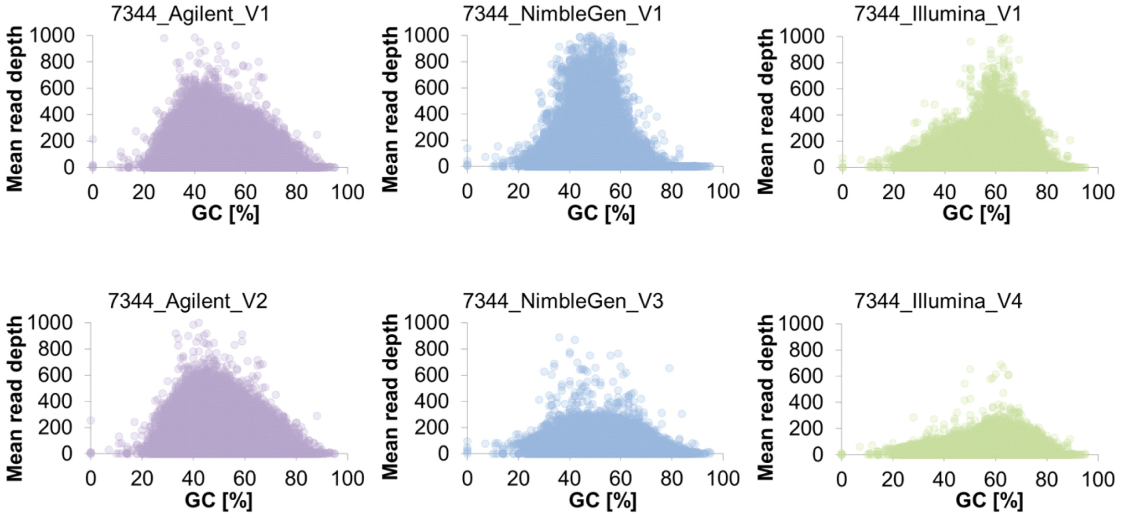
- Depth variance MUCH higher for exome
- ~2-fold more variants with GQ < 20 for exome
- Read ratio for heterozygous variants significantly skewed for exome
  - Especially pronounced for INDELs

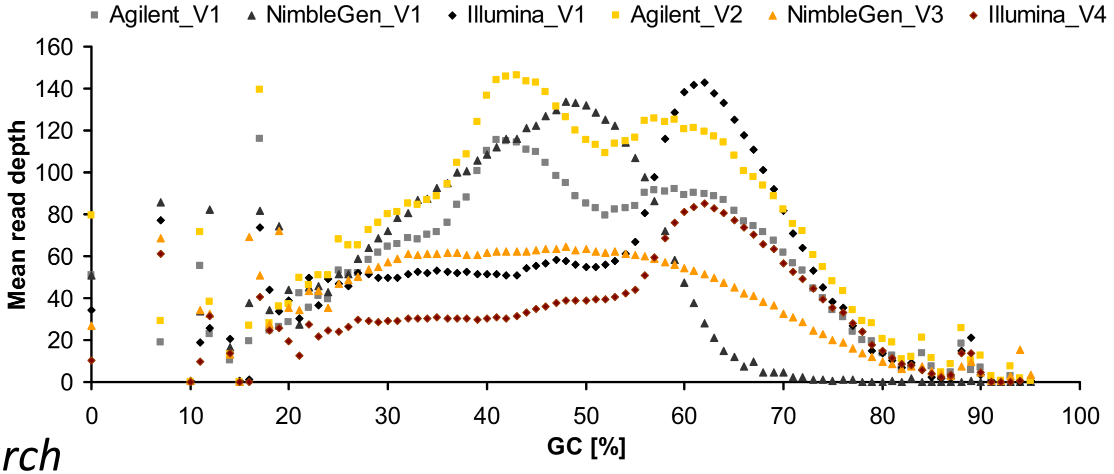


Belkadi et al., 2015 *PNAS*

Exome Capture Considerations

- Significant capture and enrichment biases for different kits
- Illustrates issue with combining samples from multiple kits
- For germline-only analysis, WGS strongly preferred

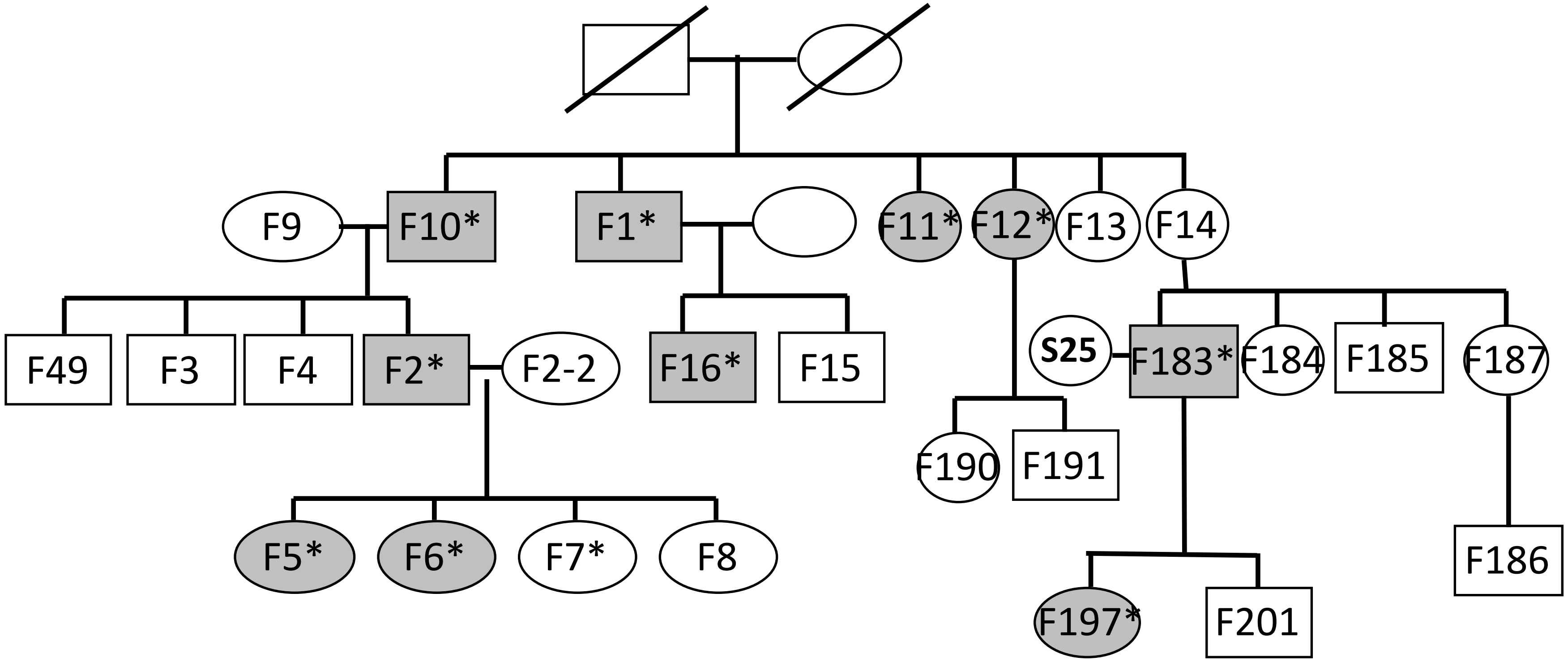




Meienberg et al., 2015 *Nucleic Acids Research*

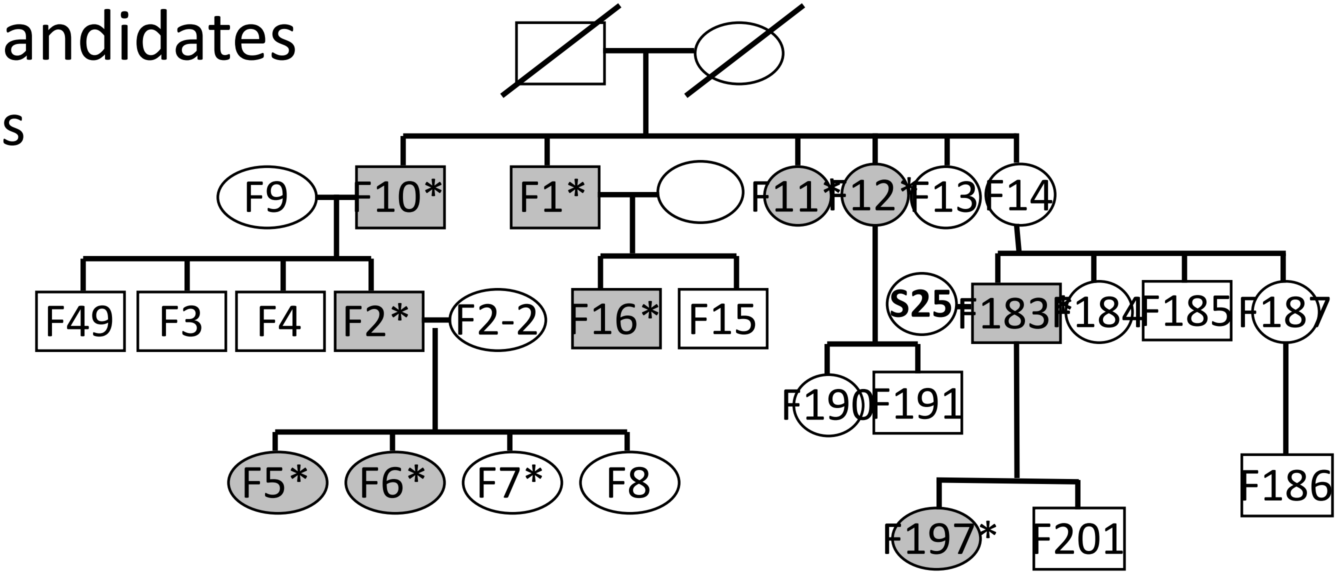
Familial Sequencing Design

- Power is the primary limiting factor
- When budgets are limited, decisions have to be made about who to sequence



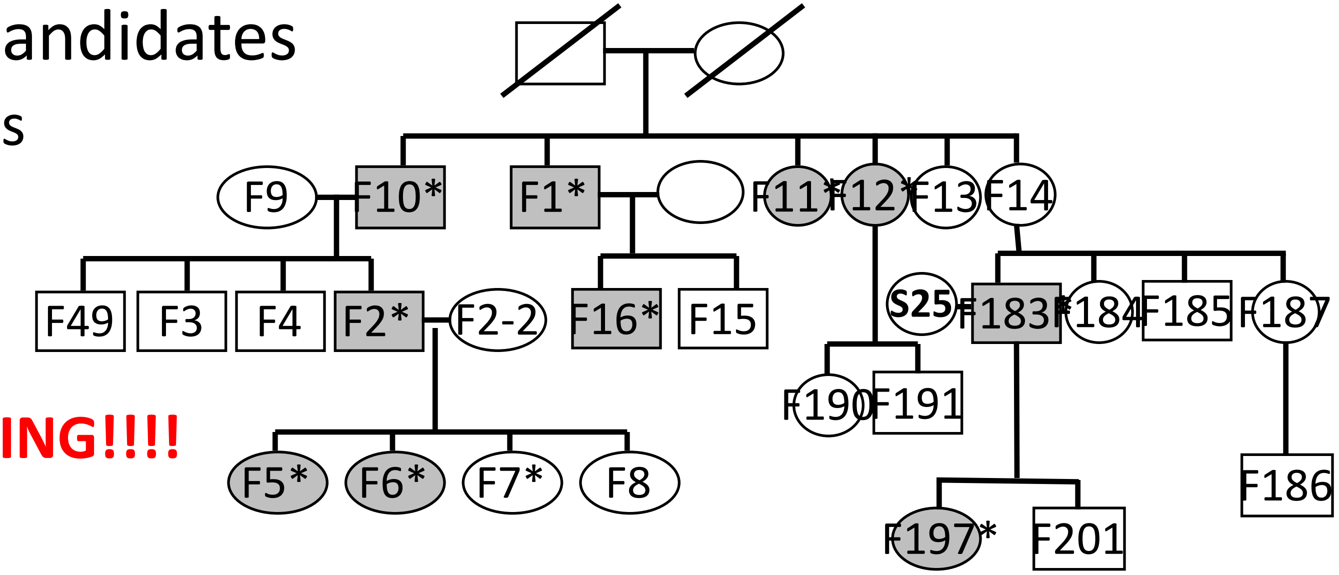
Familial Sequencing Design

- 3 cases, no controls
  - 3,176 candidates
- 3 cases, 1 spousal control (ethinicity matched) - 1542 candidates
  - +1 spouse controls - 1121 candidates
  - +1 case - 525 candidates
- 3 cases, 1 related control - 854 candidates
  - +1 related control - 307 candidates
  - +1 case – 284 candidates



Familial Sequencing Design

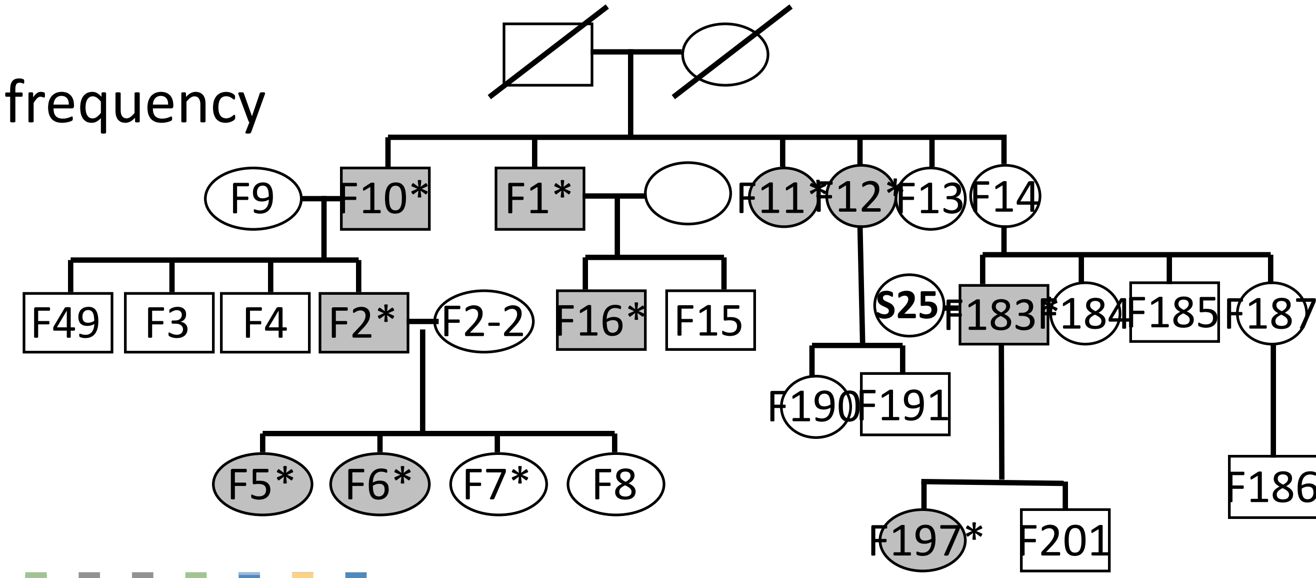
- 3 cases, no controls
  - 3,176 candidates
- 3 cases, 1 spousal control (ethinicity matched) - 1542 candidates
  - +1 spouse controls - 1121 candidates
  - +1 case - 525 candidates
- 3 cases, 1 related control - 854 candidates
  - +1 related control - 307 candidates
  - +1 case – 284 candidates



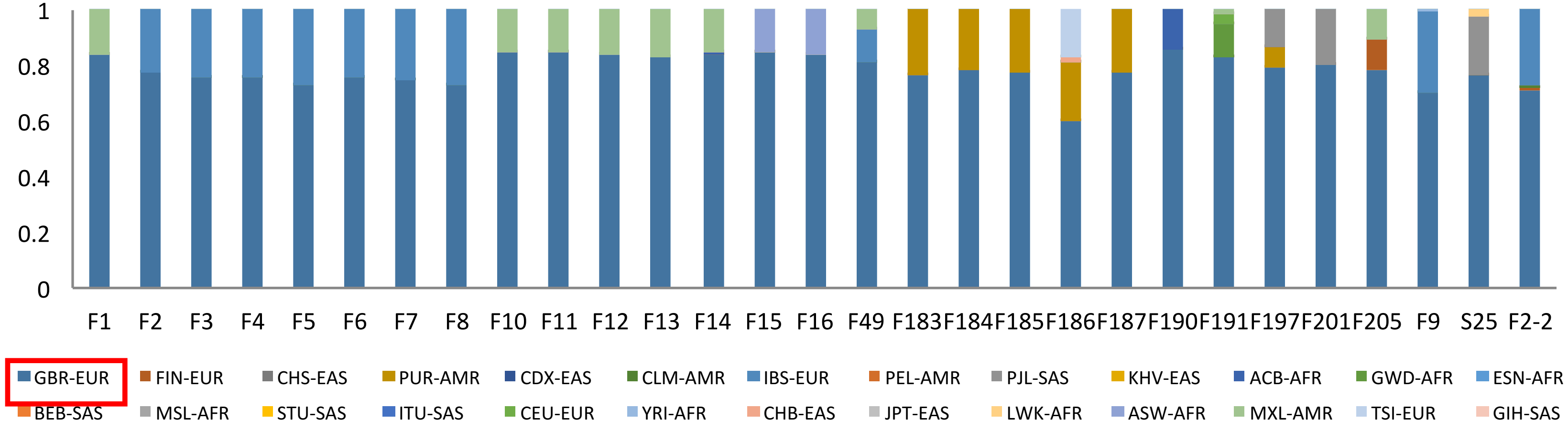
ALWAYS PERFORM ETHNICITY-AWARE FILTERING!!!!

Familial Sequencing Design

- 3 cases, no controls
  - 3,176 candidates with global allele frequency threshold of ≤0.01
  - 2,923 candidates with EUR-only!

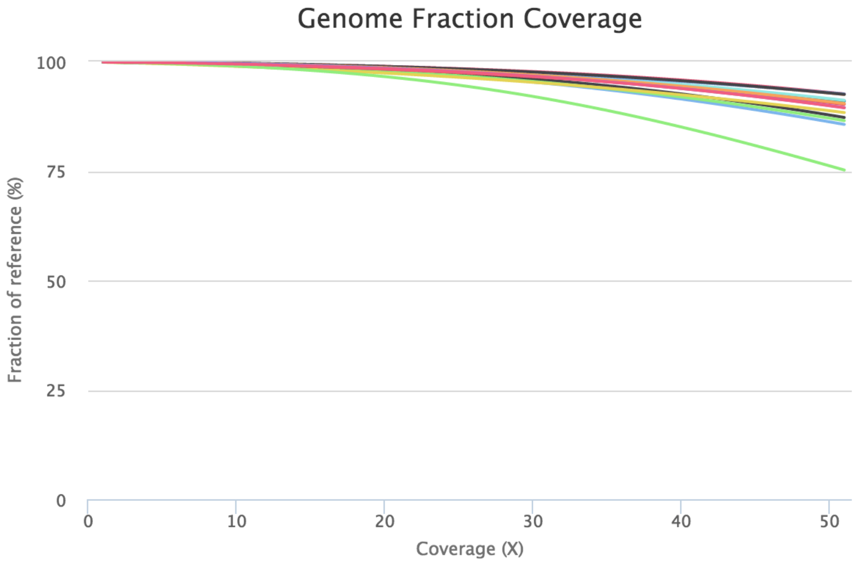


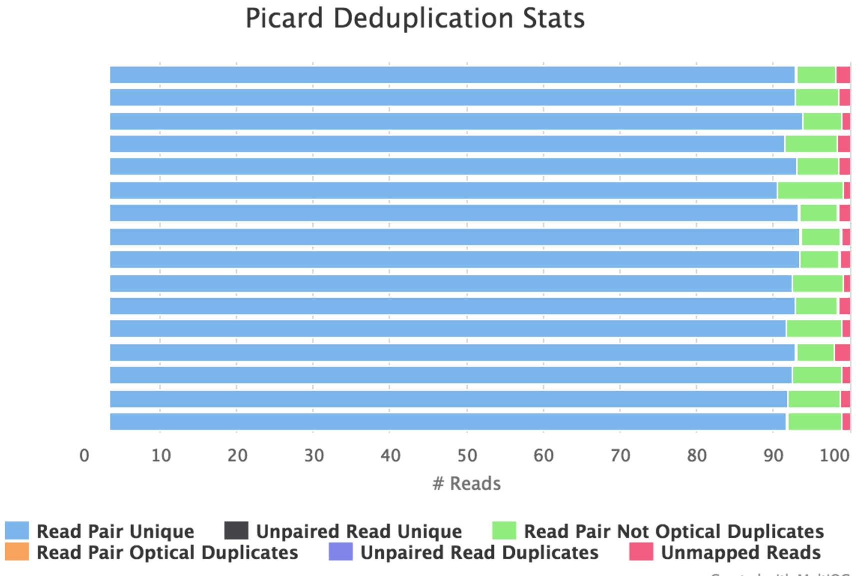
Family 1 Admixture



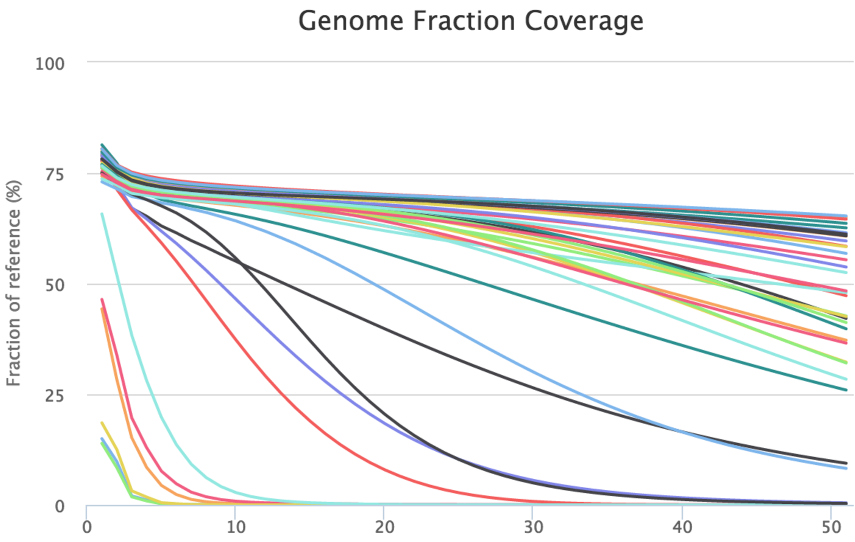
FFPE vs Fresh/Frozen Tissue – 50X target depth

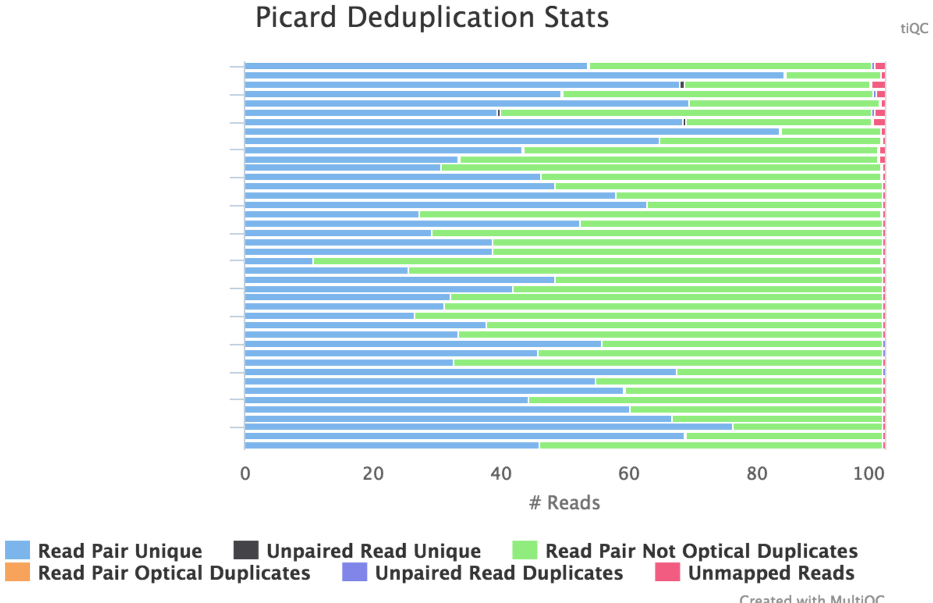
Fresh/Frozen





FFPE



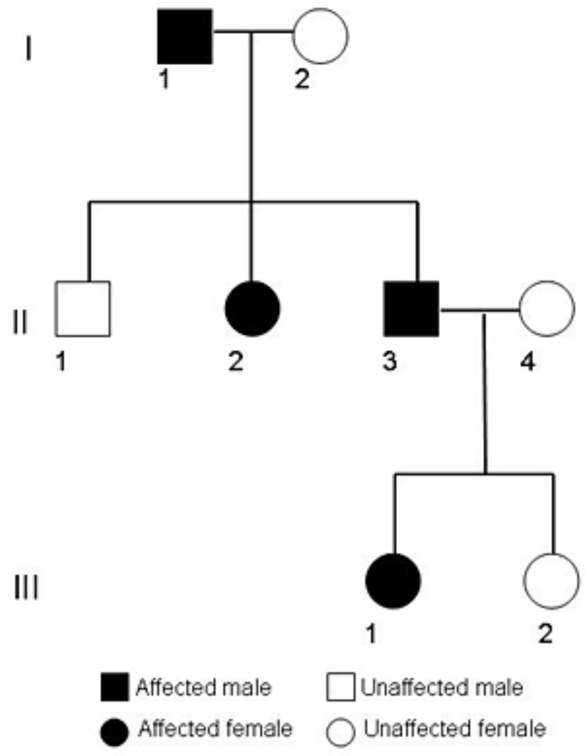


Somatic Variant Calling – Best Practices

- 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

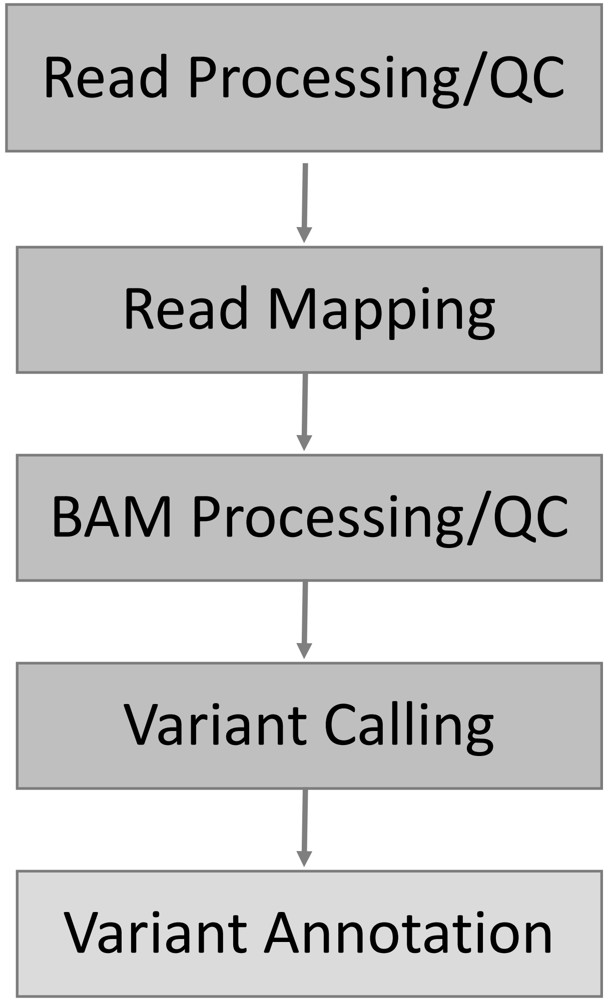
- Whole genome strongly preferred
  - >=30X mean target depth
  - Superior to exome for structural variants, copy number analysis
- Germline exome
  - >=50X mean depth
- For familial/trio analyses, we strongly encourage early consultation
  - Selection of samples for sequencing can be CRUCIAL to maximizing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



Pipeline Details…

Pipeline Details…

- All variant calling follows the same basic approach



Read Processing/QC

Read Mapping

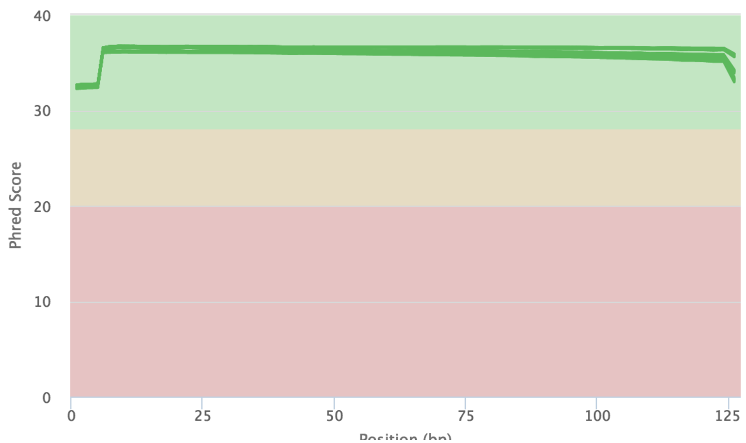
BAM Processing/QC

Variant Calling

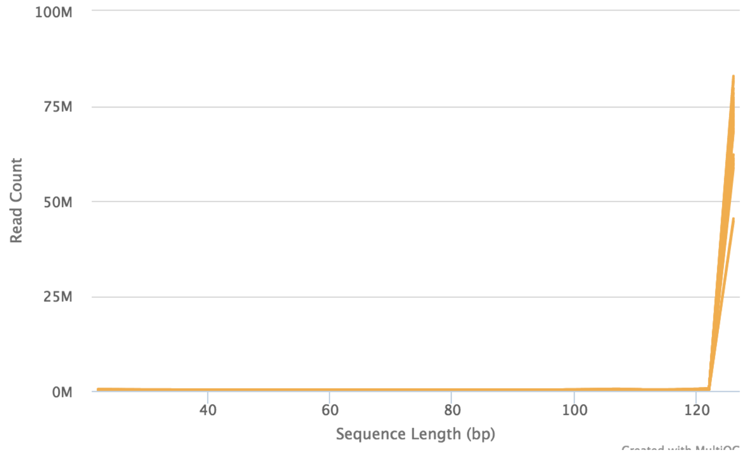
Variant Annotation

Pipeline Details…

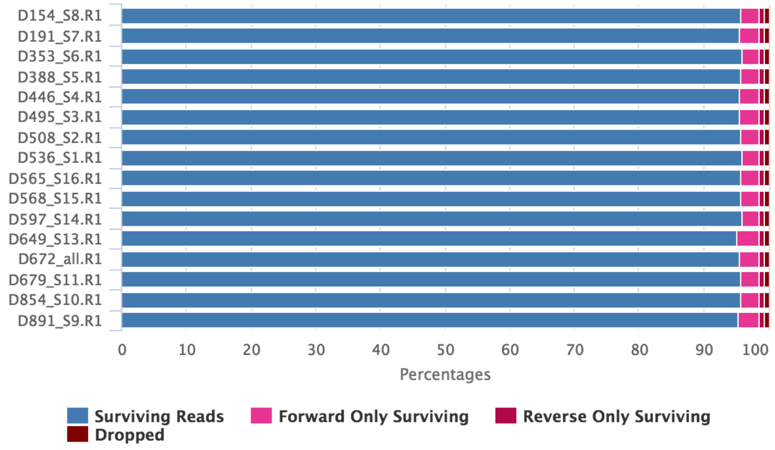
Mean Quality Scores



Sequence Length Distribution

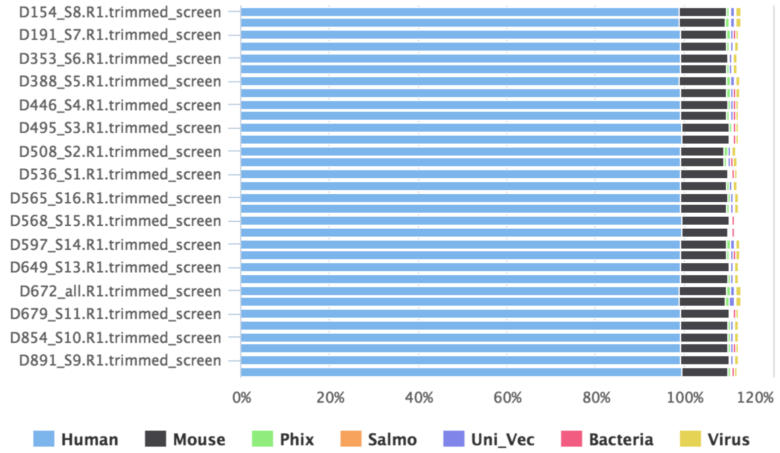


Trimmomatic

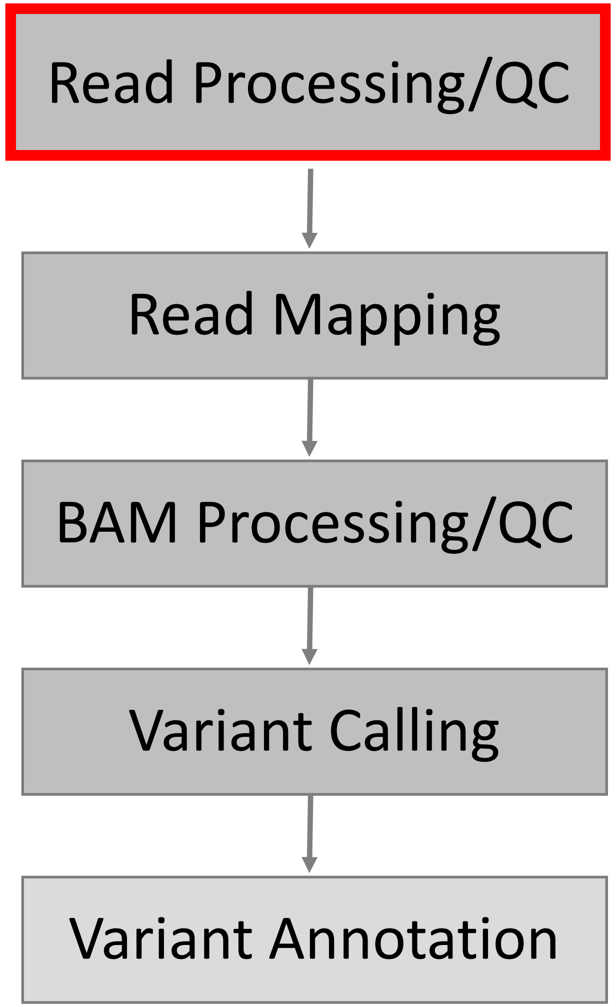


Created with MultiQC

FastQ Screen



Created with MultiQC



Read Processing/QC

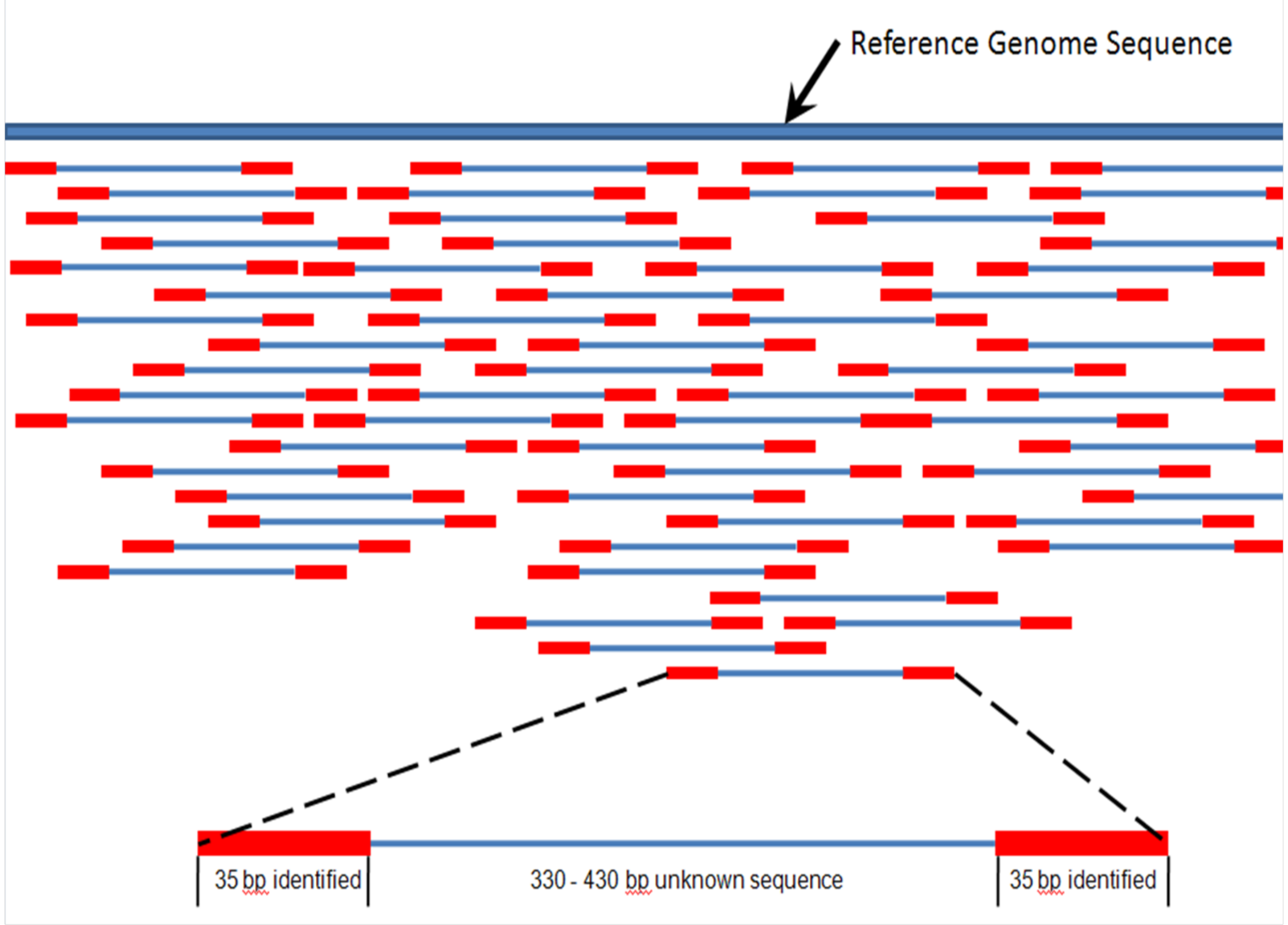
Read Mapping

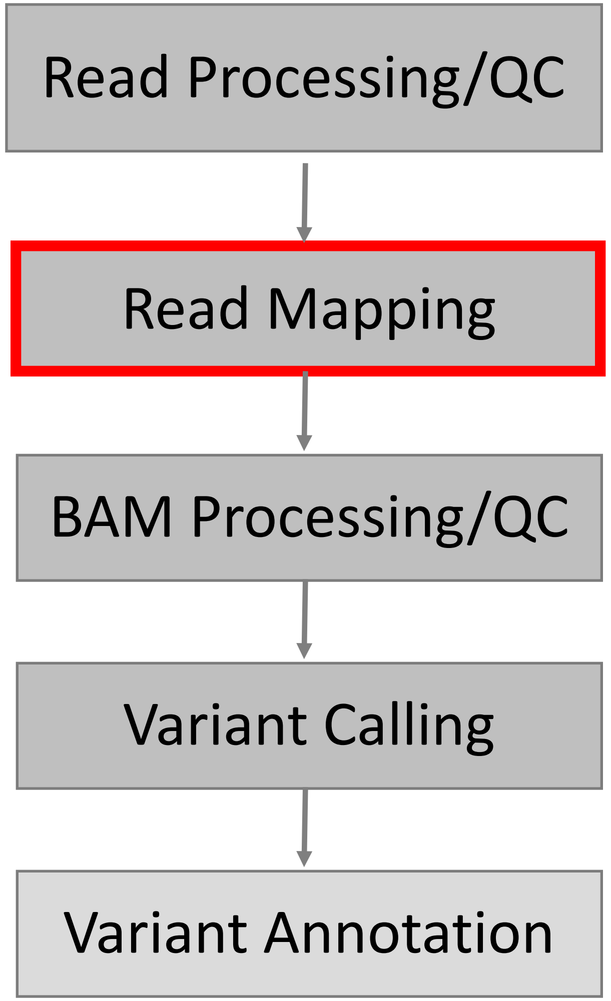
BAM Processing/QC

Variant Calling

Variant Annotation

Pipeline Details…





Read Processing/QC

Read Mapping

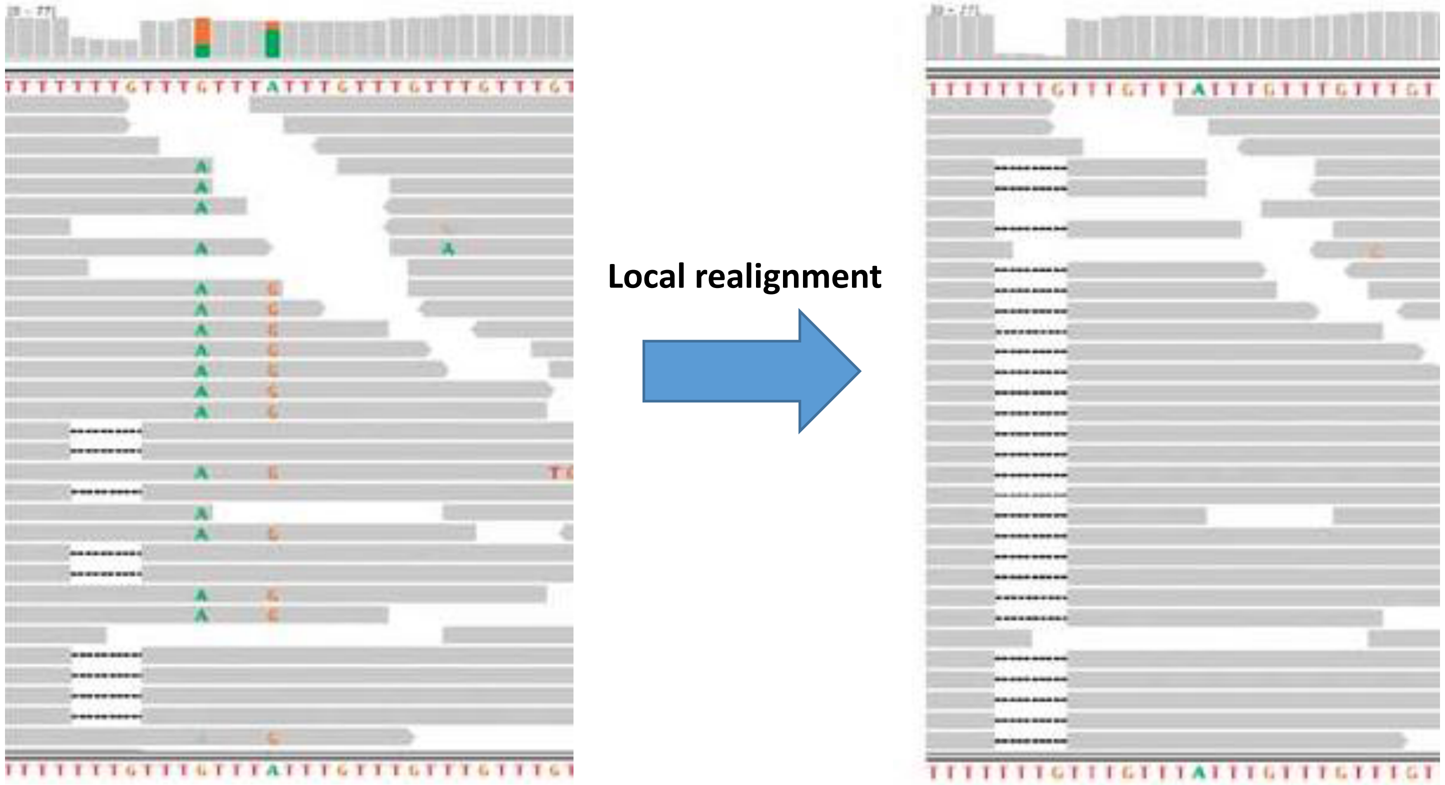
BAM Processing/QC

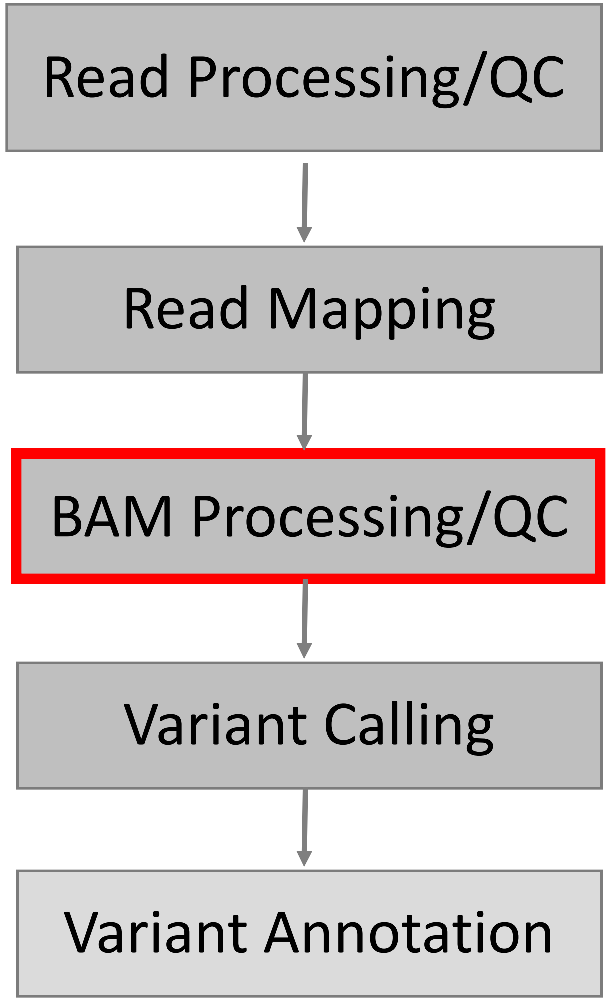
Variant Calling

Variant Annotation

Pipeline Details…

- Indel realignment




Read Processing/QC

Read Mapping

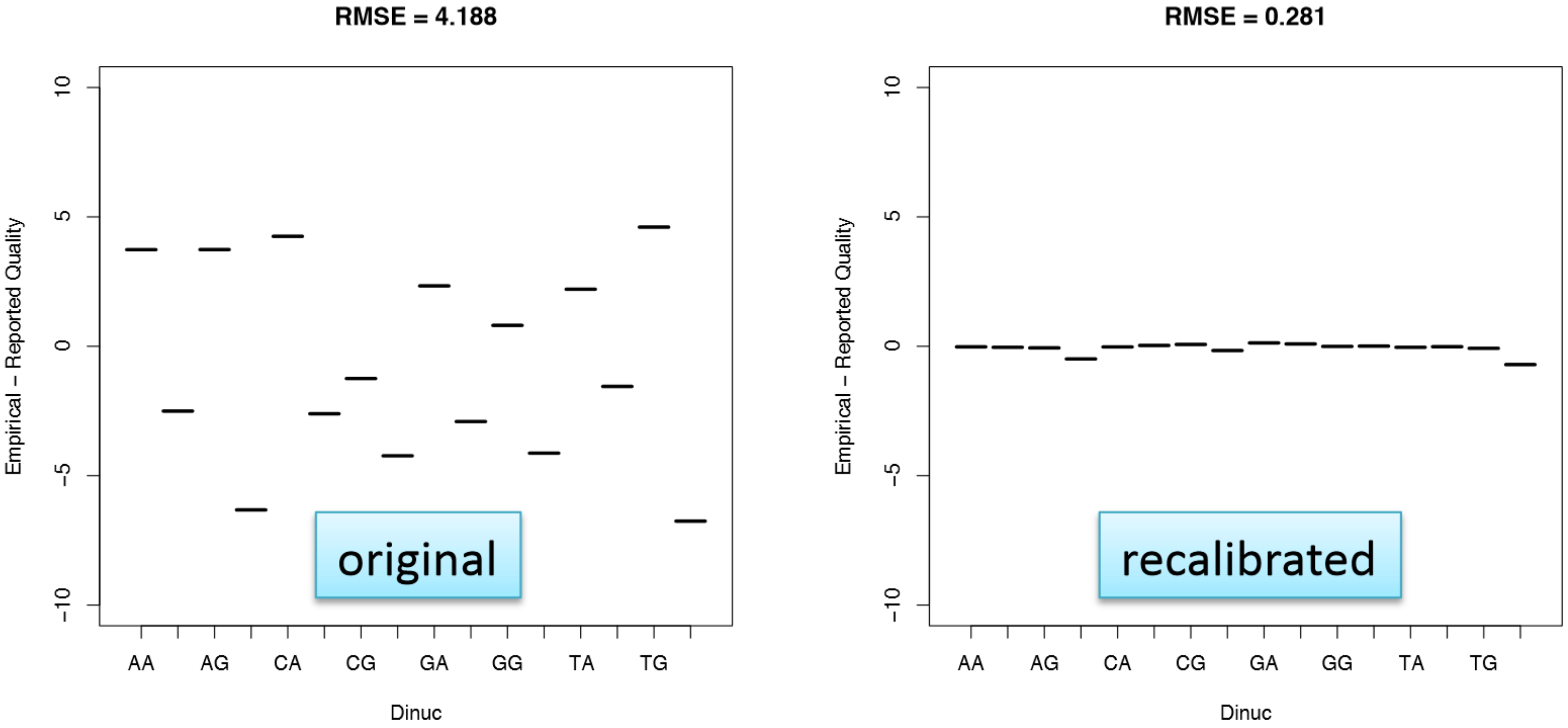
BAM Processing/QC

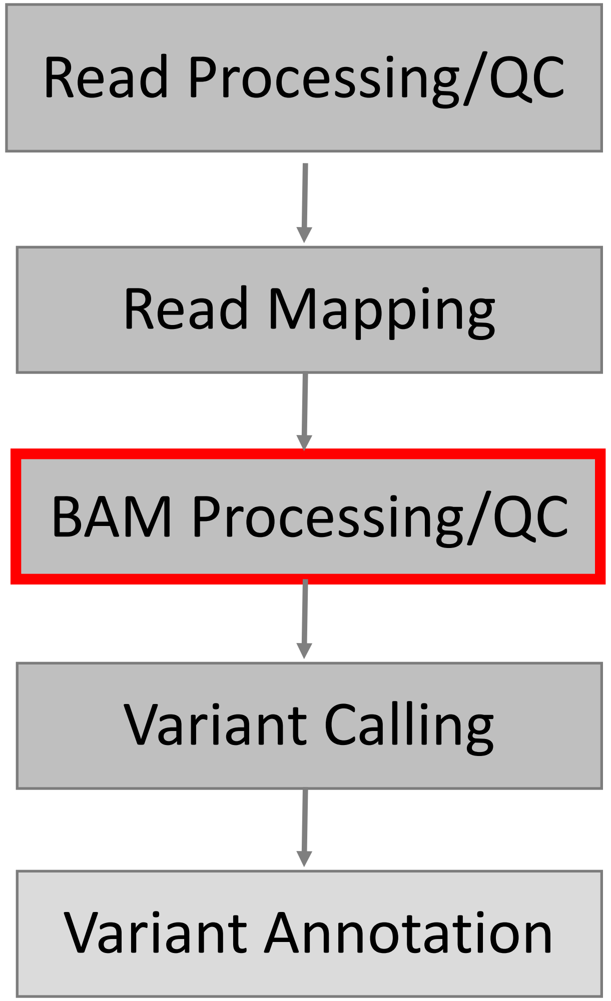
Variant Calling

Variant Annotation

Pipeline Details…

- Multiple sources of quality score bias





Read Processing/QC

Read Mapping

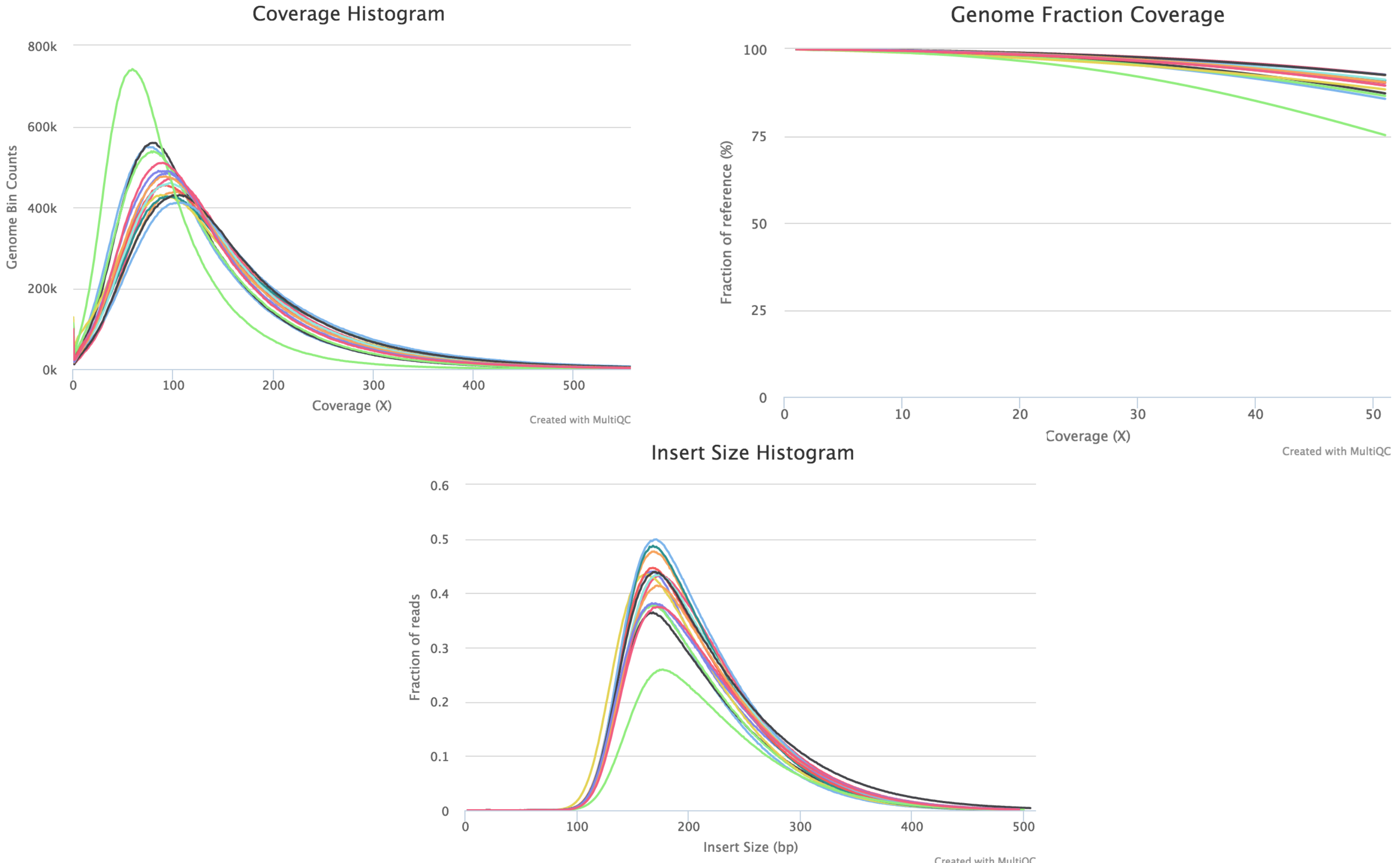
BAM Processing/QC

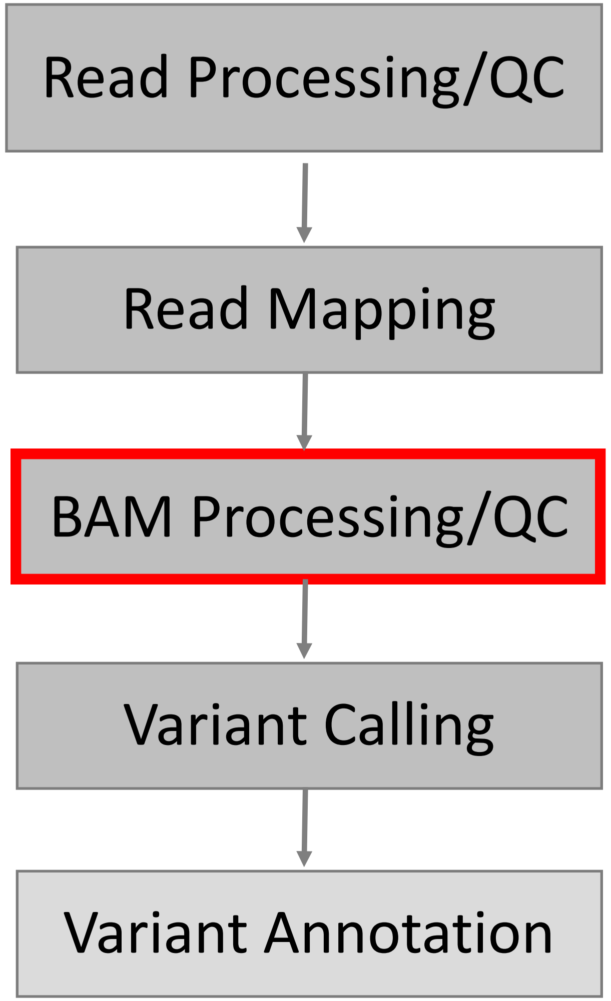
Variant Calling

Variant Annotation

Pipeline Details…

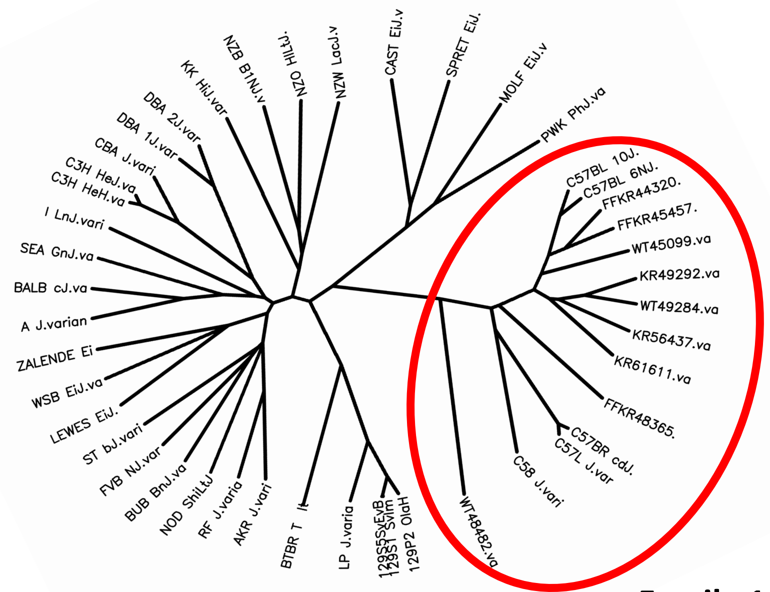
- Alignment QC

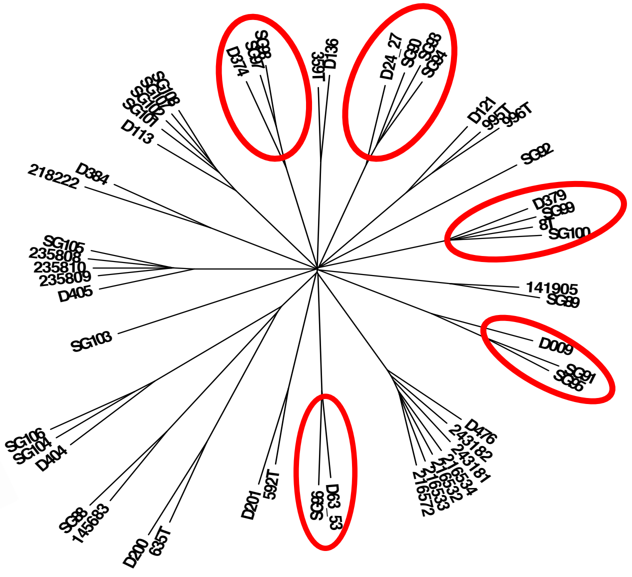


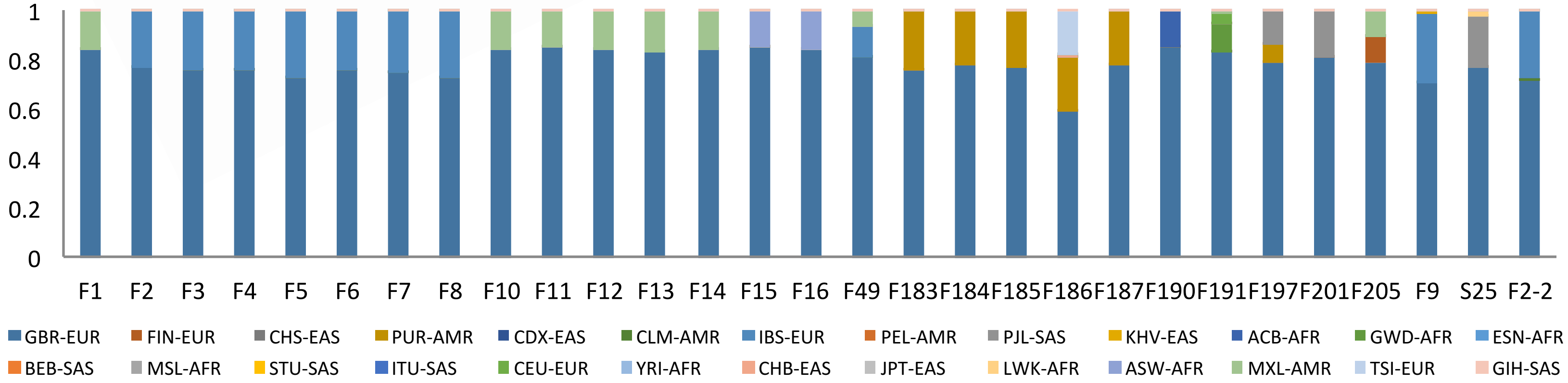


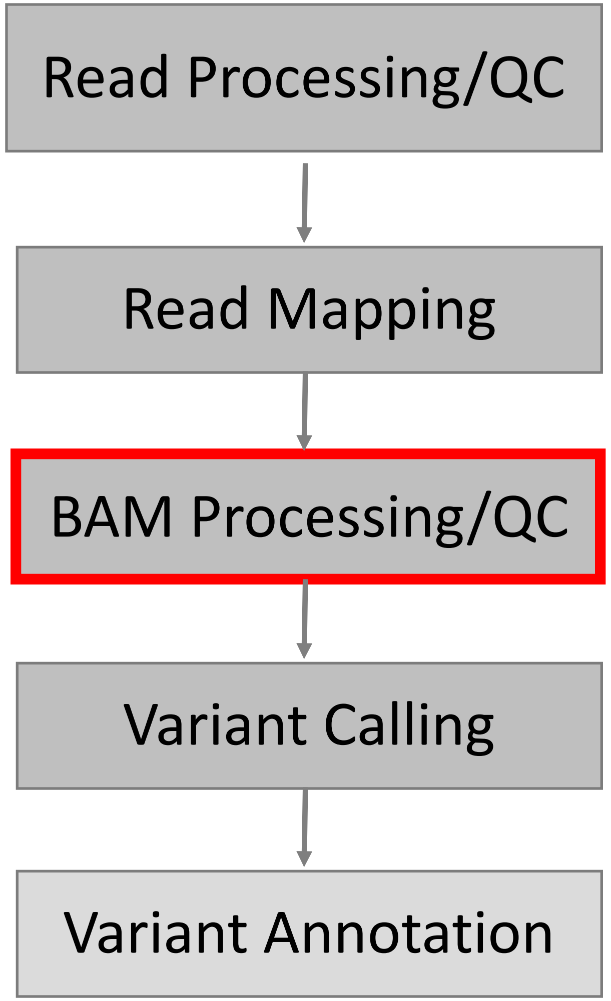
Variant Calling at CCBR

- Additional QC






Family 1 Admixture


Read Processing/QC

Read Mapping

BAM Processing/QC

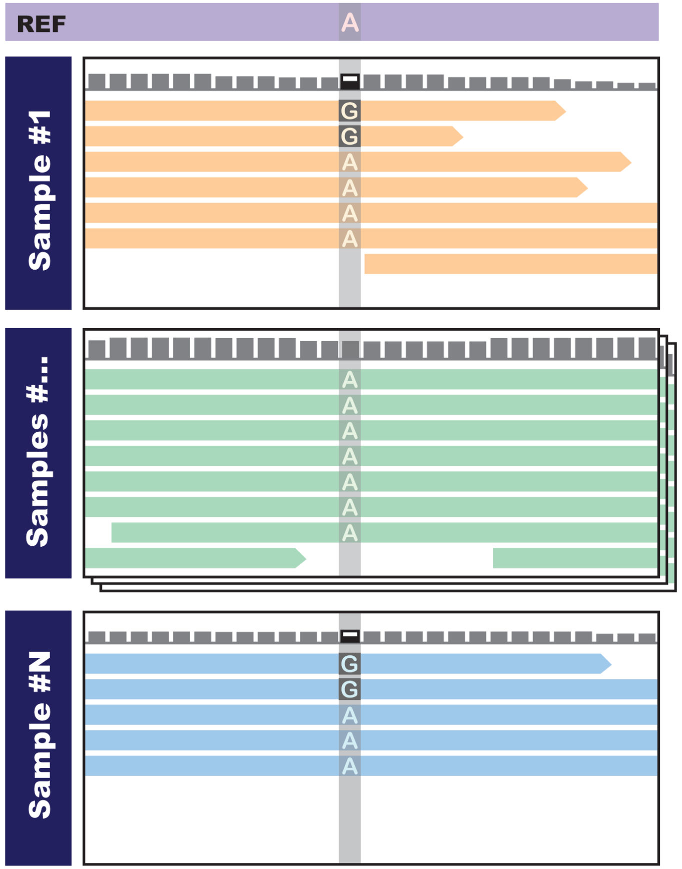
Variant Calling

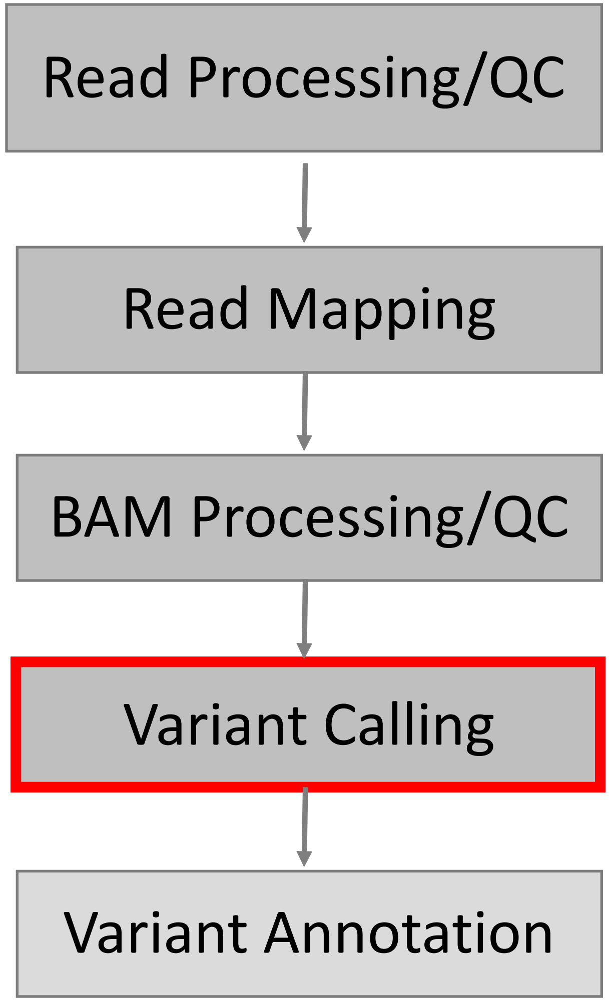
Variant Annotation

Variant Calling at CCBR

Germline

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





Read Processing/QC

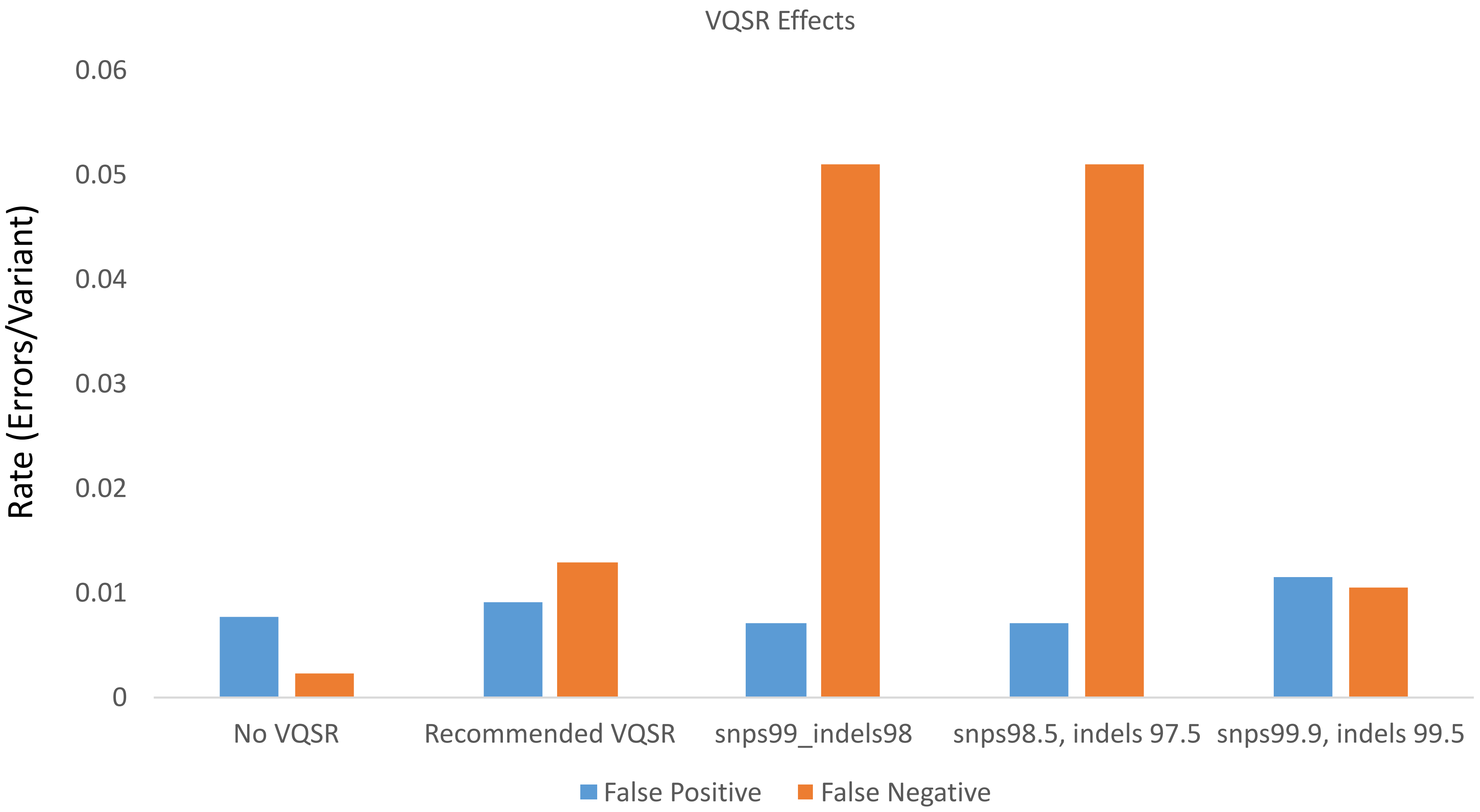
Read Mapping

BAM Processing/QC

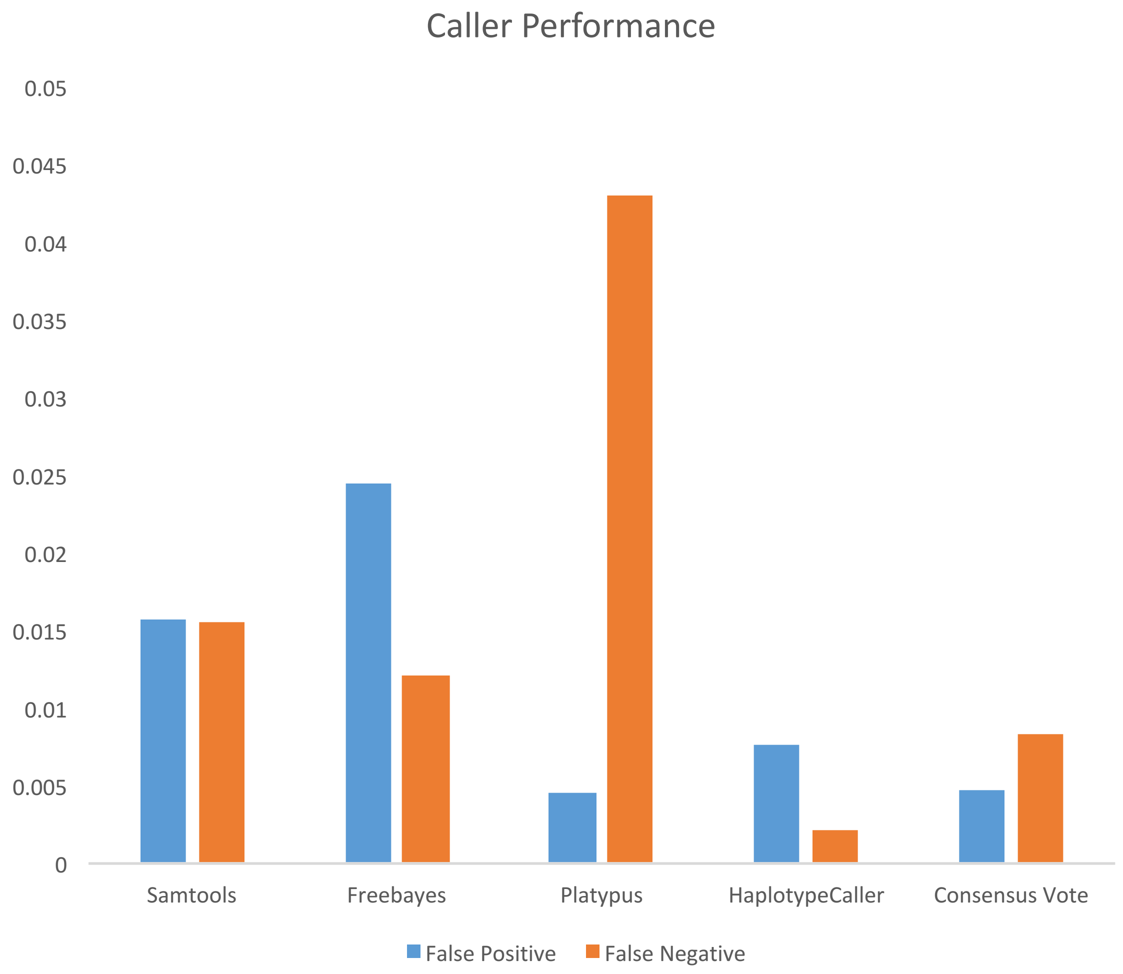
Variant Calling

Variant Annotation

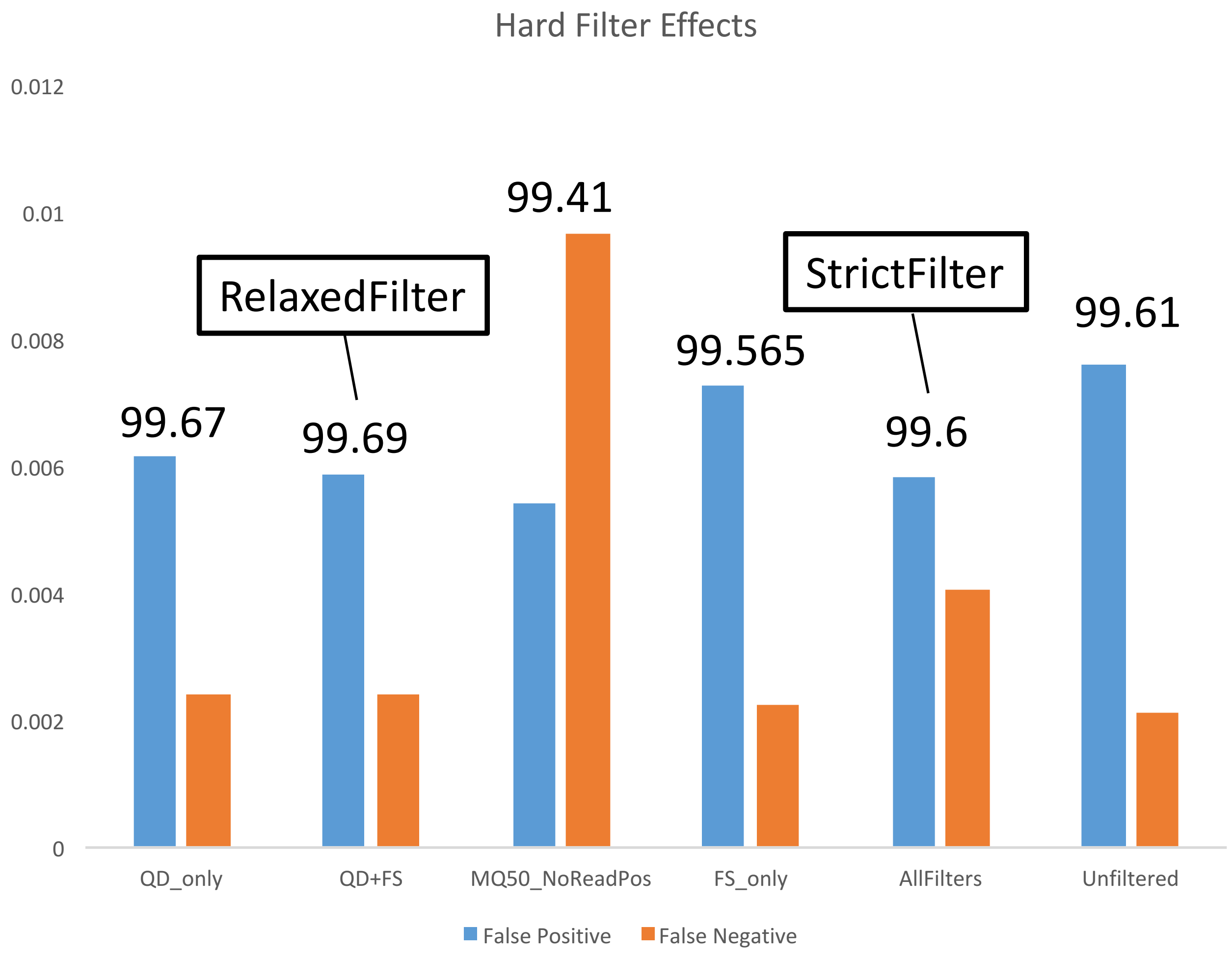
GATK - Variant Quality Score Recalibration (VQSR)



Variant Caller Performance and Filtering



Caller Performance

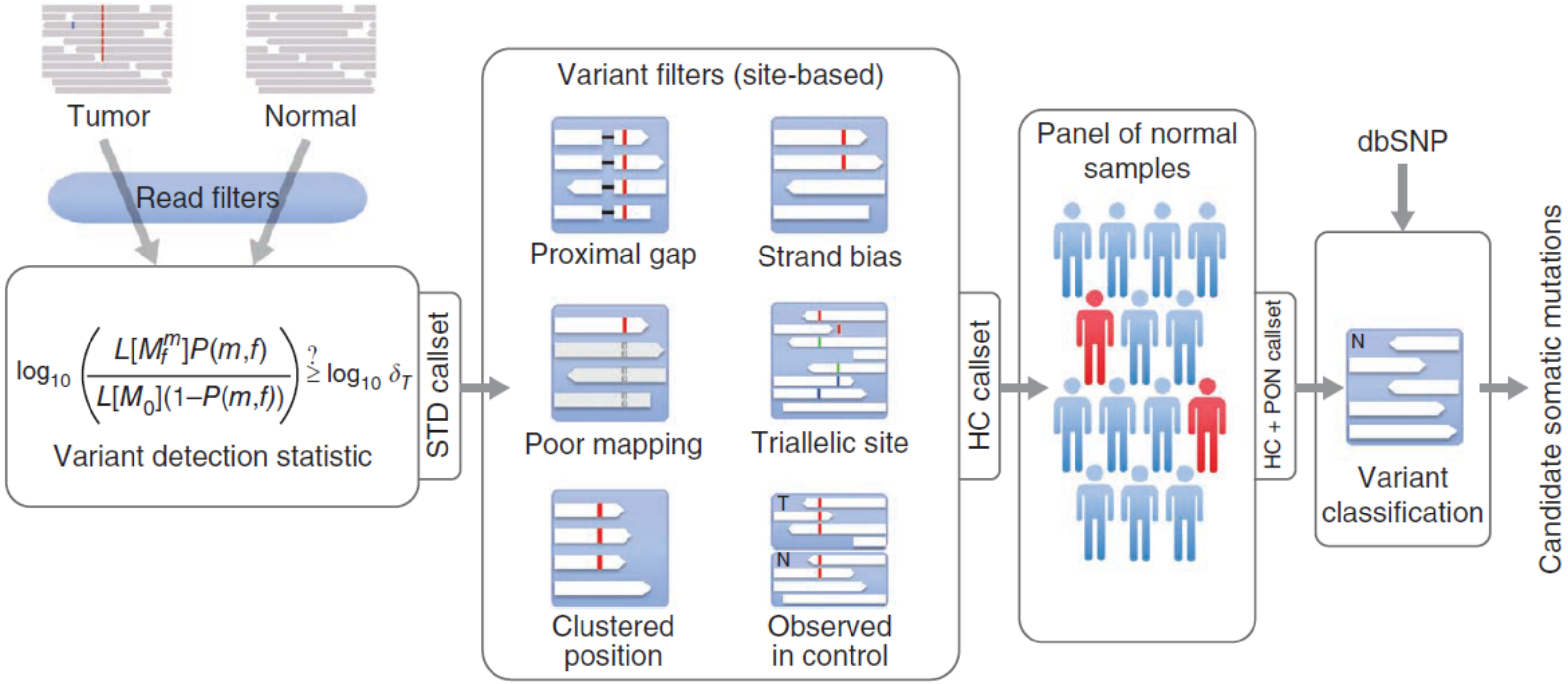


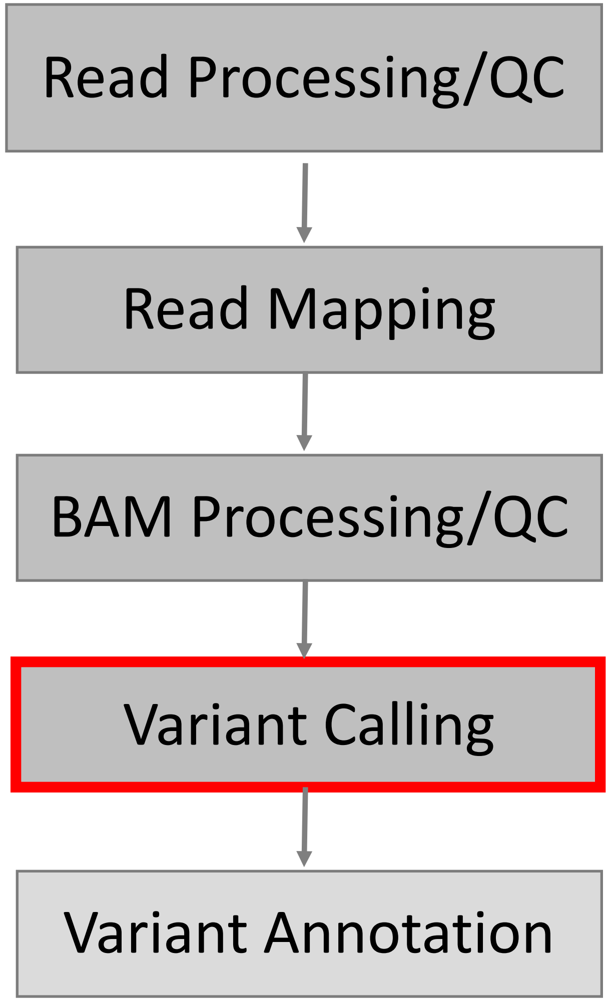
Hard Filter Effects

Variant Calling at CCBR

Somatic

- MuTect, MuTect2 (with hard filters), Strelka

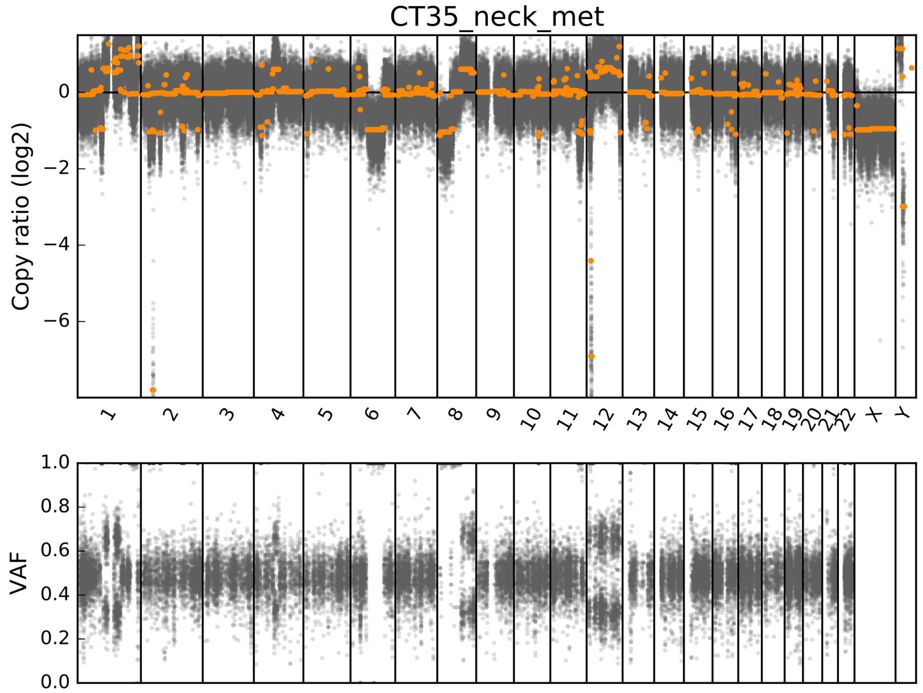


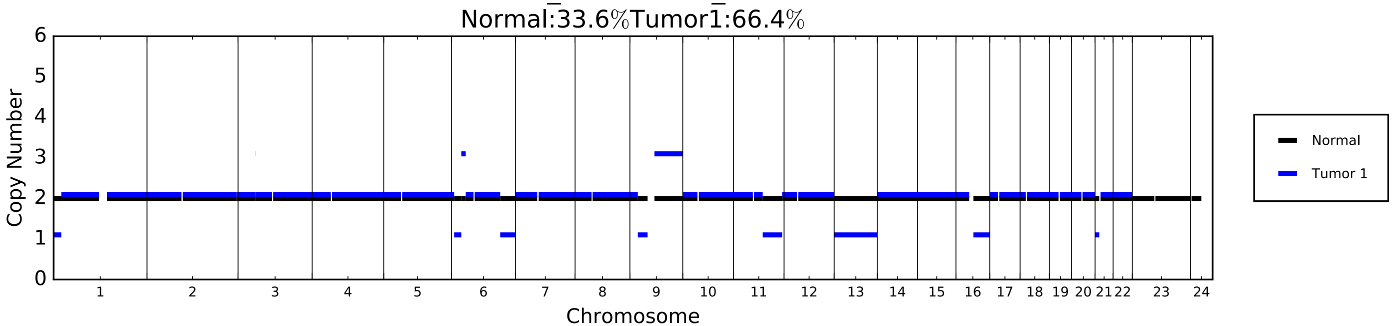


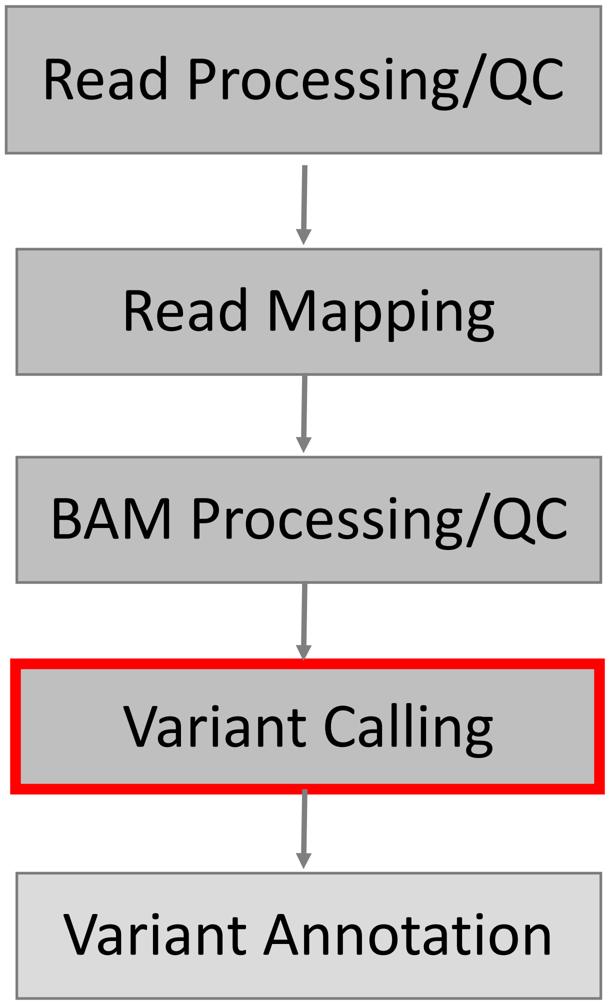
Variant Calling at CCBR

Somatic

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







Read Processing/QC

Read Mapping

BAM Processing/QC

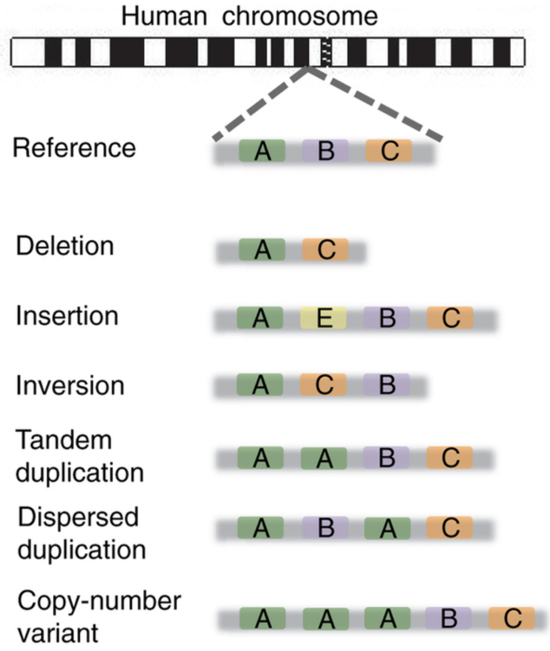
Variant Calling

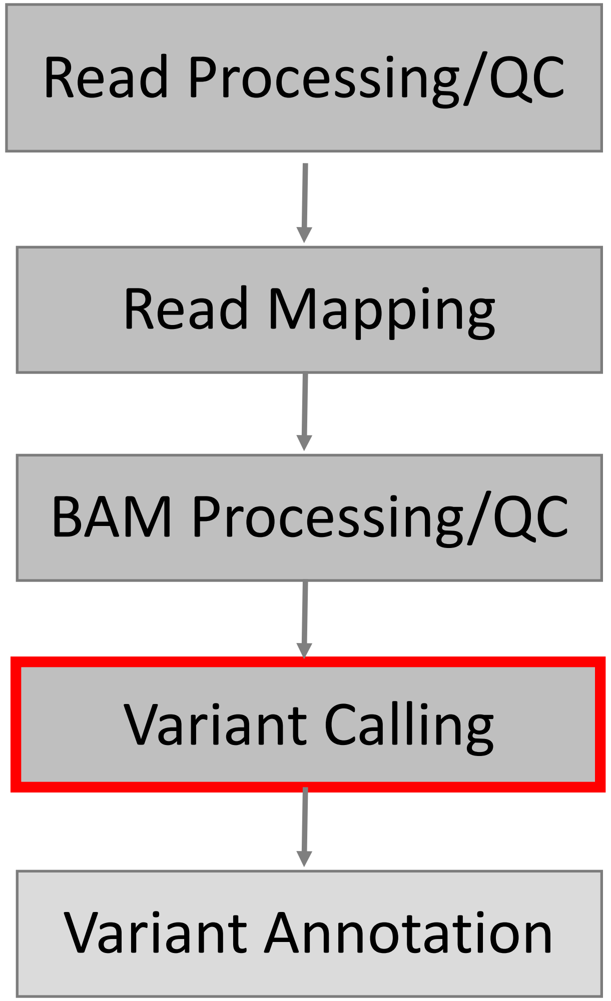
Variant Annotation

Variant Calling at CCBR

Somatic

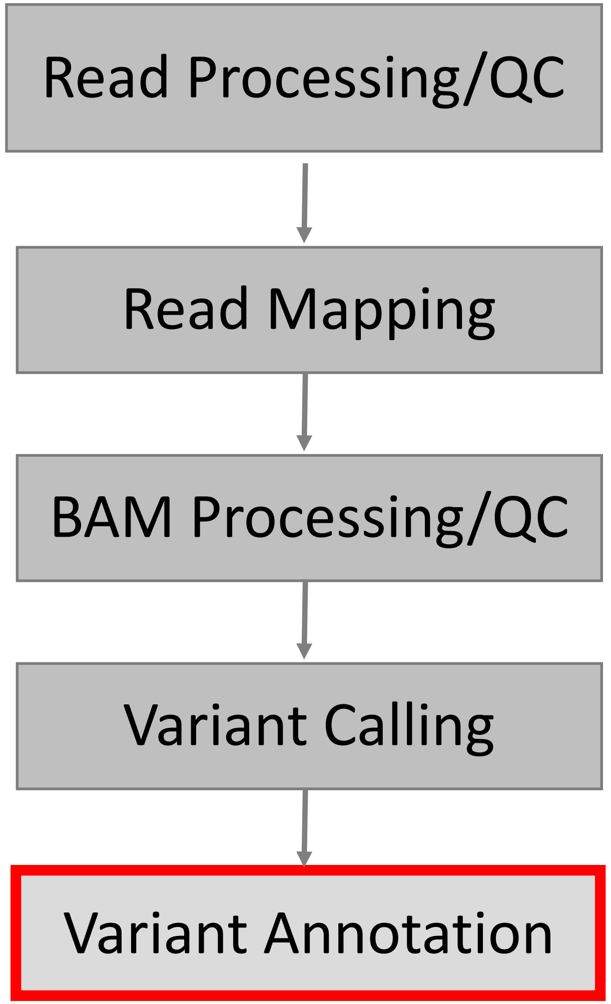
- MuTect, MuTect2, Strelka
- Copy number – CNVkit, THetA2
- Structural Variation
  - MANTA
  - DELLY



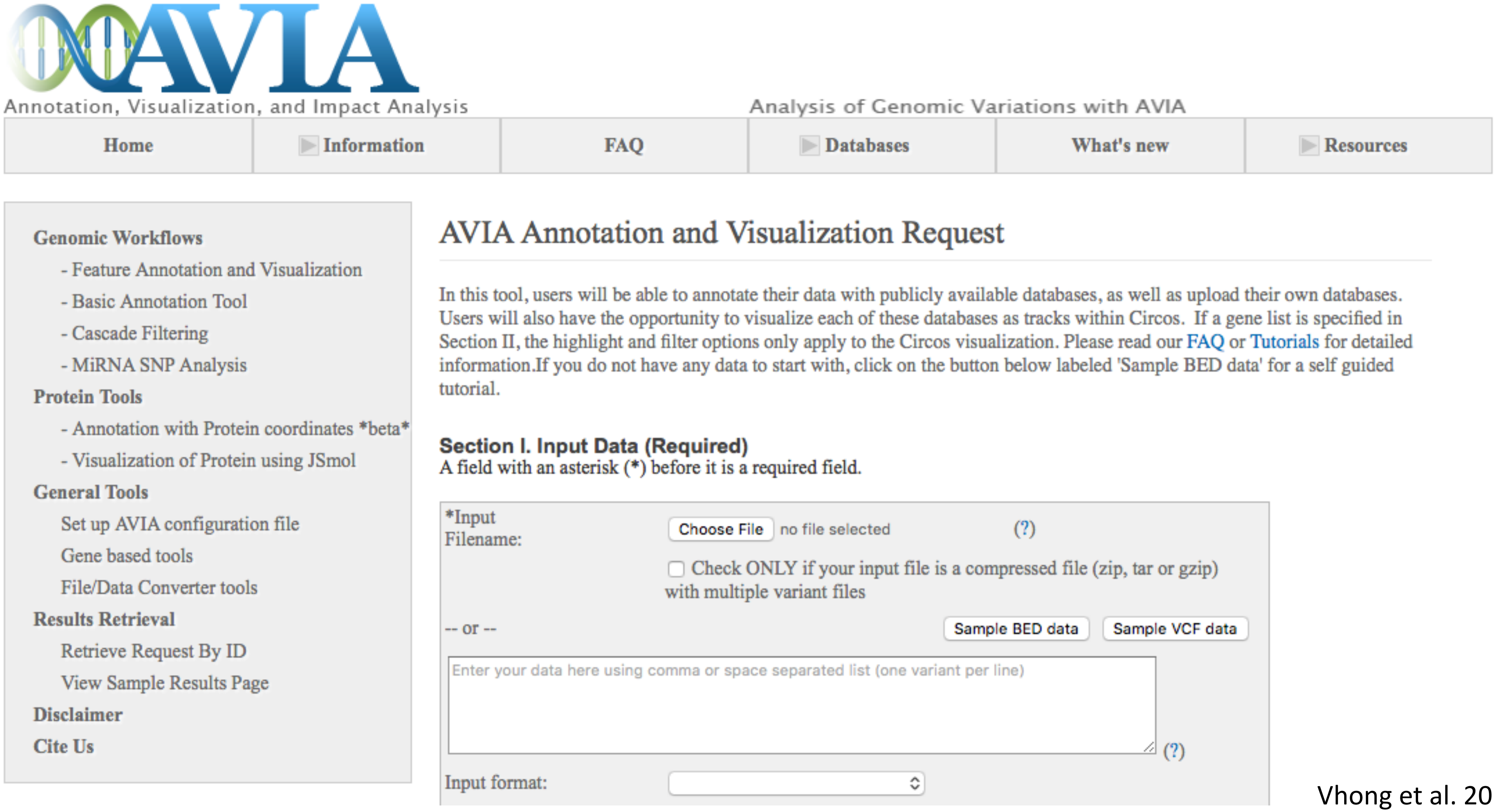


Variant Calling at CCBR

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



Variant Annotation – AVIA



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.

*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) (?)

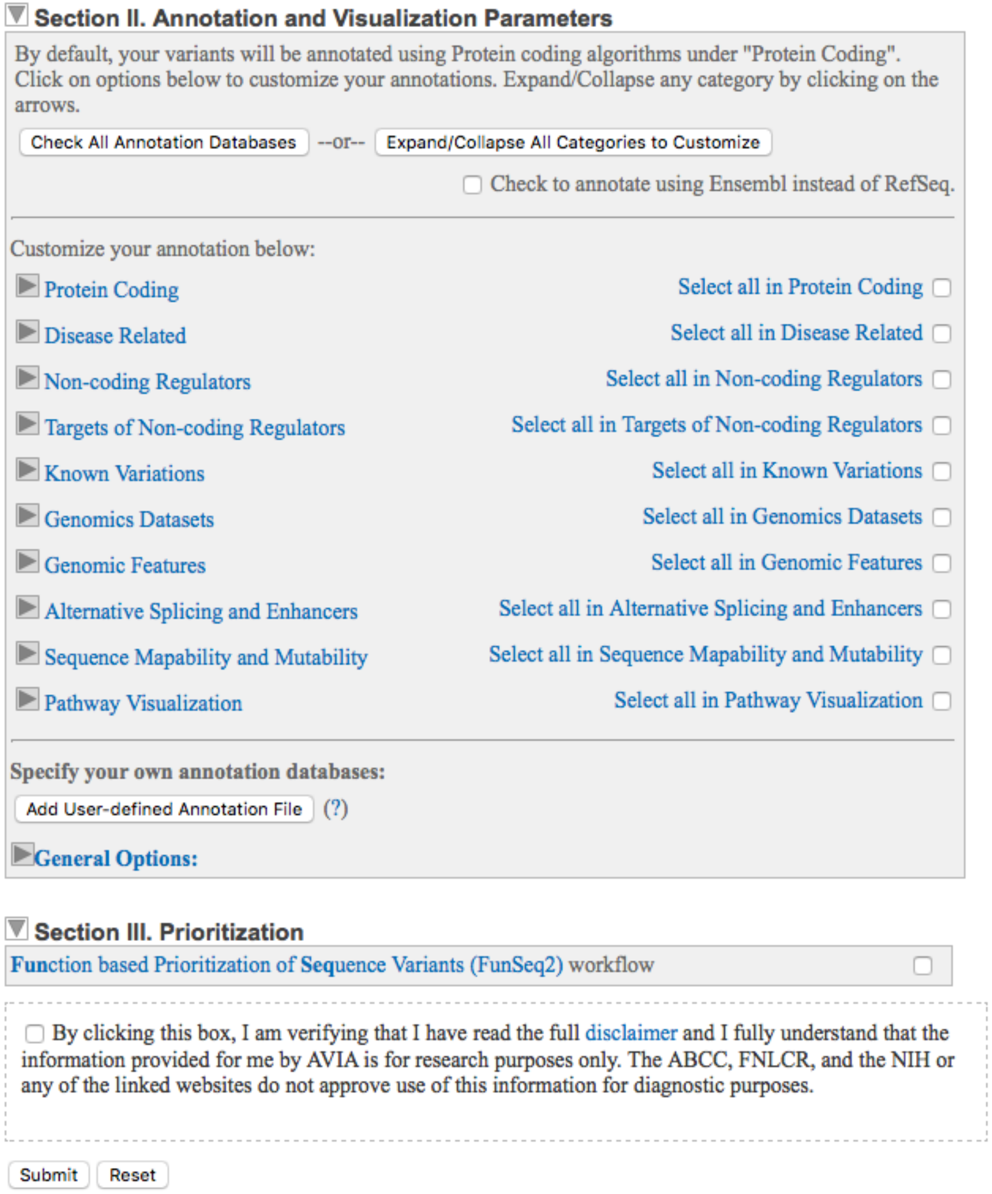
Input format:

Vhong et al. 2015, Bioinformatics

Variant Annotation - AVIA

- Created and maintained at NCI-Frederick by ABCC team members
    - Hue Vhong and Uma Mudunuri
- Comprehensive annotation of human and mouse genomes
- Flexible input/output format
    - VCF and BED inputs
    - Tabular and annotated VCF outputs
- Highly customizable annotations
- hg19/GRCh37, hg18, mm10 currently available
- hg38 available in the very near future

AVIA Annotations



Section II. Annotation and Visualization Parameters

By default, your variants will be annotated using Protein coding algorithms under "Protein Coding". Click on options below to customize your annotations. Expand/Collapse any category by clicking on the arrows.

Check All Annotation Databases --or-- Expand/Collapse All Categories to Customize

□ Check to annotate using Ensembl instead of RefSeq.

Customize your annotation below:

▶ Protein Coding	Select all in Protein Coding □
▶ Disease Related	Select all in Disease Related □
▶ Non-coding Regulators	Select all in Non-coding Regulators □
▶ Targets of Non-coding Regulators	Select all in Targets of Non-coding Regulators □
▶ Known Variations	Select all in Known Variations □
▶ Genomics Datasets	Select all in Genomics Datasets □
▶ Genomic Features	Select all in Genomic Features □
▶ Alternative Splicing and Enhancers	Select all in Alternative Splicing and Enhancers □
▶ Sequence Mapability and Mutability	Select all in Sequence Mapability and Mutability □
▶ Pathway Visualization	Select all in Pathway Visualization □

Specify your own annotation databases:

Add User-defined Annotation File (?)

▶**General Options:**

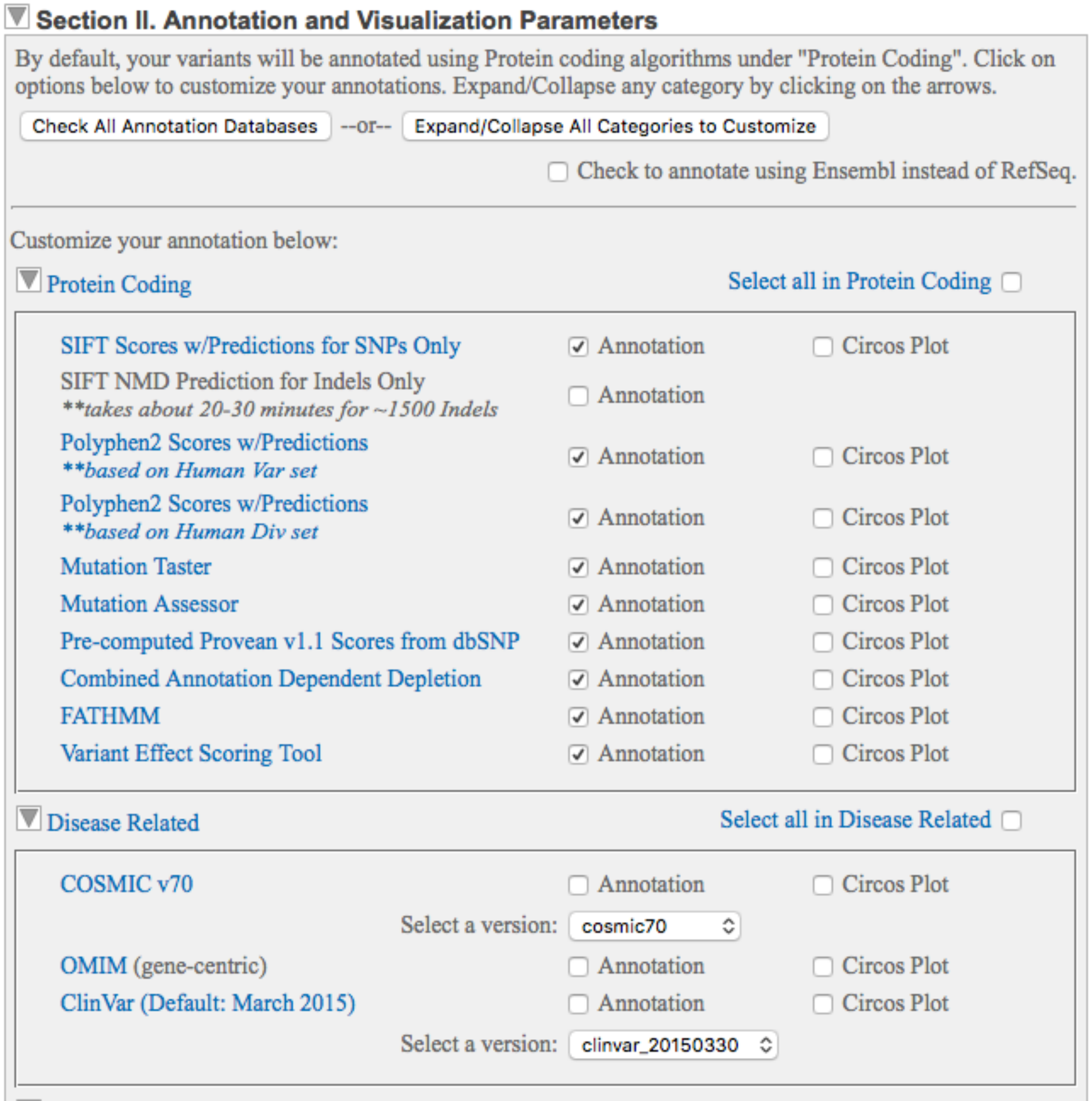
Section III. Prioritization

Function based Prioritization of Sequence Variants (FunSeq2) workflow □

□ By clicking this box, I am verifying that I have read the full disclaimer and I fully understand that the information provided for me by AVIA is for research purposes only. The ABCC, FNLCR, and the NIH or any of the linked websites do not approve use of this information for diagnostic purposes.

Submit Reset

AVIA Annotations



Section II. Annotation and Visualization Parameters

By default, your variants will be annotated using Protein coding algorithms under "Protein Coding". Click on options below to customize your annotations. Expand/Collapse any category by clicking on the arrows.

Check All Annotation Databases
--or--
Expand/Collapse All Categories to Customize

□ Check to annotate using Ensembl instead of RefSeq.

Customize your annotation below:

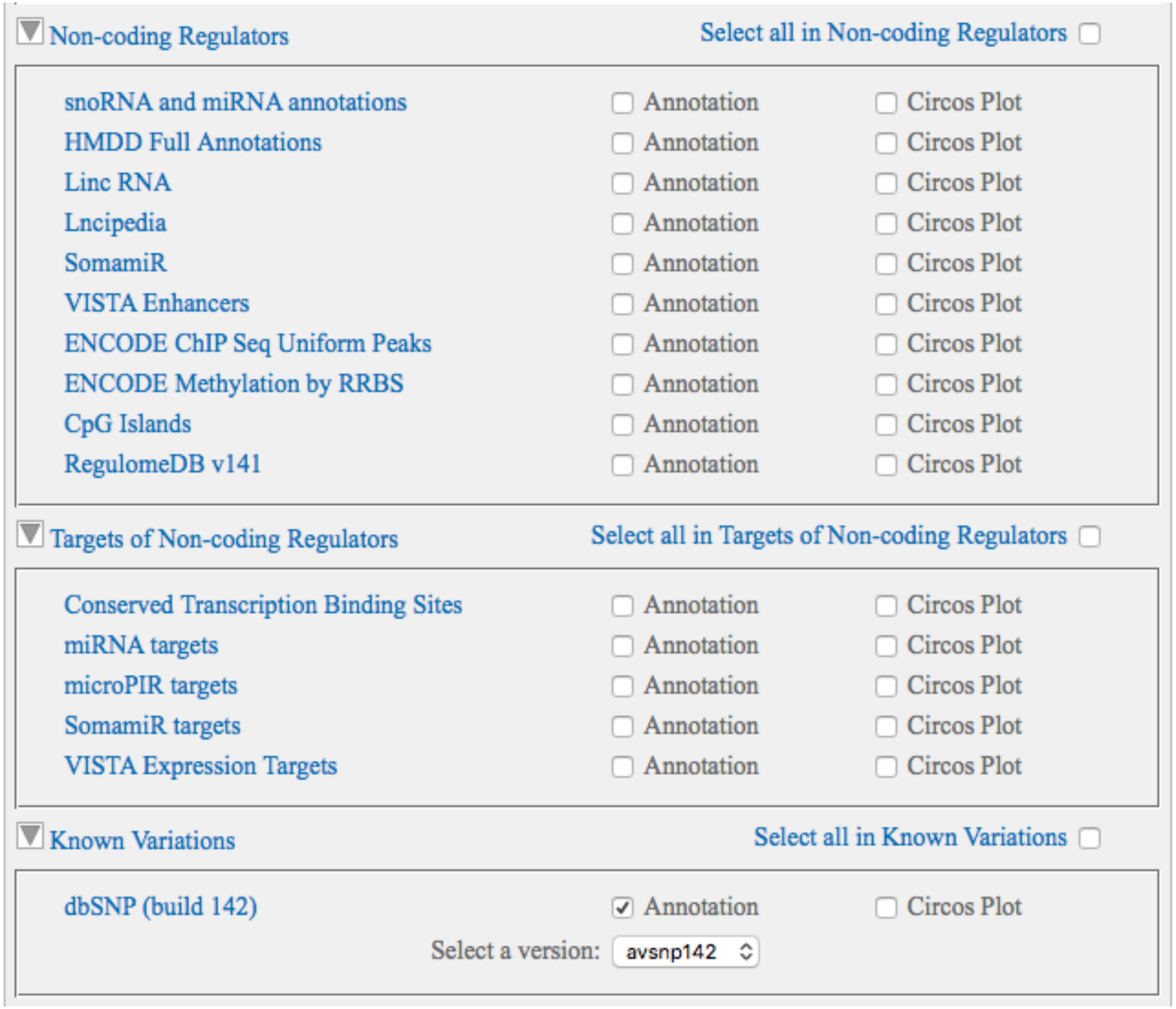
Protein Coding — Select all in Protein Coding □

SIFT Scores w/Predictions for SNPs Only	☑ Annotation	□ Circos Plot
SIFT NMD Prediction for Indels Only ***takes about 20-30 minutes for ~1500 Indels*	□ Annotation	
Polyphen2 Scores w/Predictions ***based on Human Var set*	☑ Annotation	□ Circos Plot
Polyphen2 Scores w/Predictions ***based on Human Div set*	☑ Annotation	□ Circos Plot
Mutation Taster	☑ Annotation	□ Circos Plot
Mutation Assessor	☑ Annotation	□ Circos Plot
Pre-computed Provean v1.1 Scores from dbSNP	☑ Annotation	□ Circos Plot
Combined Annotation Dependent Depletion	☑ Annotation	□ Circos Plot
FATHMM	☑ Annotation	□ Circos Plot
Variant Effect Scoring Tool	☑ Annotation	□ Circos Plot

Disease Related — Select all in Disease Related □

COSMIC v70	□ Annotation	□ Circos Plot
Select a version:	cosmic70	
OMIM (gene-centric)	□ Annotation	□ Circos Plot
ClinVar (Default: March 2015)	□ Annotation	□ Circos Plot
Select a version:	clinvar_20150330	

AVIA Annotations



▼ Non-coding Regulators — Select all in Non-coding Regulators □

snoRNA and miRNA annotations	□ Annotation	□ Circos Plot
HMDD Full Annotations	□ Annotation	□ Circos Plot
Linc RNA	□ Annotation	□ Circos Plot
Lncipedia	□ Annotation	□ Circos Plot
SomamiR	□ Annotation	□ Circos Plot
VISTA Enhancers	□ Annotation	□ Circos Plot
ENCODE ChIP Seq Uniform Peaks	□ Annotation	□ Circos Plot
ENCODE Methylation by RRBS	□ Annotation	□ Circos Plot
CpG Islands	□ Annotation	□ Circos Plot
RegulomeDB v141	□ Annotation	□ Circos Plot

▼ Targets of Non-coding Regulators — Select all in Targets of Non-coding Regulators □

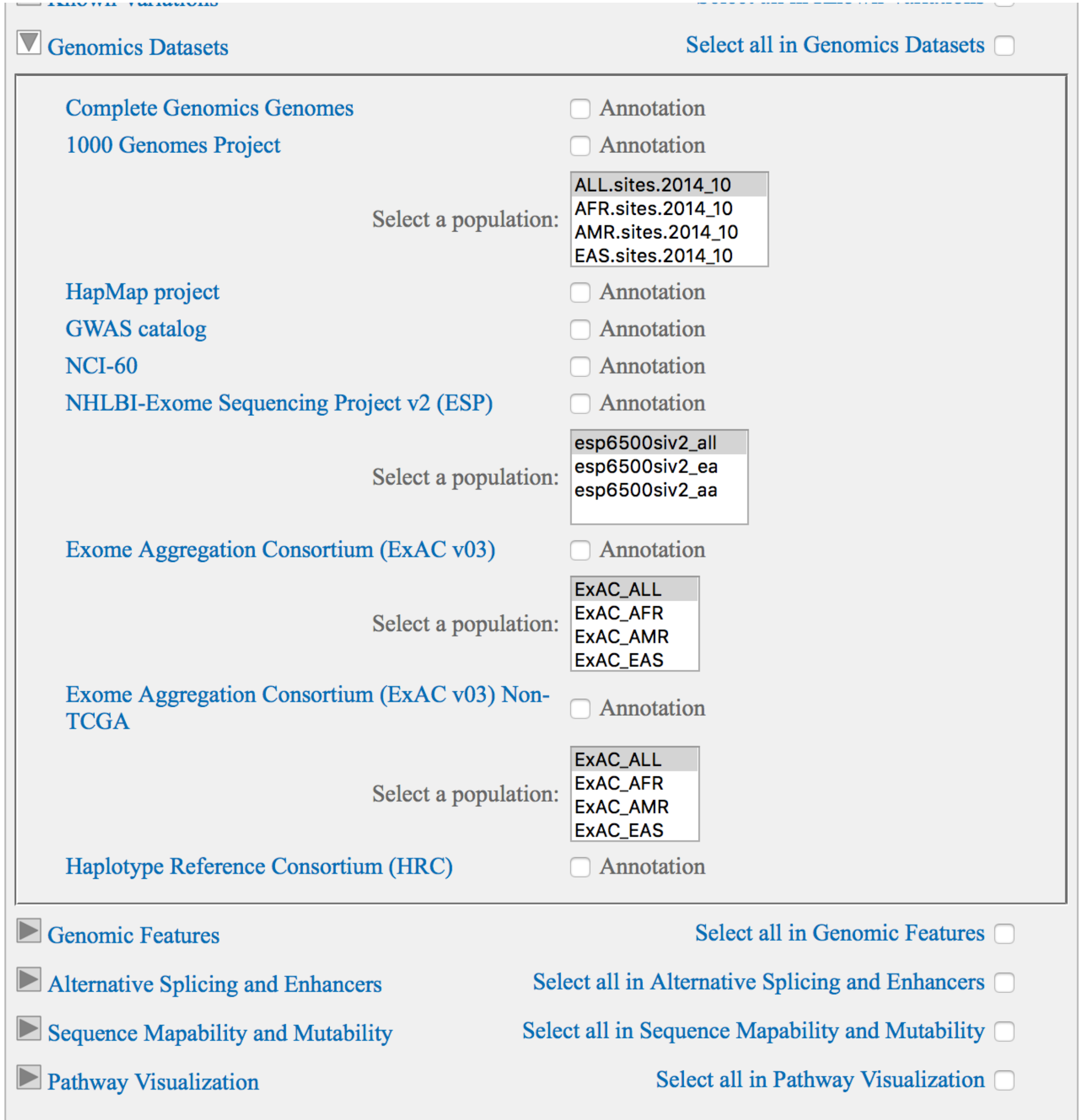
Conserved Transcription Binding Sites	□ Annotation	□ Circos Plot
miRNA targets	□ Annotation	□ Circos Plot
microPIR targets	□ Annotation	□ Circos Plot
SomamiR targets	□ Annotation	□ Circos Plot
VISTA Expression Targets	□ Annotation	□ Circos Plot

▼ Known Variations — Select all in Known Variations □

dbSNP (build 142)	☑ Annotation	□ Circos Plot

Select a version: avsnp142

AVIA Annotations



▼ Genomics Datasets — Select all in Genomics Datasets □

Complete Genomics Genomes	□ Annotation
1000 Genomes Project	□ Annotation
Select a population:	ALL.sites.2014_10 AFR.sites.2014_10 AMR.sites.2014_10 EAS.sites.2014_10
HapMap project	□ Annotation
GWAS catalog	□ Annotation
NCI-60	□ Annotation
NHLBI-Exome Sequencing Project v2 (ESP)	□ Annotation
Select a population:	esp6500siv2_all esp6500siv2_ea esp6500siv2_aa
Exome Aggregation Consortium (ExAC v03)	□ Annotation
Select a population:	ExAC_ALL ExAC_AFR ExAC_AMR ExAC_EAS
Exome Aggregation Consortium (ExAC v03) Non-TCGA	□ Annotation
Select a population:	ExAC_ALL ExAC_AFR ExAC_AMR ExAC_EAS
Haplotype Reference Consortium (HRC)	□ Annotation

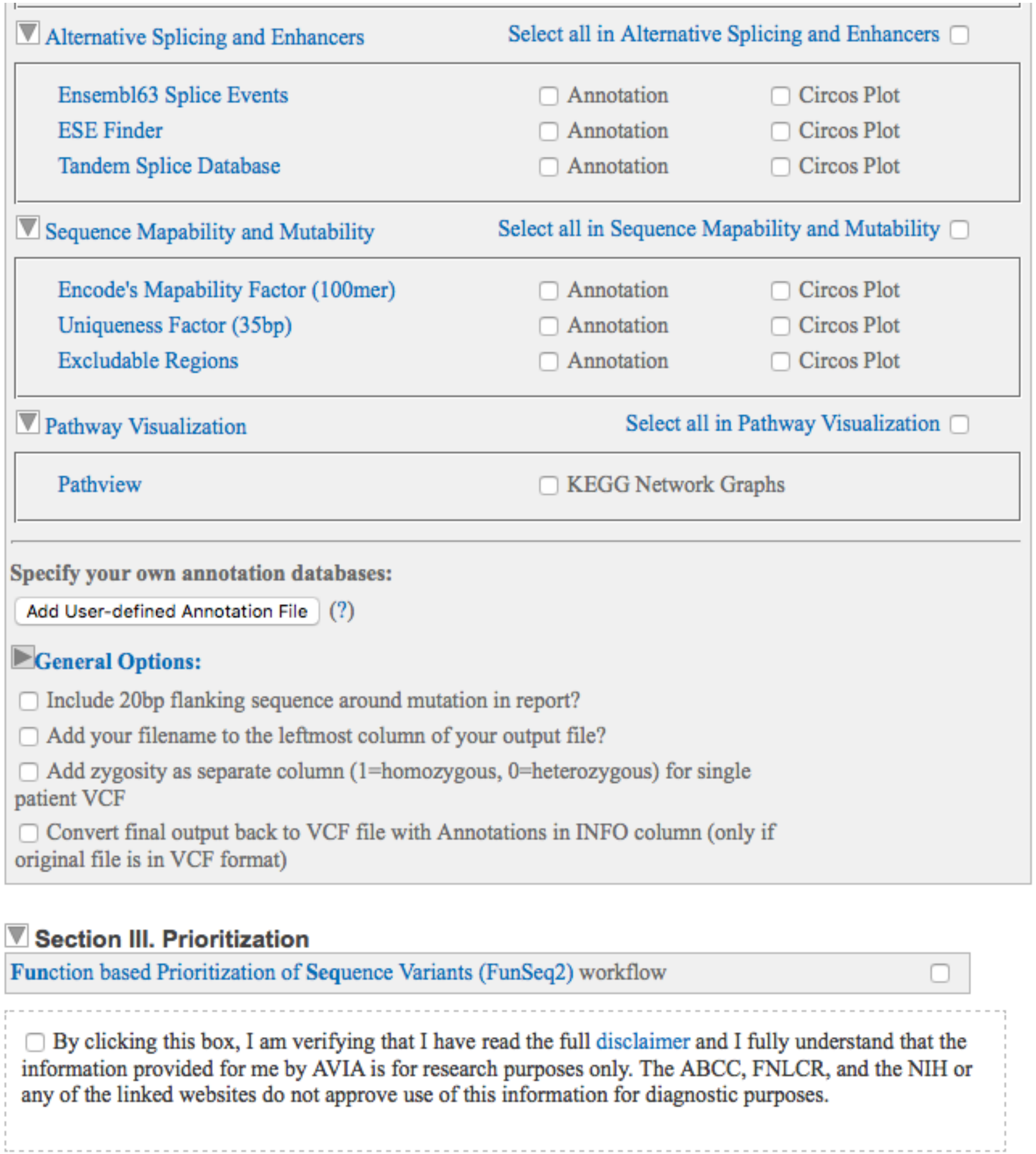
▶ Genomic Features — Select all in Genomic Features □

▶ Alternative Splicing and Enhancers — Select all in Alternative Splicing and Enhancers □

▶ Sequence Mapability and Mutability — Select all in Sequence Mapability and Mutability □

▶ Pathway Visualization — Select all in Pathway Visualization □

AVIA Annotations



▼ Alternative Splicing and Enhancers — Select all in Alternative Splicing and Enhancers □

Ensembl63 Splice Events	□ Annotation	□ Circos Plot
ESE Finder	□ Annotation	□ Circos Plot
Tandem Splice Database	□ Annotation	□ Circos Plot

▼ Sequence Mapability and Mutability — Select all in Sequence Mapability and Mutability □

Encode's Mapability Factor (100mer)	□ Annotation	□ Circos Plot
Uniqueness Factor (35bp)	□ Annotation	□ Circos Plot
Excludable Regions	□ Annotation	□ Circos Plot

▼ Pathway Visualization — Select all in Pathway Visualization □

Pathview	□ KEGG Network Graphs

Specify your own annotation databases:

▶ **General Options:**

□ Include 20bp flanking sequence around mutation in report?

□ Add your filename to the leftmost column of your output file?

□ Add zygosity as separate column (1=homozygous, 0=heterozygous) for single patient VCF

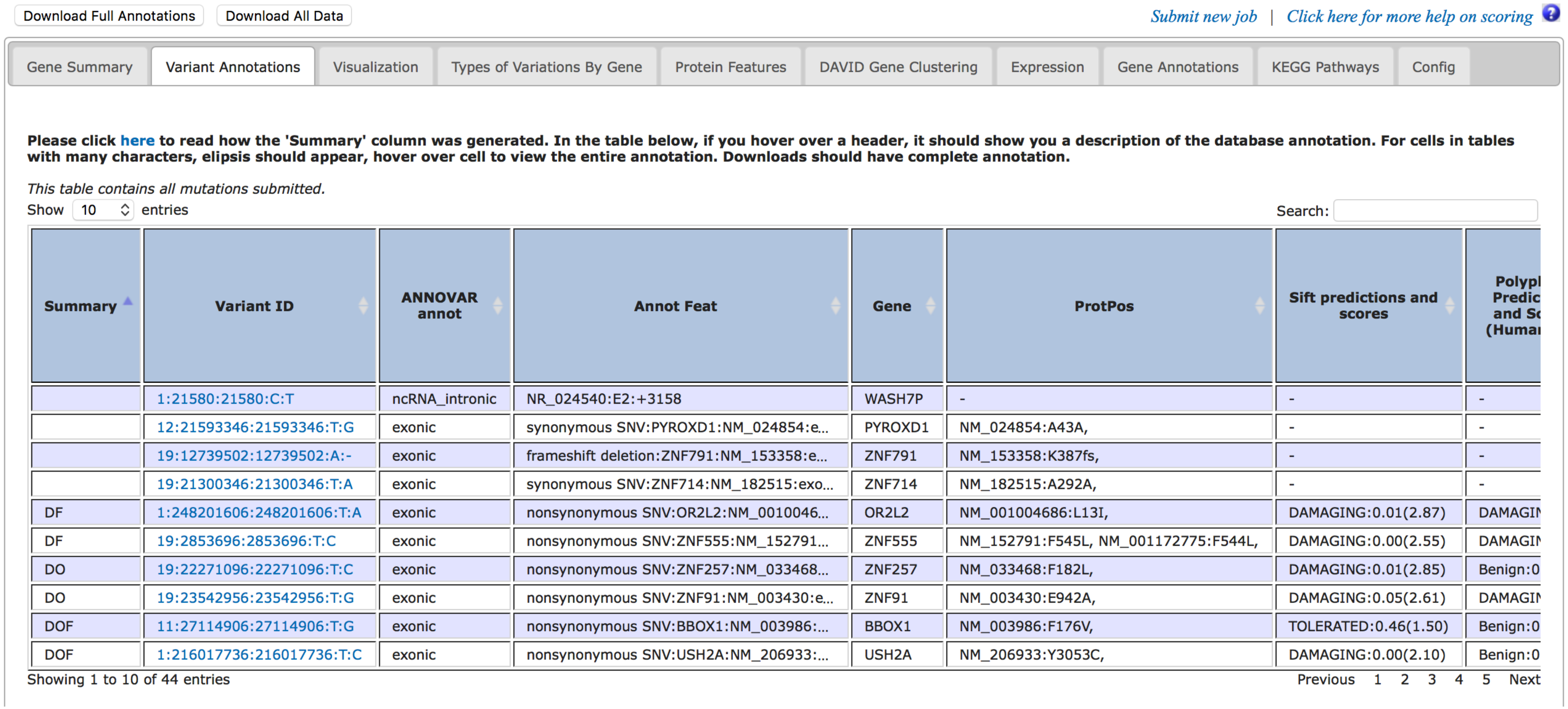
□ Convert final output back to VCF file with Annotations in INFO column (only if original file is in VCF format)

▼ **Section III. Prioritization**

Function based Prioritization of Sequence Variants (FunSeq2) workflow □

□ By clicking this box, I am verifying that I have read the full disclaimer and I fully understand that the information provided for me by AVIA is for research purposes only. The ABCC, FNLCR, and the NIH or any of the linked websites do not approve use of this information for diagnostic purposes.

Example Results - Web



Please click here to read how the 'Summary' column was generated. In the table below, if you hover over a header, it should show you a description of the database annotation. For cells in tables with many characters, elipsis should appear, hover over cell to view the entire annotation. Downloads should have complete annotation.

This table contains all mutations submitted.

Show 10 entries

Search:

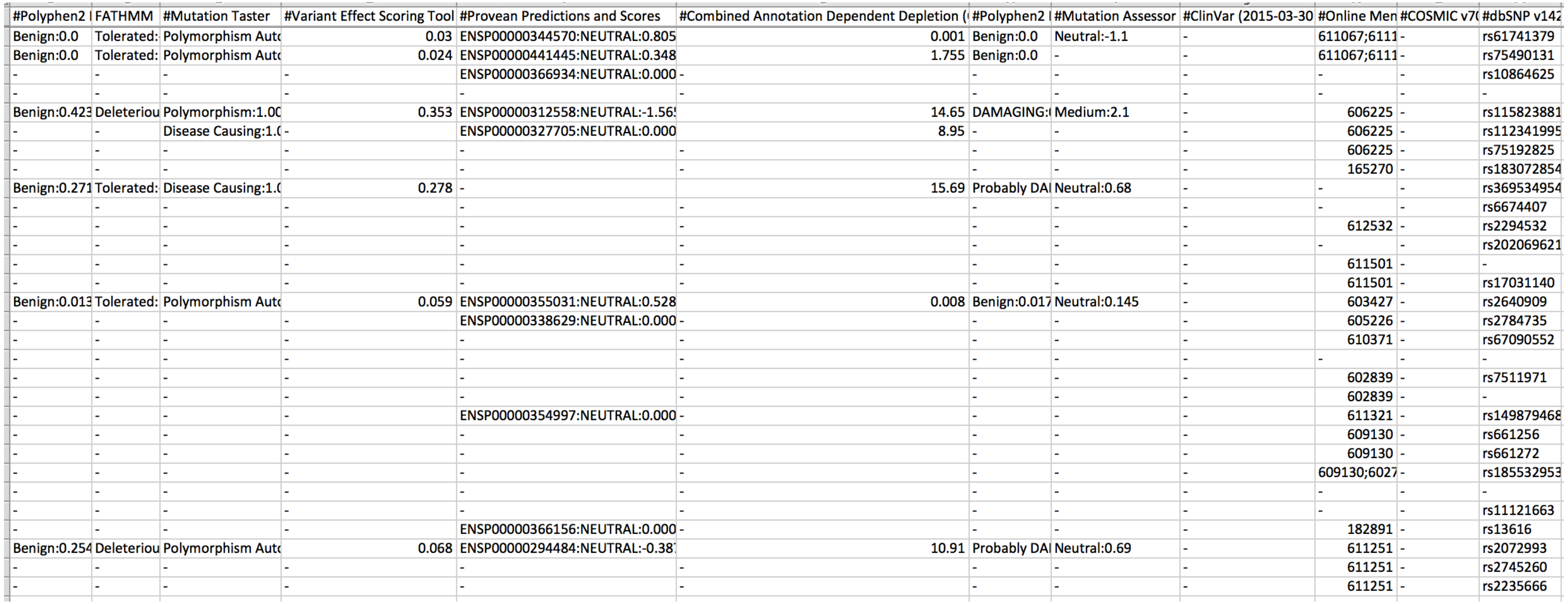
Summary	Variant ID	ANNOVAR annot	Annot Feat	Gene	ProtPos	Sift predictions and scores	Polypl Predic and Sc (Humai
	1:21580:21580:C:T	ncRNA_intronic	NR_024540:E2:+3158	WASH7P	-	-	-
	12:21593346:21593346:T:G	exonic	synonymous SNV:PYROXD1:NM_024854:e...	PYROXD1	NM_024854:A43A,	-	-
	19:12739502:12739502:A:-	exonic	frameshift deletion:ZNF791:NM_153358:e...	ZNF791	NM_153358:K387fs,	-	-
	19:21300346:21300346:T:A	exonic	synonymous SNV:ZNF714:NM_182515:exo...	ZNF714	NM_182515:A292A,	-	-
DF	1:248201606:248201606:T:A	exonic	nonsynonymous SNV:OR2L2:NM_0010046...	OR2L2	NM_001004686:L13I,	DAMAGING:0.01(2.87)	DAMAGIN
DF	19:2853696:2853696:T:C	exonic	nonsynonymous SNV:ZNF555:NM_152791...	ZNF555	NM_152791:F545L, NM_001172775:F544L,	DAMAGING:0.00(2.55)	DAMAGIN
DO	19:22271096:22271096:T:C	exonic	nonsynonymous SNV:ZNF257:NM_033468...	ZNF257	NM_033468:F182L,	DAMAGING:0.01(2.85)	Benign:0
DO	19:23542956:23542956:T:G	exonic	nonsynonymous SNV:ZNF91:NM_003430:e...	ZNF91	NM_003430:E942A,	DAMAGING:0.05(2.61)	DAMAGIN
DOF	11:27114906:27114906:T:G	exonic	nonsynonymous SNV:BBOX1:NM_003986:...	BBOX1	NM_003986:F176V,	TOLERATED:0.46(1.50)	Benign:0
DOF	1:216017736:216017736:T:C	exonic	nonsynonymous SNV:USH2A:NM_206933:...	USH2A	NM_206933:Y3053C,	DAMAGING:0.00(2.10)	Benign:0

Showing 1 to 10 of 44 entries

Previous 1 2 3 4 5 Next

Example Results - Text

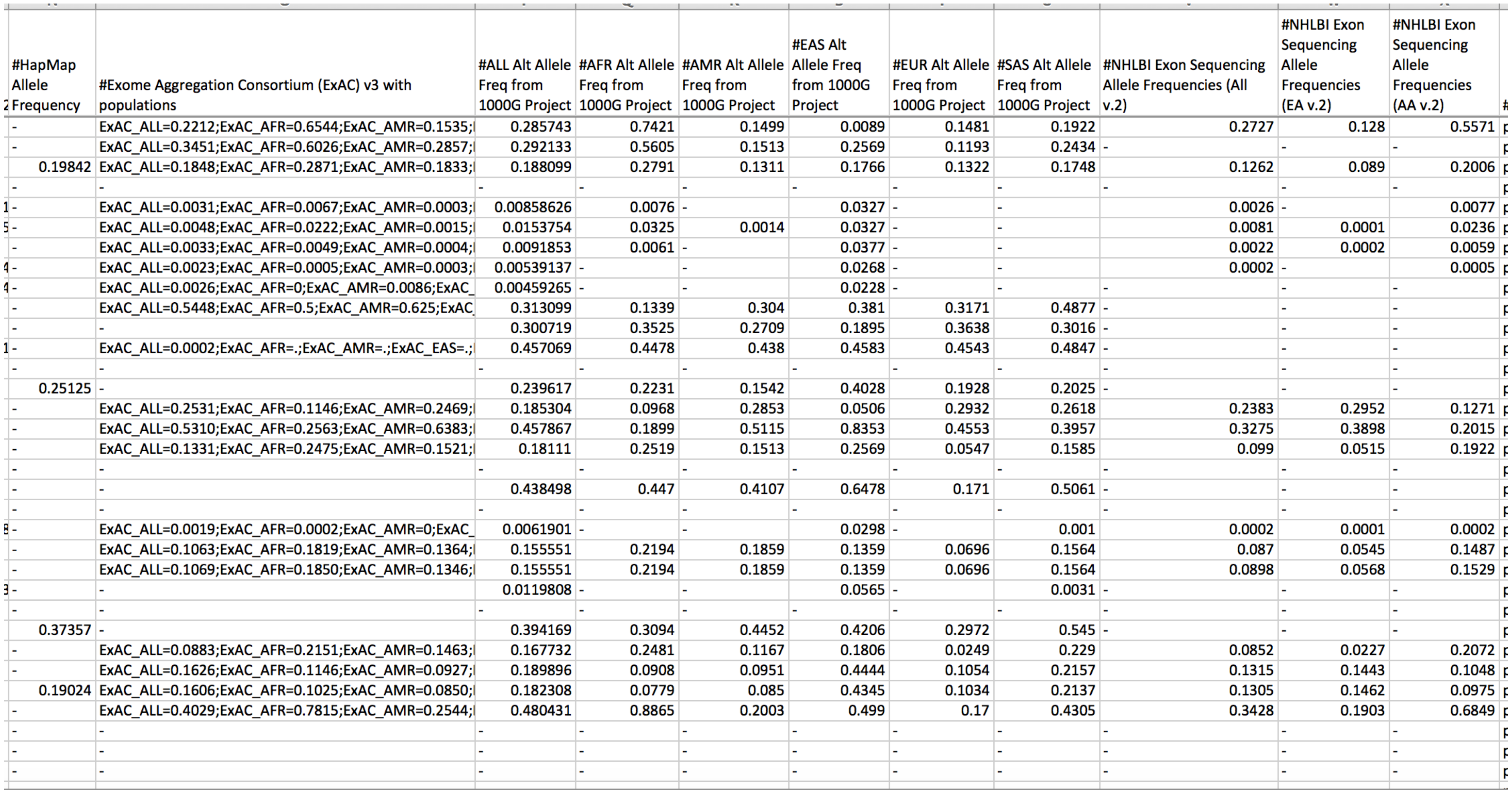
Effect Annotations



#Polyphen2 I	FATHMM	#Mutation Taster	#Variant Effect Scoring Tool	#Provean Predictions and Scores	#Combined Annotation Dependent Depletion (	#Polyphen2 I	#Mutation Assessor	#ClinVar (2015-03-30	#Online Men	#COSMIC v7	#dbSNP v142
Benign:0.0	Tolerated:	Polymorphism Aut	0.03	ENSP00000344570:NEUTRAL:0.805	0.001	Benign:0.0	Neutral:-1.1	-	611067;6111	-	rs61741379
Benign:0.0	Tolerated:	Polymorphism Aut	0.024	ENSP00000441445:NEUTRAL:0.348	1.755	Benign:0.0	-	-	611067;6111	-	rs75490131
-	-	-	-	ENSP00000366934:NEUTRAL:0.000	-	-	-	-	-	-	rs10864625
-	-	-	-	-	-	-	-	-	-	-	-
Benign:0.423	Deleteriou	Polymorphism:1.0	0.353	ENSP00000312558:NEUTRAL:-1.56	14.65	DAMAGING:	Medium:2.1	-	606225	-	rs115823881
-	-	Disease Causing:1.	-	ENSP00000327705:NEUTRAL:0.000	8.95	-	-	-	606225	-	rs112341995
-	-	-	-	-	-	-	-	-	606225	-	rs75192825
-	-	-	-	-	-	-	-	-	165270	-	rs183072854
Benign:0.271	Tolerated:	Disease Causing:1.	0.278	-	15.69	Probably DA	Neutral:0.68	-	-	-	rs369534954
-	-	-	-	-	-	-	-	-	-	-	rs6674407
-	-	-	-	-	-	-	-	-	612532	-	rs2294532
-	-	-	-	-	-	-	-	-	-	-	rs202069621
-	-	-	-	-	-	-	-	-	611501	-	-
-	-	-	-	-	-	-	-	-	611501	-	rs17031140
Benign:0.013	Tolerated:	Polymorphism Aut	0.059	ENSP00000355031:NEUTRAL:0.528	0.008	Benign:0.017	Neutral:0.145	-	603427	-	rs2640909
-	-	-	-	ENSP00000338629:NEUTRAL:0.000	-	-	-	-	605226	-	rs2784735
-	-	-	-	-	-	-	-	-	610371	-	rs67090552
-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	602839	-	rs7511971
-	-	-	-	-	-	-	-	-	602839	-	-
-	-	-	-	ENSP00000354997:NEUTRAL:0.000	-	-	-	-	611321	-	rs149879468
-	-	-	-	-	-	-	-	-	609130	-	rs661256
-	-	-	-	-	-	-	-	-	609130	-	rs661272
-	-	-	-	-	-	-	-	-	609130;6027	-	rs185532953
-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	rs11121663
-	-	-	-	ENSP00000366156:NEUTRAL:0.000	-	-	-	-	182891	-	rs13616
Benign:0.254	Deleteriou	Polymorphism Aut	0.068	ENSP00000294484:NEUTRAL:-0.38	10.91	Probably DA	Neutral:0.69	-	611251	-	rs2072993
-	-	-	-	-	-	-	-	-	611251	-	rs2745260
-	-	-	-	-	-	-	-	-	611251	-	rs2235666

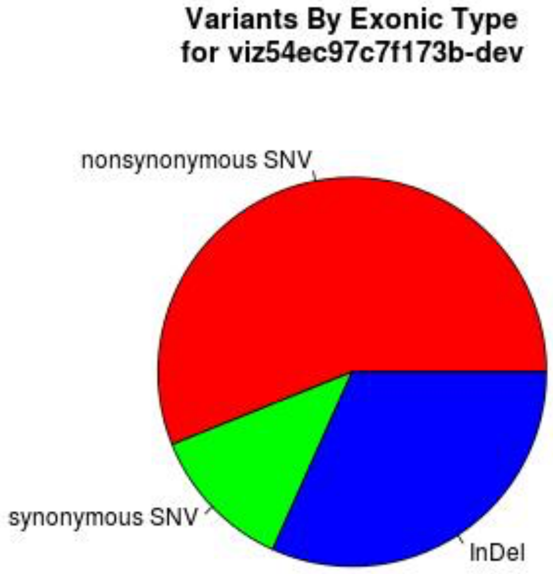
Example Results - Text

Allele frequencies

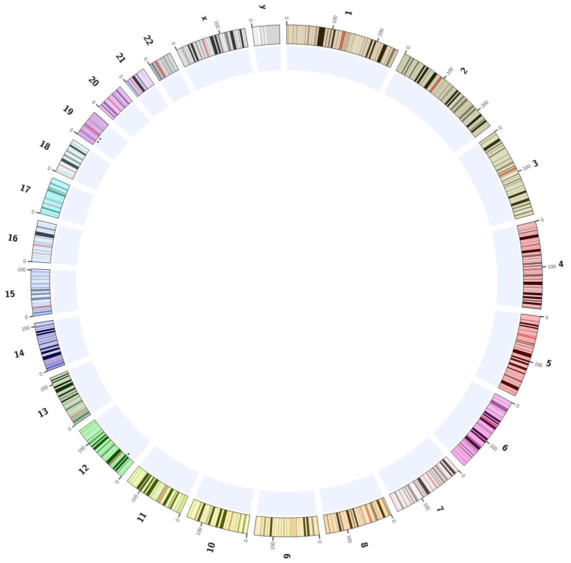


2	#HapMap Allele Frequency	#Exome Aggregation Consortium (ExAC) v3 with populations	#ALL Alt Allele Freq from 1000G Project	#AFR Alt Allele Freq from 1000G Project	#AMR Alt Allele Freq from 1000G Project	#EAS Alt Allele Freq from 1000G Project	#EUR Alt Allele Freq from 1000G Project	#SAS Alt Allele Freq from 1000G Project	#NHLBI Exon Sequencing Allele Frequencies (All v.2)	#NHLBI Exon Sequencing Allele Frequencies (EA v.2)	#NHLBI Exon Sequencing Allele Frequencies (AA v.2)	#
	-	ExAC_ALL=0.2212;ExAC_AFR=0.6544;ExAC_AMR=0.1535;	0.285743	0.7421	0.1499	0.0089	0.1481	0.1922	0.2727	0.128	0.5571	p
	-	ExAC_ALL=0.3451;ExAC_AFR=0.6026;ExAC_AMR=0.2857;	0.292133	0.5605	0.1513	0.2569	0.1193	0.2434	-	-	-	p
	0.19842	ExAC_ALL=0.1848;ExAC_AFR=0.2871;ExAC_AMR=0.1833;	0.188099	0.2791	0.1311	0.1766	0.1322	0.1748	0.1262	0.089	0.2006	p
	-	-	-	-	-	-	-	-	-	-	-	p
1	-	ExAC_ALL=0.0031;ExAC_AFR=0.0067;ExAC_AMR=0.0003;	0.00858626	0.0076	-	0.0327	-	-	0.0026	-	0.0077	p
5	-	ExAC_ALL=0.0048;ExAC_AFR=0.0222;ExAC_AMR=0.0015;	0.0153754	0.0325	0.0014	0.0327	-	-	0.0081	0.0001	0.0236	p
	-	ExAC_ALL=0.0033;ExAC_AFR=0.0049;ExAC_AMR=0.0004;	0.0091853	0.0061	-	0.0377	-	-	0.0022	0.0002	0.0059	p
4	-	ExAC_ALL=0.0023;ExAC_AFR=0.0005;ExAC_AMR=0.0003;	0.00539137	-	-	0.0268	-	-	0.0002	-	0.0005	p
4	-	ExAC_ALL=0.0026;ExAC_AFR=0;ExAC_AMR=0.0086;ExAC_	0.00459265	-	-	0.0228	-	-	-	-	-	p
	-	ExAC_ALL=0.5448;ExAC_AFR=0.5;ExAC_AMR=0.625;ExAC_	0.313099	0.1339	0.304	0.381	0.3171	0.4877	-	-	-	p
	-	-	0.300719	0.3525	0.2709	0.1895	0.3638	0.3016	-	-	-	p
1	-	ExAC_ALL=0.0002;ExAC_AFR=.;ExAC_AMR=.;ExAC_EAS=.;	0.457069	0.4478	0.438	0.4583	0.4543	0.4847	-	-	-	p
	-	-	-	-	-	-	-	-	-	-	-	p
	0.25125	-	0.239617	0.2231	0.1542	0.4028	0.1928	0.2025	-	-	-	p
	-	ExAC_ALL=0.2531;ExAC_AFR=0.1146;ExAC_AMR=0.2469;	0.185304	0.0968	0.2853	0.0506	0.2932	0.2618	0.2383	0.2952	0.1271	p
	-	ExAC_ALL=0.5310;ExAC_AFR=0.2563;ExAC_AMR=0.6383;	0.457867	0.1899	0.5115	0.8353	0.4553	0.3957	0.3275	0.3898	0.2015	p
	-	ExAC_ALL=0.1331;ExAC_AFR=0.2475;ExAC_AMR=0.1521;	0.18111	0.2519	0.1513	0.2569	0.0547	0.1585	0.099	0.0515	0.1922	p
	-	-	-	-	-	-	-	-	-	-	-	p
	-	-	0.438498	0.447	0.4107	0.6478	0.171	0.5061	-	-	-	p
	-	-	-	-	-	-	-	-	-	-	-	p
8	-	ExAC_ALL=0.0019;ExAC_AFR=0.0002;ExAC_AMR=0;ExAC_	0.0061901	-	-	0.0298	-	0.001	0.0002	0.0001	0.0002	p
	-	ExAC_ALL=0.1063;ExAC_AFR=0.1819;ExAC_AMR=0.1364;	0.155551	0.2194	0.1859	0.1359	0.0696	0.1564	0.087	0.0545	0.1487	p
	-	ExAC_ALL=0.1069;ExAC_AFR=0.1850;ExAC_AMR=0.1346;	0.155551	0.2194	0.1859	0.1359	0.0696	0.1564	0.0898	0.0568	0.1529	p
3	-	-	0.0119808	-	-	0.0565	-	0.0031	-	-	-	p
	-	-	-	-	-	-	-	-	-	-	-	p
	0.37357	-	0.394169	0.3094	0.4452	0.4206	0.2972	0.545	-	-	-	p
	-	ExAC_ALL=0.0883;ExAC_AFR=0.2151;ExAC_AMR=0.1463;	0.167732	0.2481	0.1167	0.1806	0.0249	0.229	0.0852	0.0227	0.2072	p
	-	ExAC_ALL=0.1626;ExAC_AFR=0.1146;ExAC_AMR=0.0927;	0.189896	0.0908	0.0951	0.4444	0.1054	0.2157	0.1315	0.1443	0.1048	p
	0.19024	ExAC_ALL=0.1606;ExAC_AFR=0.1025;ExAC_AMR=0.0850;	0.182308	0.0779	0.085	0.4345	0.1034	0.2137	0.1305	0.1462	0.0975	p
	-	ExAC_ALL=0.4029;ExAC_AFR=0.7815;ExAC_AMR=0.2544;	0.480431	0.8865	0.2003	0.499	0.17	0.4305	0.3428	0.1903	0.6849	p
	-	-	-	-	-	-	-	-	-	-	-	p
	-	-	-	-	-	-	-	-	-	-	-	p
	-	-	-	-	-	-	-	-	-	-	-	p

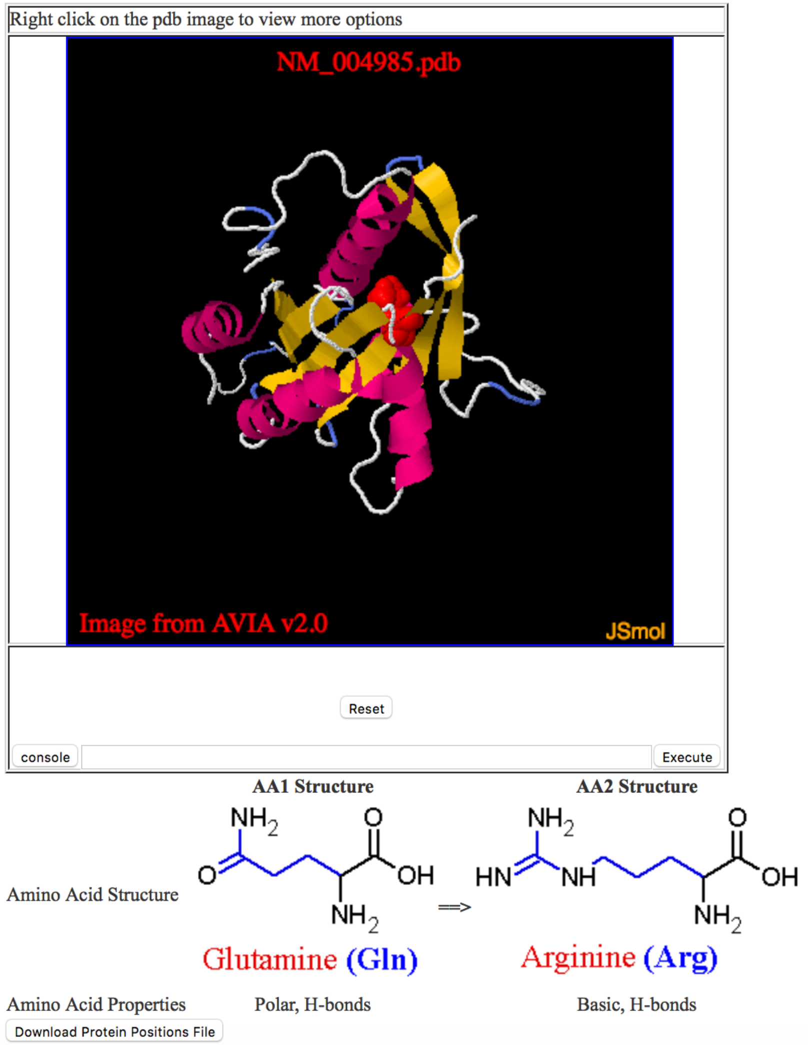
Example Results - Visualizations



Circos plots



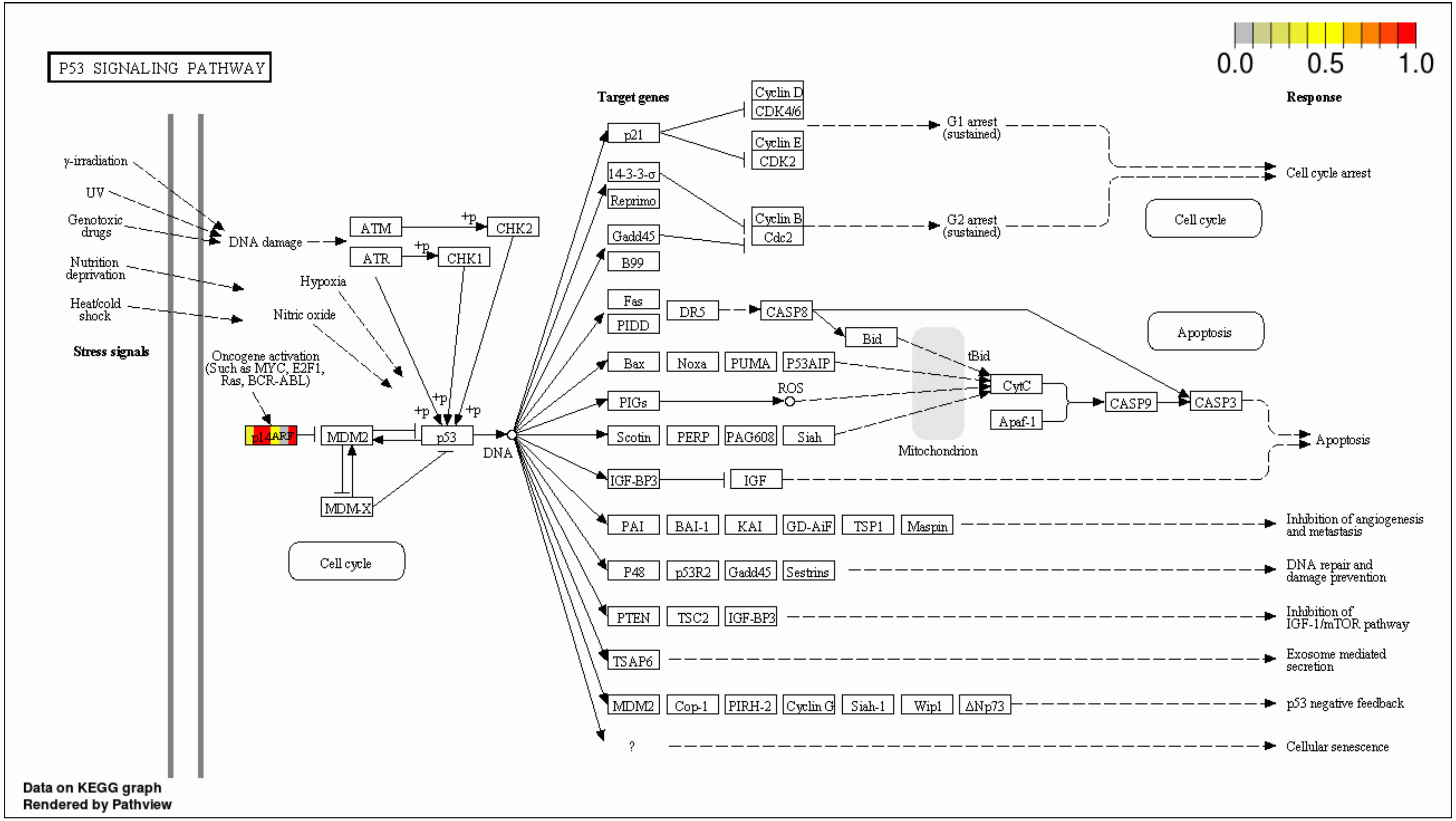
Protein Mutation Model



Right click on the pdb image to view more options

	AA1 Structure	AA2 Structure
Amino Acid Structure	Glutamine (Gln)	Arginine (Arg)
Amino Acid Properties	Polar, H-bonds	Basic, H-bonds

Example Results - Pathways

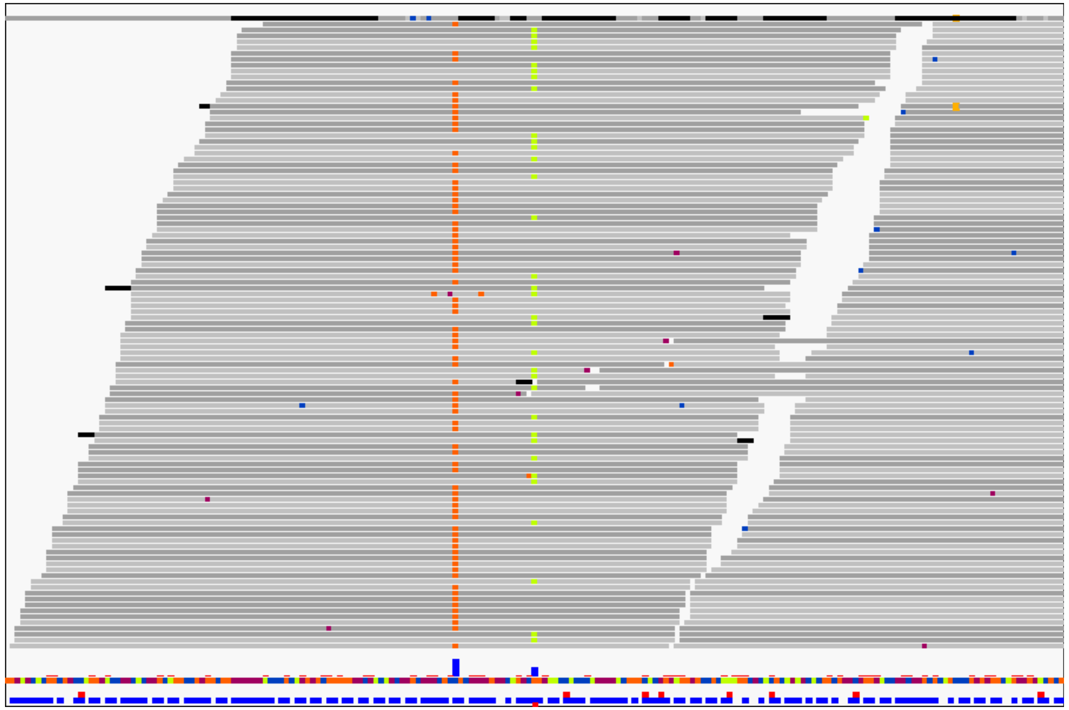


AVIA v2.0 Legend

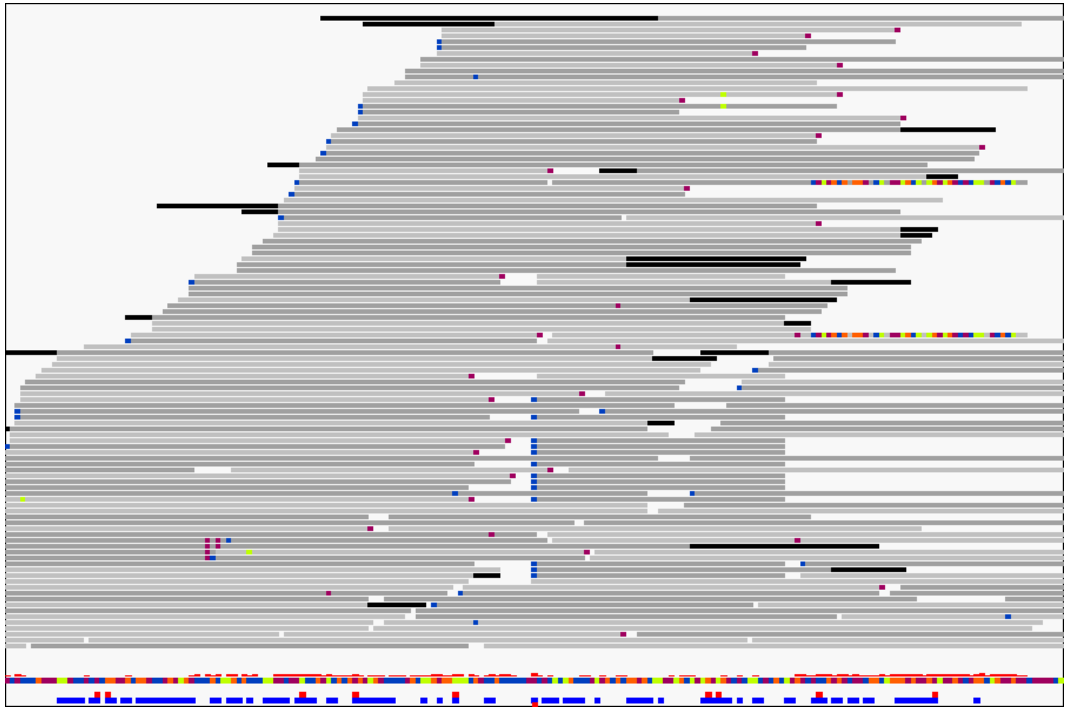
Variant Verification

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

True Positive

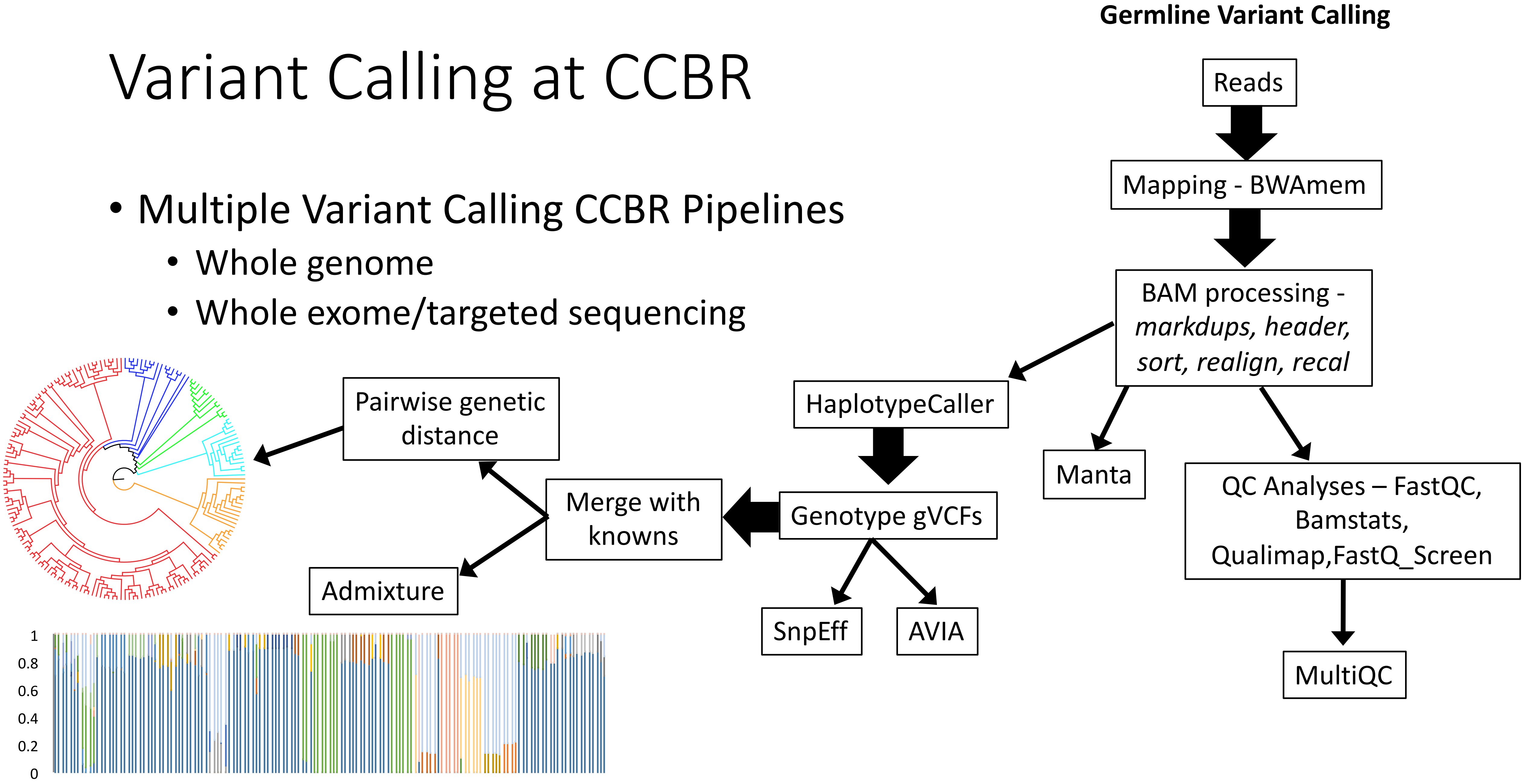


False Positive



Variant Calling at CCBR

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



Germline Variant Calling

Reads

Mapping - BWAmem

BAM processing - *markdups, header, sort, realign, recal*

HaplotypeCaller

Manta

QC Analyses – FastQC, Bamstats, Qualimap,FastQ_Screen

MultiQC

Genotype gVCFs

SnpEff

AVIA

Merge with knowns

Pairwise genetic distance

Admixture

Germline Final Outputs

- multiqc_report.html – final report after initialQC AND after variant calling
- Merged VCFs (with and without SNPeff)
  - combined.vcf – completely unfiltered variants
  - combined.relaxedFilter.vcf**
  - combined.strictFilter.vcf

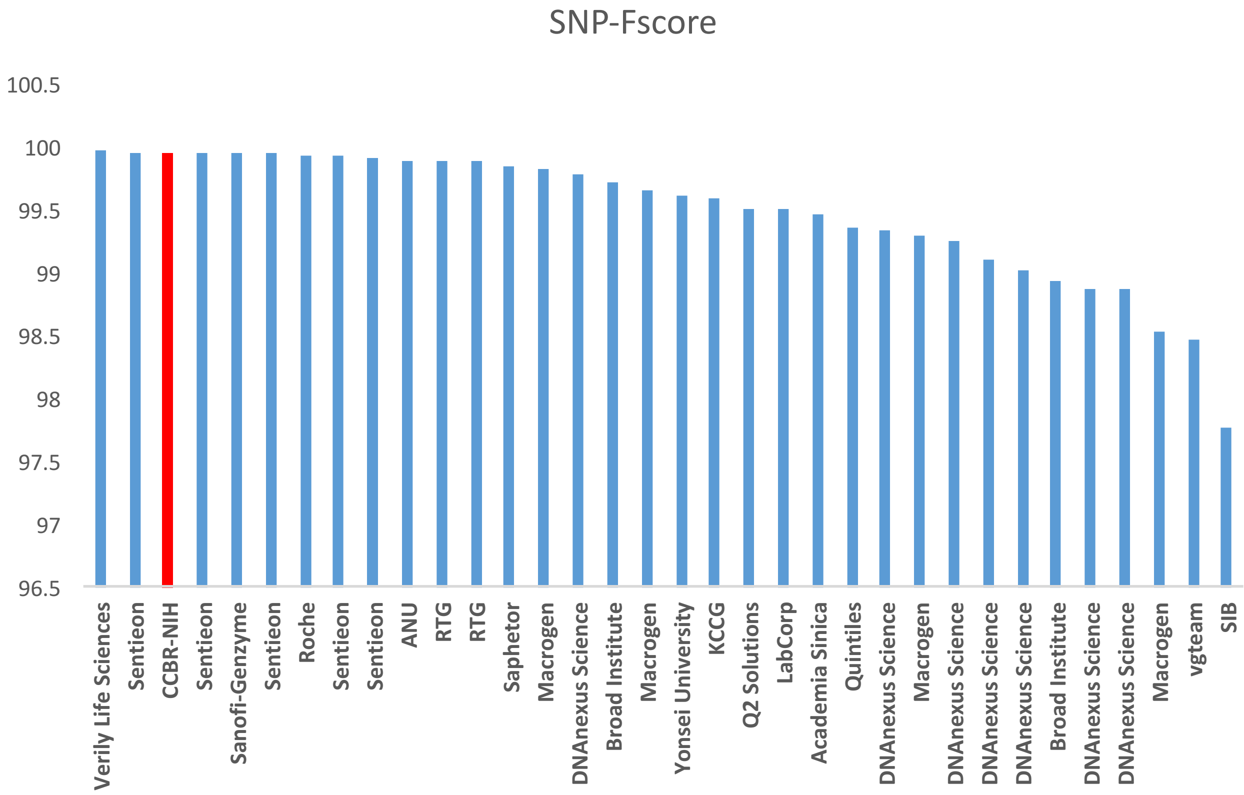
filtered for on-target variants, in addition to hard quality filters

- Structural Variants –manta_out/results/variants/
- Sample VCFs -sample_vcfs/
- sample_network.bmp
- full_annot.txt.zip – full AVIA annotation table
- variants.database – AVIA annotation table with sample genotypes added
- *recal.bam files – final BAM for each sample

Variant Calling at CCBR

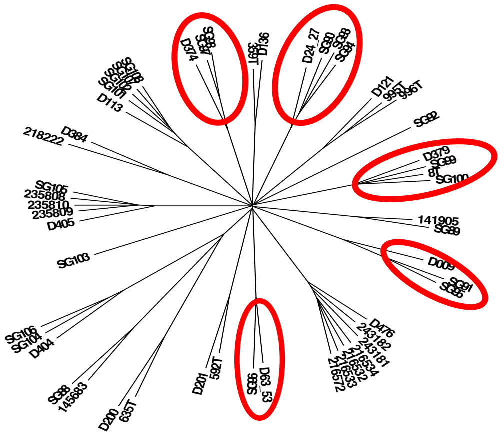
- Multiple Variant Calling CCBR Pipelines
  - Whole genome
  - Whole exome/targeted sequencing
  - Excellent performance in Precision FDA Challenge

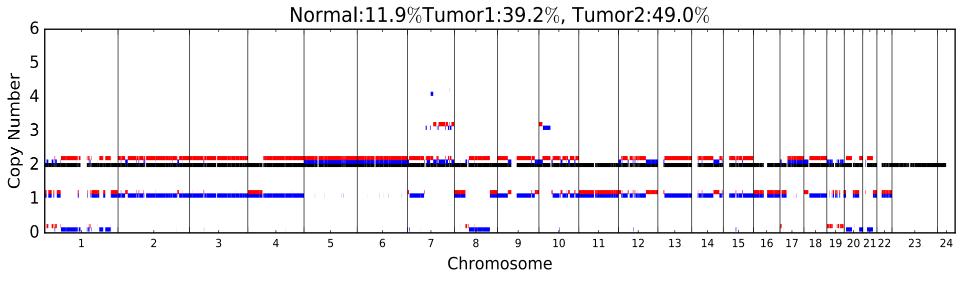
Germline Variant Calling

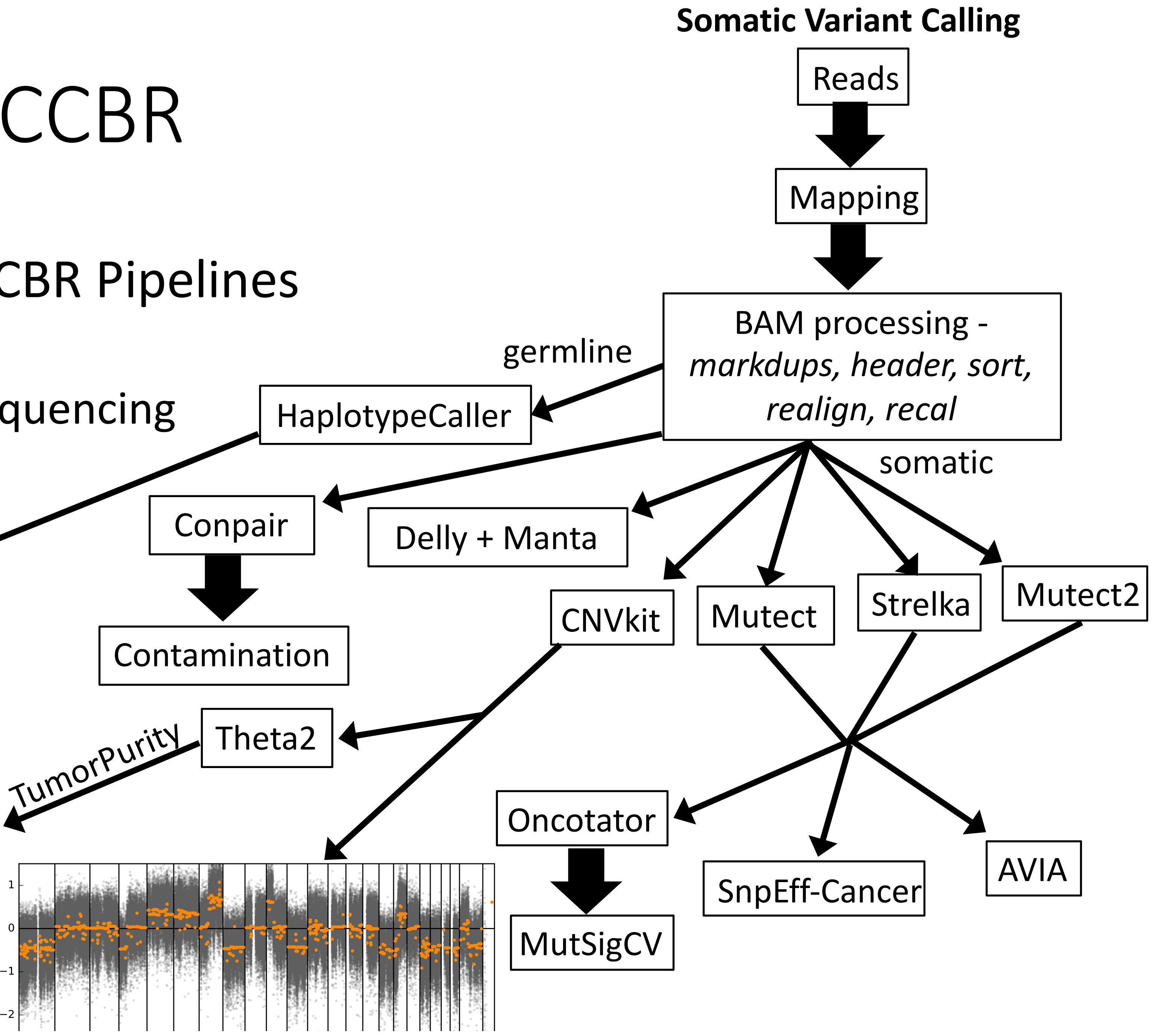


Variant Calling at CCBR

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







Somatic Variant Calling

Reads

Mapping

BAM processing - *markdups, header, sort, realign, recal*

germline

HaplotypeCaller

somatic

Conpair

Contamination

Delly + Manta

CNVkit

Mutect

Strelka

Mutect2

Theta2

TumorPurity

Oncotator

MutSigCV

SnpEff-Cancer

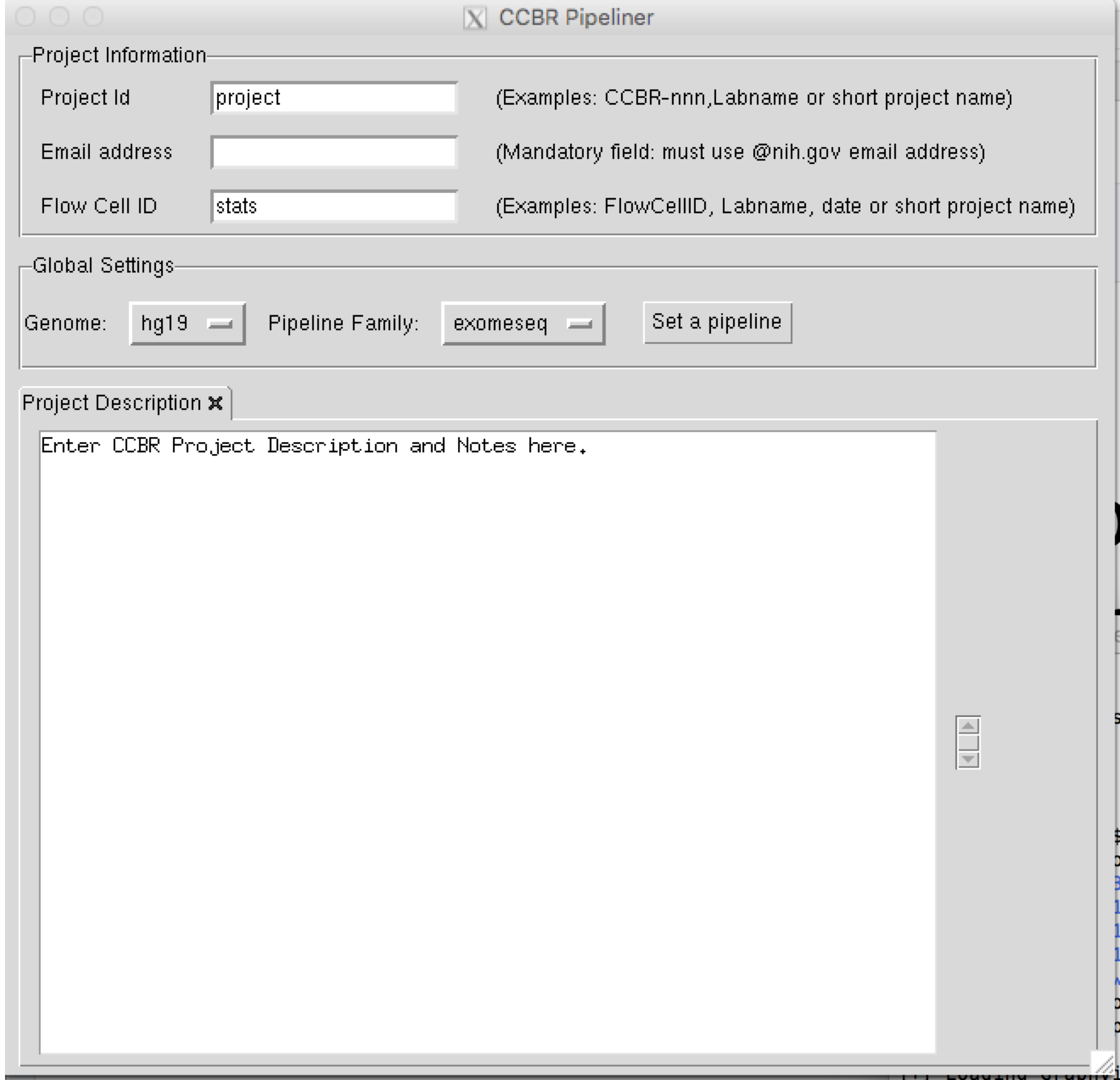
AVIA

Somatic Final Outputs

- multiqc_report.html – final report after initialQC AND after variant calling
- Merged and sample VCFs (with and without SNPeff)
  - strelka_out/*.vcf
  - mutect_out/*.vcf
  - mutect2_out/*.vcf
- sample_network.bmp
- full_annot.txt.zip – full AVIA annotation table for MuTect2 final VCF
- variants.database – AVIA annotation table with sample genotypes added
- *recal.bam files – final BAM for each sample
- Oncotator annotated sample MAFs and merged MAFs for each caller
  - mutect_out/oncotator_out/
  - mutect2_out/oncotator_out/
  - strelka_out/oncotator_out/
- MutSigCV results for each caller
  - mutect_out/mutsigCV_out/
  - mutect2_out/mutsigCV_out/
  - strelka_out/mutsigCV_out/
- Tumor purity/clonality – theta2_out/sample_dir/*.BEST.results
- Contamination – conpair_out/*.conpair
- Copy-number results – cnvkit_out/sample_dir/*
- Structural variant results
  - delly_out/*bcf
  - manta_out/*

Variant Calling at CCBR

- All pipelines (and several others) available through CCBR_Pipeliner app
  - Just need Biowulf account
  - https://github.com/CCBR/Pipeliner
  - ***module load ccbrpipeliner (enter)***
  - ***ccbrpipe.sh (enter)***



CCBR Pipeliner

Project Information

Project Id: project (Examples: CCBR-nnn,Labname or short project name)

Email address: (Mandatory field: must use @nih.gov email address)

Flow Cell ID: stats (Examples: FlowCellID, Labname, date or short project name)

Global Settings

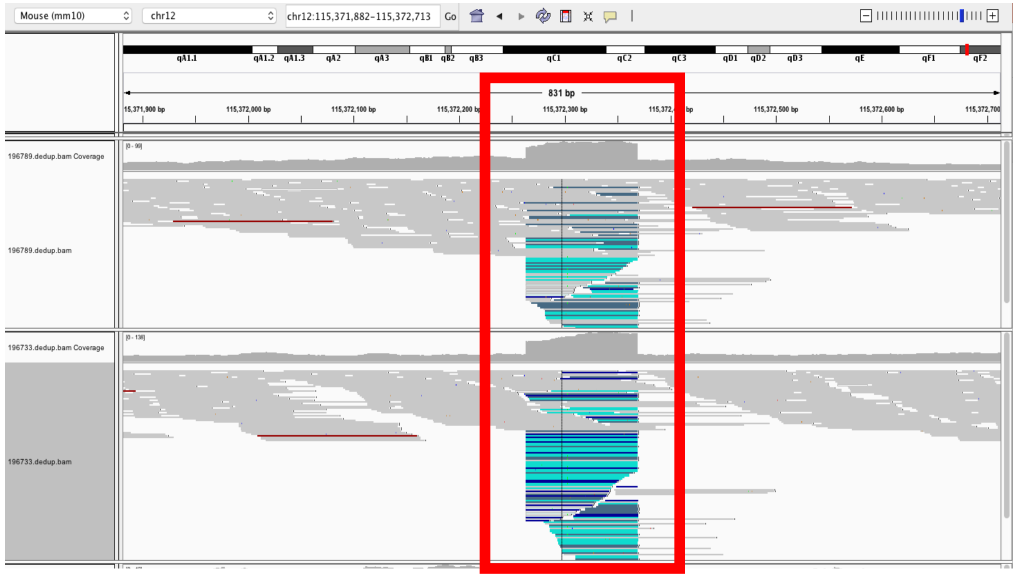
Genome: hg19 Pipeline Family: exomeseq

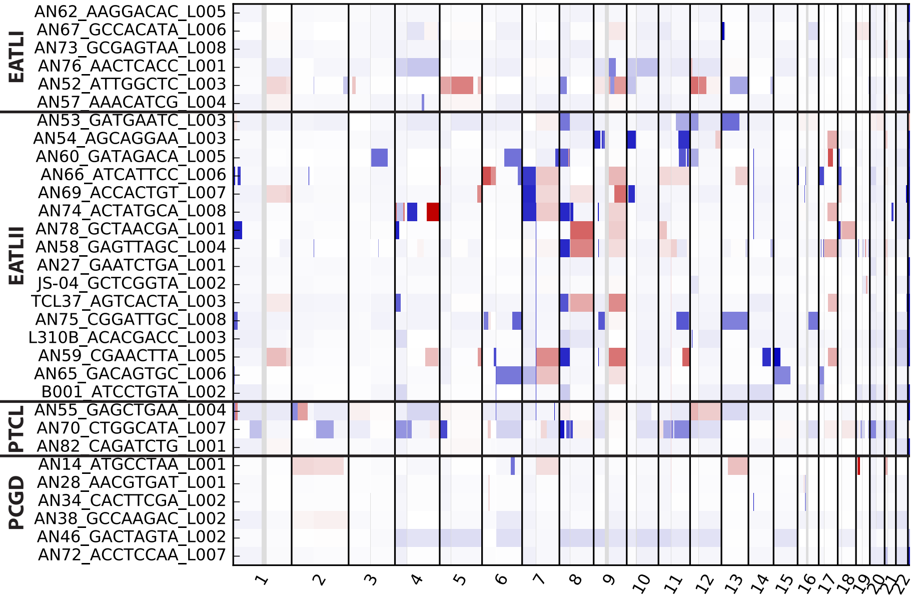
Enter CCBR Project Description and Notes here.

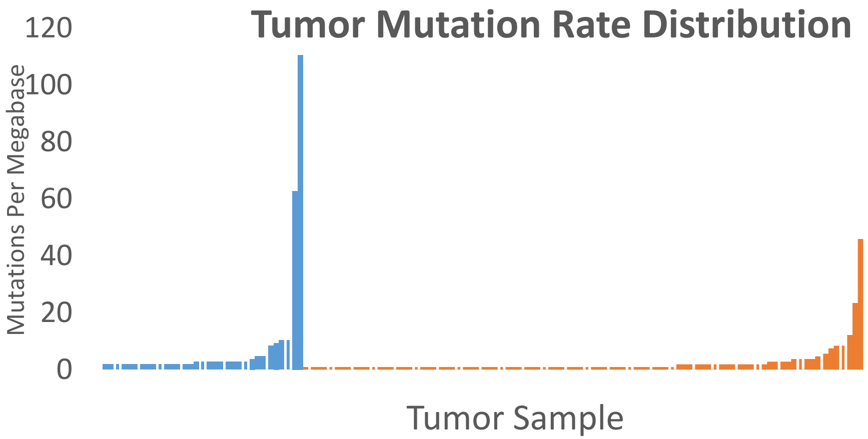
Now lets look at Exome-seq Pipeline Output

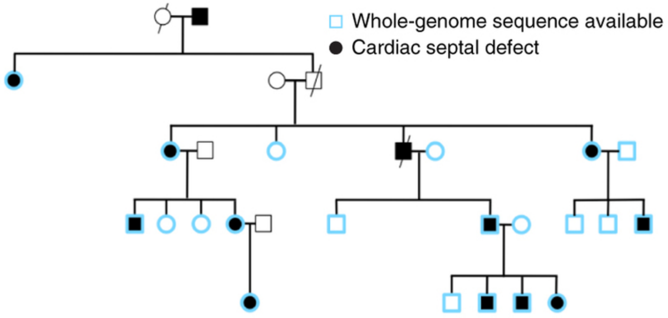
- test reads: /data/CCBR/datashare/BTEP/reads
- example pipeline: /data/CCBR/datashare/pipe_example2/exome_test3

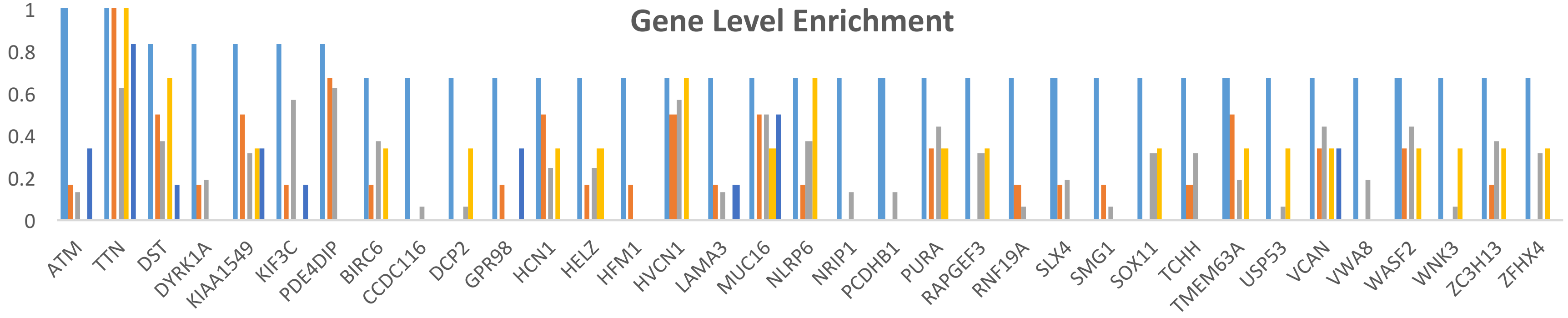
Downstream Analysis

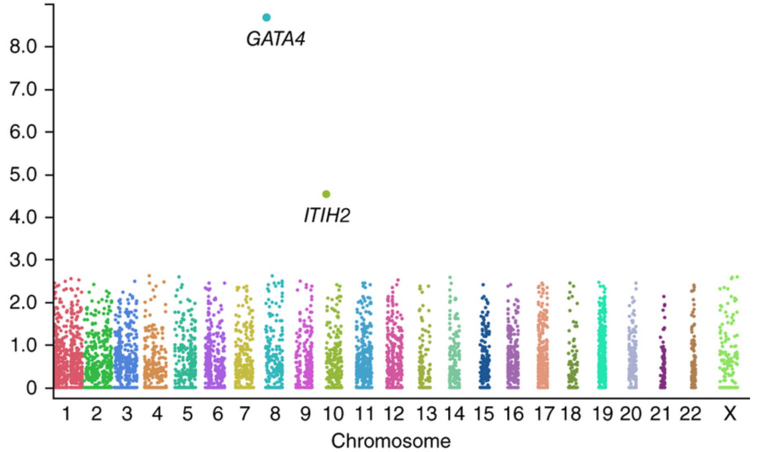








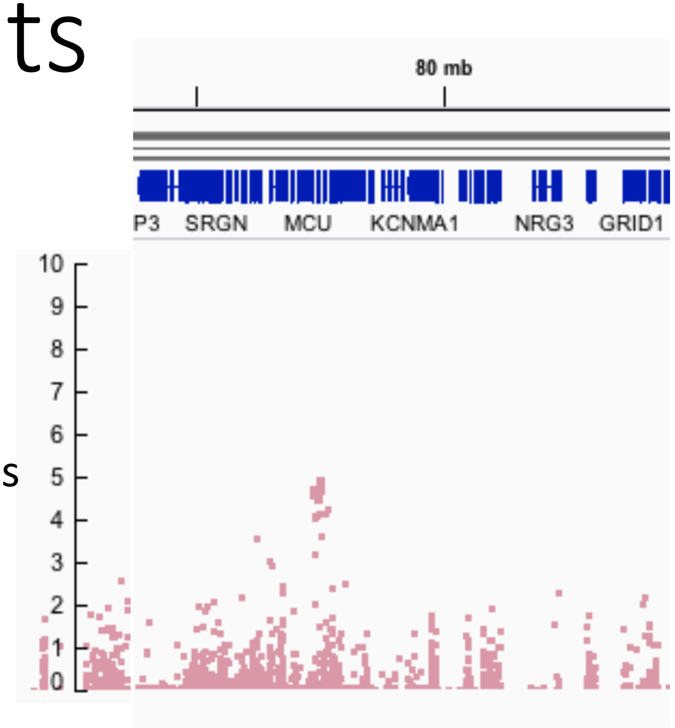


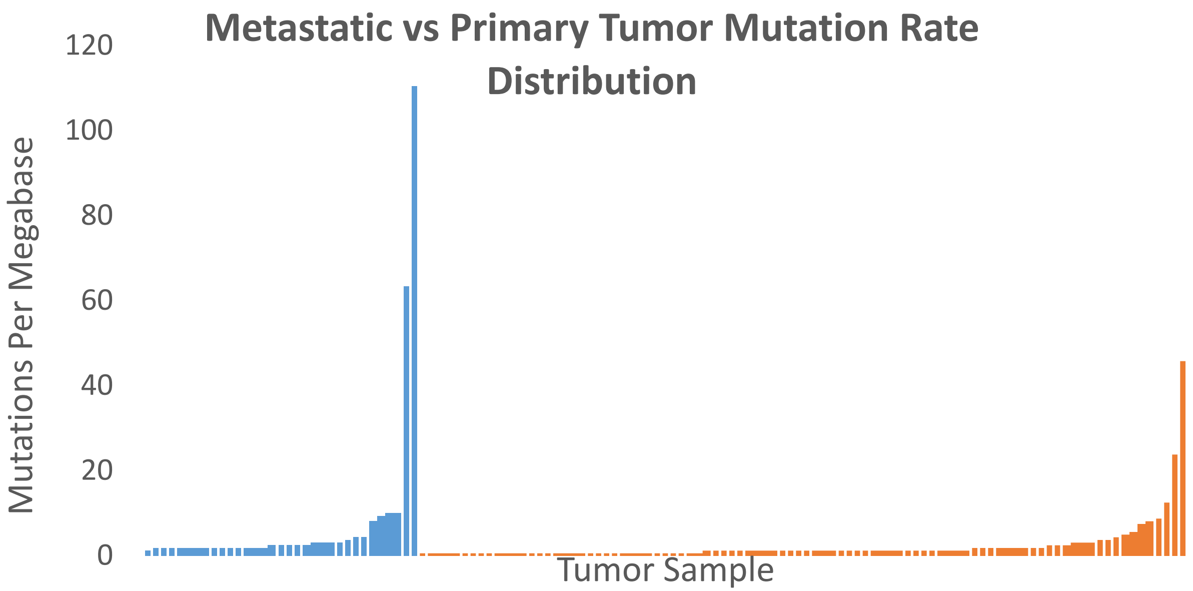


Analysis of Publicly Available Datasets

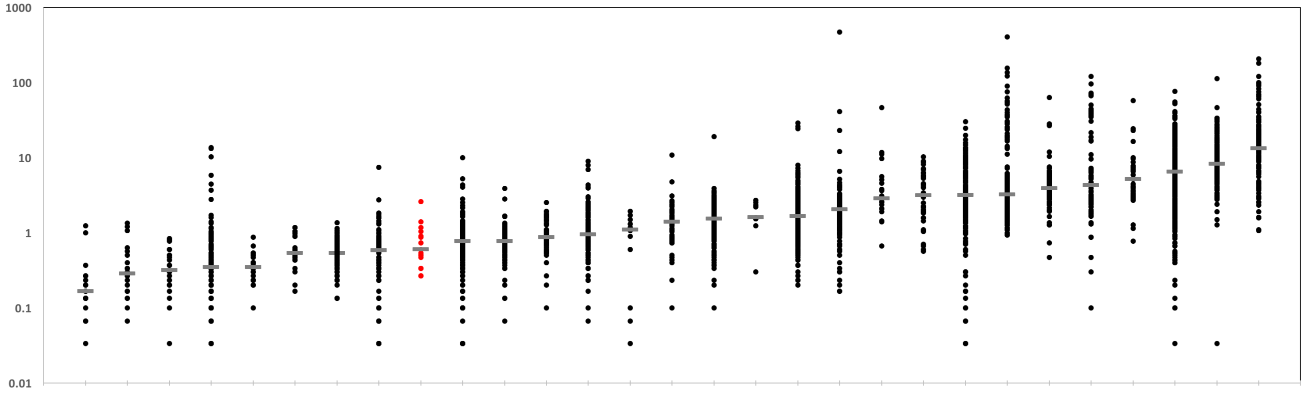
- In-depth analysis of large, public datasets
  - 1k Genomes, ExAC
  - TCGA

TCGA Germline Association Analysis





TCGA Mutational Load Analysis


1000
100
10
1
0.1
0.01