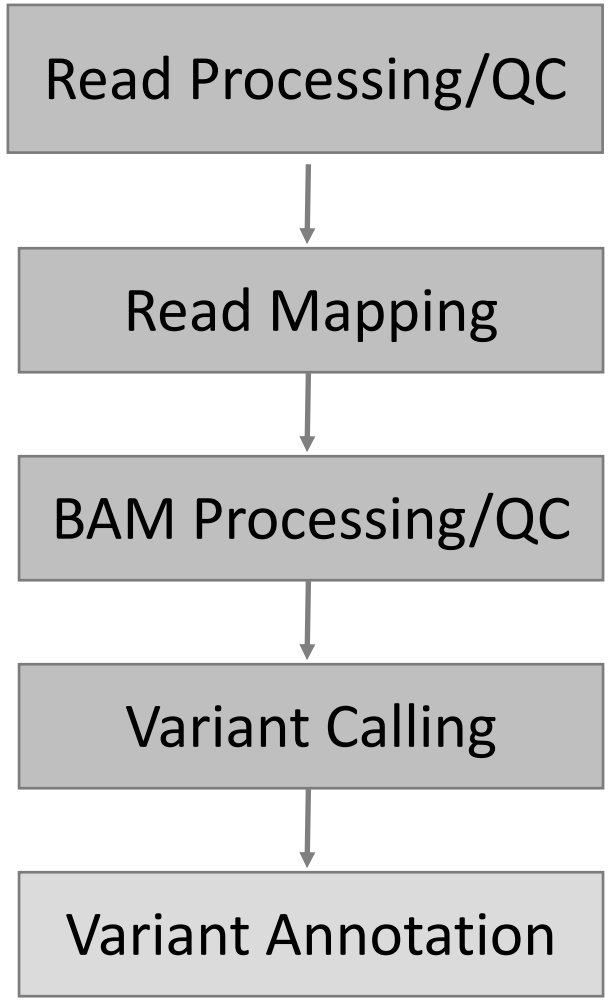
Germline and Somatic Mutation Analysis: Experimental Design, Variant Calling, and Analysis

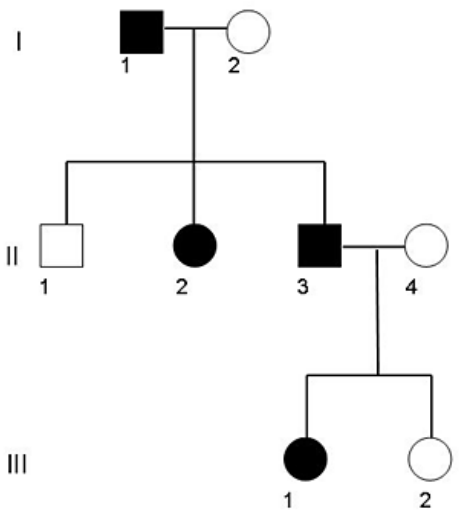
Justin Lack
NIAID Collaborative Bioinformatics Resource (NCBR)
Frederick National Laboratory for Cancer Research
October 11, 2018

Presentation Outline

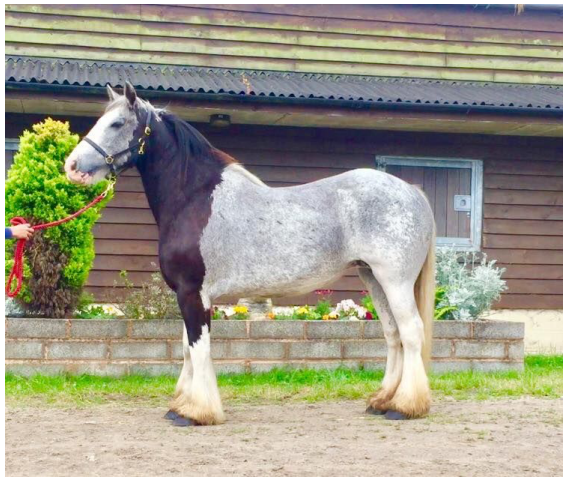
- Experimental Design Considerations in Variant Analysis
  - Germline vs Somatic Variant Detection
  - Whole Genome vs Whole Exome Sequencing
  - Best Practices
- Variant Calling Pipelines from NCBR/CCBR
  - Pipeline Performance
  - Using Pipeliner on Biowulf



Germline vs Somatic Variation

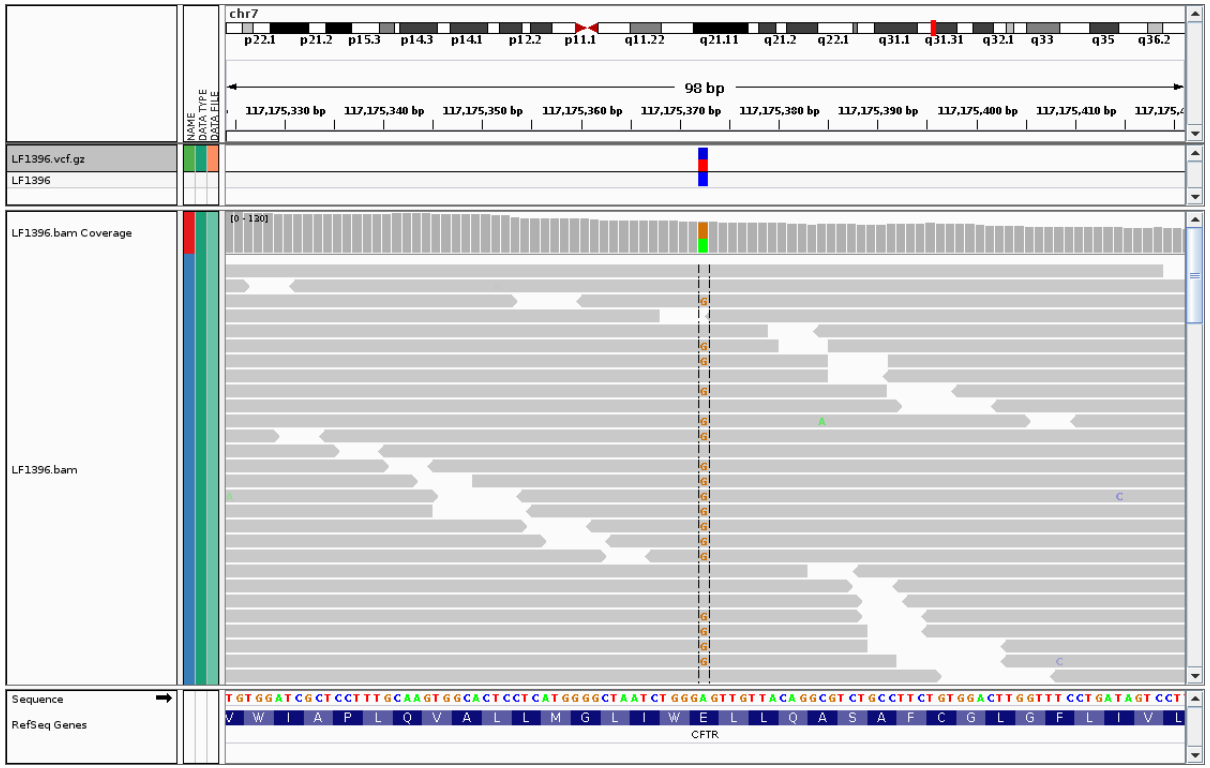


- Germline Variant Analysis
  - Heritable genetic variation
  - Large cohort analysis -> GWAS/Burden testing (quantitative)
  - Small cohort analysis -> Candidate gene identification (qualitative)
  - Pedigree analysis -> variant/disease co-segregation
- Somatic Variant Analysis
  - Non-heritable genetic variation arising in non-germ cells
  - Tumor/Normal or tumor-only analysis
  - Somatic mosaicism (e.g., Neurofibromatosis)
- ***Very different expectations in terms of allele frequency distribution***

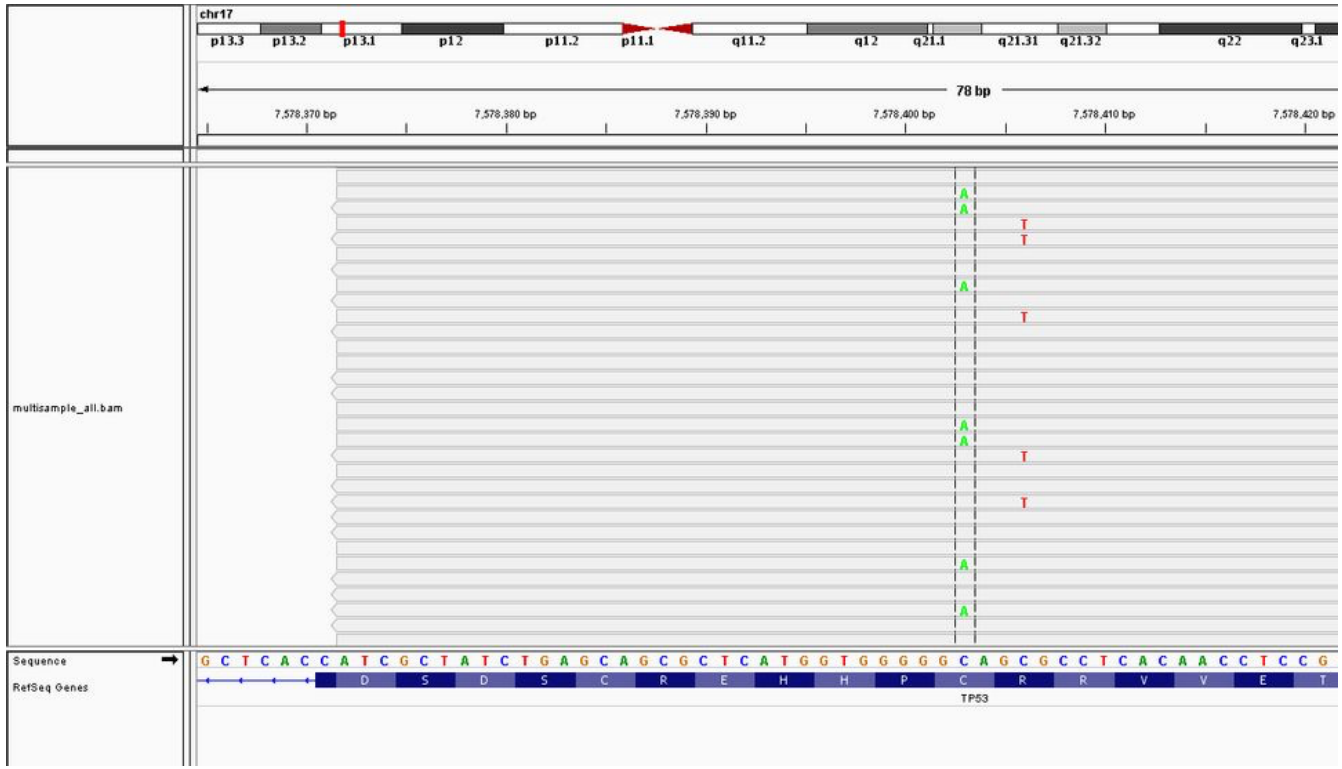


Germline vs Somatic Variant Calling

- Potentially very different allele frequency expectations



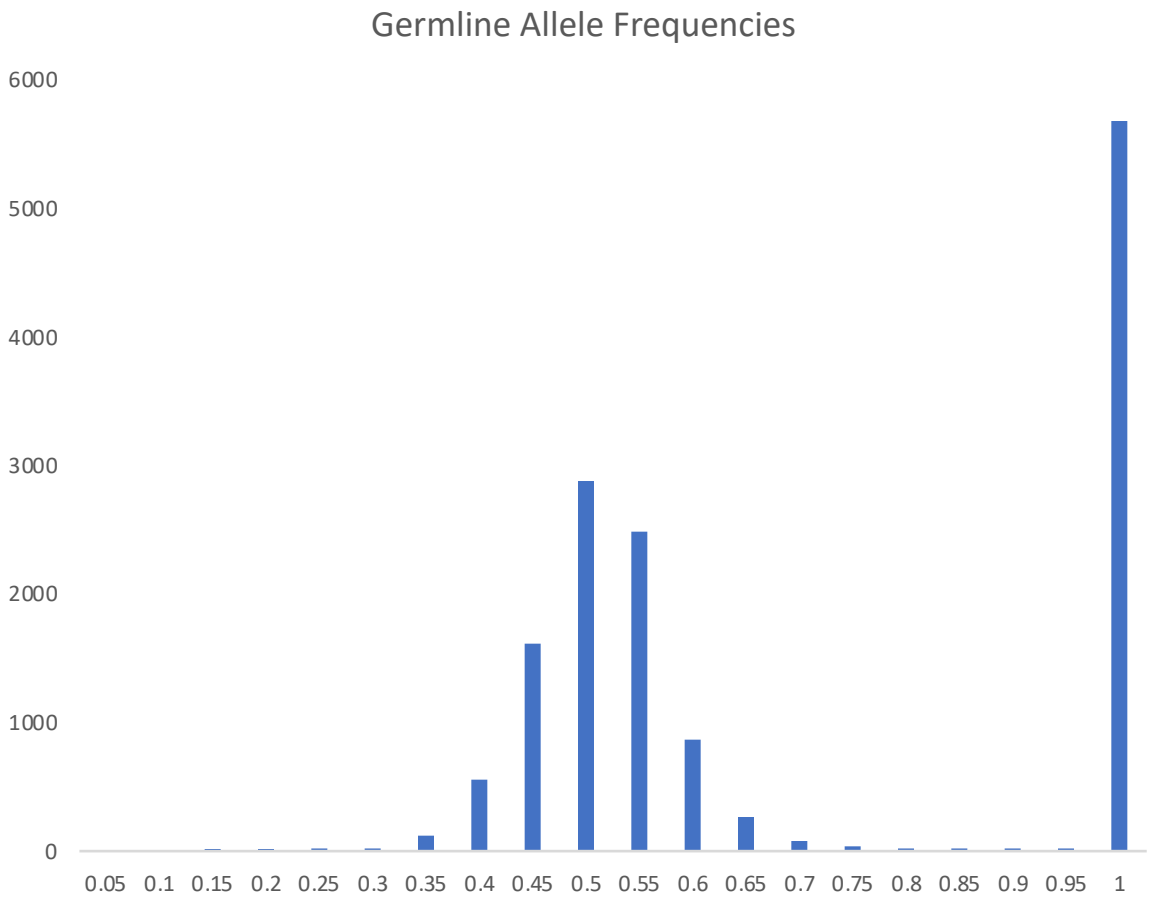
Germline - ~0.5 read proportions

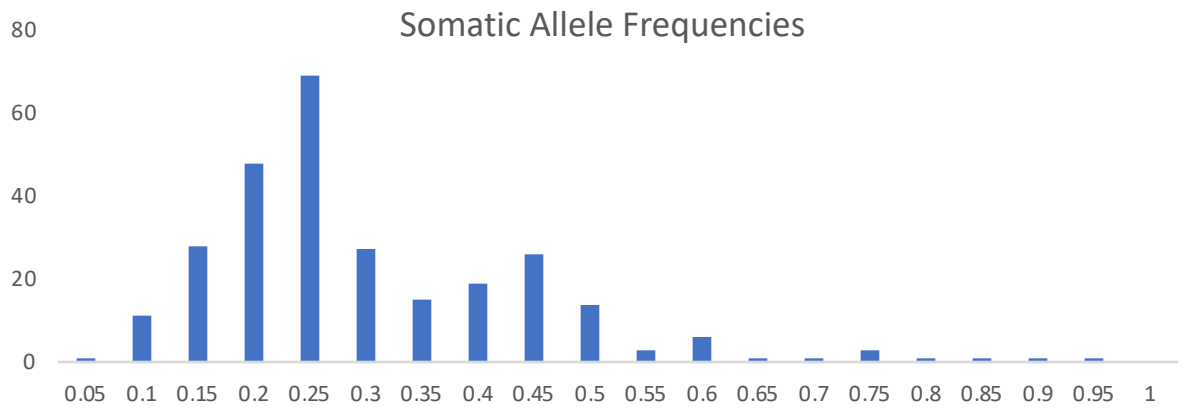


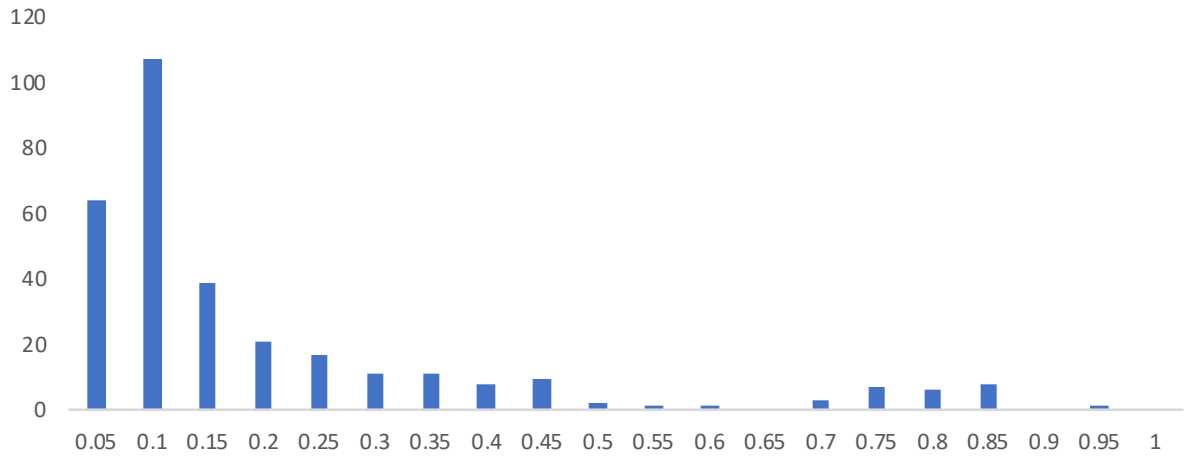
Somatic - ~0.3 read proportions

Germline vs Somatic Variant Calling

- Potentially very different allele frequency expectations



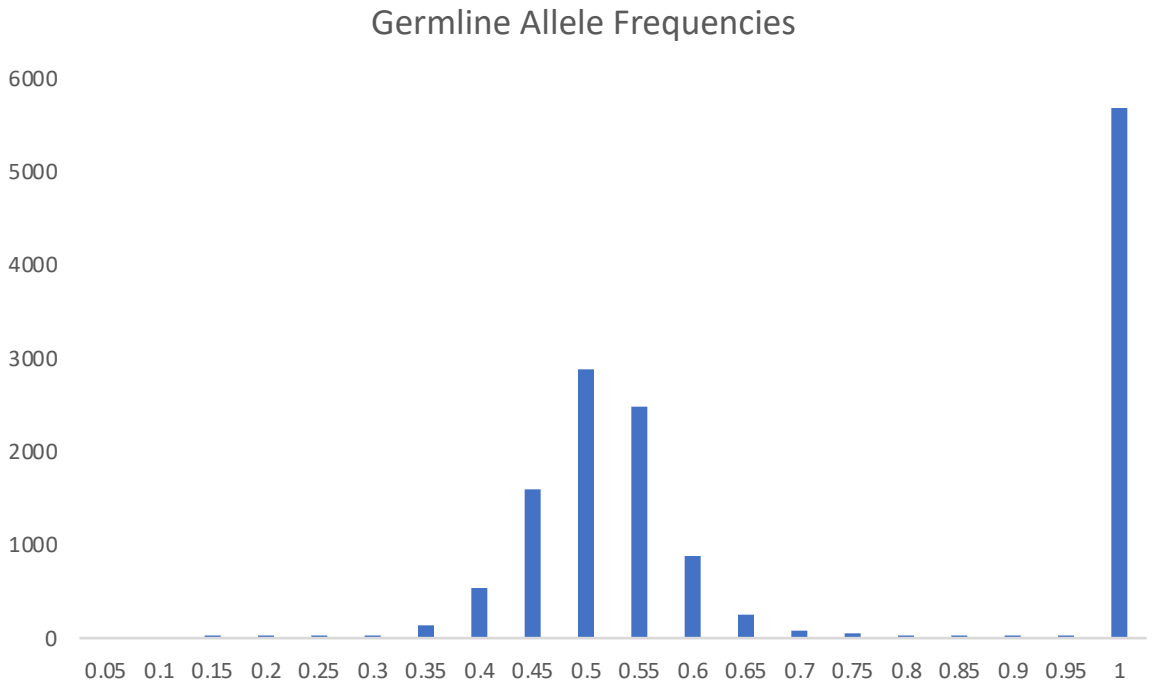




Germline vs Somatic Variant Calling

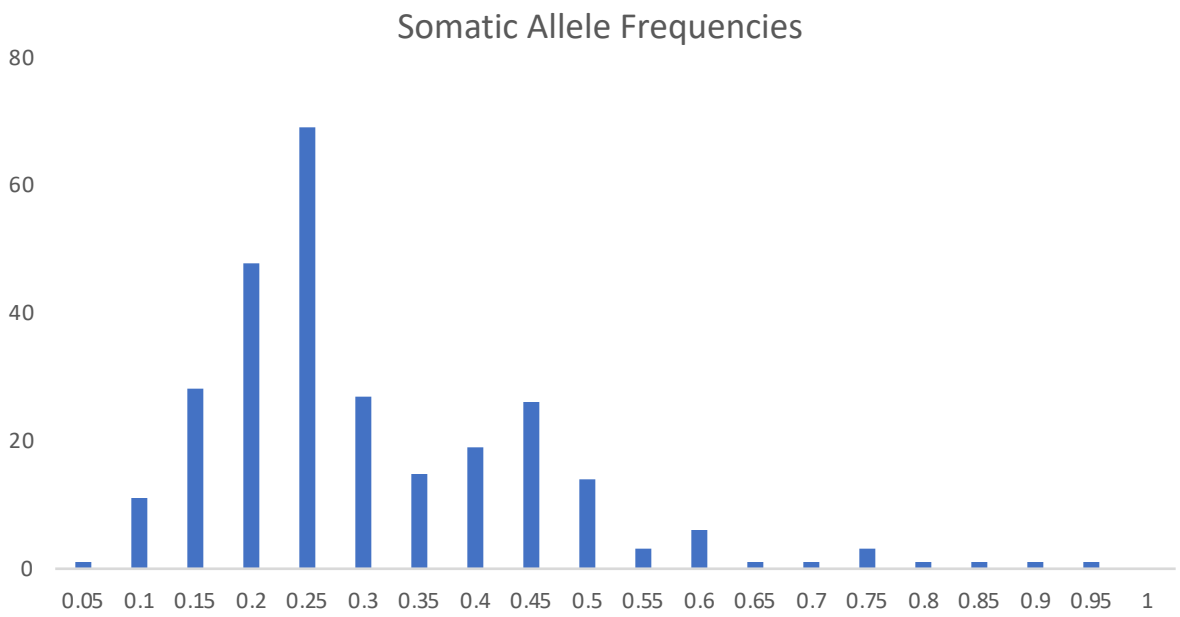
Within-sample germline allele frequency variance is driven by sample quality and sequencing quality

- **Amount of sample input**
- **Sequencing depth**
- **Read/base quality**



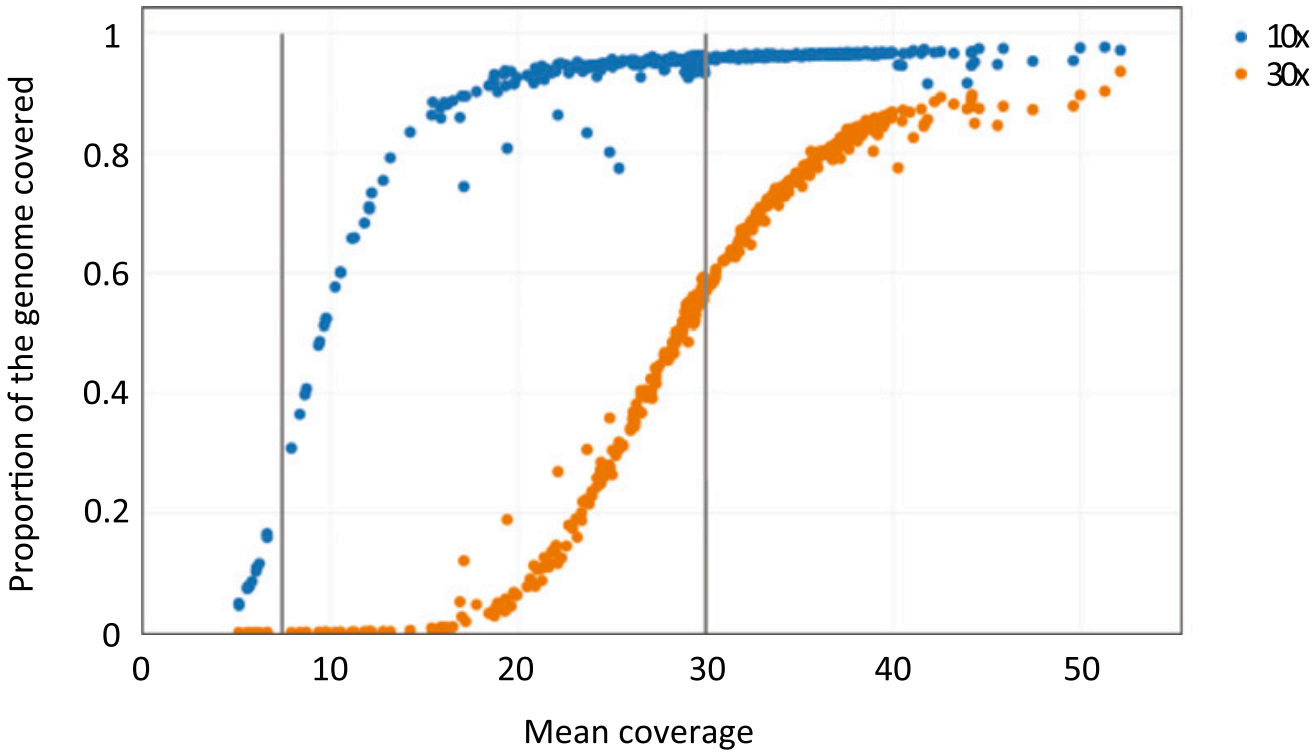
Somatic allele frequency variance is driven by MANY more factors:

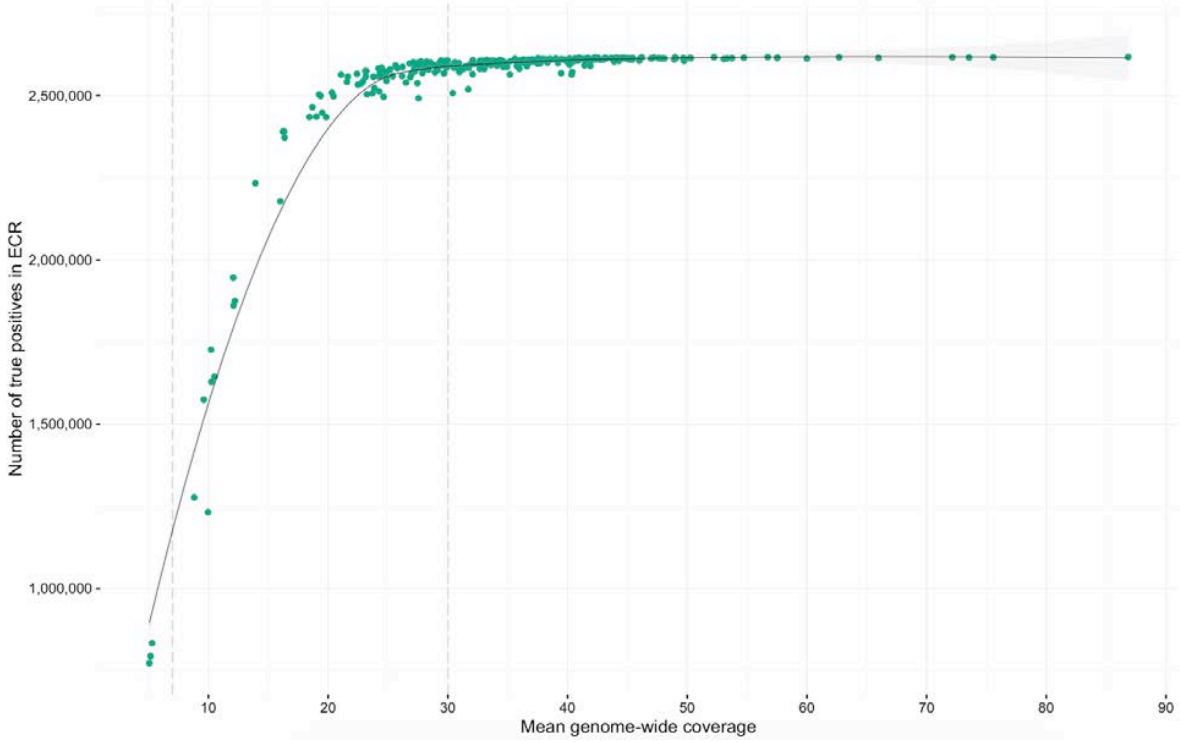
- **Sample quality and sequencing quality**
- **Subclonality/heterogeneity**
- **Copy number variation**
- **Tumor purity**



Depth Effects - Germline

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

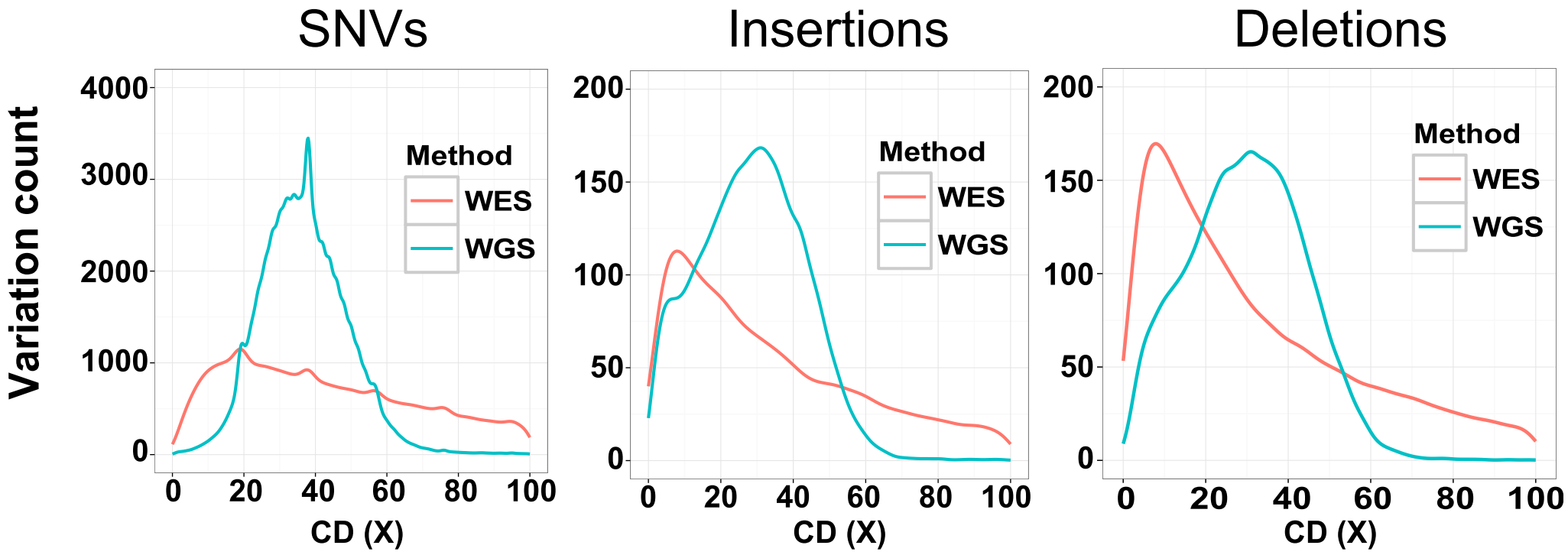




Telenti et al., 2016 *PNAS*

Depth Effects - Germline

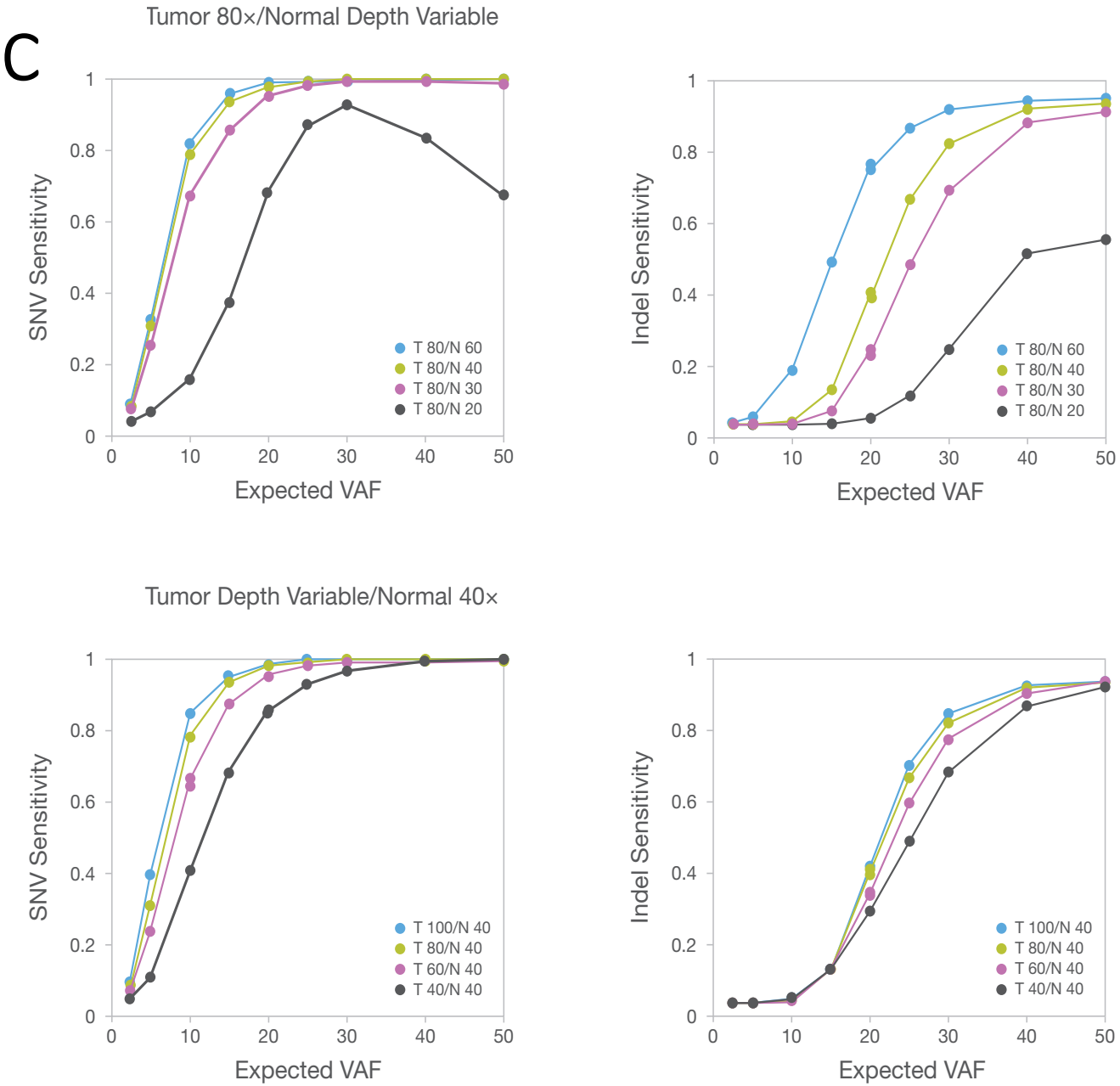
- ~30X target for genome data (below)
- ~50X target for exome, due to increased depth variance



Belkadi et al., 2015 *PNAS*

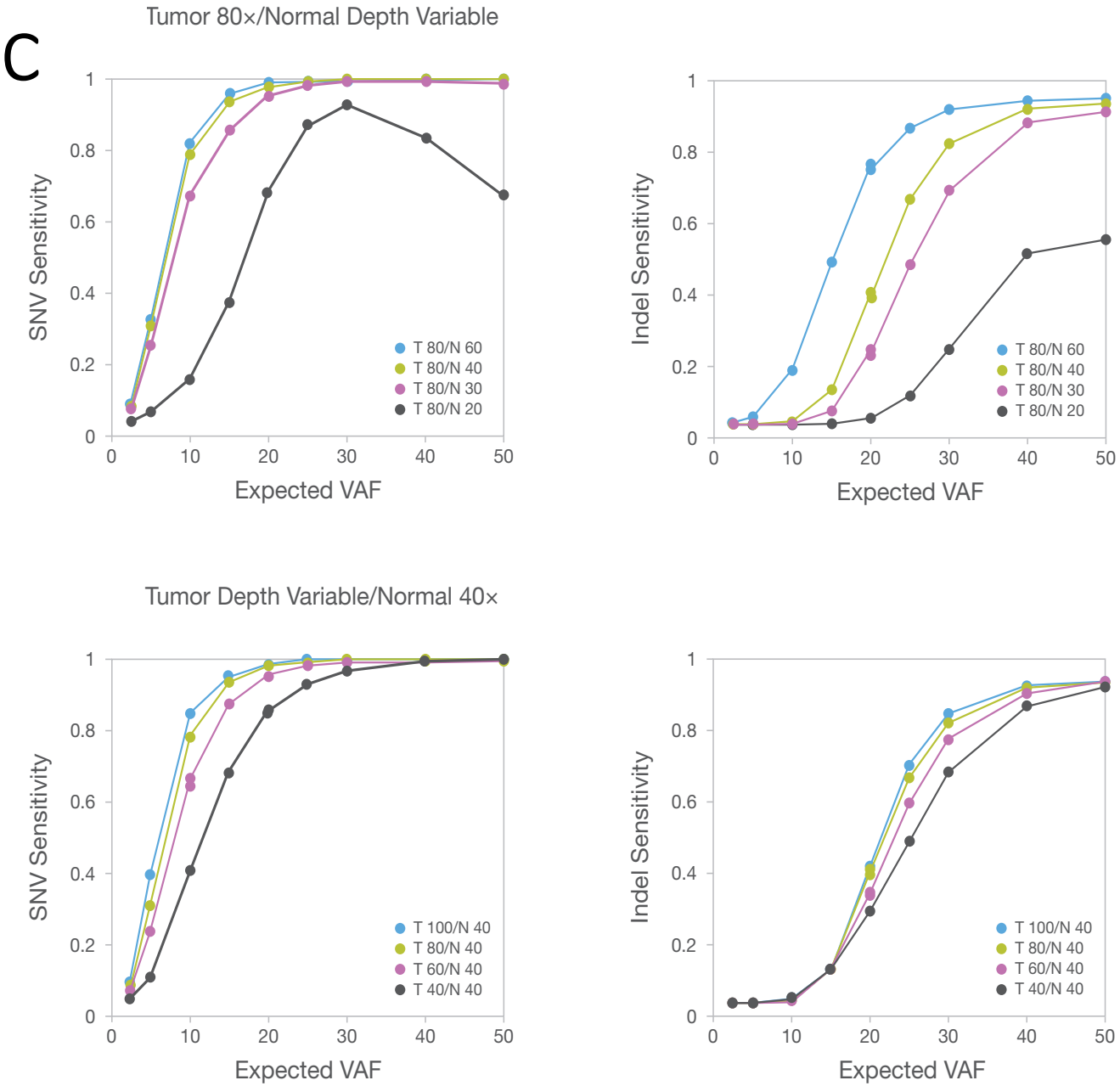
Depth Effects - Somatic

- "Spike-in" simulation analysis of impact of tumor purity and sequencing depth
- Simulated somatic variants into real germline whole exome sequencing on the NA12878 sample



Depth Effects - Somatic

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



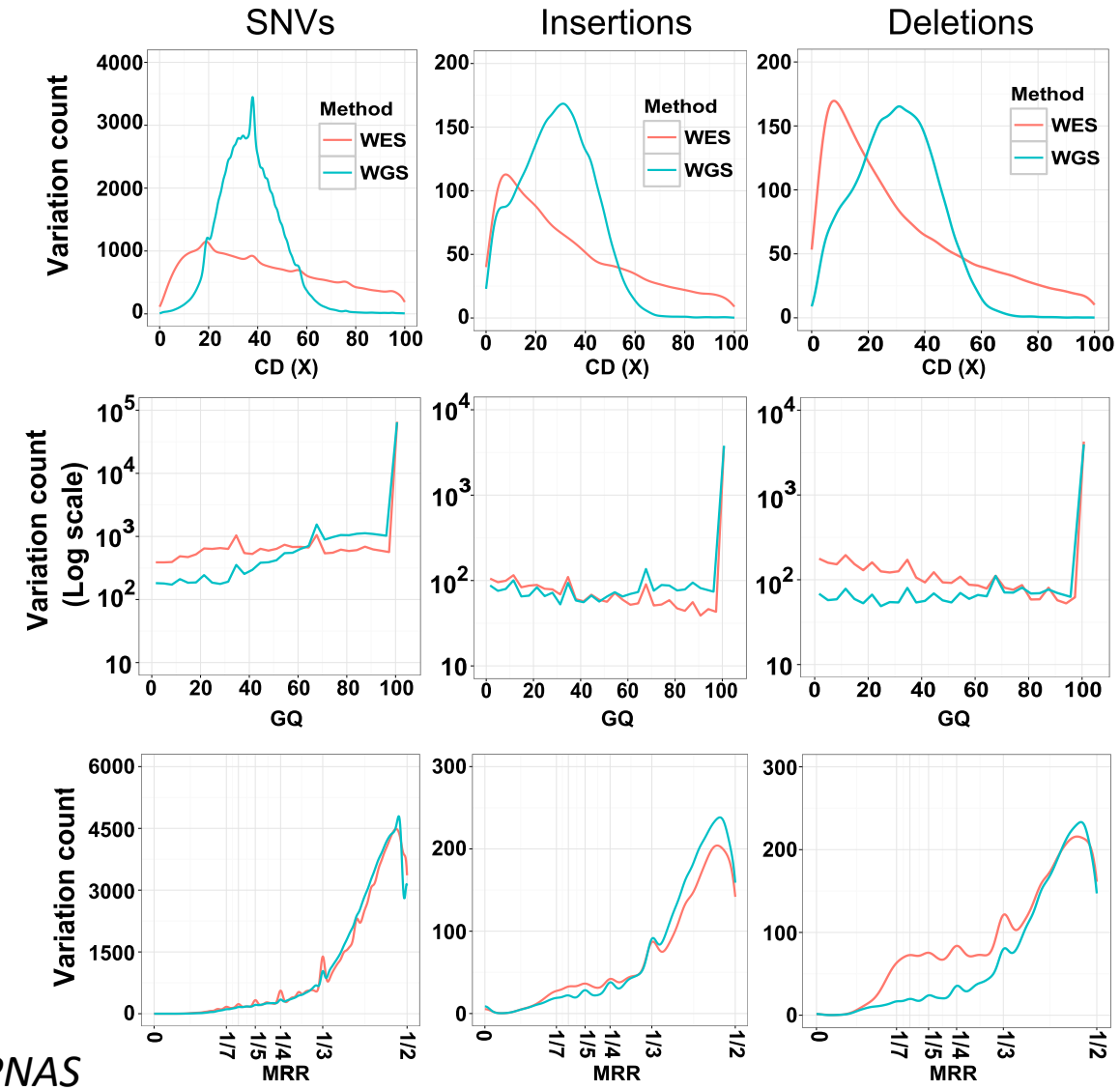
Exome vs Whole Genome Sequencing

Exome vs Whole Genome Sequencing

- Exome Sequencing
  - Covers ~5% of genome (depending on capture kit)
  - Allows for high depth targeting
  - Most reasonable option for somatic variant analysis
  - Poor 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

Exome vs Whole Genome Sequencing

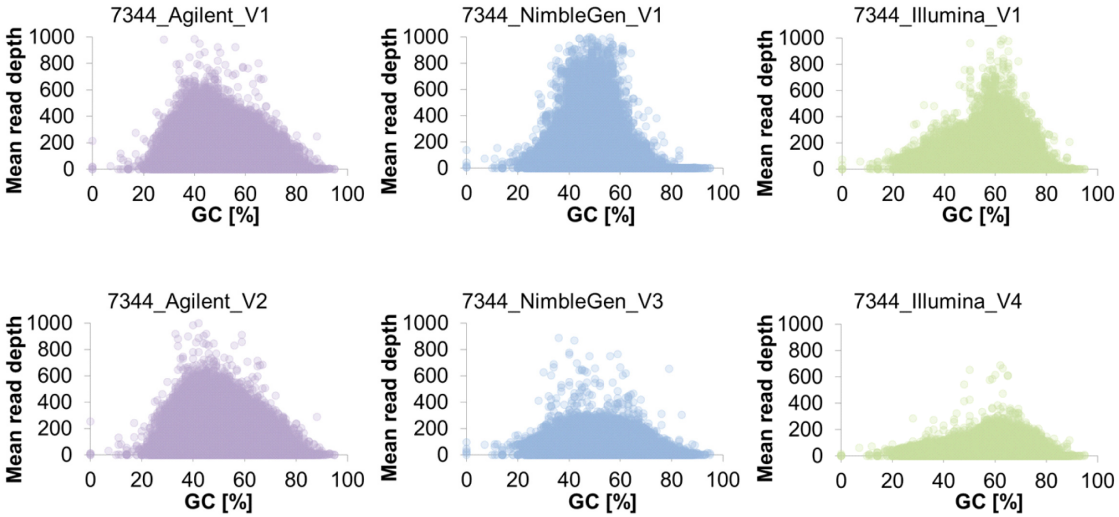
- Depth variance MUCH higher for exome
- ~2-fold more variants with GQ < 20 for exome
- Read ratio for heterozygous variants significantly skewed for exome
  - Especially pronounced for INDELs

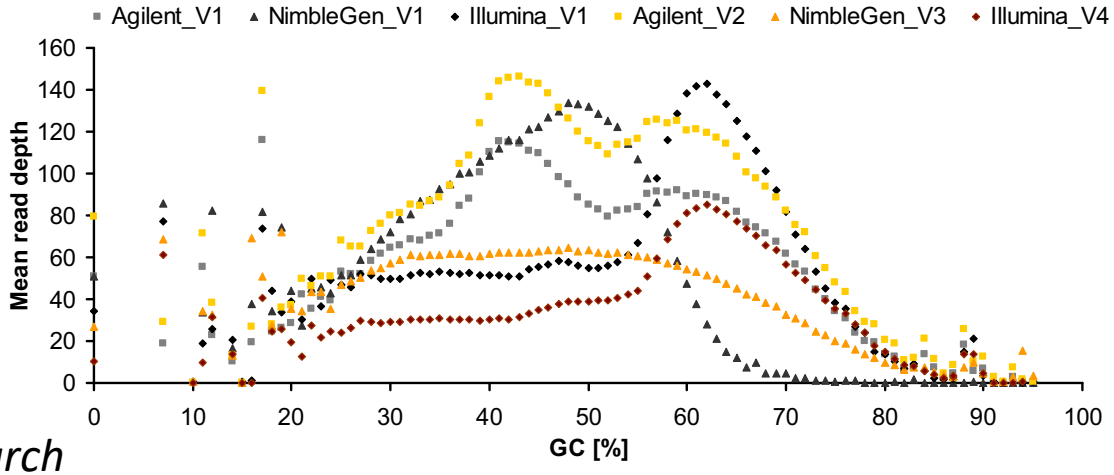


Belkadi et al., 2015 *PNAS*

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*

Exome vs Whole Genome Sequencing

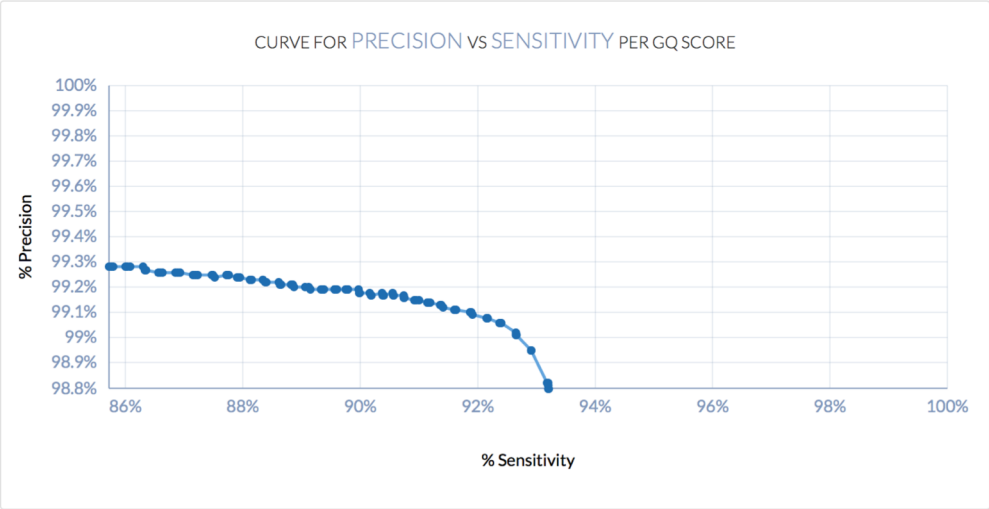
- Sure, there's bias in WES introduced due to capture, but does it significantly affect variant calling?
- Genome in a bottle (GIAB) truth sample (NA12878)
  - 50X WES and 30 WGS available from exact same sample
  - Processed both WES and WGS through identical pipelines
  - Compared both variant sets to GIAB truth set
  - Used only exonic sites targeted in WES capture for performance assessment

Exome vs Whole Genome Sequencing

Exome (50X WES)

PRECISION	98.8%
RECALL	93.21%
F-MEASURE	95.92%

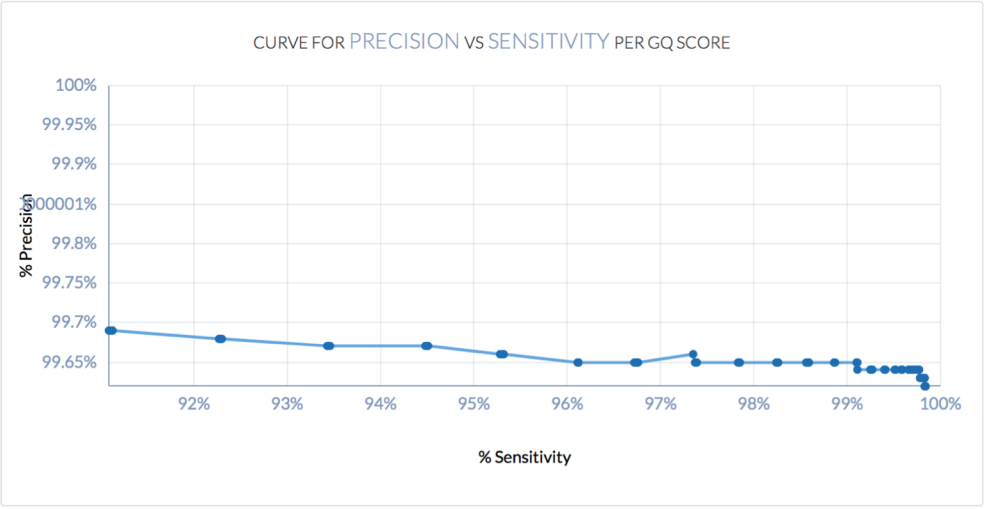
TRUE-POSITIVES	35,768
FALSE-POSITIVES	436
FALSE-NEGATIVES	2,607



Genome (30X WGS)

PRECISION	99.62%
RECALL	99.84%
F-MEASURE	99.73%

TRUE-POSITIVES	38,314
FALSE-POSITIVES	147
FALSE-NEGATIVES	61



Exome vs Whole Genome Sequencing

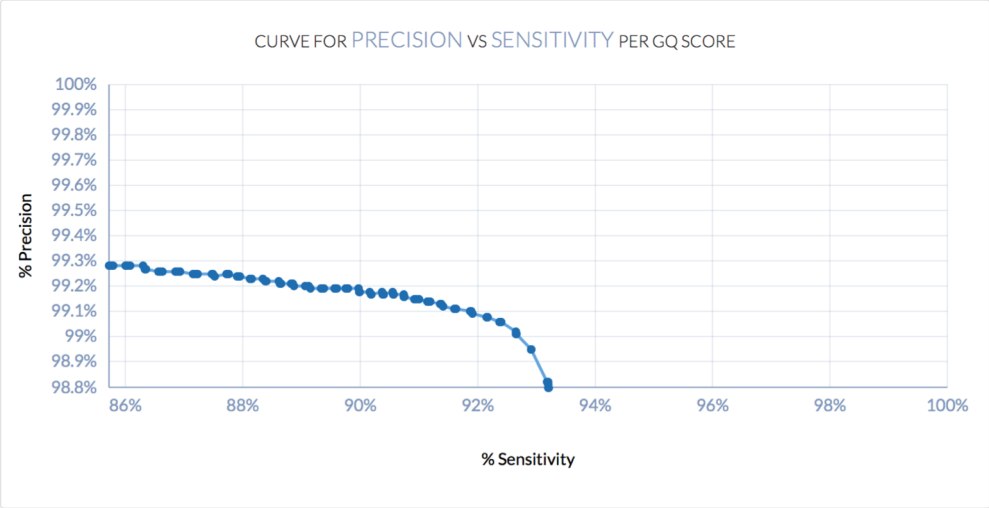
3X higher False Positive rate, and >40X higher False Negative rate for exome!!!

	False Negative Rate	False Positive Rate
50X Whole Exome	0.067934853	0.011361564
30X Genome (Exome Sites)	0.001589577	0.003830619

Exome (50X WES)

PRECISION	98.8%
RECALL	93.21%
F-MEASURE	95.92%

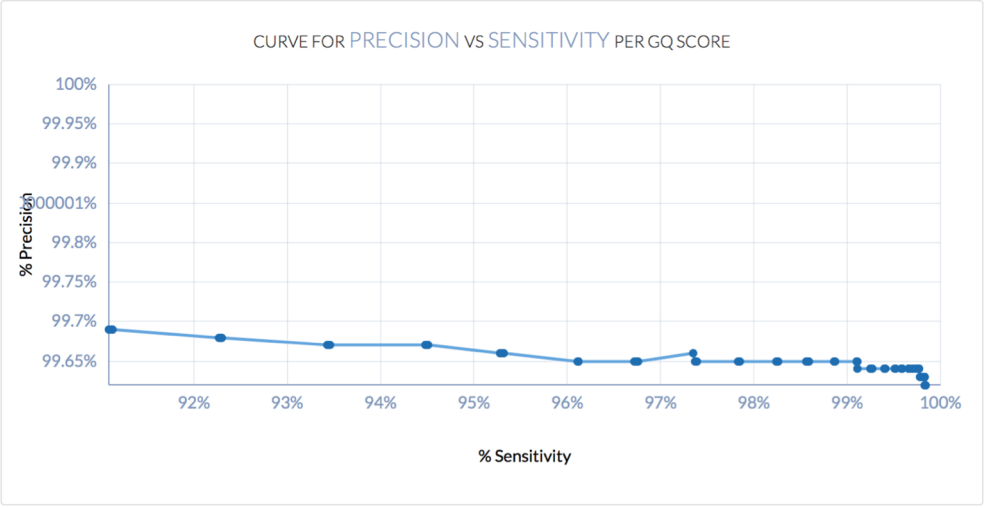
TRUE-POSITIVES	35,768
FALSE-POSITIVES	436
FALSE-NEGATIVES	2,607



Genome (30X WGS)

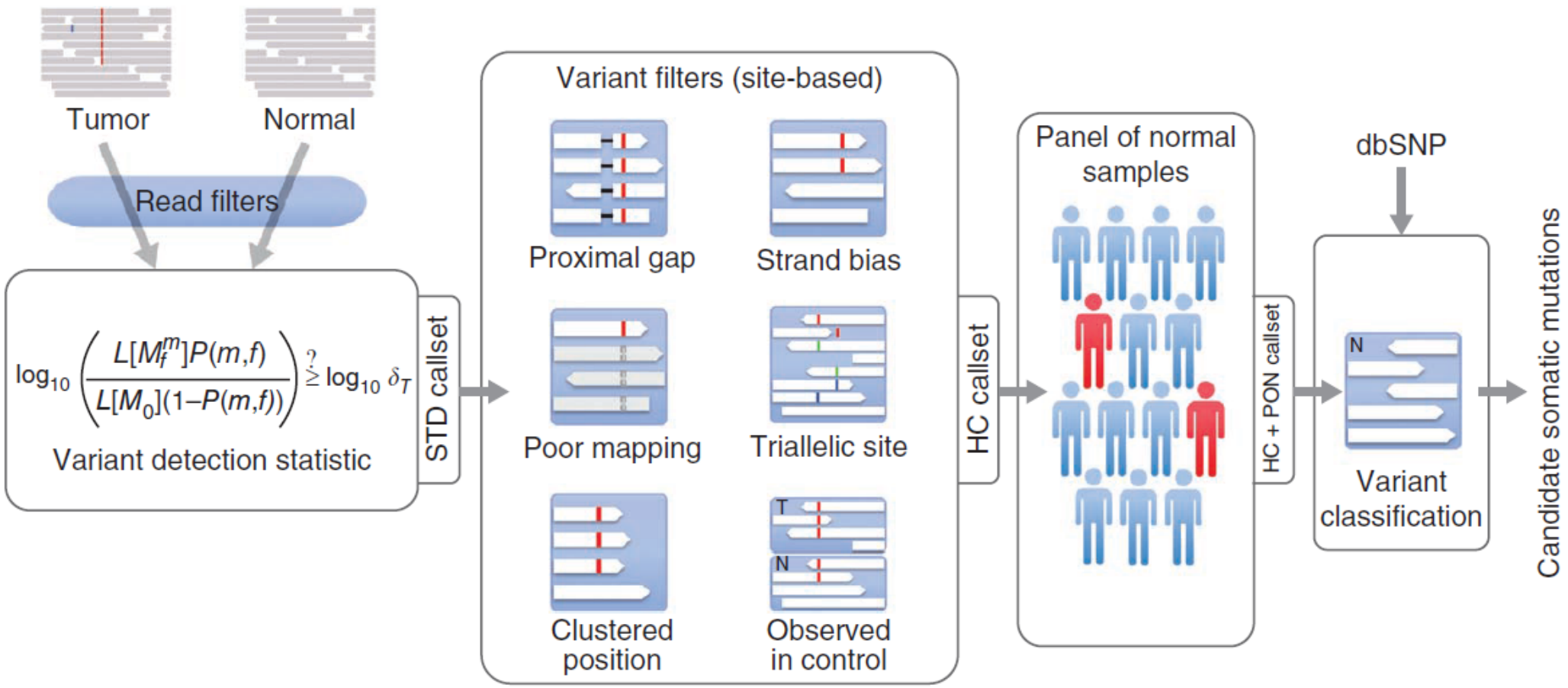
PRECISION	99.62%
RECALL	99.84%
F-MEASURE	99.73%

TRUE-POSITIVES	38,314
FALSE-POSITIVES	147
FALSE-NEGATIVES	61



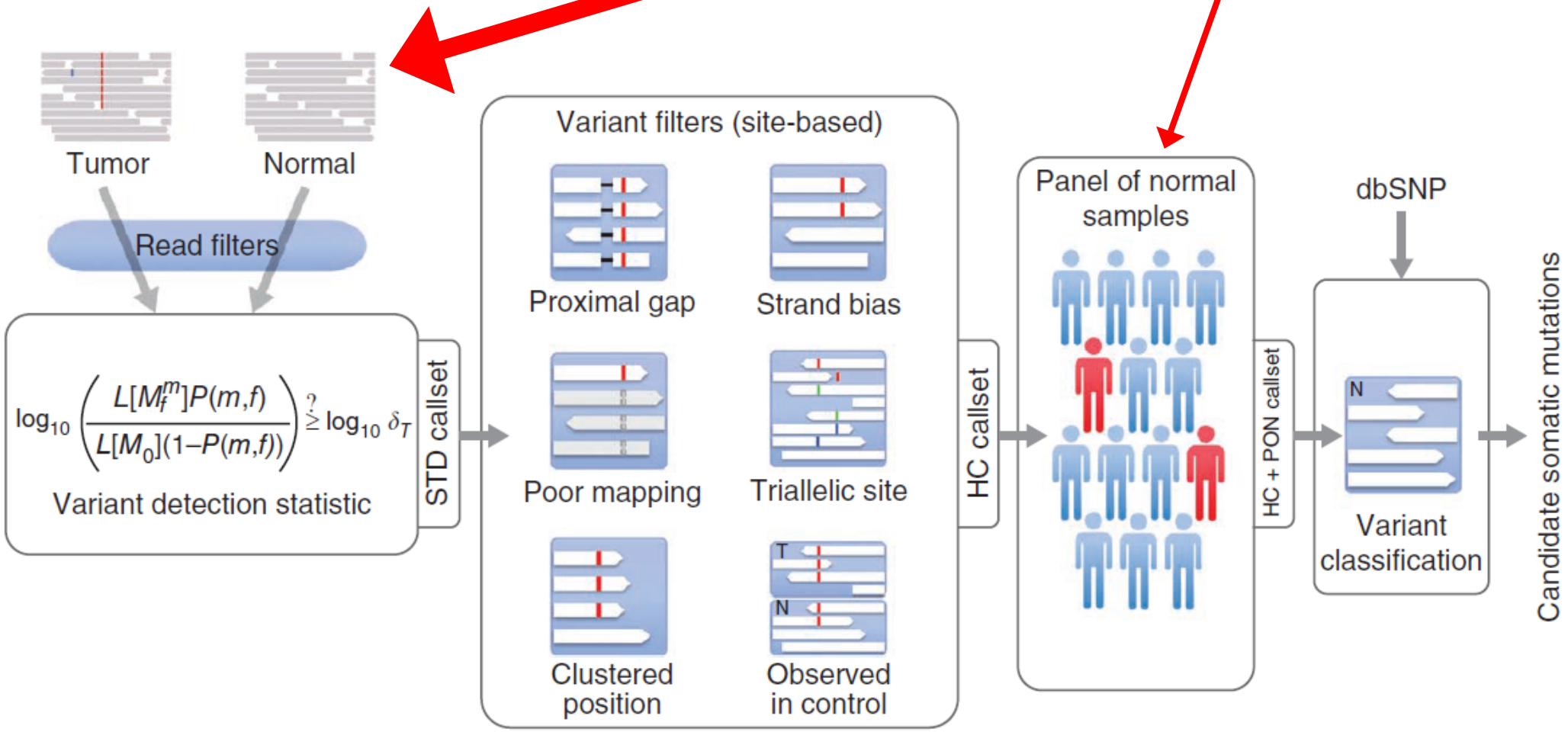
Somatic Variant Calling – Considerations and Best Practices

Paired Tumor/Normal vs Tumor-only Somatic Variant Calling



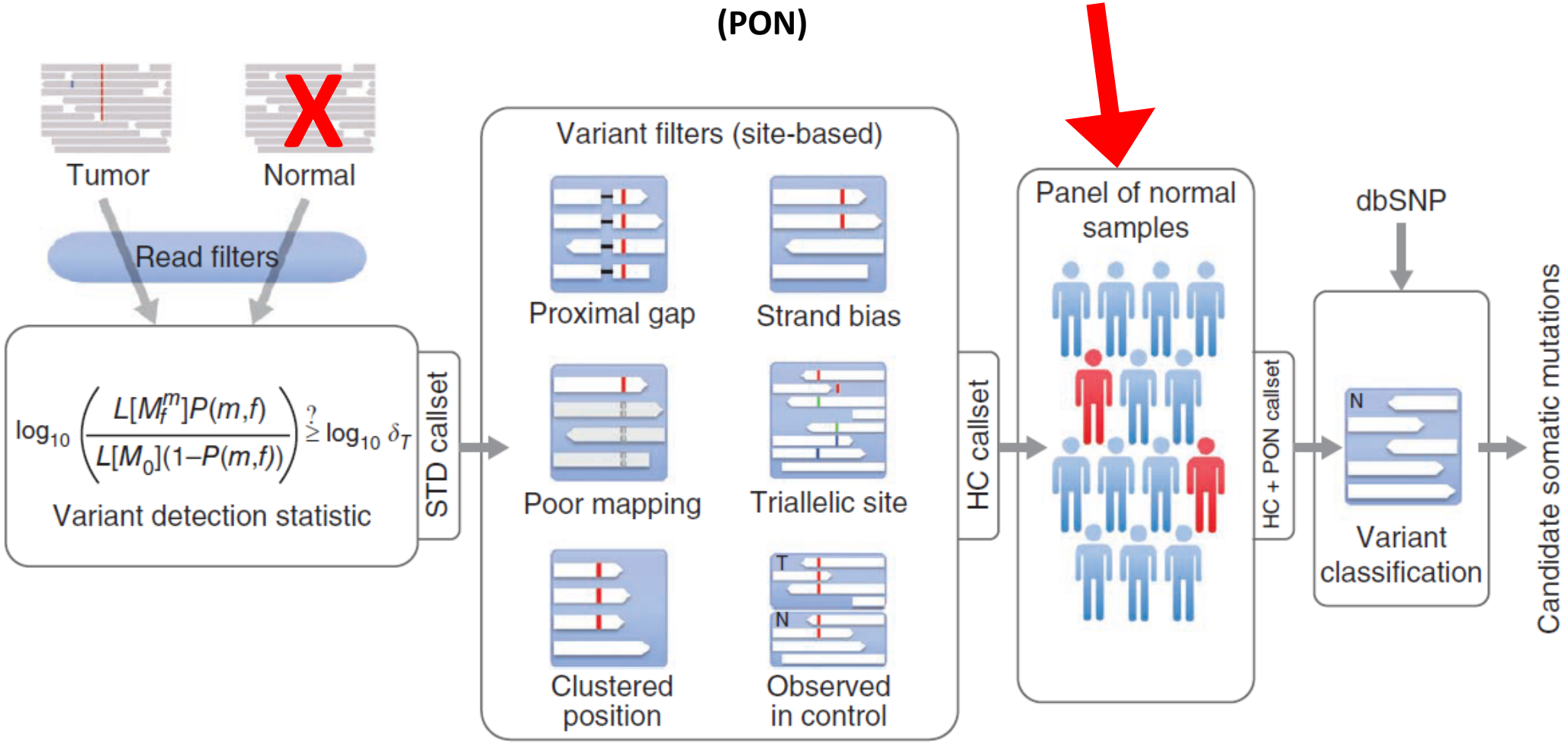
Paired Tumor/Normal vs Tumor-only Somatic Variant Calling

For tumor/normal calling, strong prior on variant evidence in germline sample



Paired Tumor/Normal vs Tumor-only Somatic Variant Calling

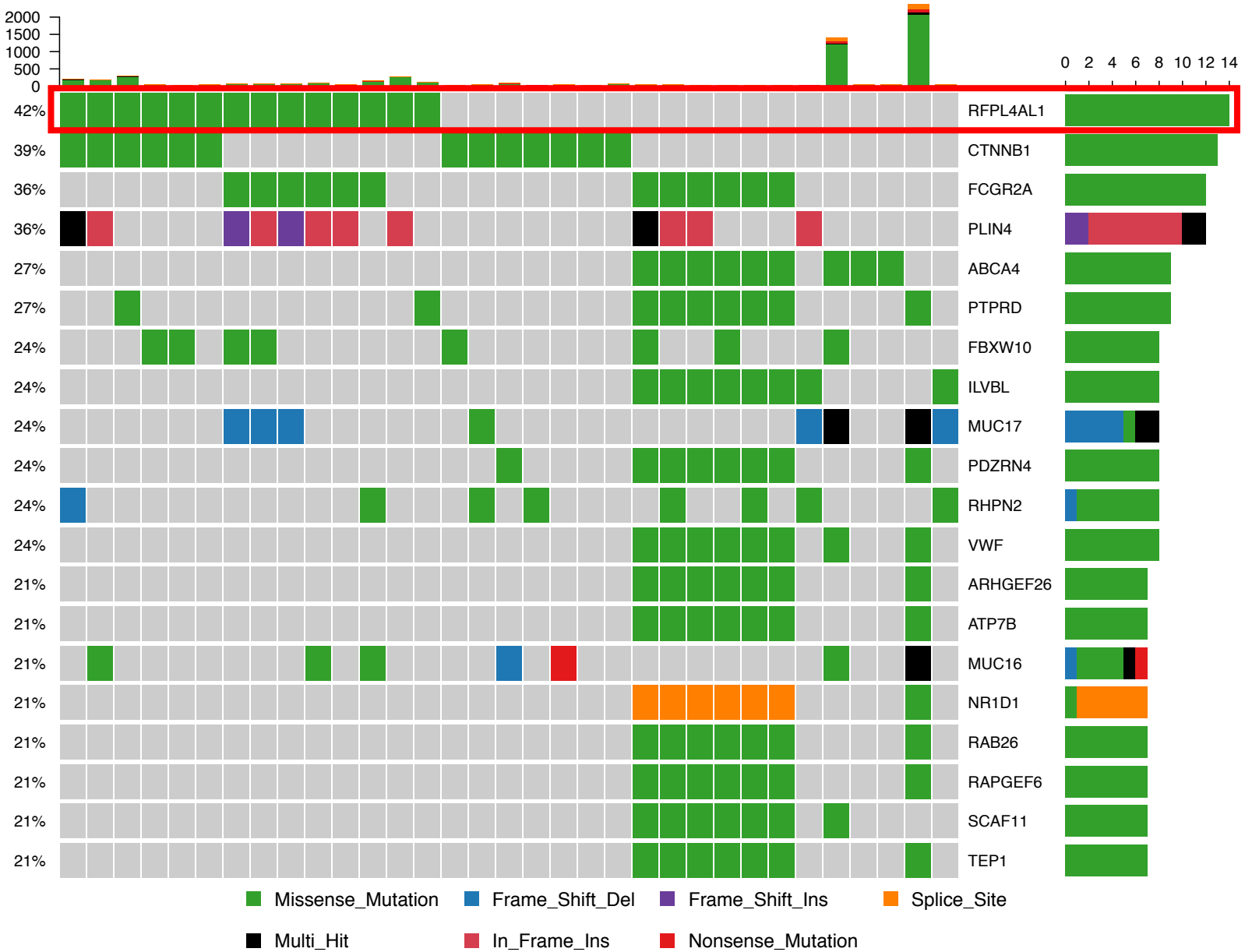
For tumor-only calling, germline contamination removed via a **panel of normal (PON)**



Panel of Normals (PON)

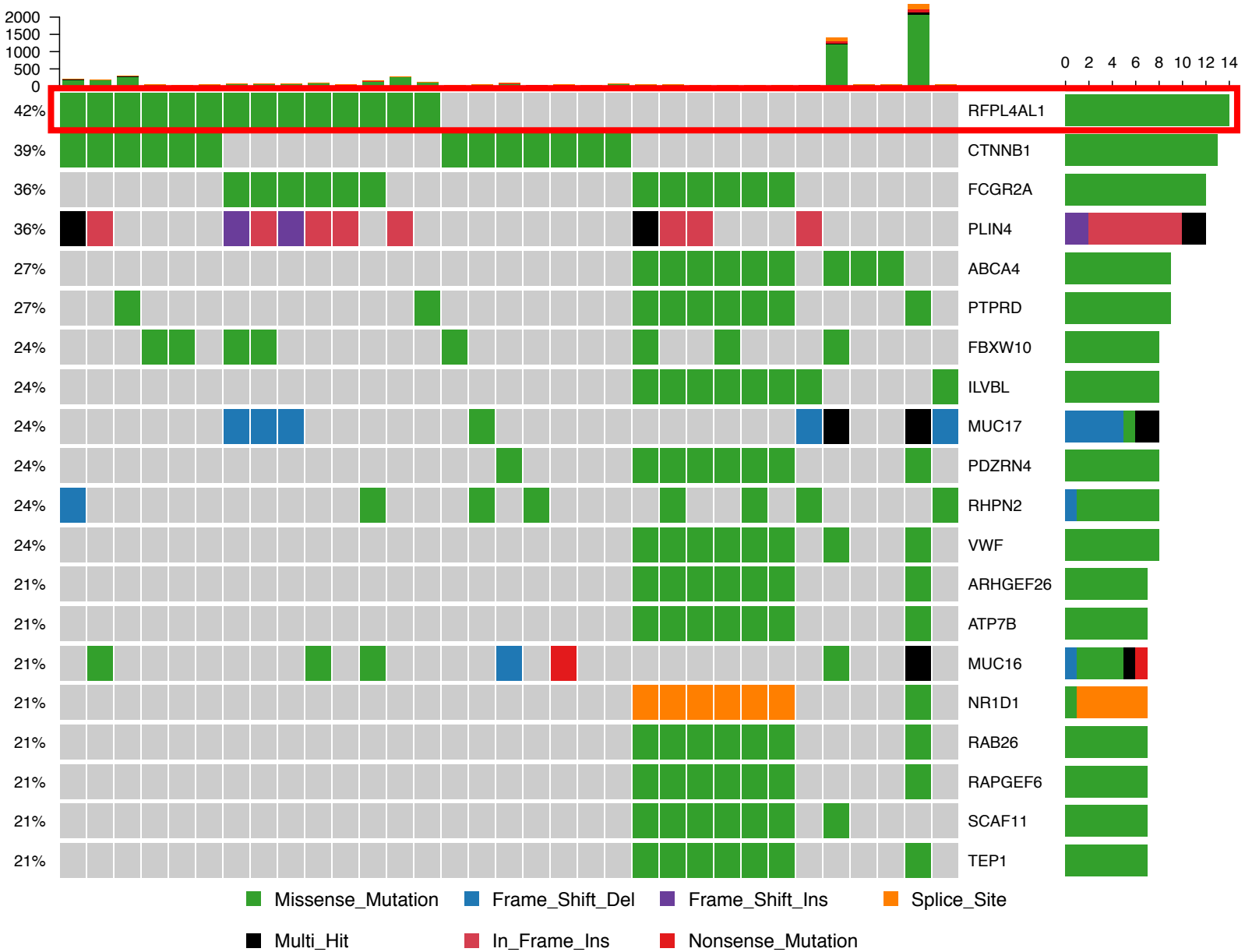
- Collection of individuals assumed to be “normal”
- Used to augment population frequency databases (e.g., ExAC, GnomAD)
  - Population databases are highly filtered and curated
  - Many segregating germline variants are missing from population databases because they occur in challenging portions of the genome to call genotypes
- Also useful for removing systematic sequencing and mapping artifacts…

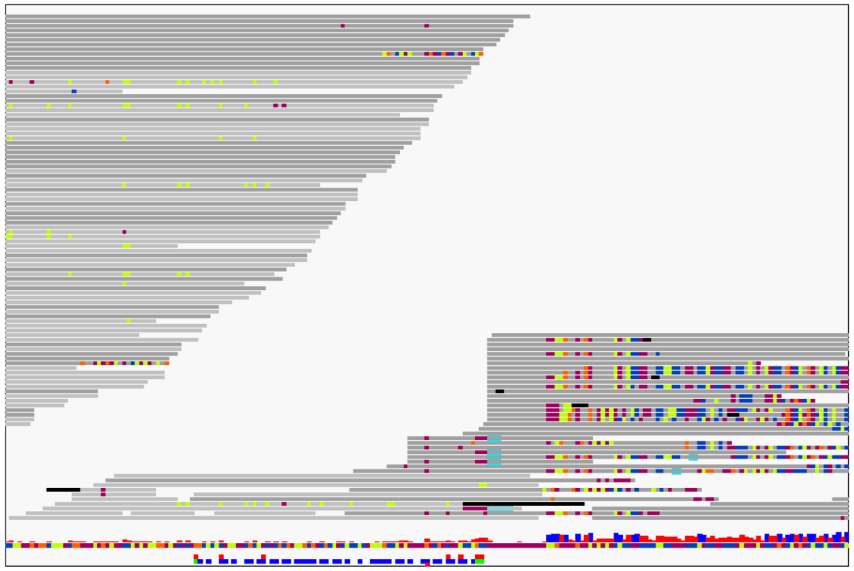
Panel of Normals (PON)



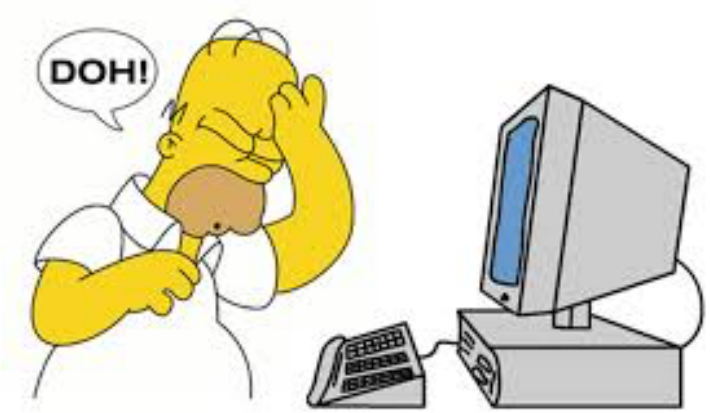
- This is a tumor/normal cohort of adrenocortical carcinoma (ACC)
- Most common driver gene for ACC is known to be beta-catenin (CTNNB1)
- Somatic variant analysis for our cohort suggests RFPL4AL1 is the most frequently mutated gene
- WE'RE GONNA BE FAMOUS! PUBLISH IN NEJM! WOOHOO!

Panel of Normals (PON)

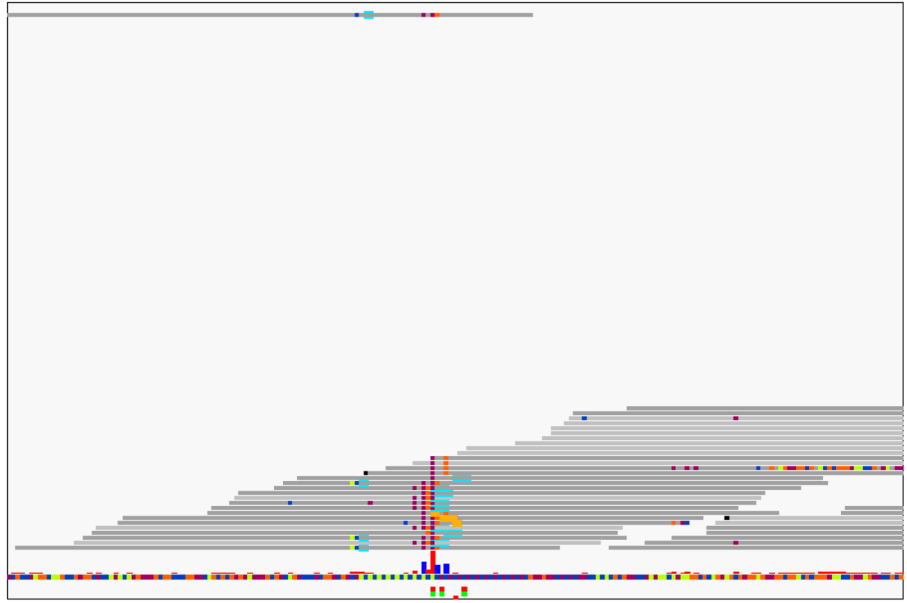




- This is a tumor/normal cohort of adrenocortical carcinoma (ACC)
- Most common driver gene for ACC is CTNNB1
- Somatic variant analysis for our cohort suggests RFPL4AL1 is the most frequently mutated gene



Panel of Normals (PON)



Panel of Normals (PON)

- Collection of individuals assumed to be “normal”
- Used to augment population frequency databases (e.g., GnomAD)
  - Population databases are highly filtered and curated
  - Many segregating germline variants are missing from population databases because they occur in challenging portions of the genome to call genotypes
- **Even with matched germline, these artifacts will be prevalent, and because they are systematic, they can be widespread**

Panel of Normals (PON)

- Collection of individuals assumed to be “normal”
- Used to augment population frequency databases (e.g., GnomAD)
  - Population databases are highly filtered and curated
  - Many segregating germline variants are missing from population databases because they occur in challenging portions of the genome to call genotypes
- **Even with matched germline, these artifacts will be prevalent, and because they are systematic, they can be widespread**
- **PON that was processed in (approximately) the same way as the case samples can remove many of these artifacts**

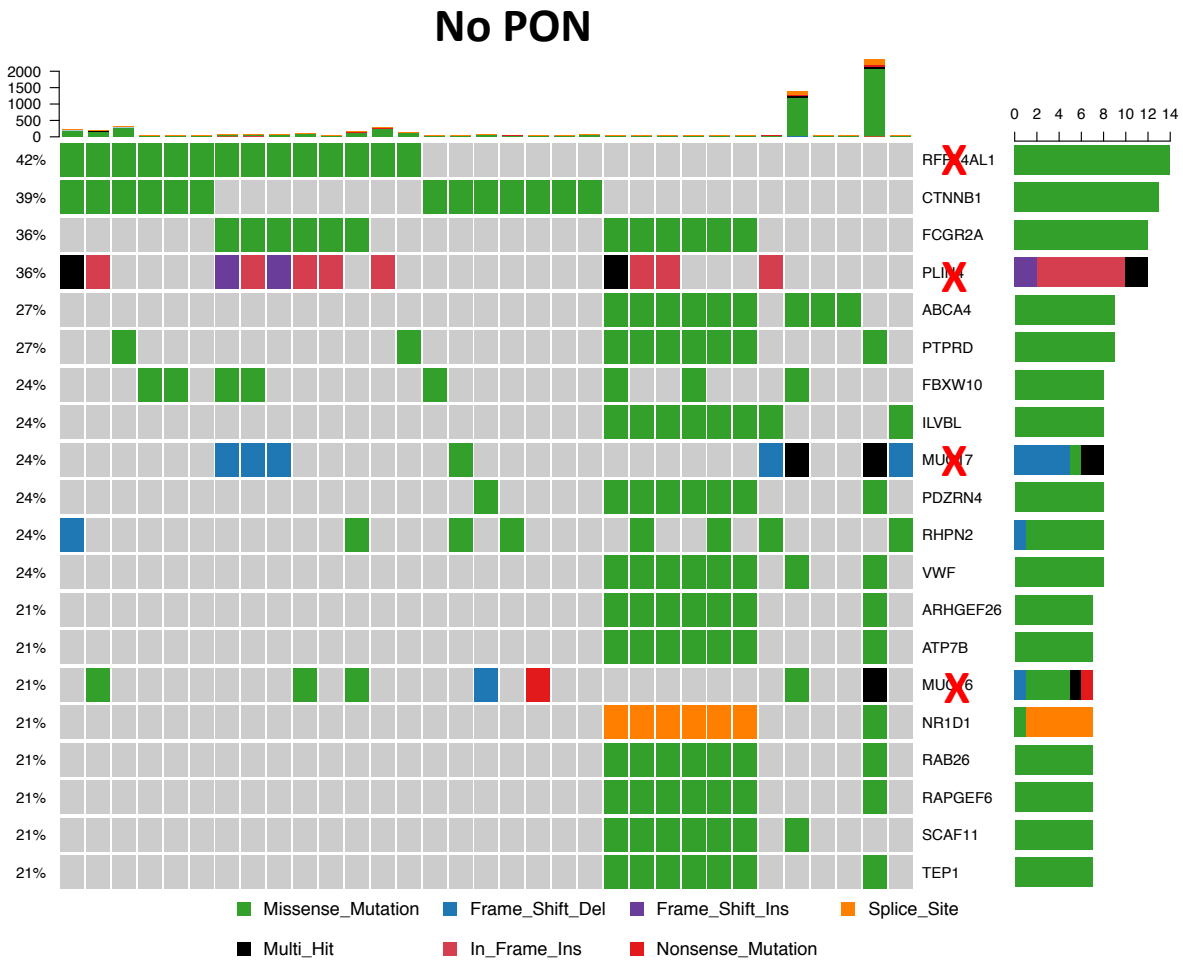
PON Development at CCBR/NCBR

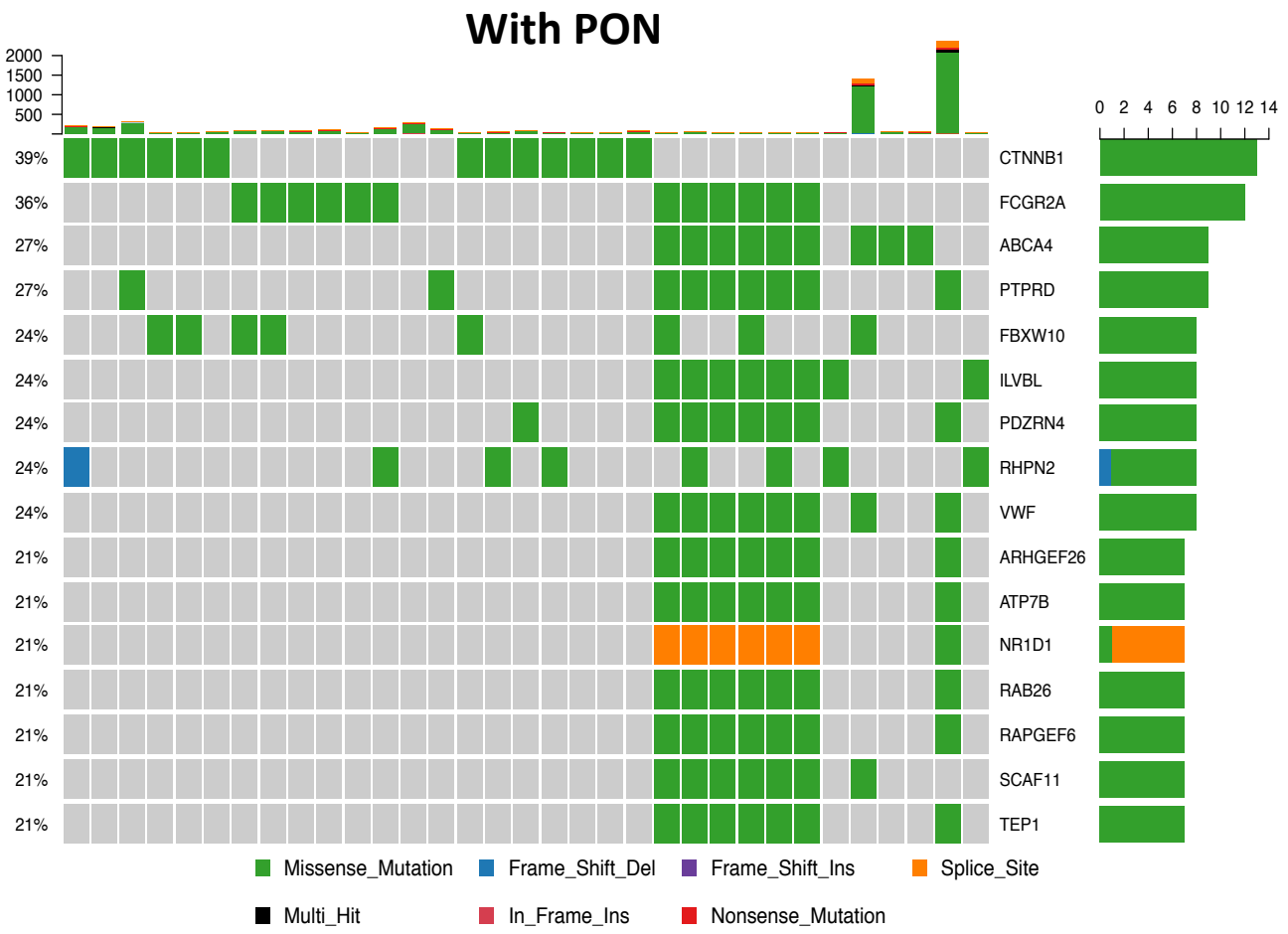
- 211 unaffected spouses from diversity of NCBR/CCBR germline projects
- 445 additional germlines from “normals” in various publicly available databases
- All WES samples
  - Processed with variety of WES capture kits (Agilent, Illumina, etc.)
  - Sequenced on multiple Illumina platforms
- Processed each sample individually in PON mode in MuTect2
- Retained only variants present in >=2 samples

PON Development

- Annotated the entire raw PON with gene information using VEP
- Removed any variant in the gene region of a gene in the COSMIC v84 database
- Removed any variant in a confirmed gene from ClinVar that was annotated as Pathogenic, Potentially Pathogenic, Drug Response, and Risk Factor
- Removed any specific variant identified in ClinVar as “Associated”
- Pooled all remaining variants into a single PON

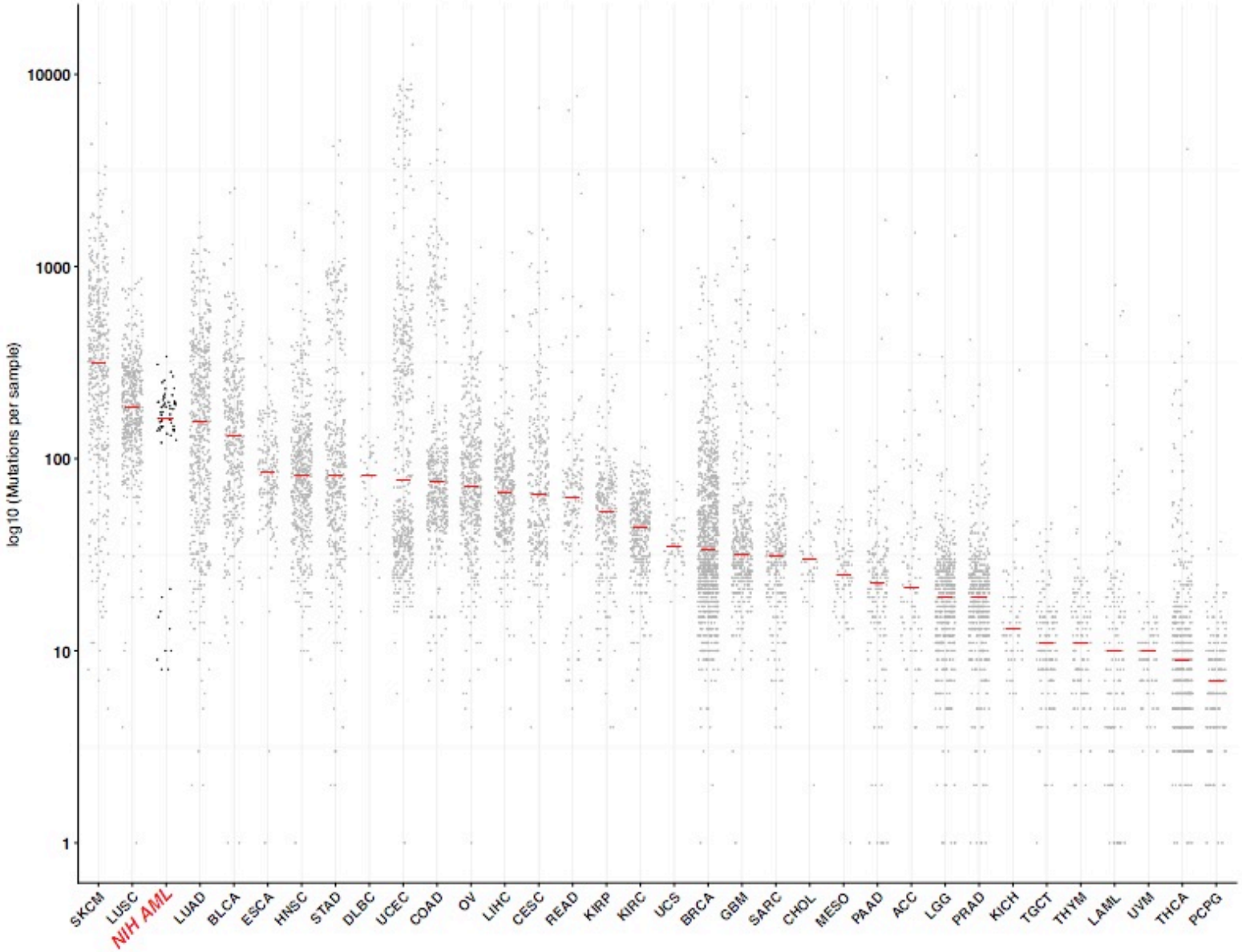
PON Performance





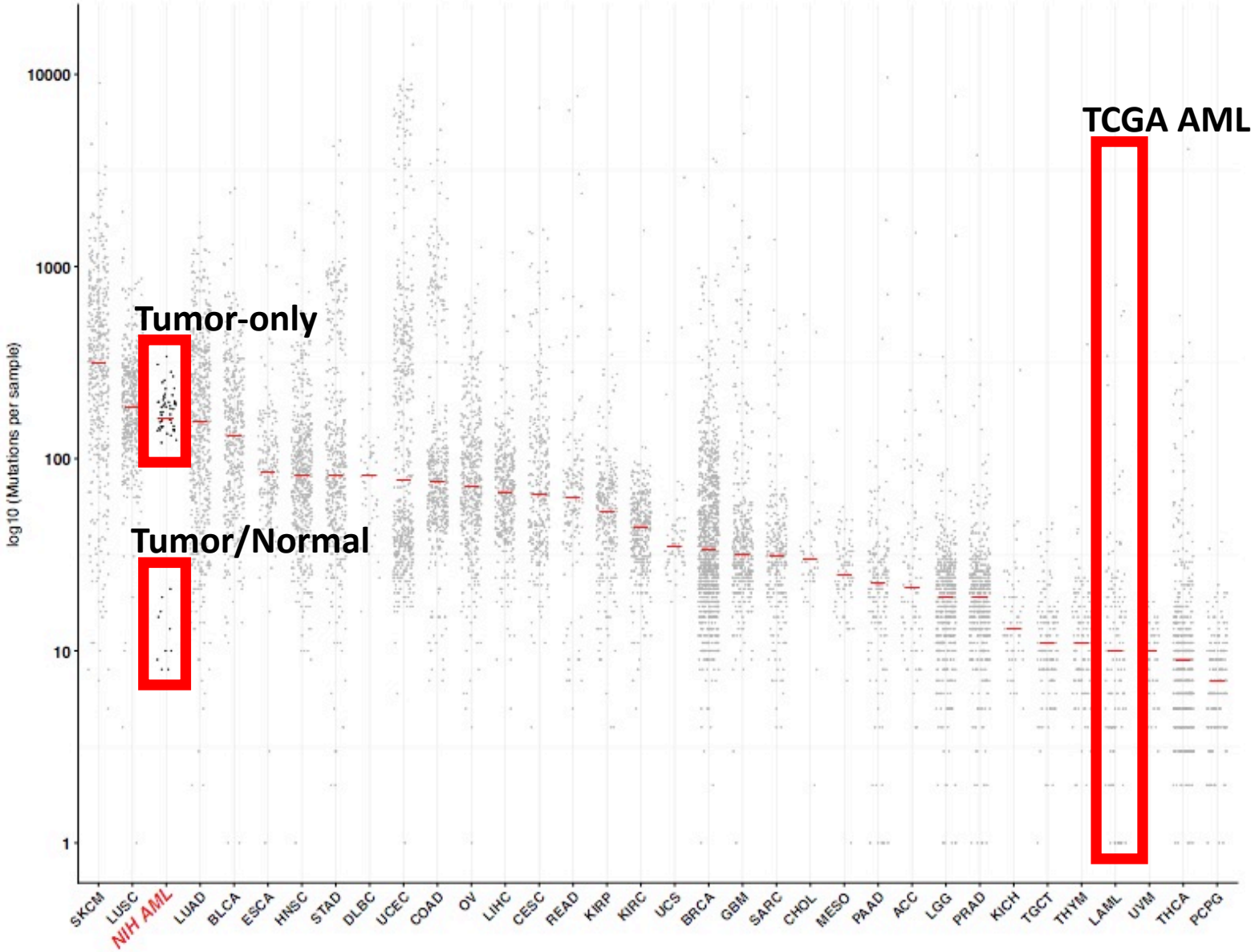
Paired Tumor/Normal vs Tumor-only Somatic Variant Calling

Even with a PON, false positive rate will be significantly higher for tumor-only relative to tumor/normal calling

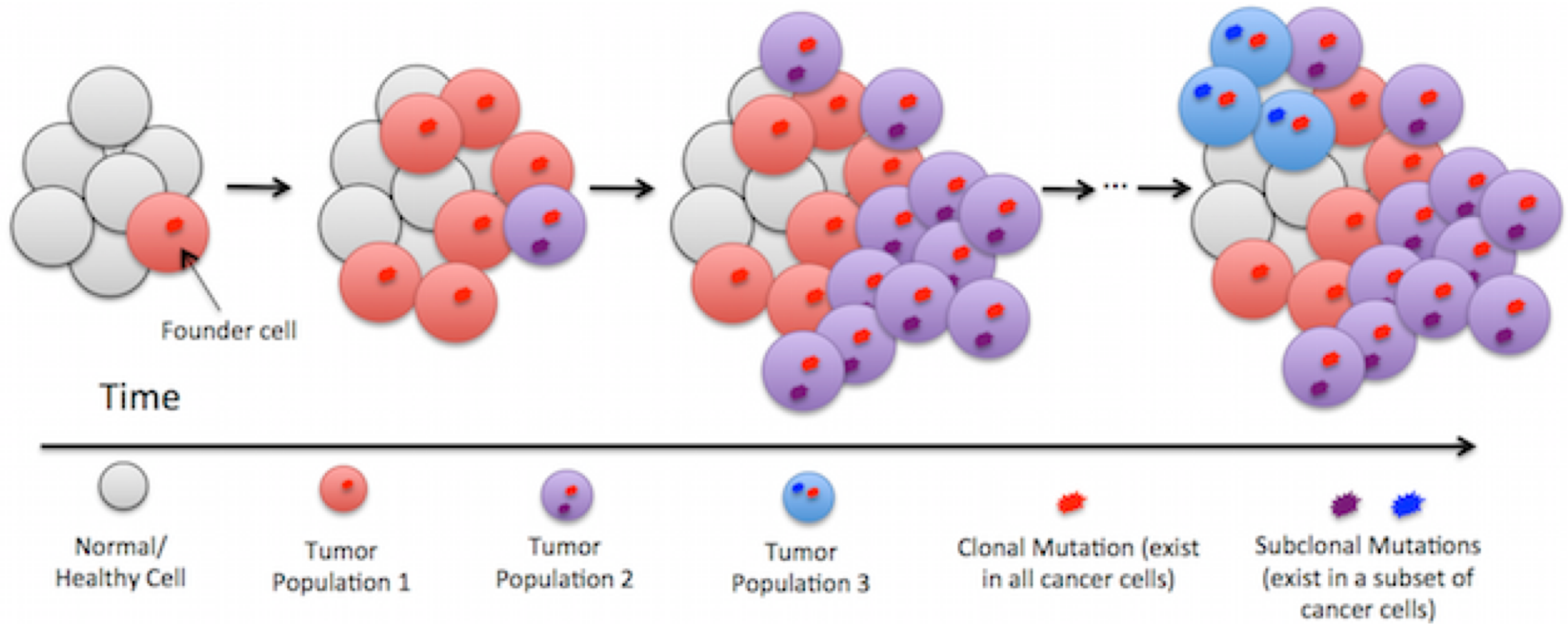


Paired Tumor/Normal vs Tumor-only Somatic Variant Calling

Even with a PON, false positive rate will be significantly higher for tumor-only relative to tumor/normal calling

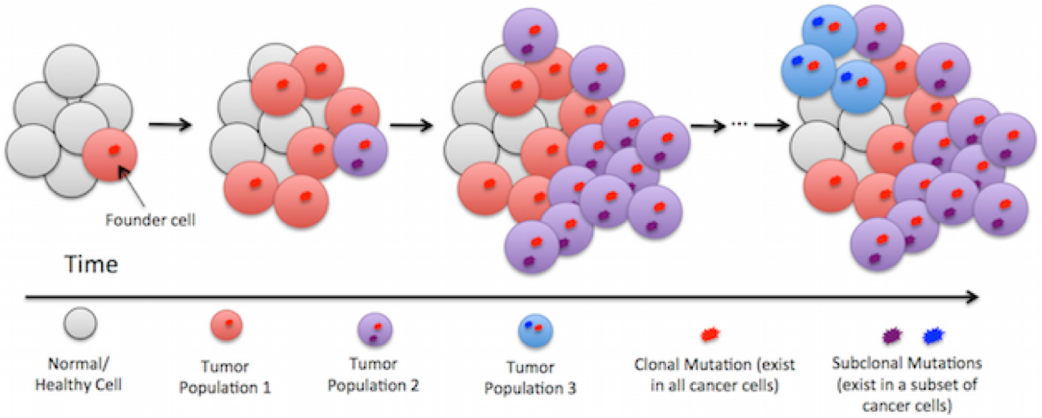


Tumor Heterogeneity/Subclonality



Tumor Heterogeneity/Subclonality

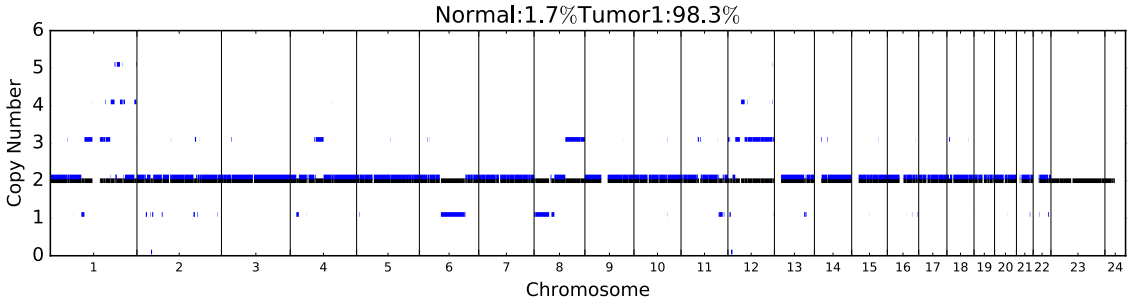
- For identification of subclones, certain conditions required:
  - High tumor purity****
  - High depth/coverage (sky's the limit!)
  - Paired tumor-normal
  - WGS essentially required for adequate sensitivity and accuracy
  - WES alone inadequate for copy number segmentation, high VAF variability, too few mutational events per subclone



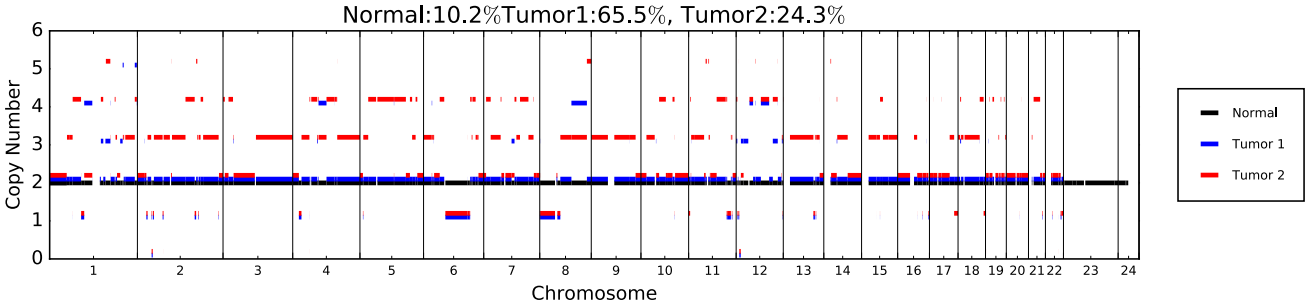
Tumor Heterogeneity/Subclonality

- For identification of subclones, certain conditions required:
  - High tumor purity****
  - High depth/coverage (sky's the limit!)
  - Paired tumor-normal
  - WGS essentially required for adequate sensitivity and accuracy
  - WES alone inadequate for copy number segmentation, high VAF variability, too few mutational events per subclone

150X WES

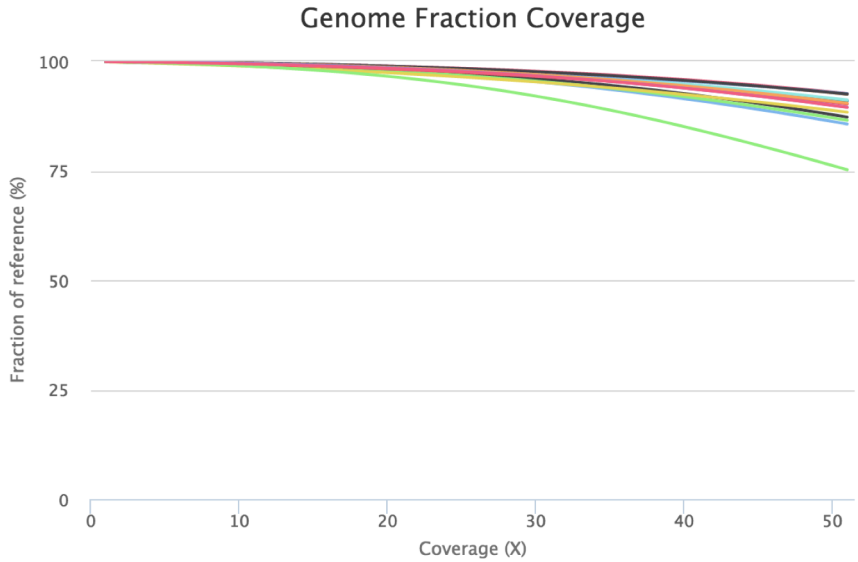


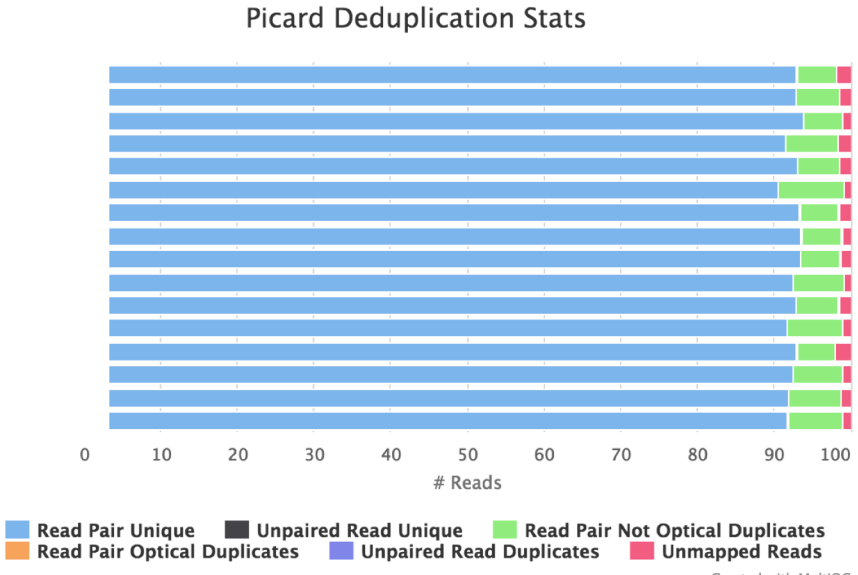
80X WGS



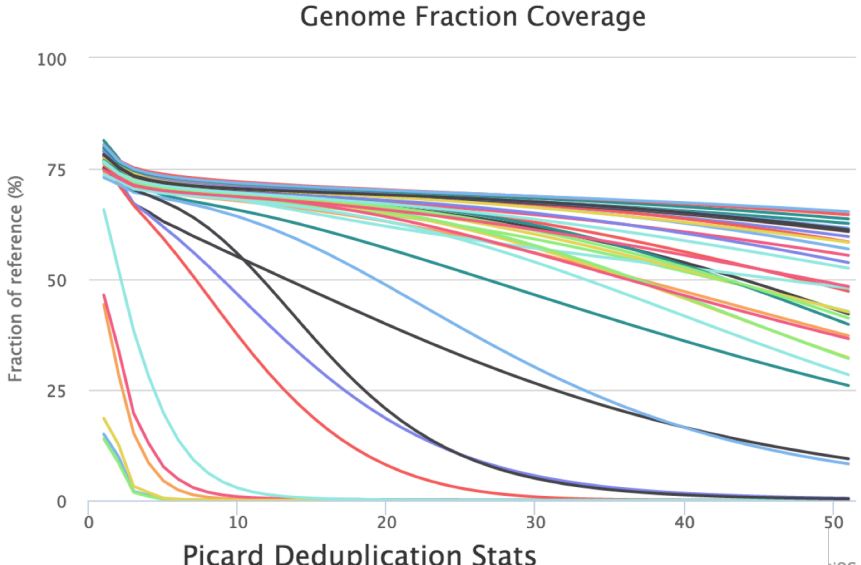
FFPE vs Fresh/Frozen Tissue – 50X target depth

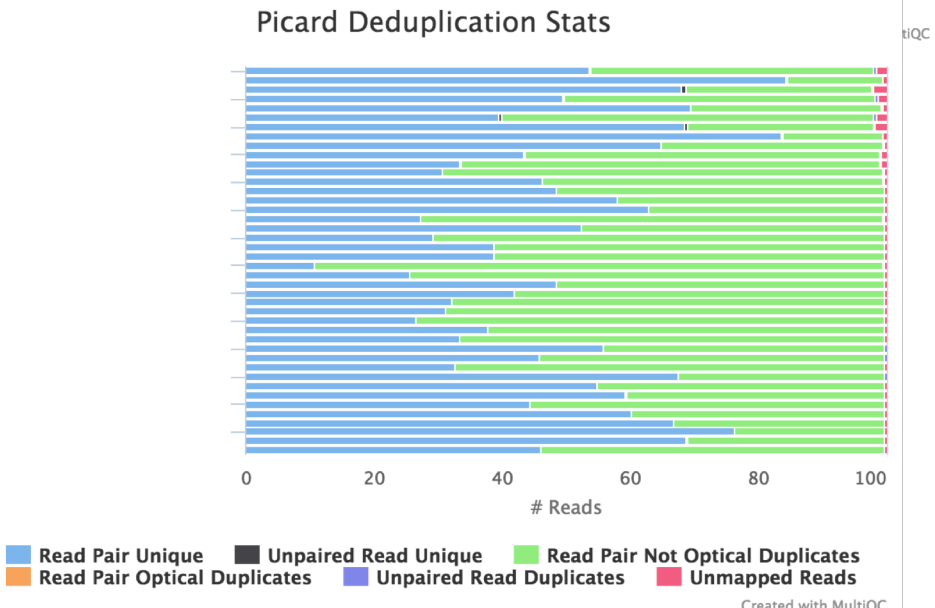
Fresh/Frozen





FFPE





Somatic Variant Calling – Best Practices

- STRONGLY favor paired tumor/normal design
- For non-human samples (e.g., mouse models) without paired somatic/germline
  - >=2 control/”germline” samples
- >=100X/50X mean depth for tumor/normal samples
- Significantly higher target depth for FFPE samples
- Tumor purity >50% (ideally, >60%) for variant calling
- MUST visually verify any somatic variant of “significance”
- Subclone analysis requires WGS, high purity, and high depth

Germline Variant Calling – Considerations and Best Practices

GWAS/Burden Testing Design

"With an odds ratio (OR) = 1.4, the sample sizes required to achieve 80% power are 6,400, 54,000, and 540,000 for a MAF = 0.1, 0.01, and 0.001, respectively, if one assumes 5% disease prevalence and a significance level of 5×10^{-8}. Because the number of rare variants is much larger than the number of common variants, more stringent significance levels might be required, further reducing power."

Lee et al., 2014, American Journal of Human Genetics

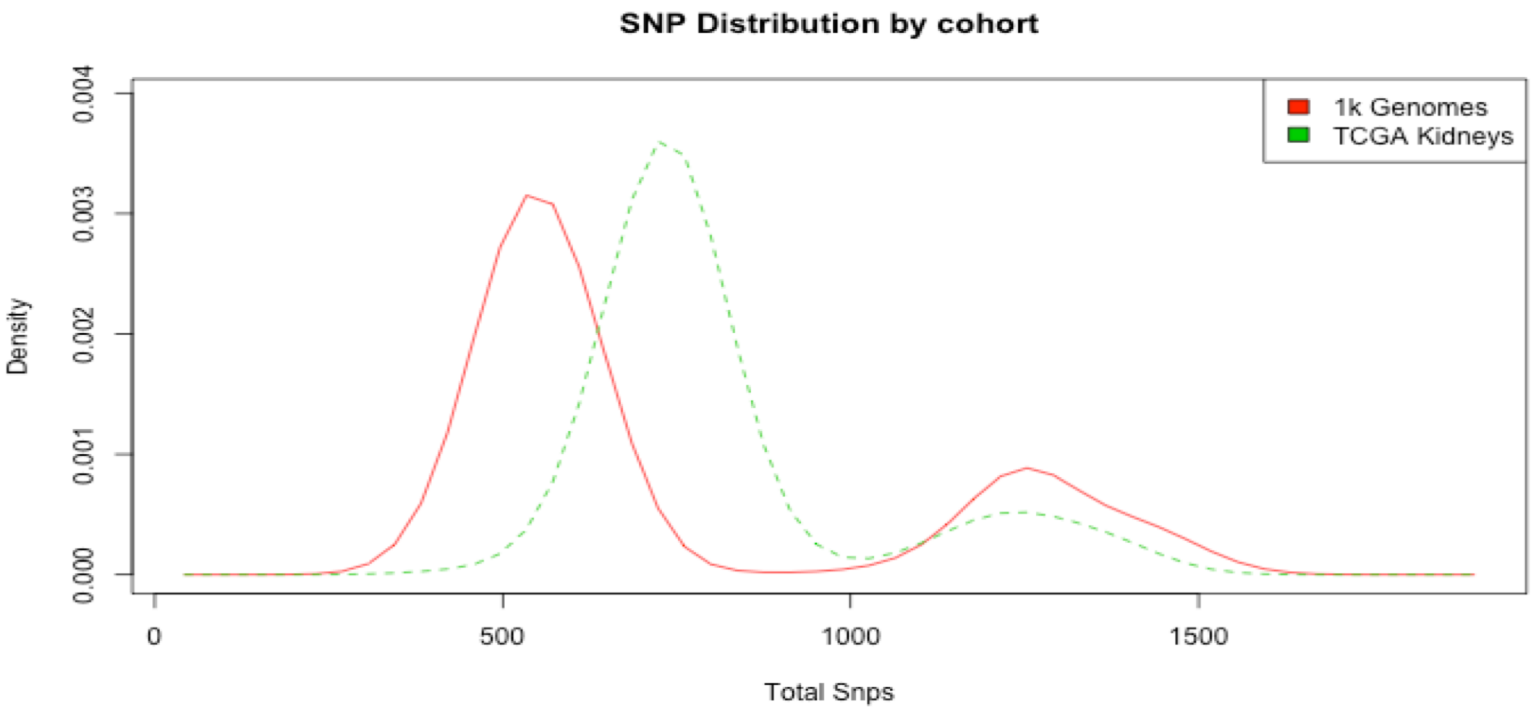
REVIEW

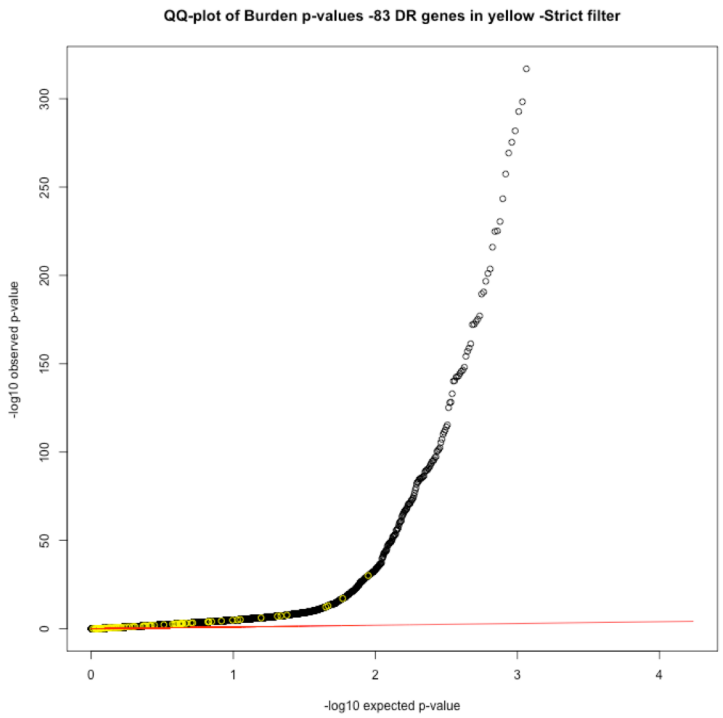
Rare-Variant Association Analysis:
Study Designs and Statistical Tests

Seunggeung Lee,[1] Gonçalo R. Abecasis,[1] Michael Boehnke,[1] and Xihong Lin[2,*]

GWAS/Burden Testing Design

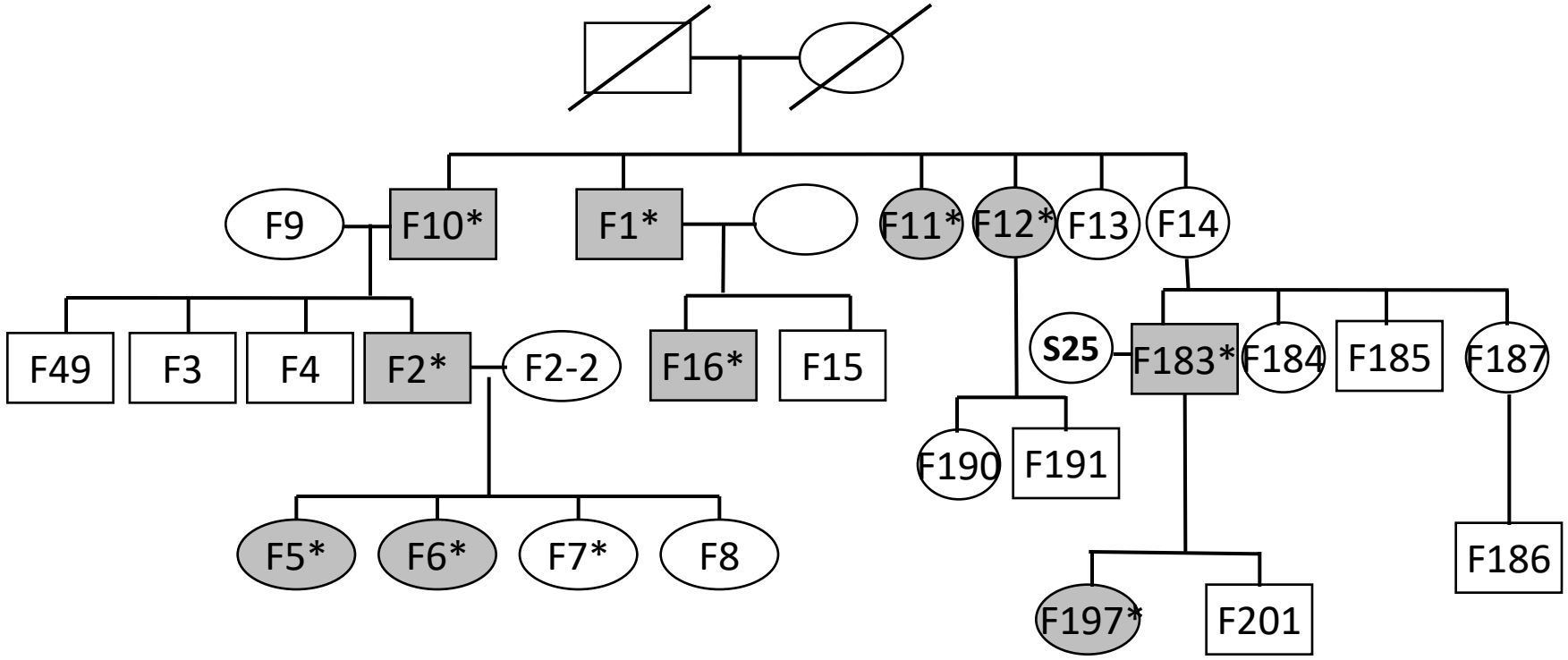
- Biggest challenge is having data that is homogenous
- Cases and controls MUST have genotype data that is generated identically





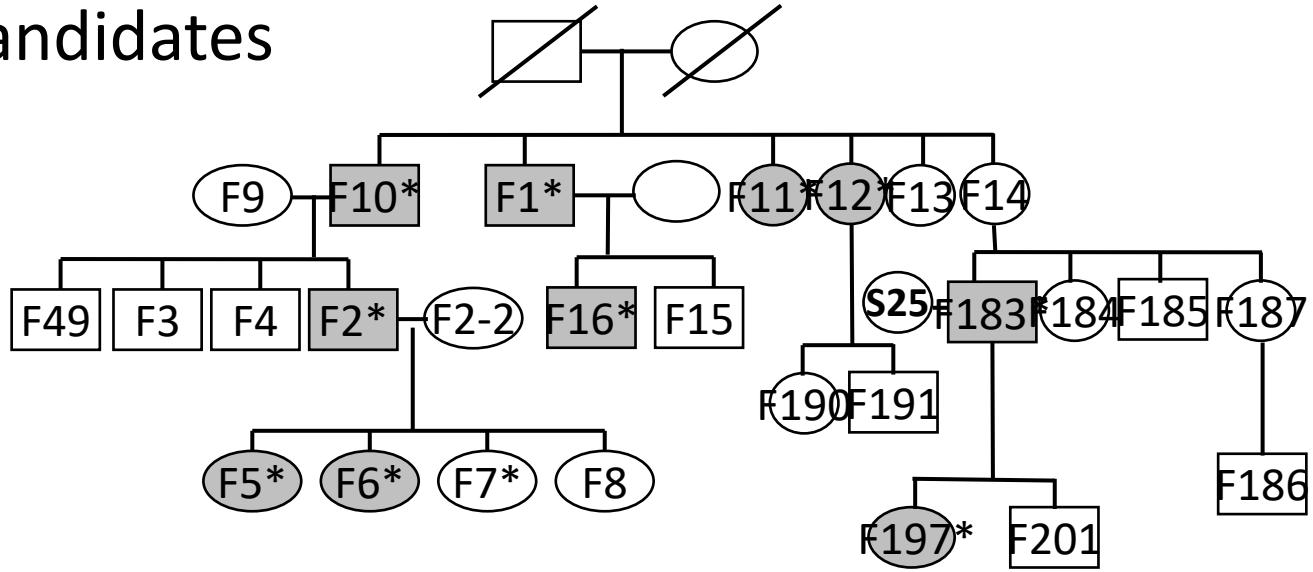
Familial Sequencing Design

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



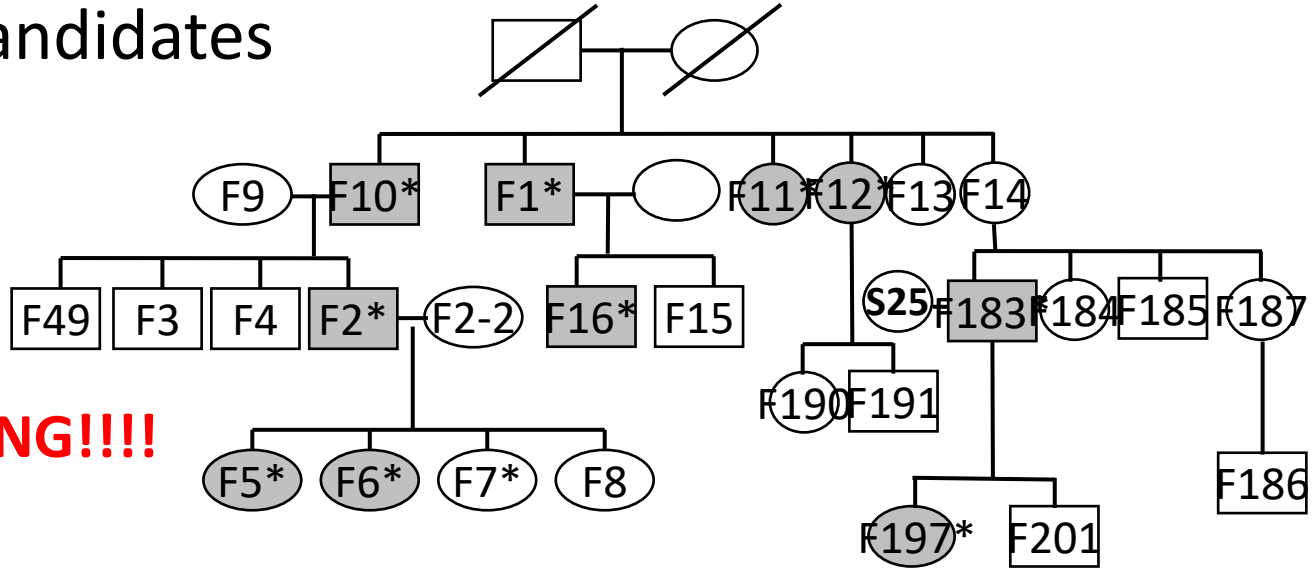
Familial Sequencing Design

- 3 cases, no controls
  - 3,176 candidates
- 3 cases, 1 spousal control (ethnicity 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



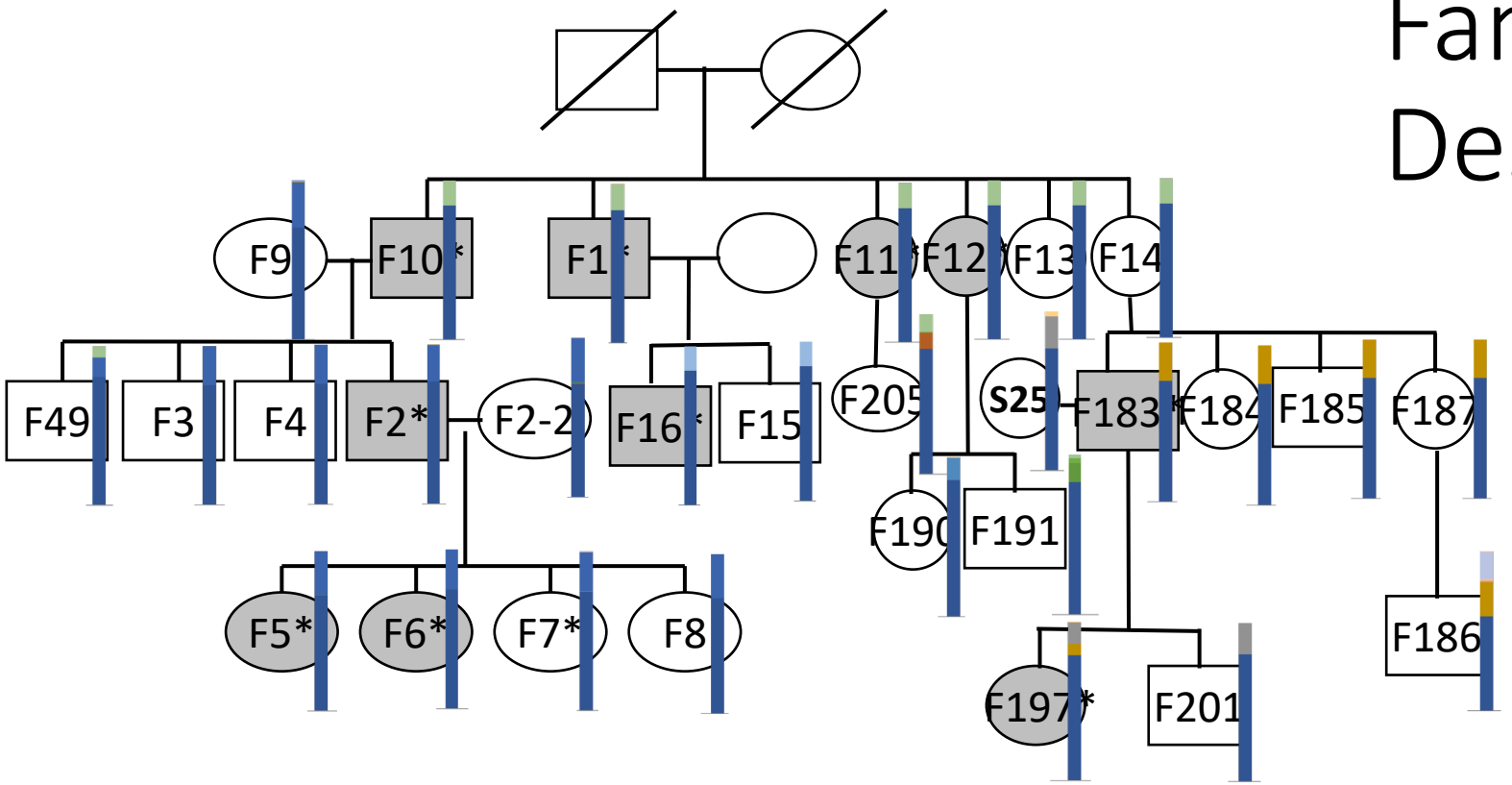
Familial Sequencing Design

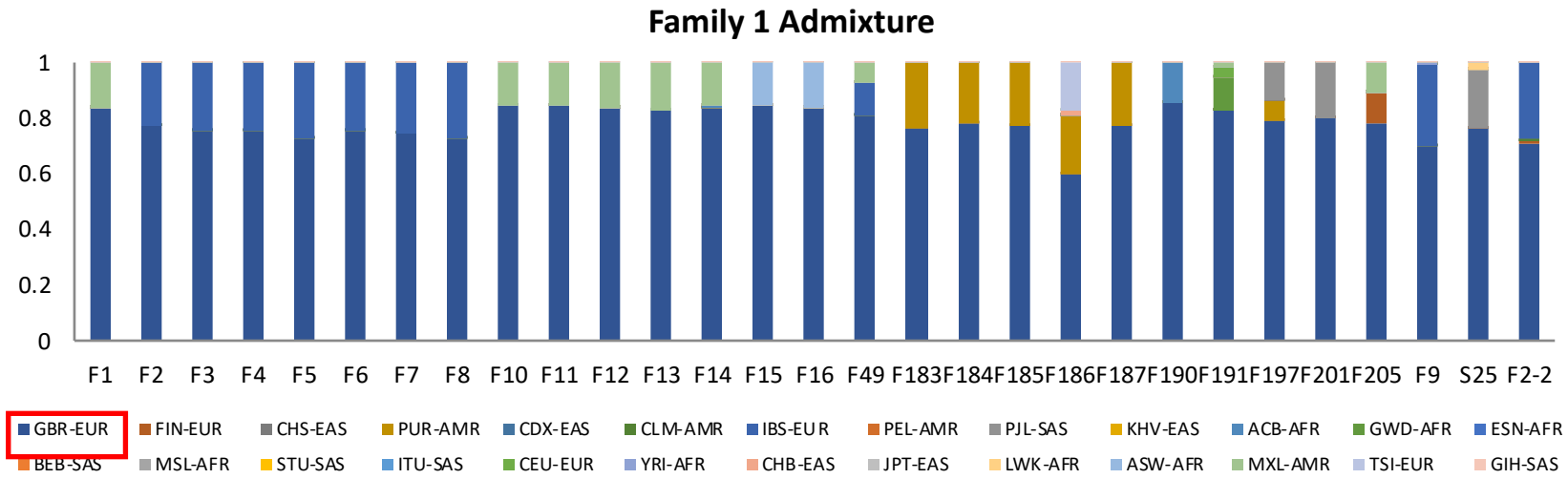
- 3 cases, no controls
  - 3,176 candidates
- 3 cases, 1 spousal control (ethnicity matched) - 1542 candidates
  - +1 spouse controls - 1121 candidates
  - +1 case - 525 candidates
- 3 cases, 1 related control - 854 candidates
  - +1 related control - 307 candidates
  - +1 case – 284 candidates



ALWAYS PERFORM ETHNICITY-AWARE FILTERING!!!!

Familial Sequencing Design

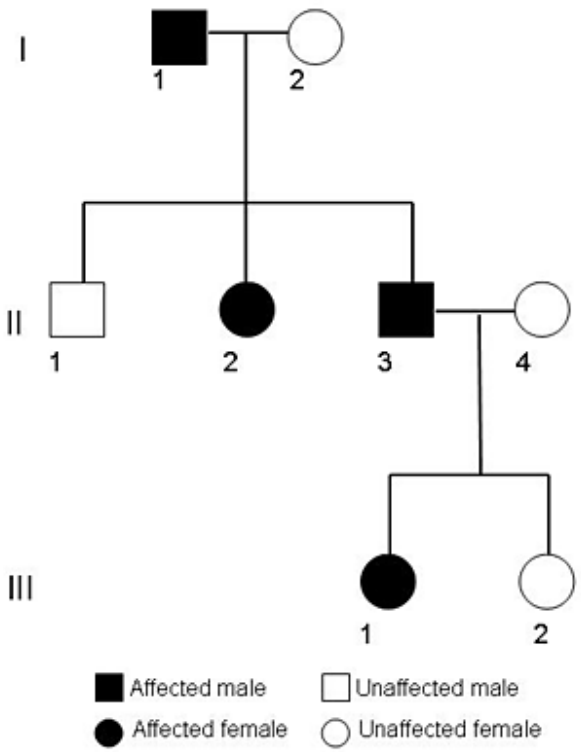




- 3 cases, no controls
  - 3,176 candidates with global allele frequency threshold of ≤0.01
  - 2,923 candidates with EUR-only!

Germline Variant Calling – Best Practices

- Whole genome strongly preferred
  - >=30X mean target depth
  - Far superior to exome for structural variants, copy number analysis, and SNP/INDEL detection
- 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

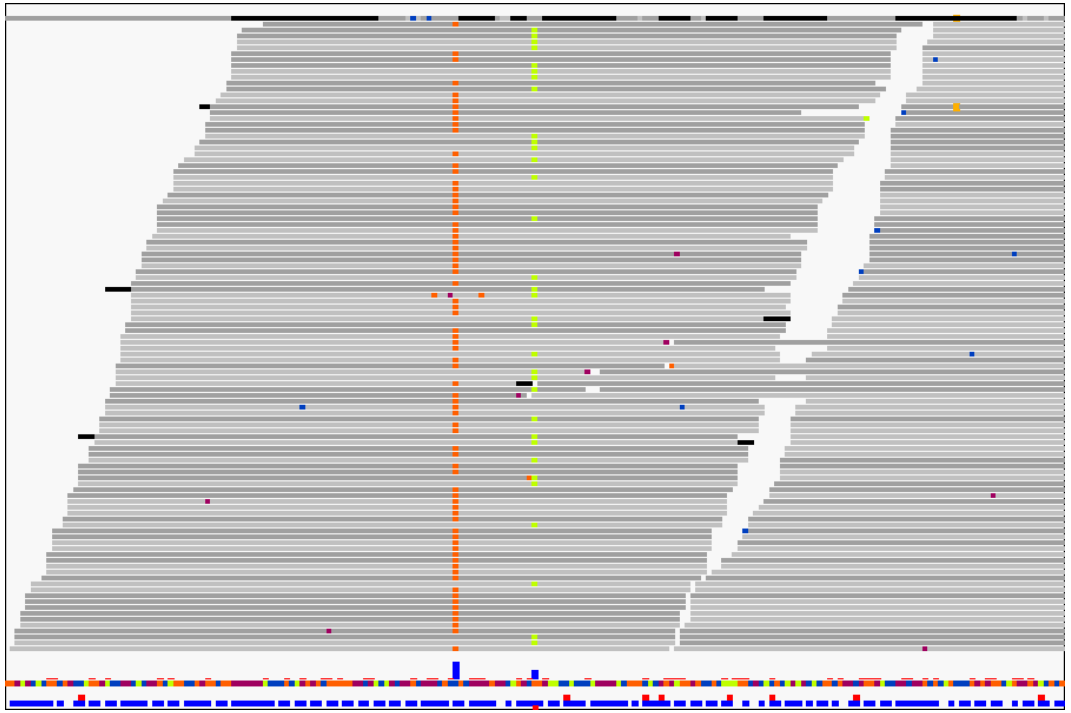


Other Considerations and Best Practices for Variant Analysis

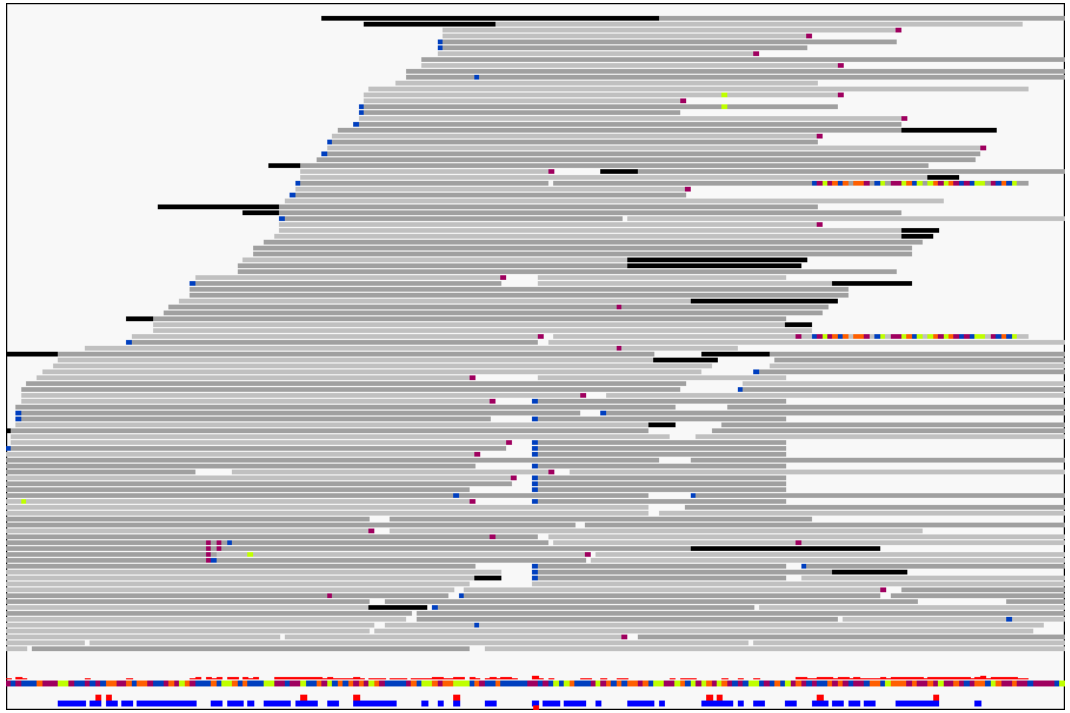
Always Visualize Significant Variants

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

True Positive

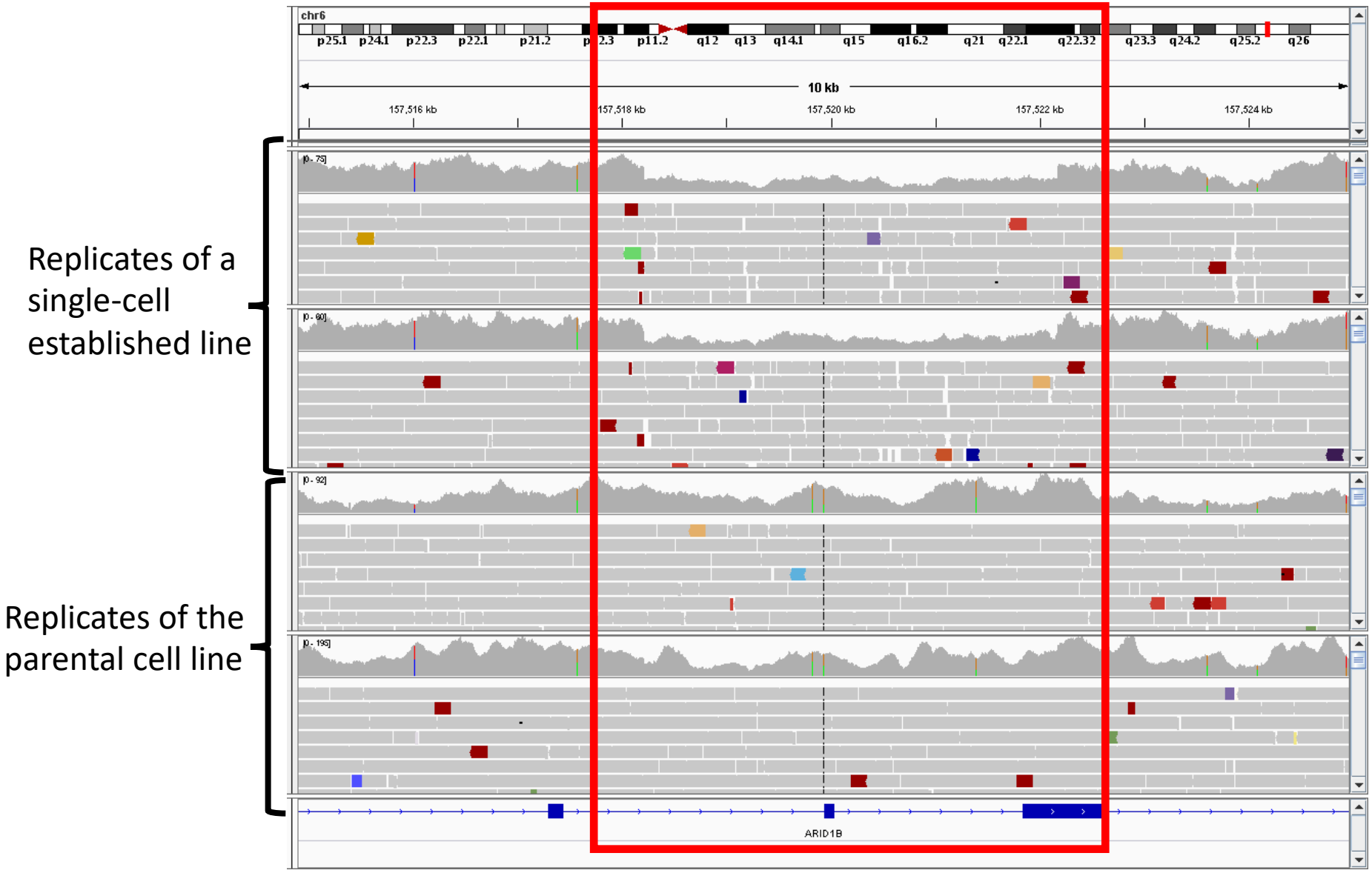


False Positive

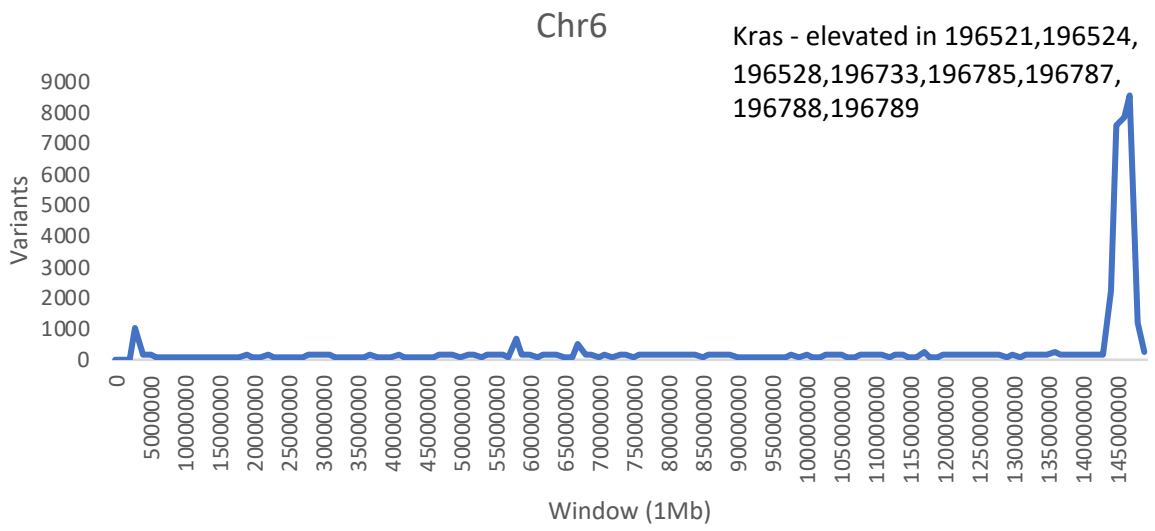


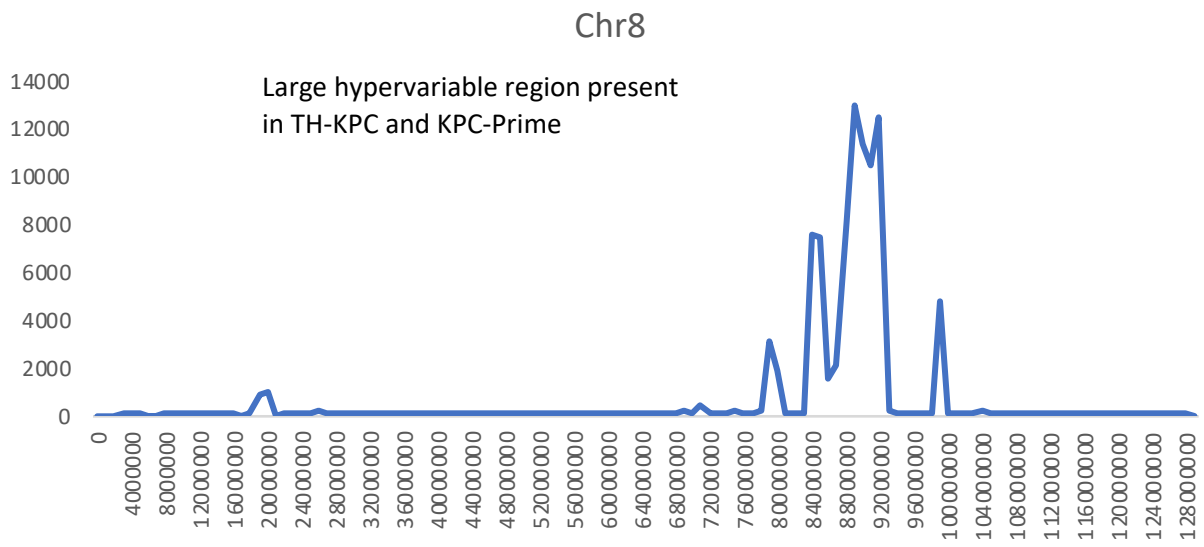
Variant Analysis in Cell Lines

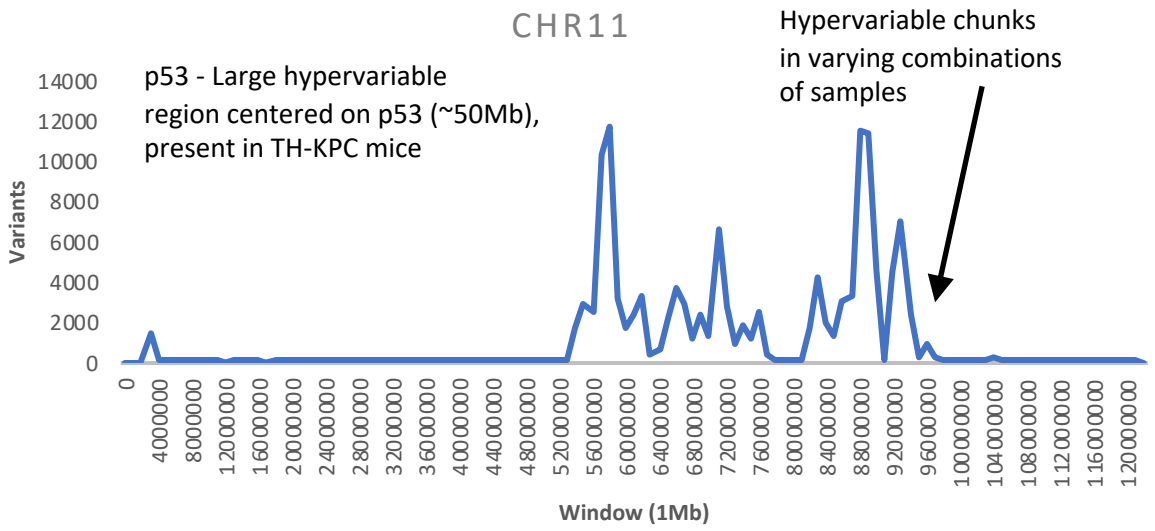
- Can never assume your cell lines are homogenous!
- This cell line had a subclone with an ARID1B loss
- Another subclone had lost the Y chromosome



Variant Analysis in Animal Models





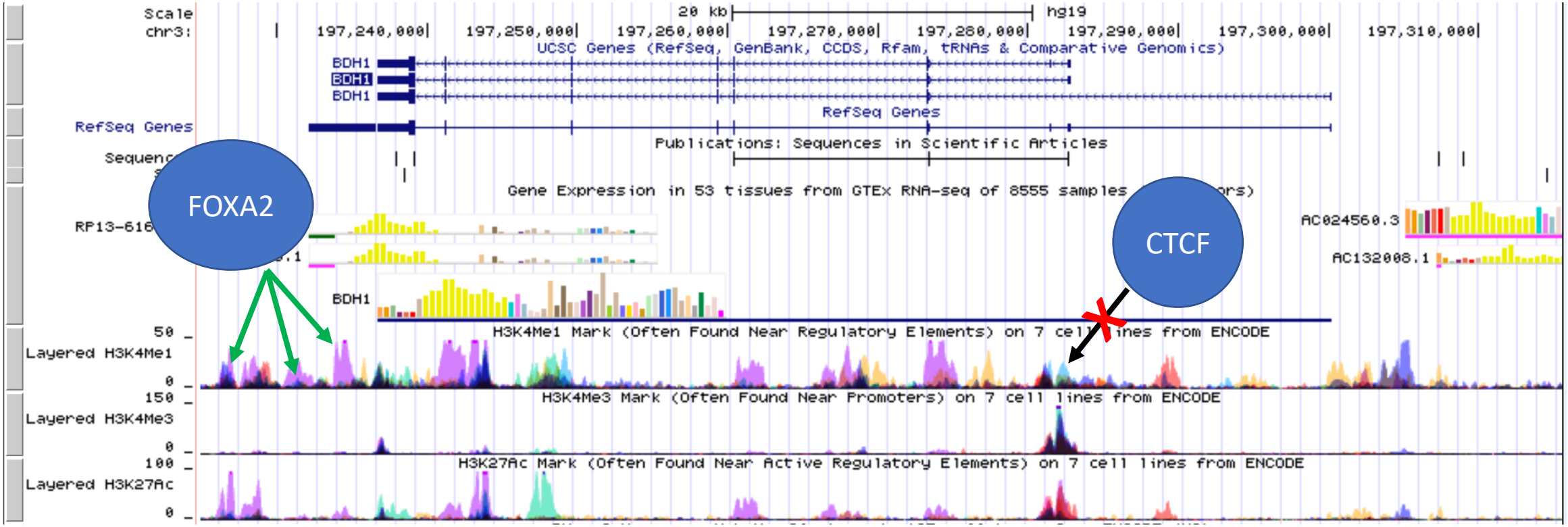


- Mice have retained significant levels of heterozygosity
- Vary considerably in regional gene expression

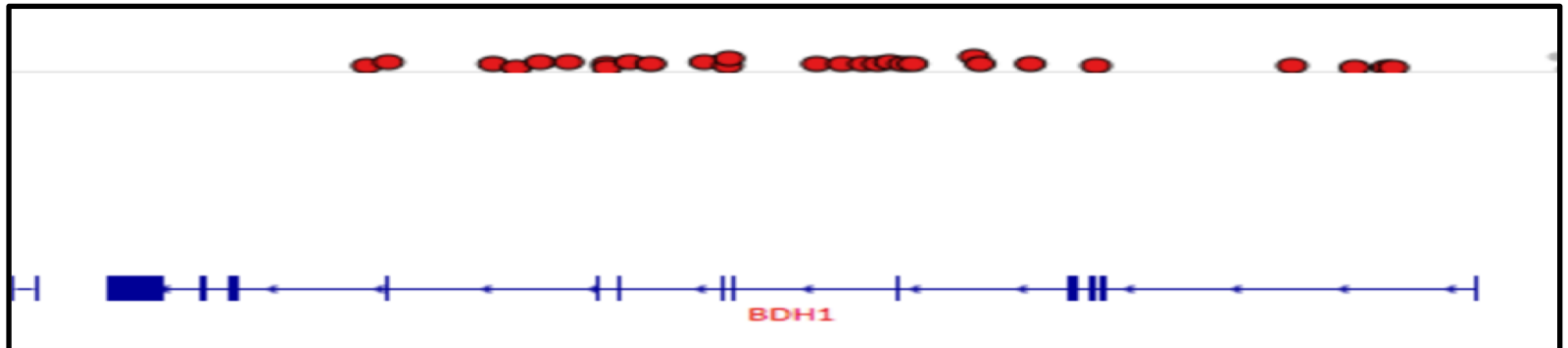
Multiomic Integration

- Gene and protein expression data can be used to prioritize variants and genes
  - Especially critical in underpowered, small cohort analyses
- Germline or somatic variants that do not have expression consequences at the gene and/or pathway level can be reduced in priority
- Powerful method for prioritizing non-coding mutations in WGS

FOXA2/FOXA3 Transcription Factor Network



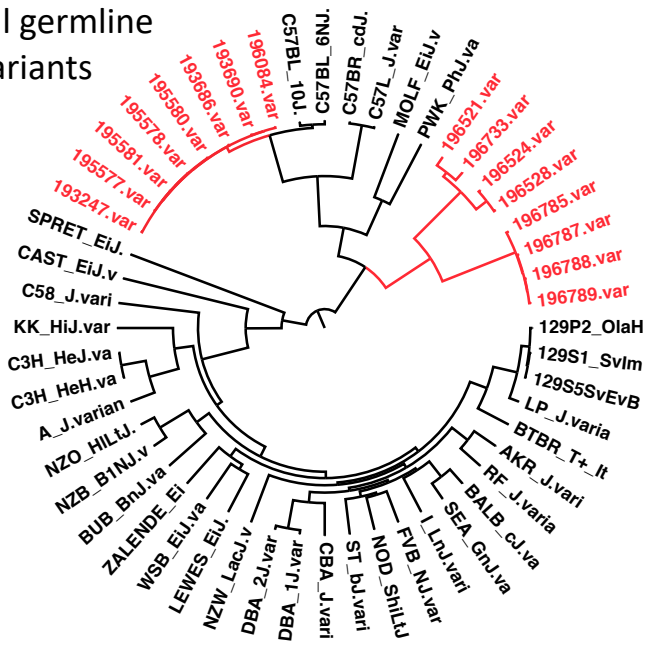
Gtext Thyroid eQTLs



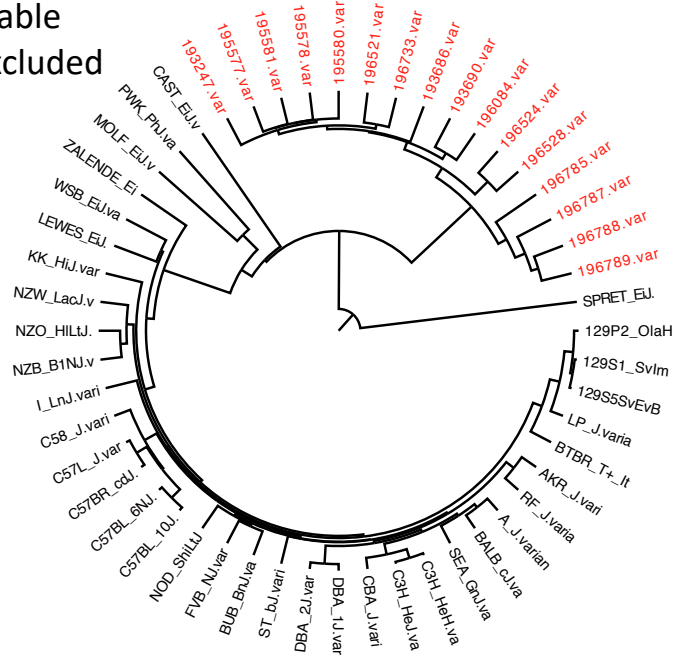
Other Considerations and Best Practices for Variant Analysis

- Visualization and validation is a necessity
- STRONGLY recommend WGS for any cell line or model organism to be used in quantitative analysis
  - Considerable variation exists among cell line replicates in gene and protein expression, and a major contributor to this is genetic heterogeneity
  - Model organisms can retain considerable levels of heterozygosity, even after long-term maintenance in colony
  - Back-crossing is a necessity
  - Drift causes cell lines and model organisms to randomly accumulate genetic differences from progenitors

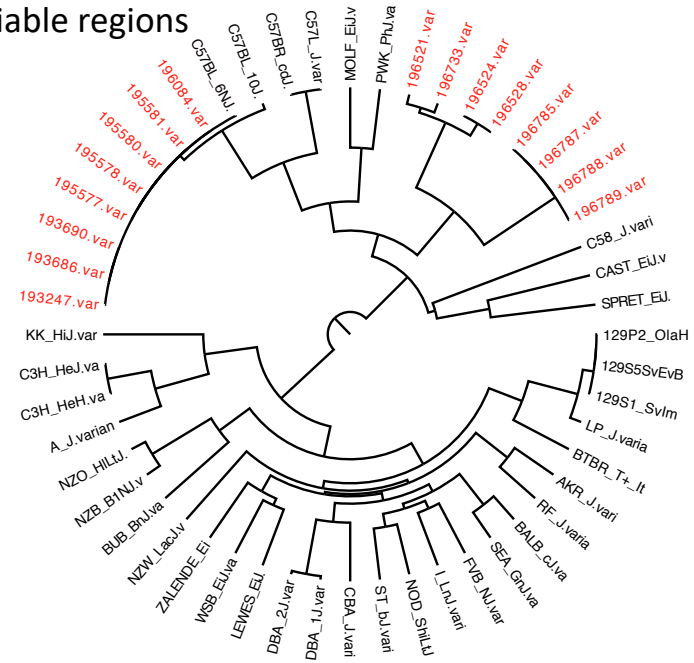
All germline variants



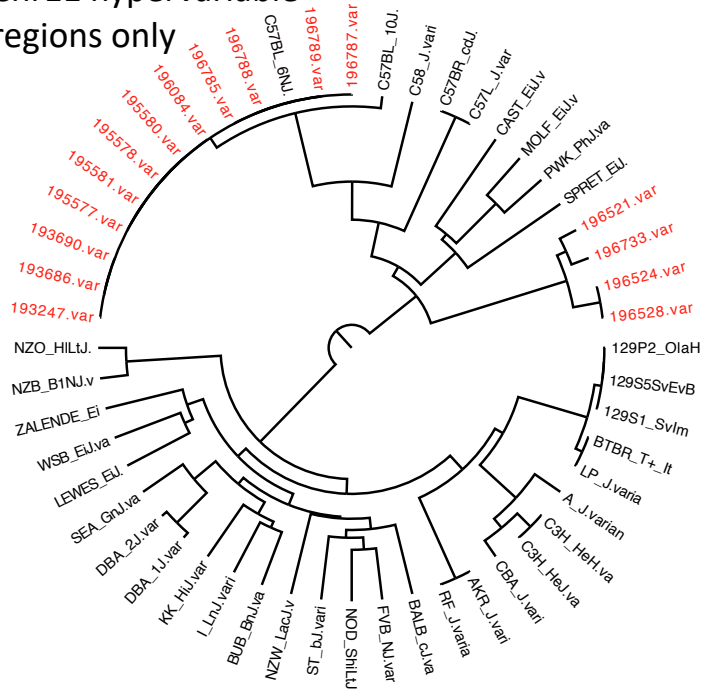
Hypervariable regions excluded



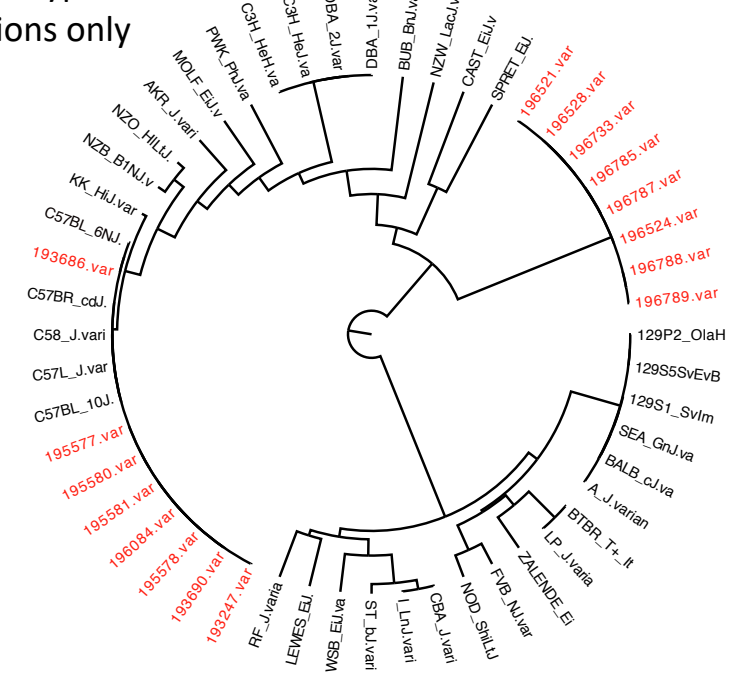
Hypervariable regions only



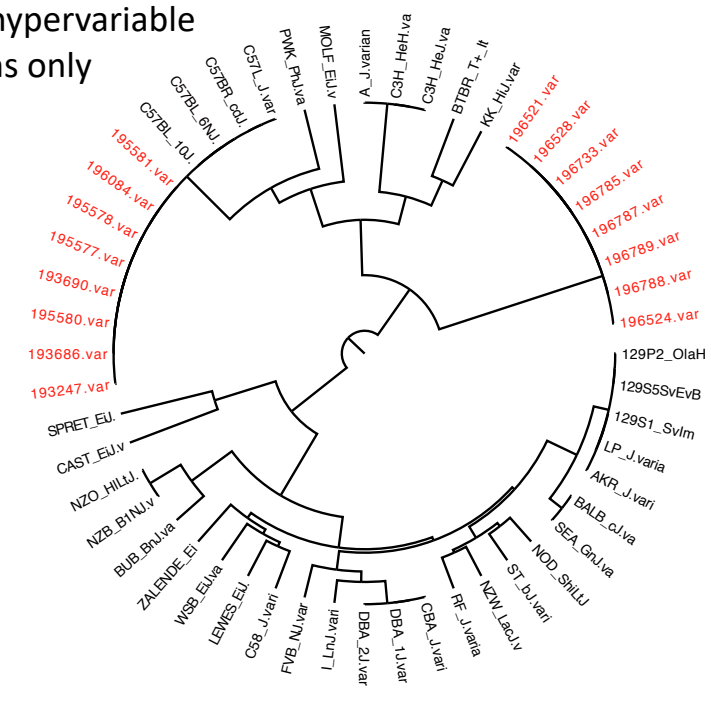
Chr11 hypervariable regions only



Chr6 hypervariable regions only



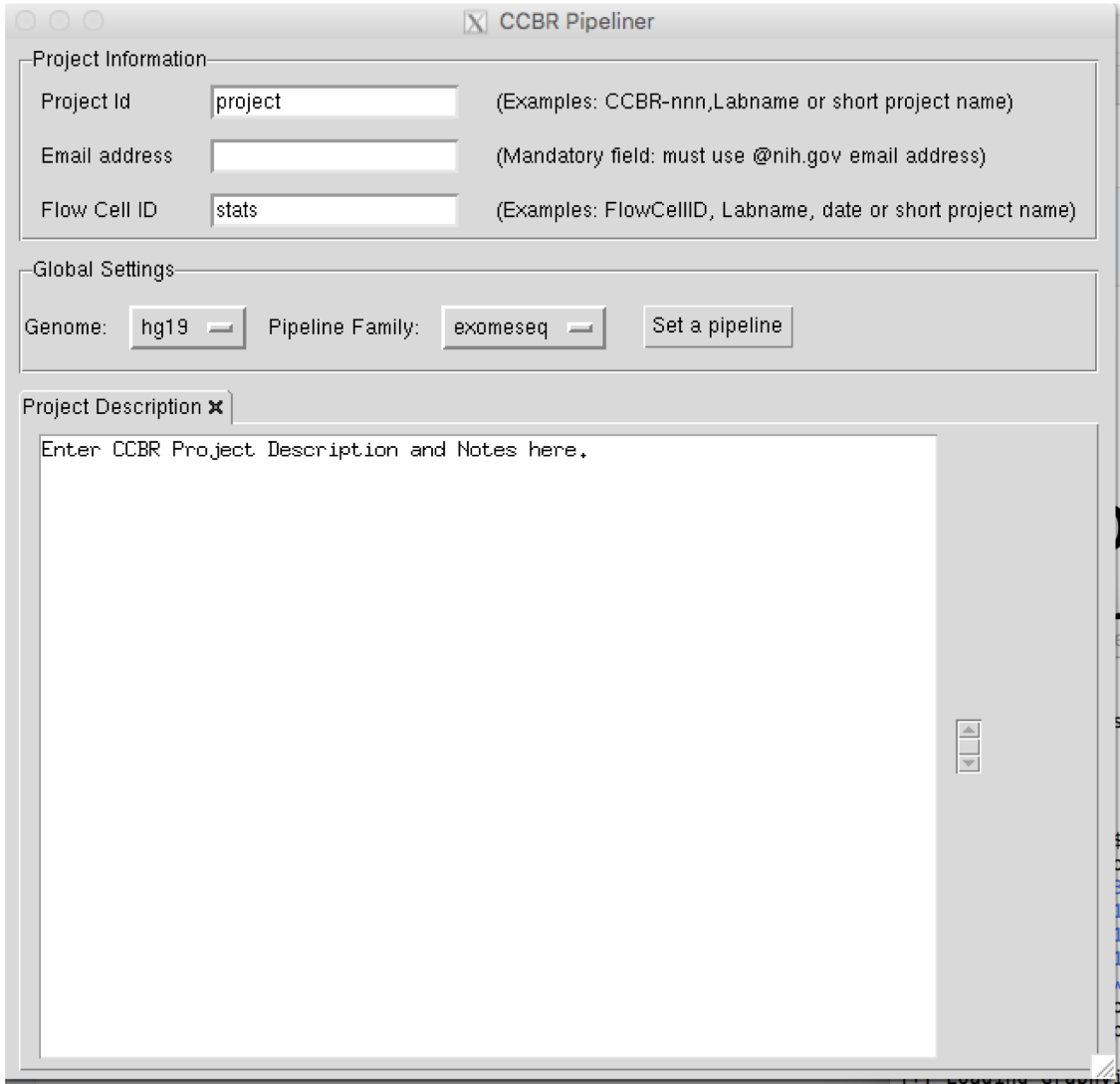
Chr8 hypervariable regions only



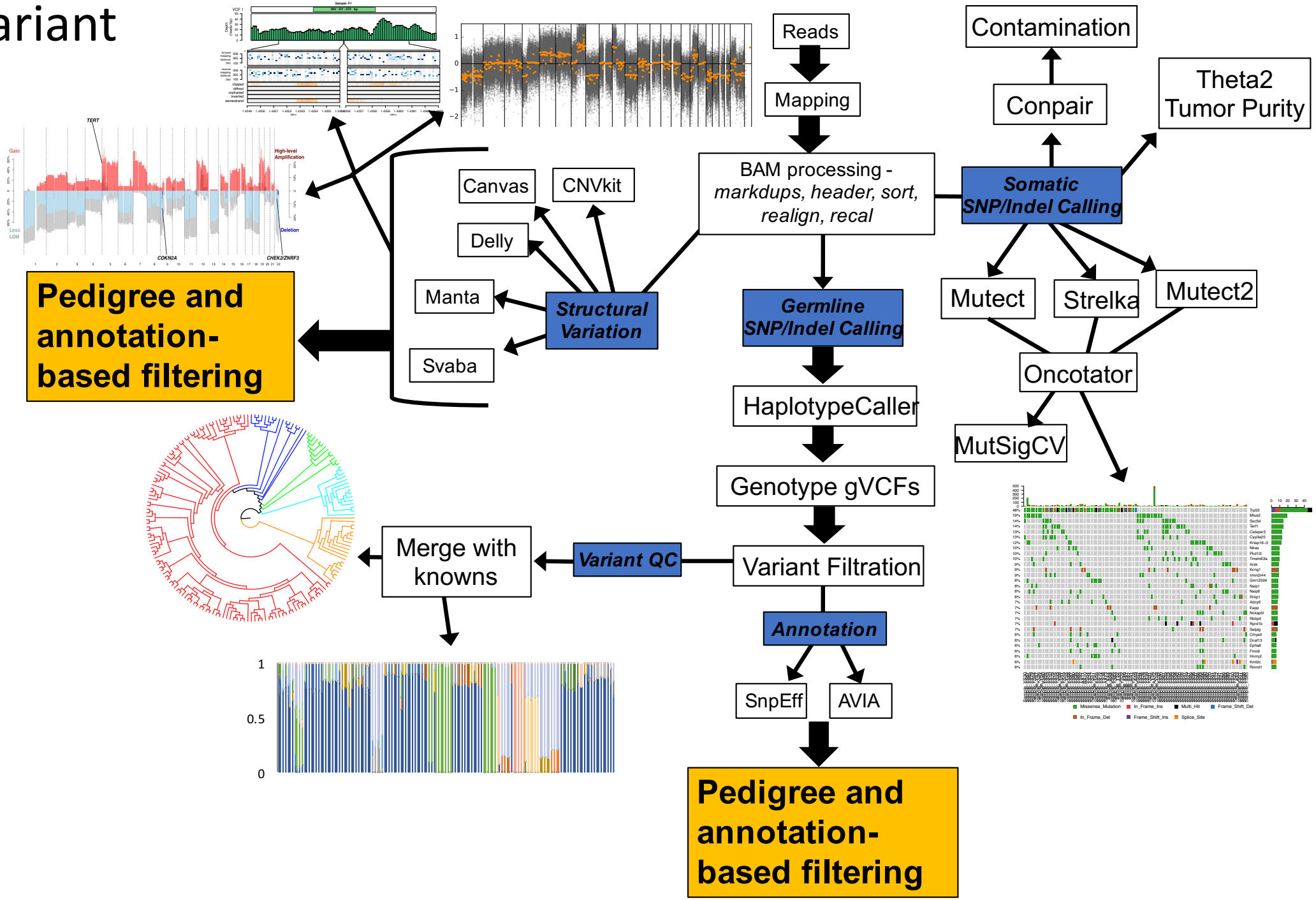
CCBR Pipeliner

Variant Calling at CCBR

- Multiple Variant Calling CCBR Pipelines
  - Whole genome (germline and somatic)
  - Whole exome/targeted sequencing (germline and somatic)
  - Variants from RNAseq
- All variant calling pipelines available through CCBR_Pipeliner app
  - https://github.com/CCBR/Pipeliner
  - Just need Biowulf account and xquartz installed on our local machine
  - ***module load ccbrpipeliner (enter)***
  - ***ccbrpipe.sh (enter)***

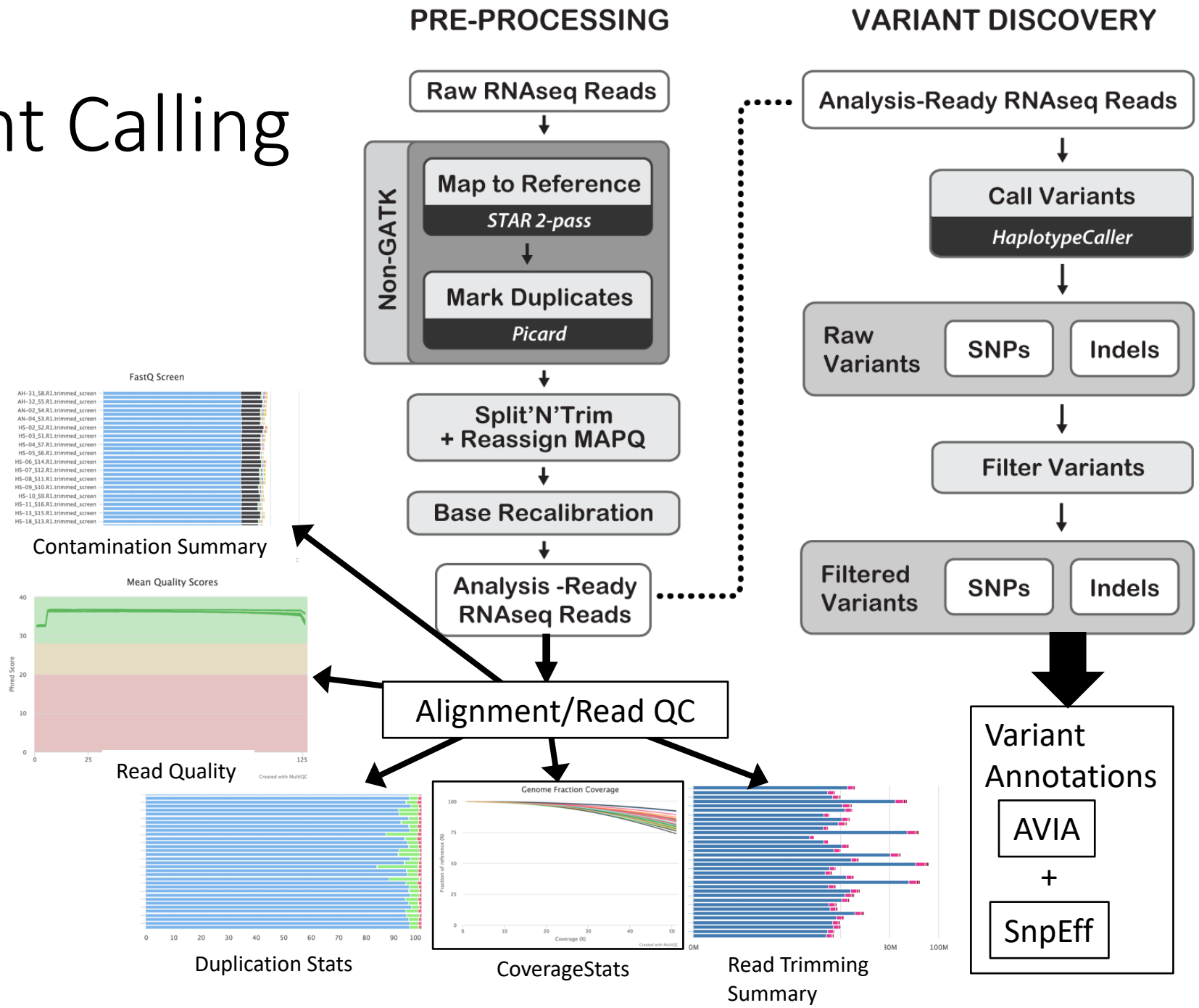


WES/WGS Variant Workflow(s)



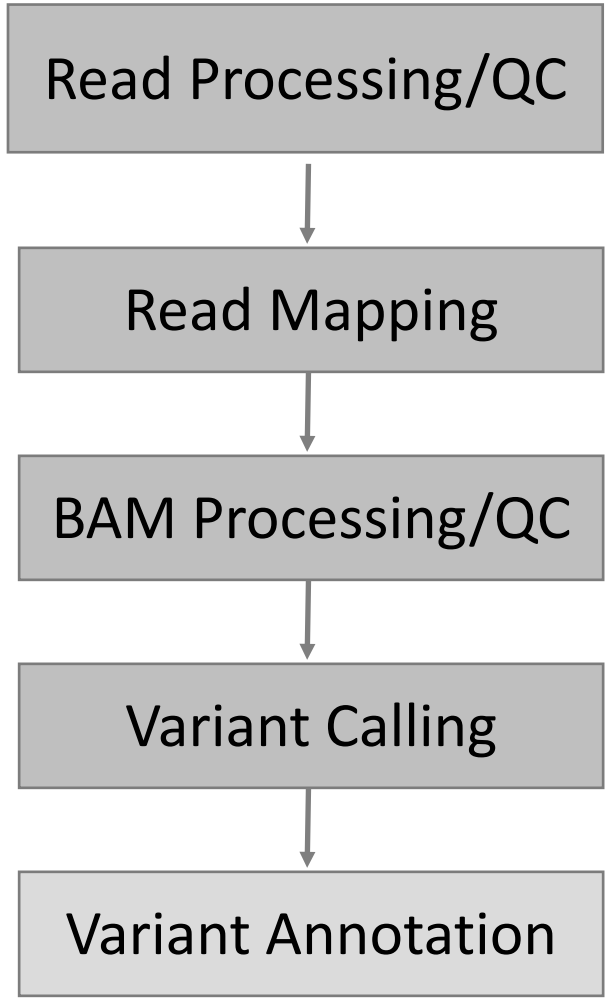
RNAseq Variant Calling

http://gatkforums.broadinstitute.org/gatk/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail

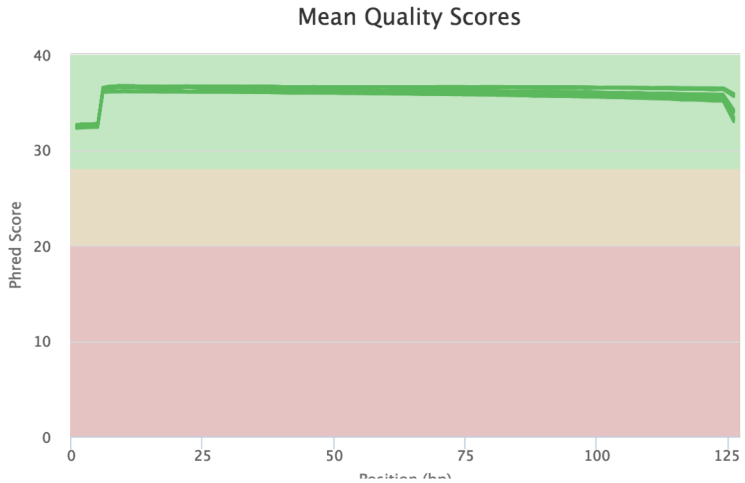


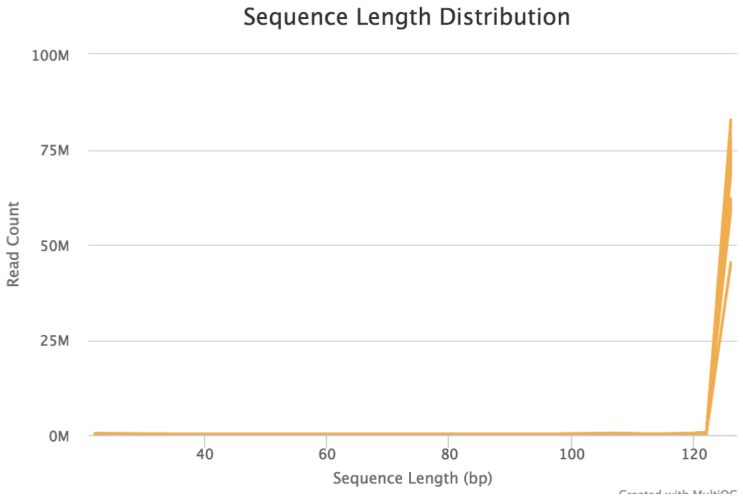
Pipeline Details…

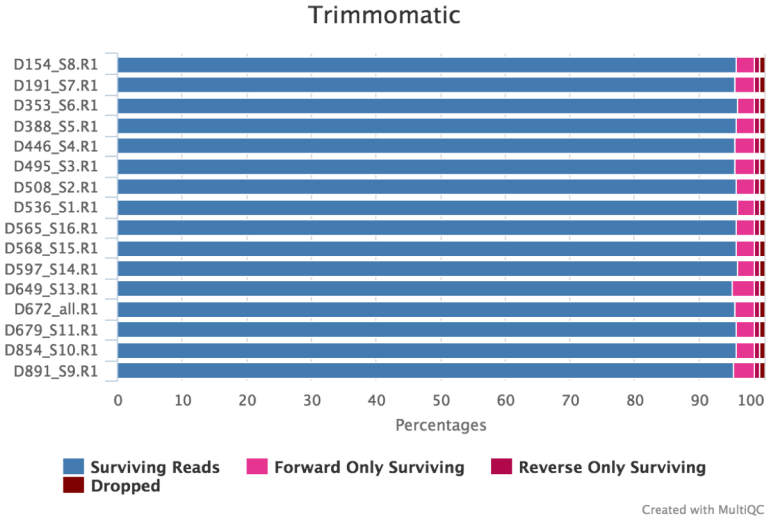
- All variant calling follows the same basic approach

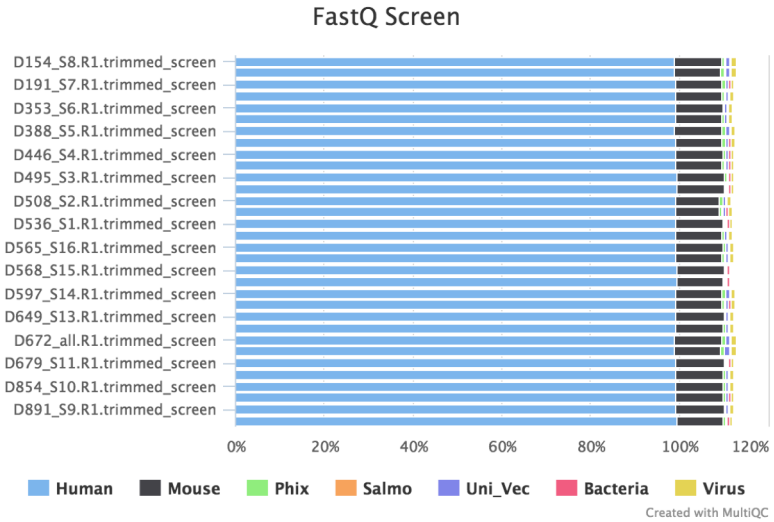


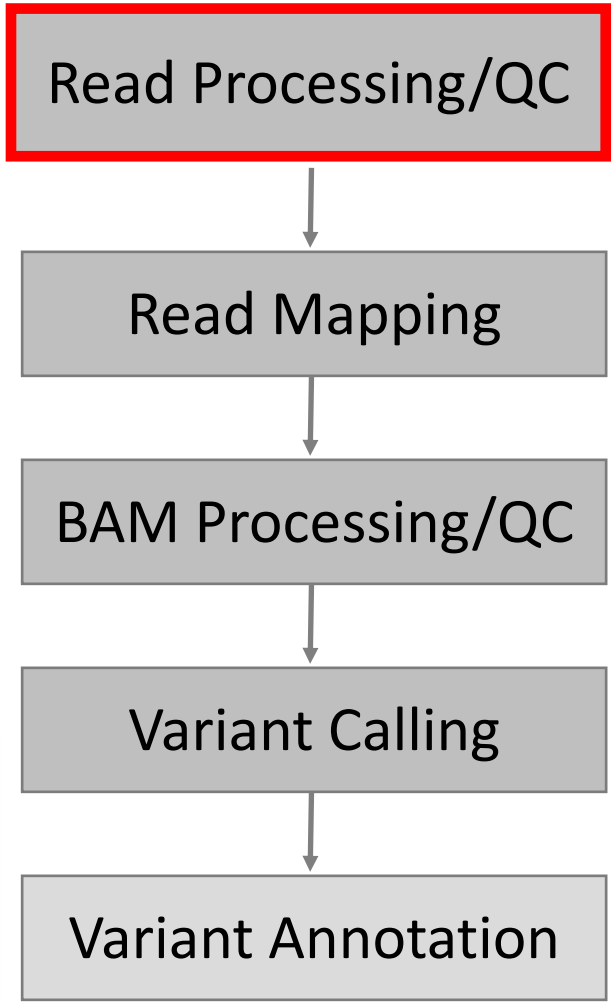
Pipeline Details…



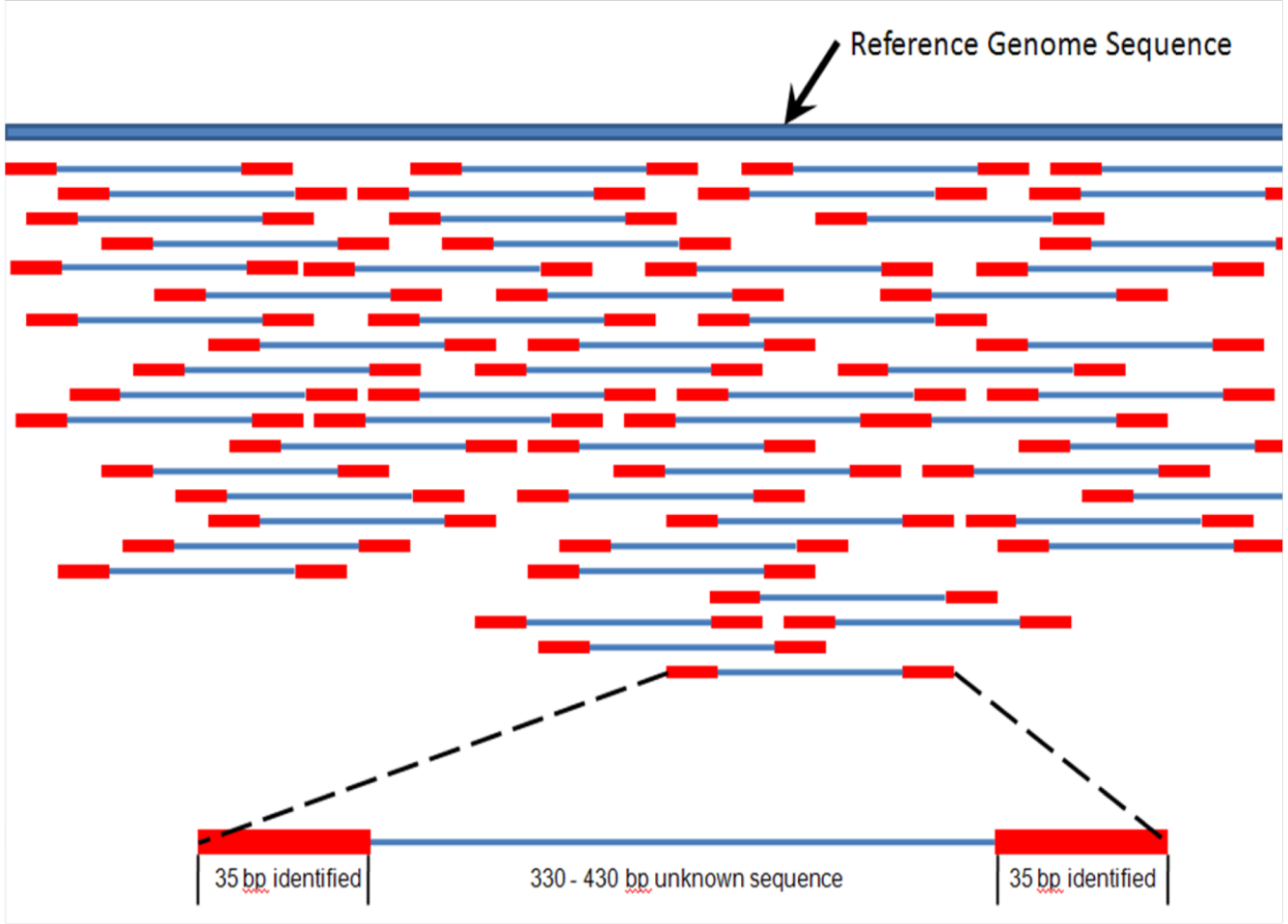


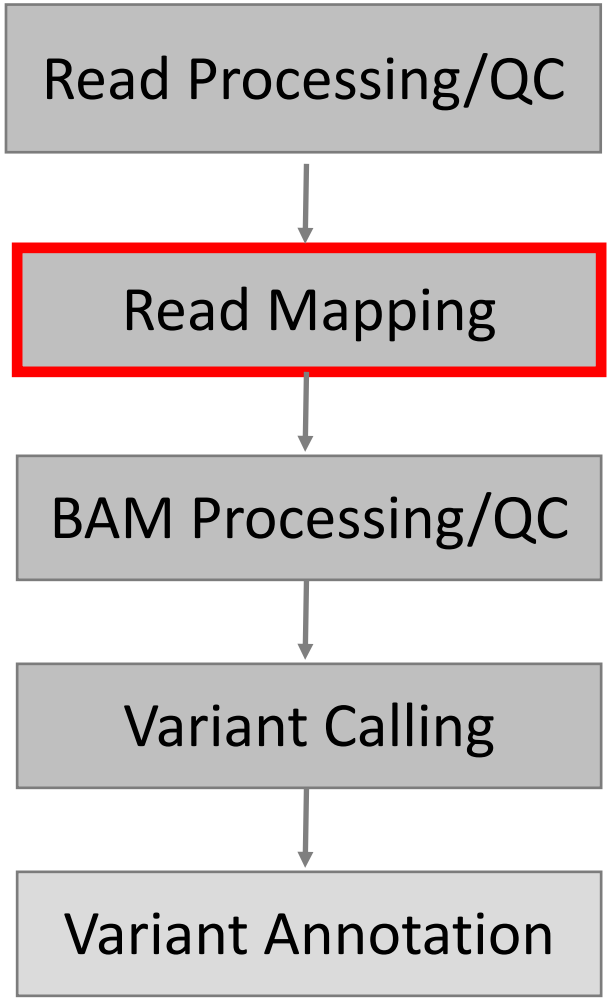






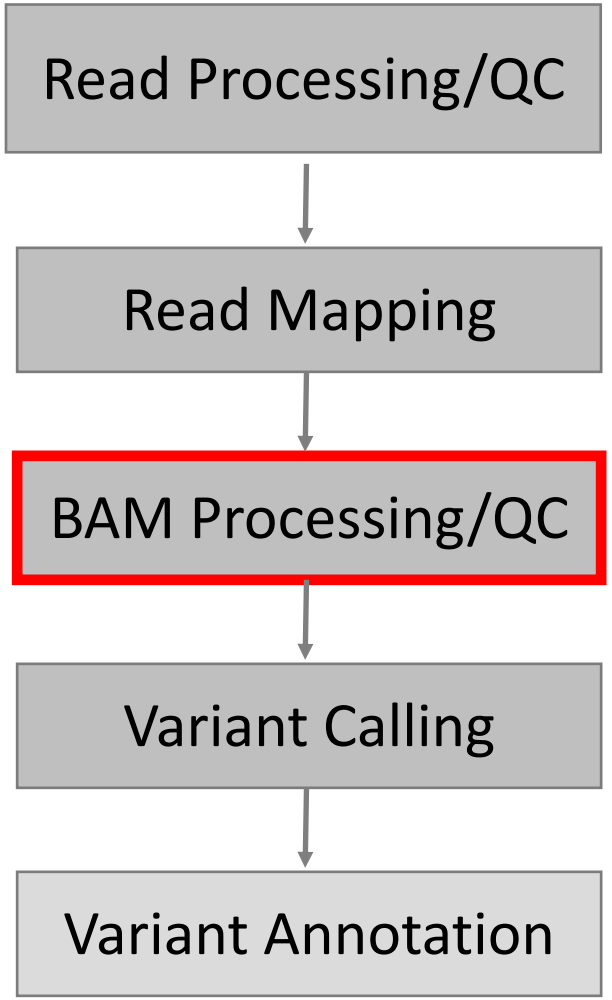
Pipeline Details…





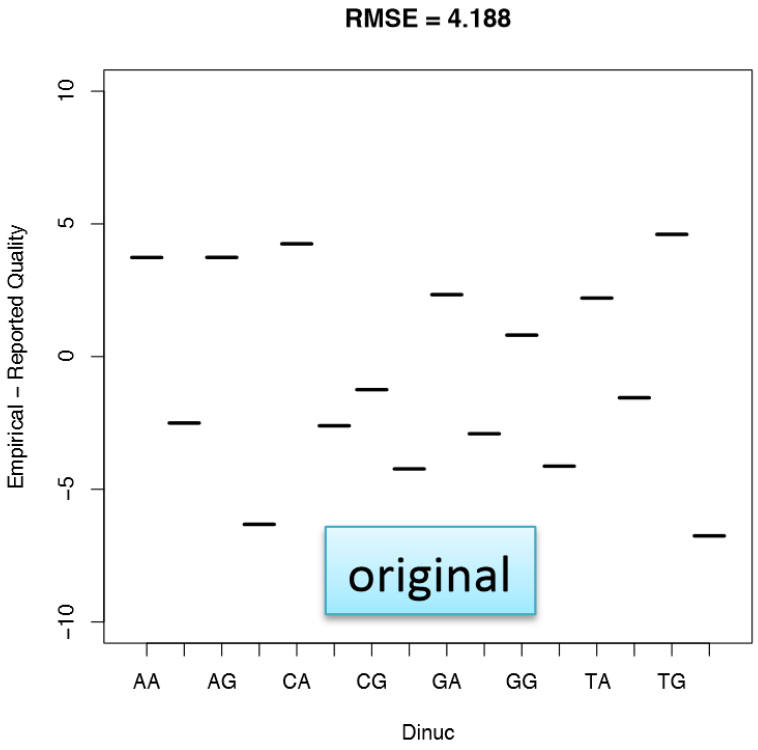
Pipeline Details…

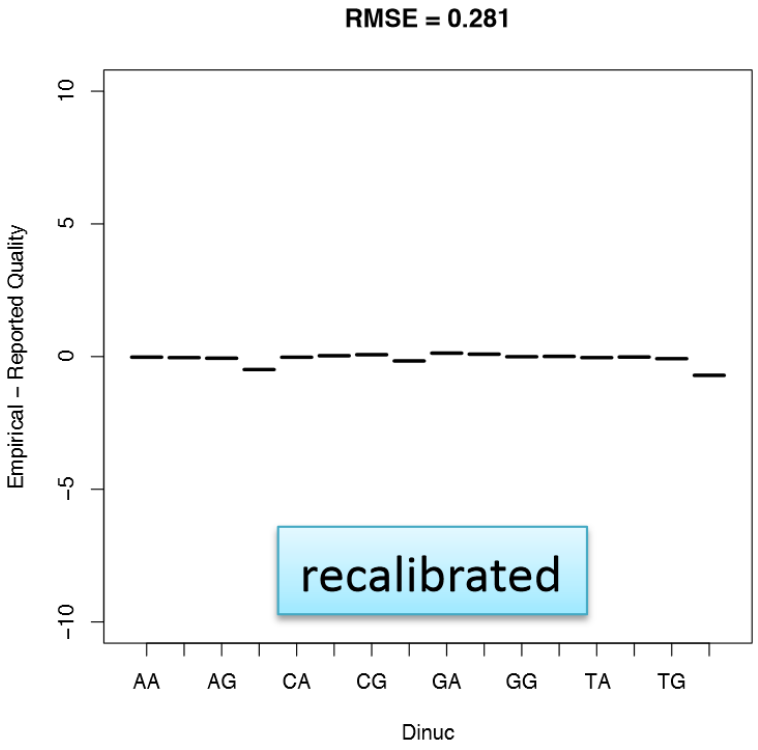
- Indel realignment

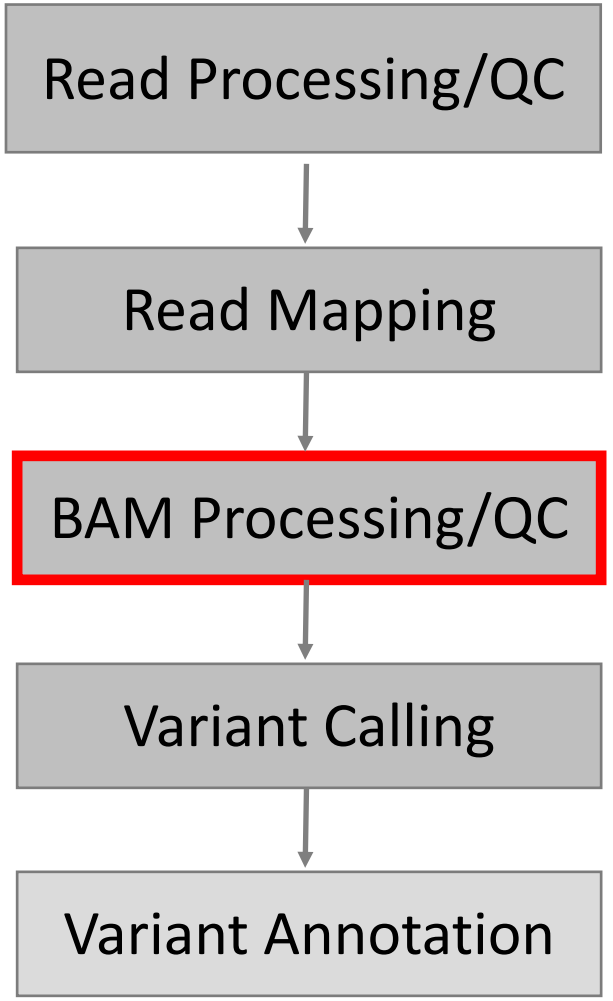


Pipeline Details…

- Multiple sources of quality score bias

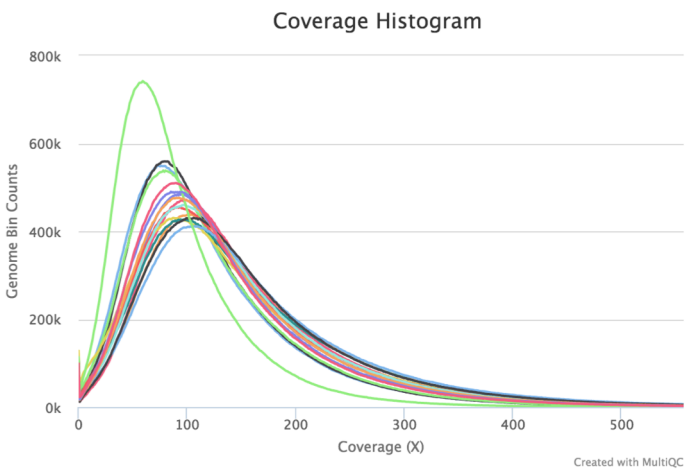


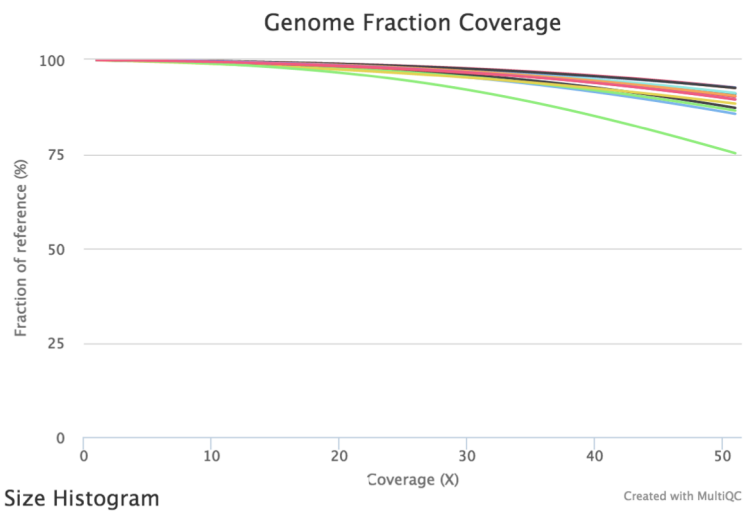


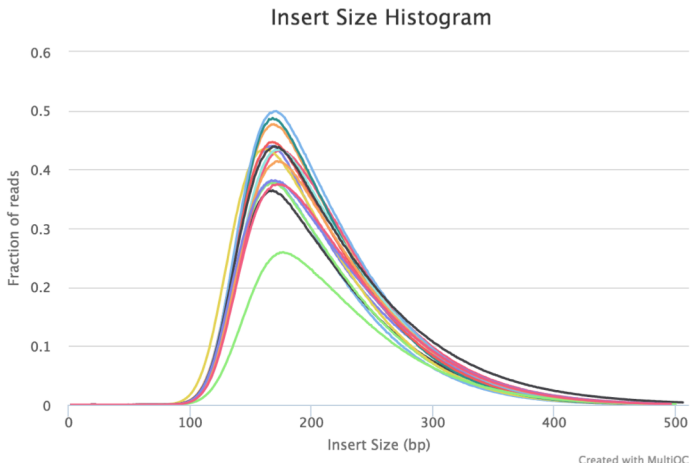


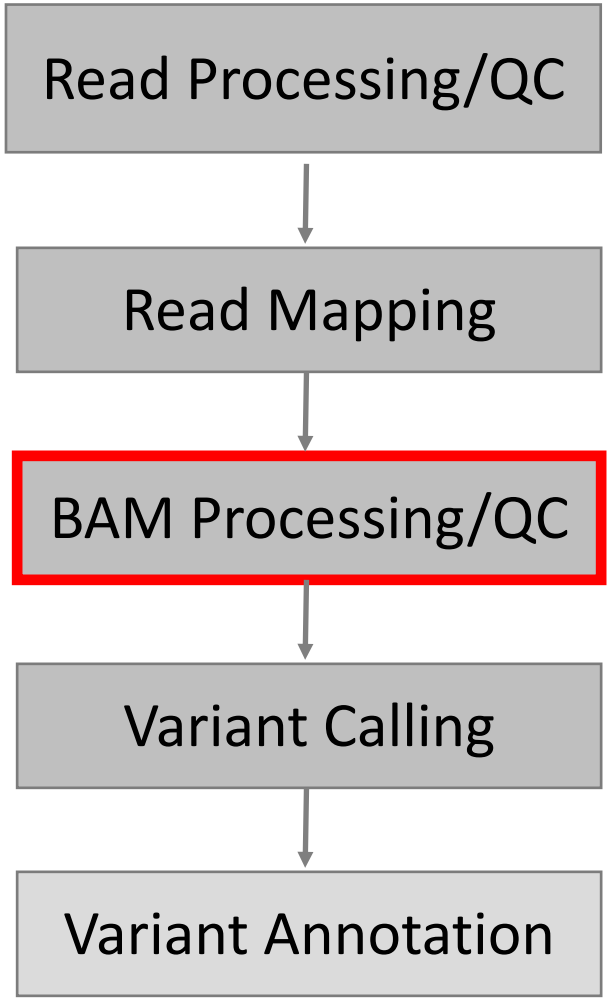
Pipeline Details…

- Alignment QC



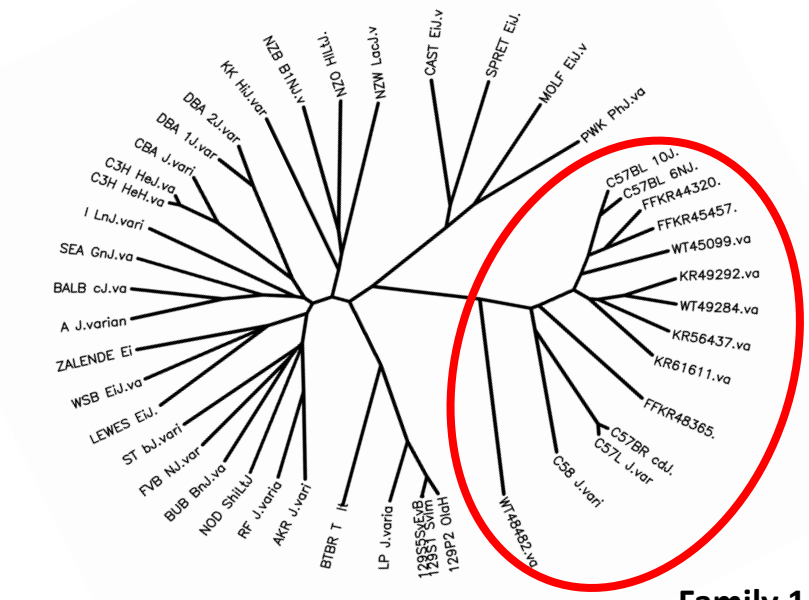


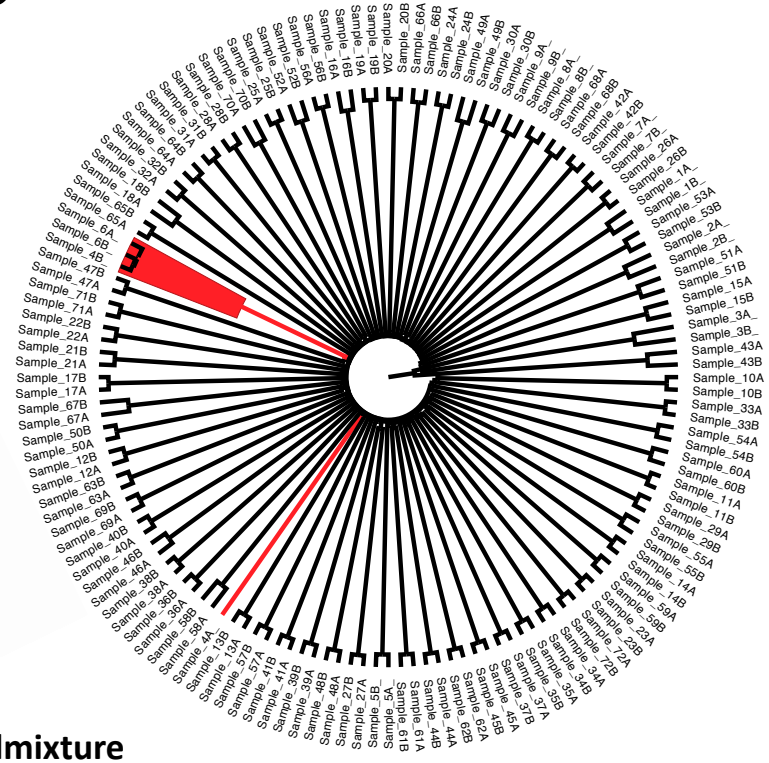


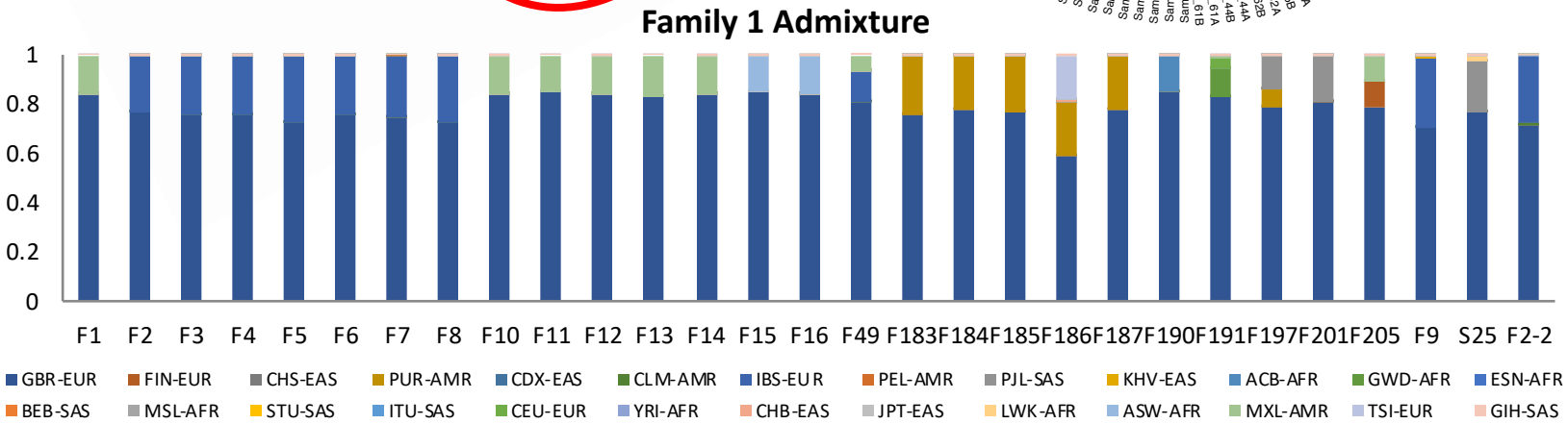


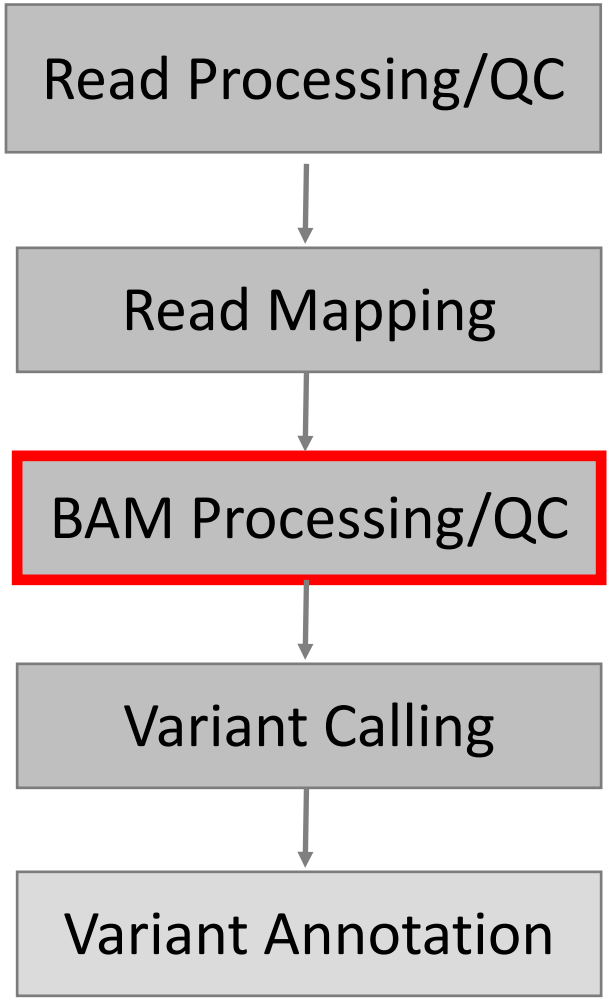
Variant Calling at CCBR

- Additional QC





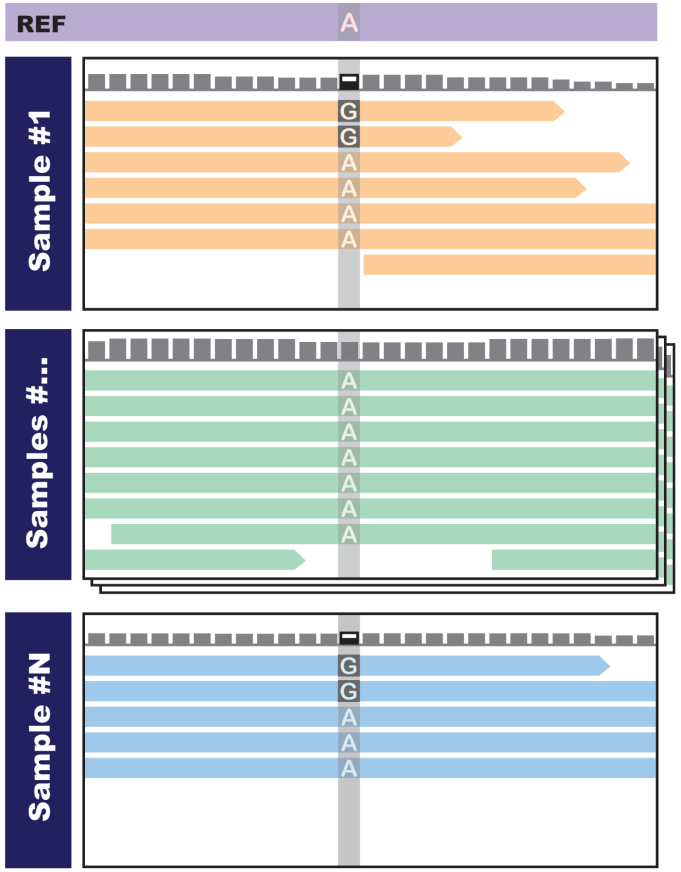


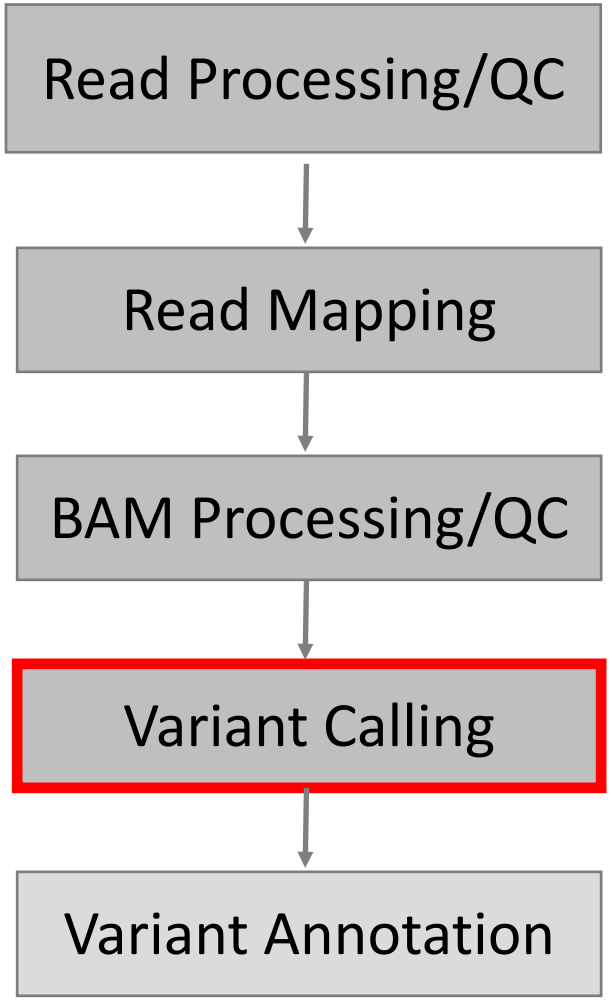


Variant Calling at CCBR

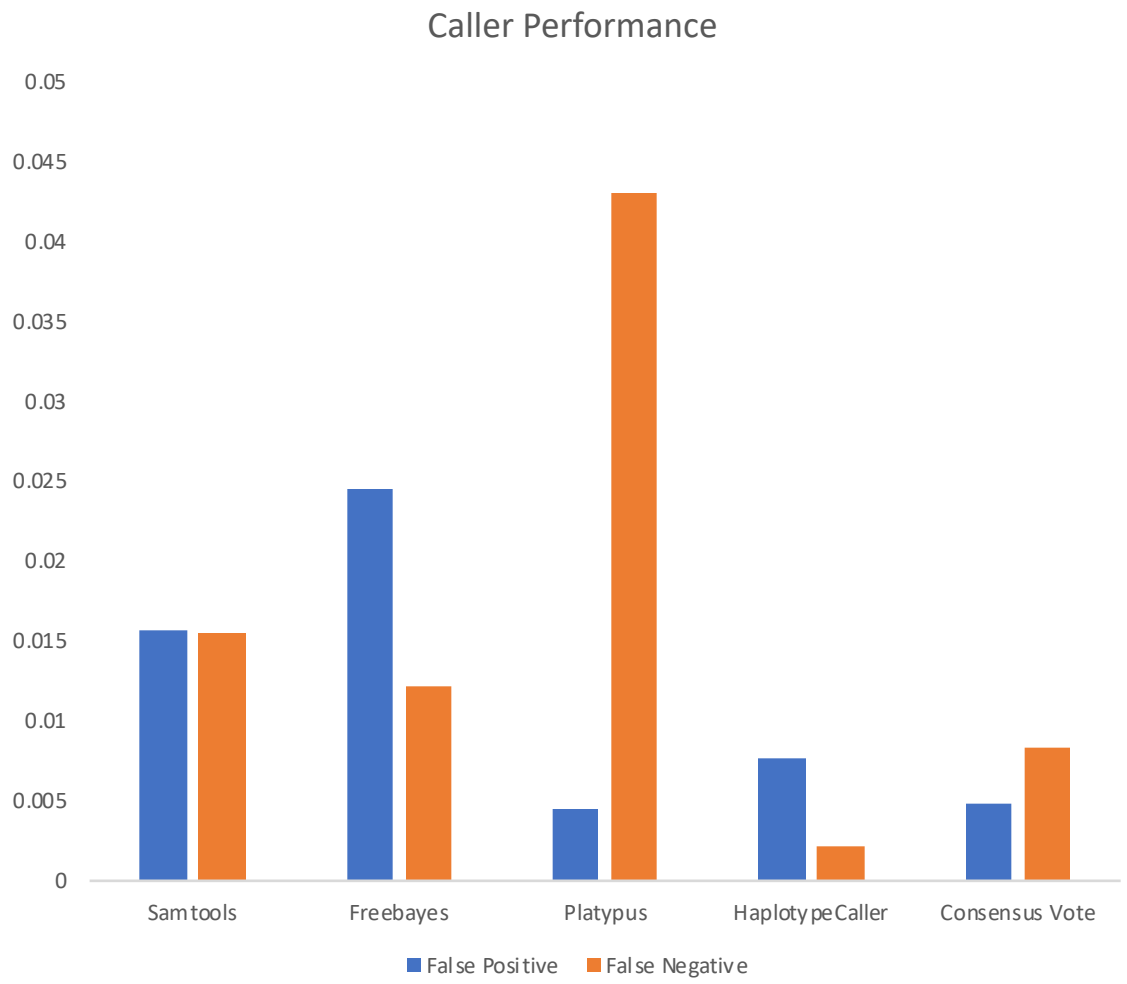
Germline

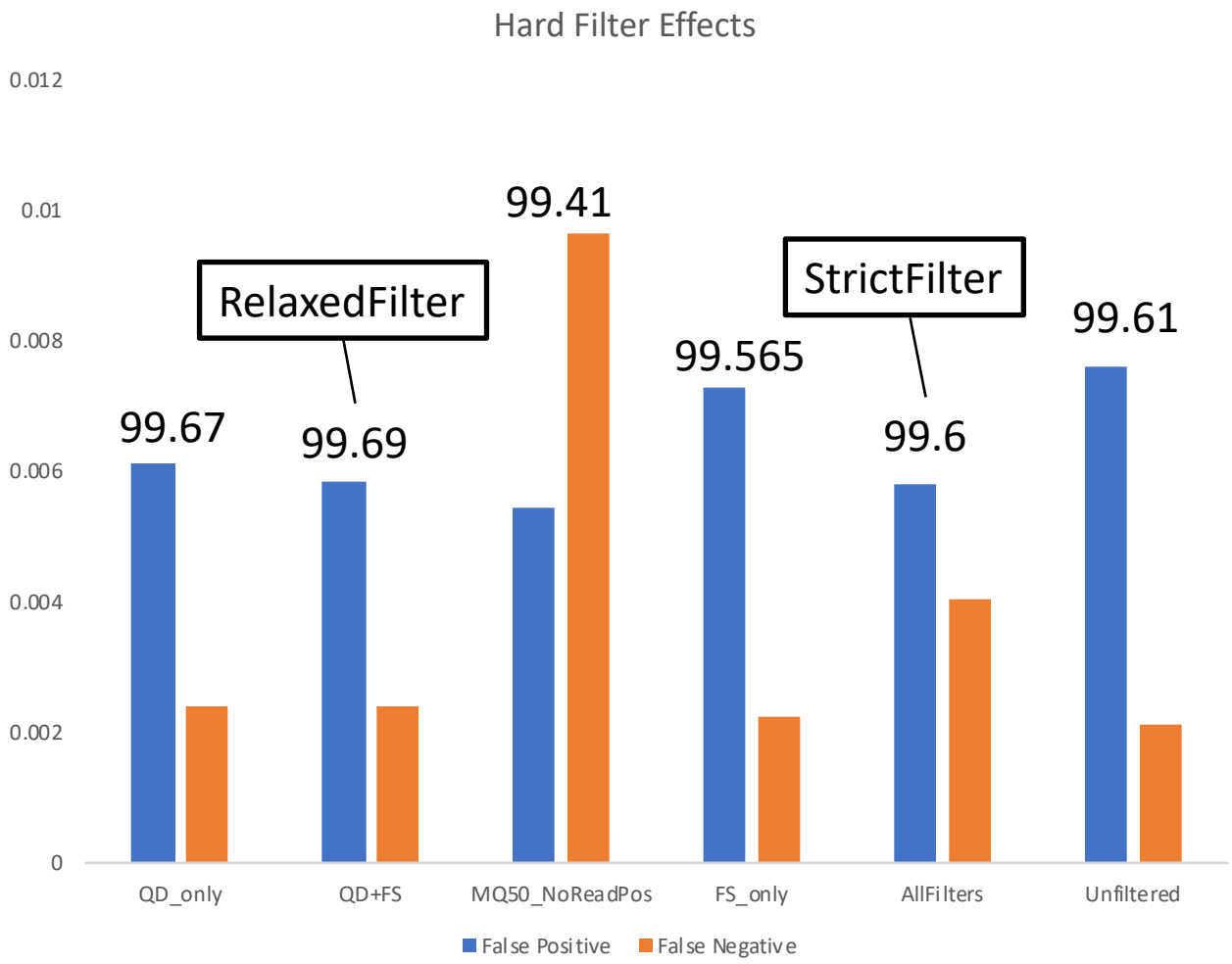
- Joint genotype with GATK HaplotypeCaller
  - SNPs/short INDELs
- We have benchmarked and optimized a series of hard filters for removing errors





Variant Caller Performance and Filtering

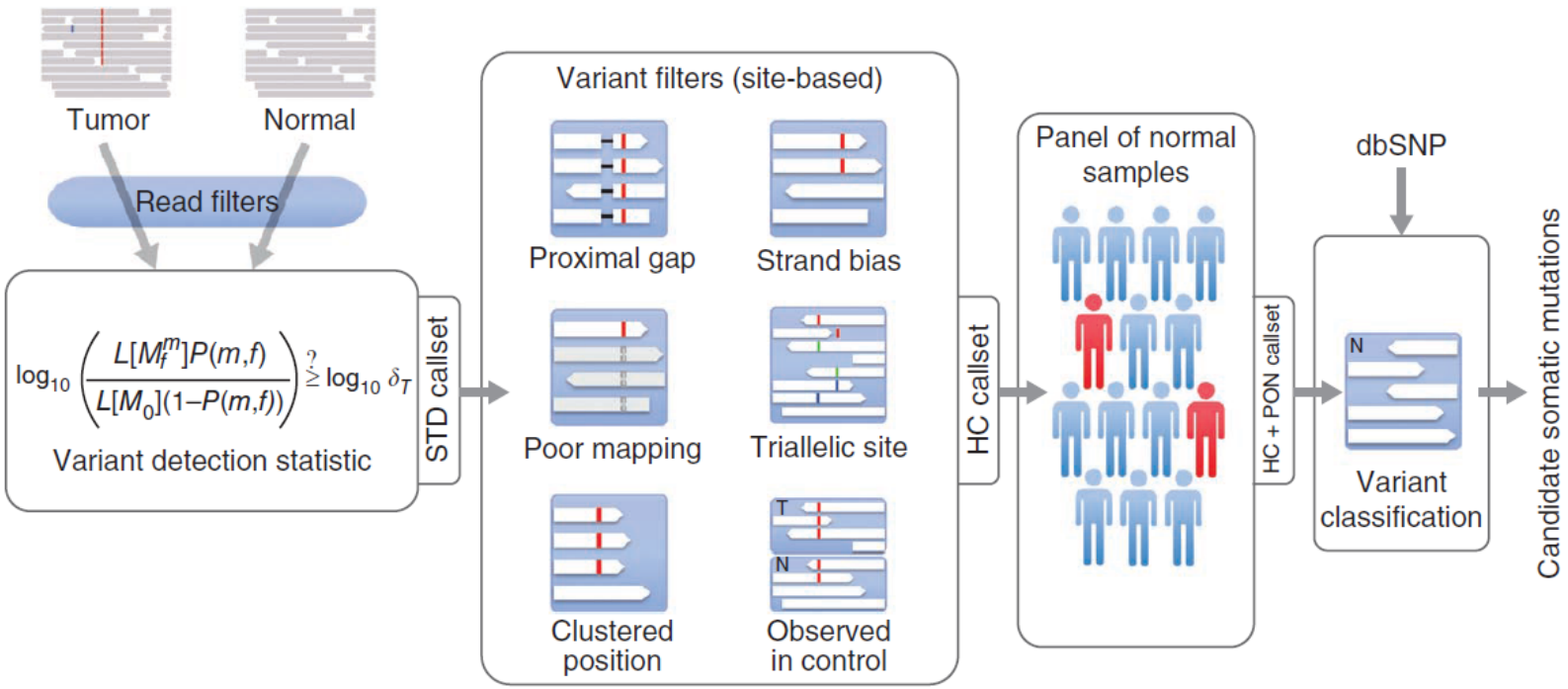


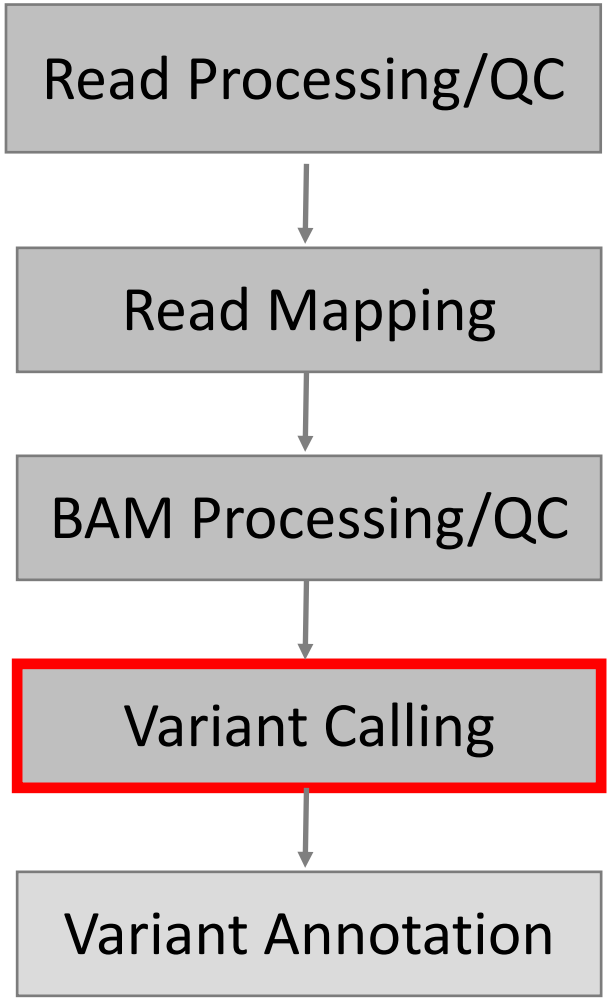


Variant Calling at CCBR

Somatic

- MuTect, MuTect2 (with hard filters), Strelka





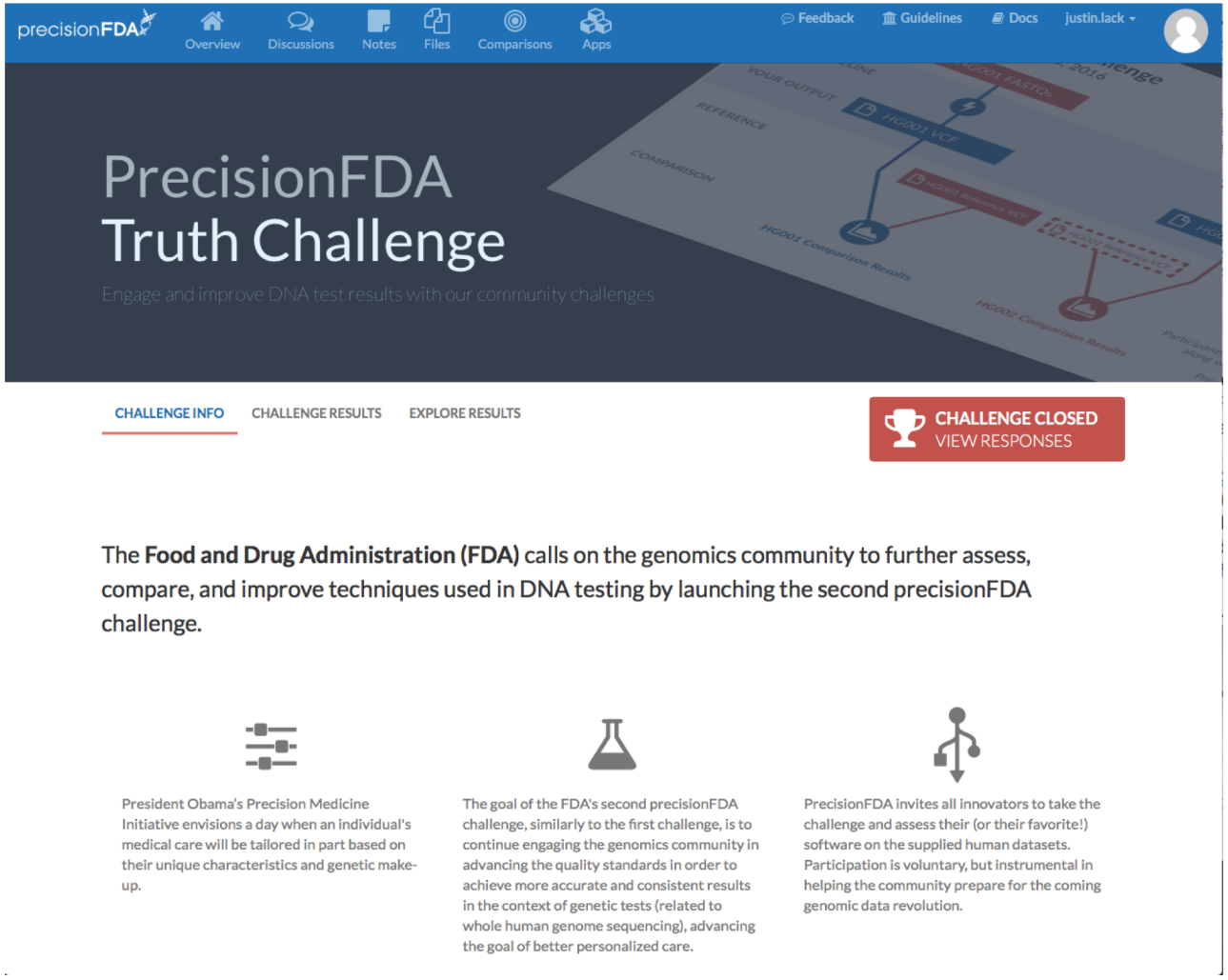
WES/WGS Pipelines with Multiple Entrypoints

- Can be run starting from fastq reads, BAMs, or gVCFs
- Setup within raw data directory
  - Raw reads all in main directory
  - ‘bams’ directory containing all BAM files from any source (e.g., Dragen)
  - ‘gvcfs’ directory containing all gVCFs from any source (e.g., Dragen)
- Initialize as usual
  - During initialization, Pipeliner automatically symlinks BAMs and gVCFs if there is a ‘bams’ and/or ‘gvcfs’ directory
- Run as usual, but skip initialQC and start from the variant calling pipeline that is appropriate

How do our Pipelines perform?

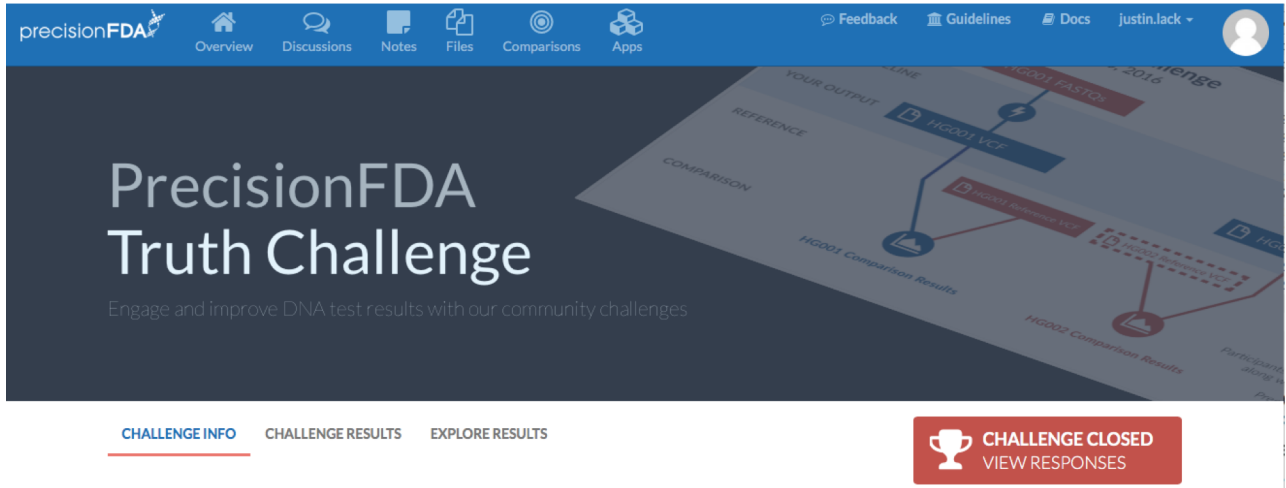
FDA Consistency/Truth Challenges

- Sought to establish best practices for germline variant calling



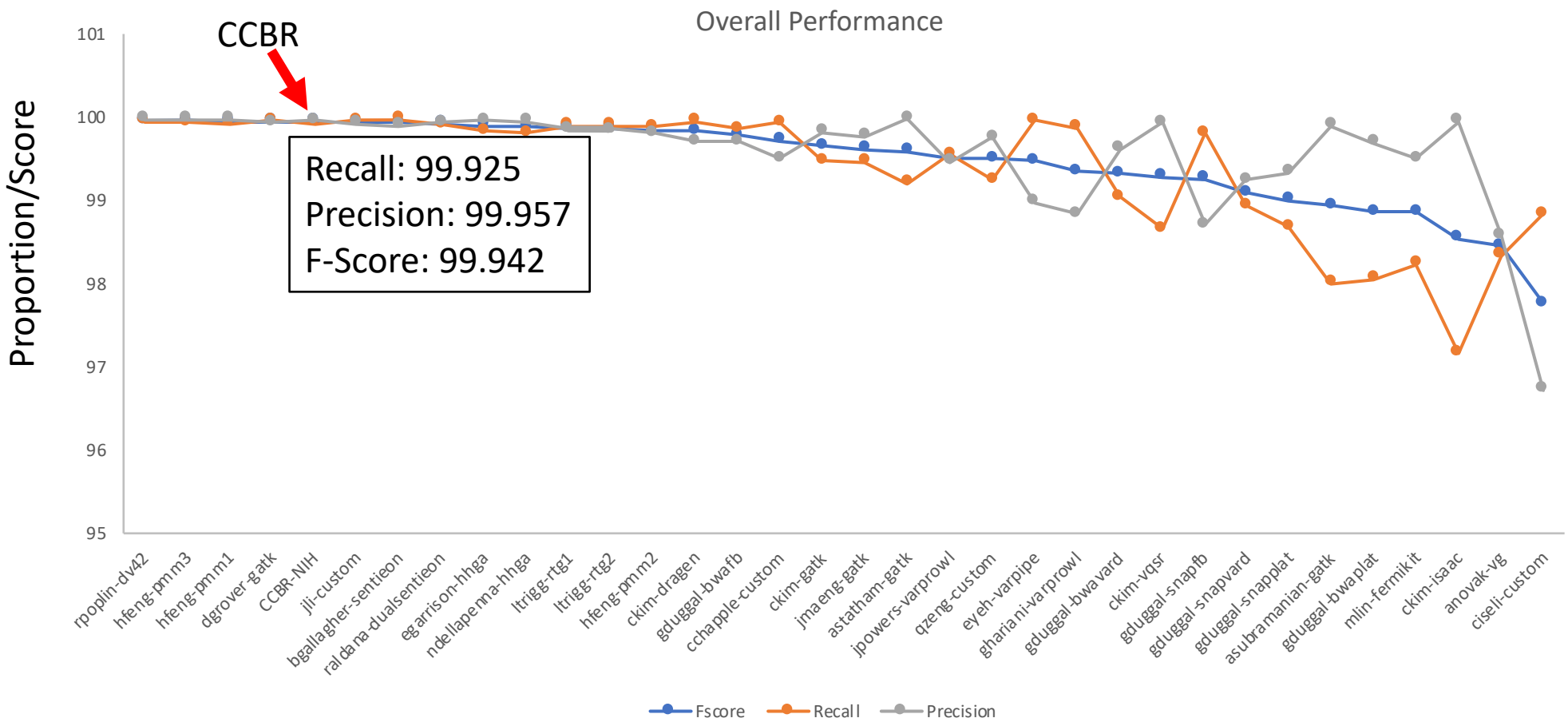
FDA Consistency/Truth Challenges

- GIAB genome (NA12878) provides known real data truth set
- Second genome (NA24385) provides unknown truth set
  - Eliminates “overtraining” problem
- Challenge(s):
  1. Pipeline determinism
  2. Precision/recall on a known truth (training set available)
  3. Precision/recall on unknown truth



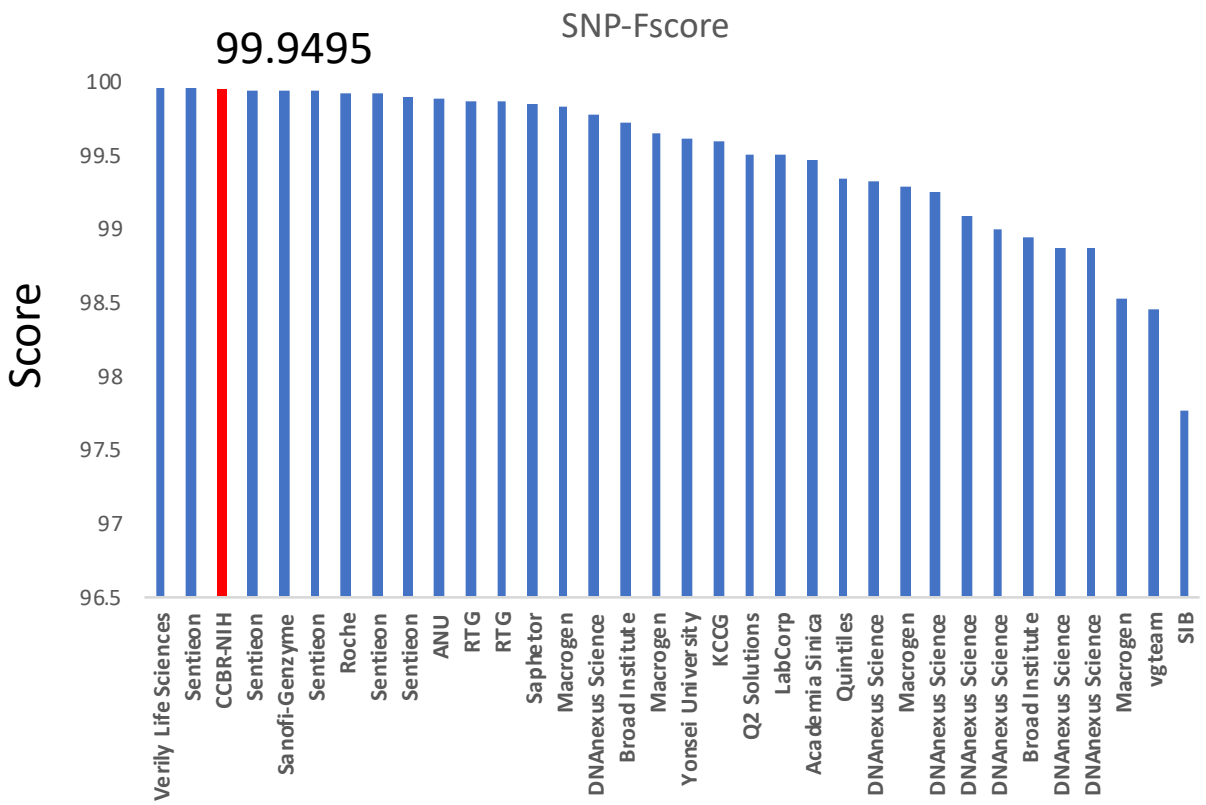
FDA Consistency/Truth Challenges

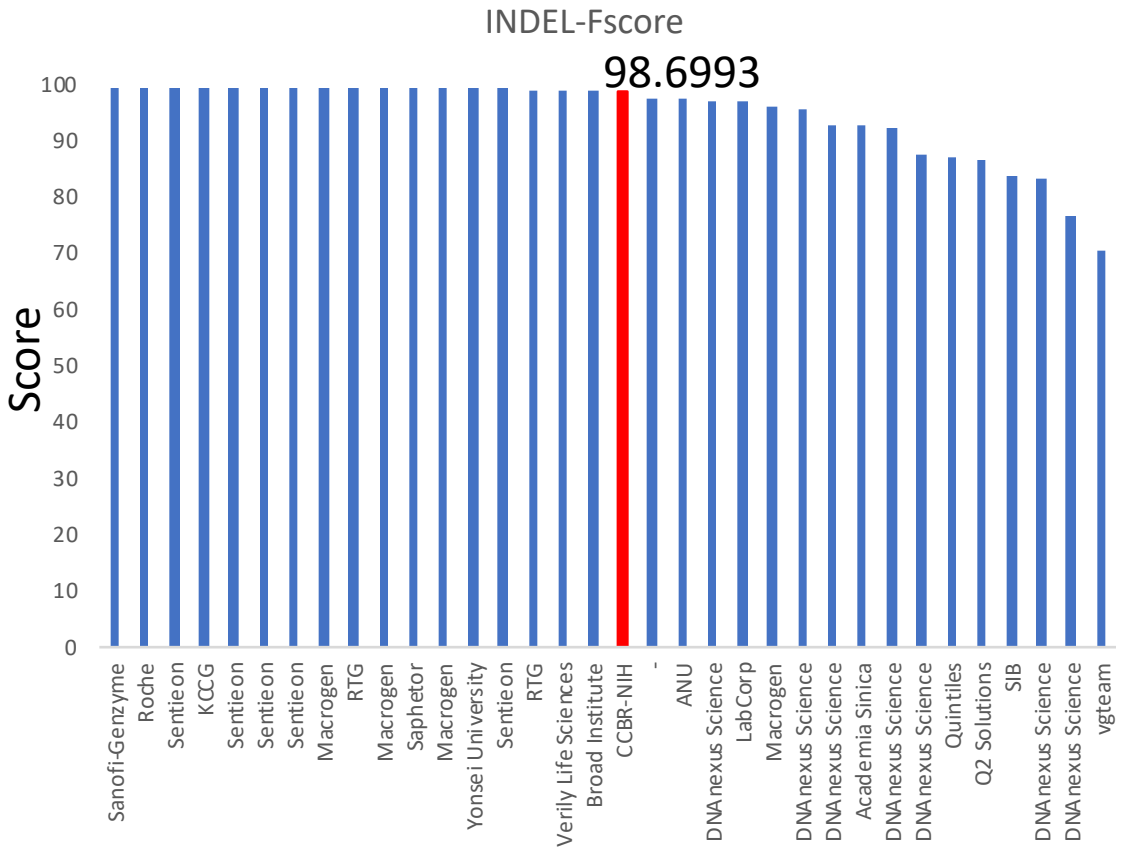
- Performance of most recent pipeline version



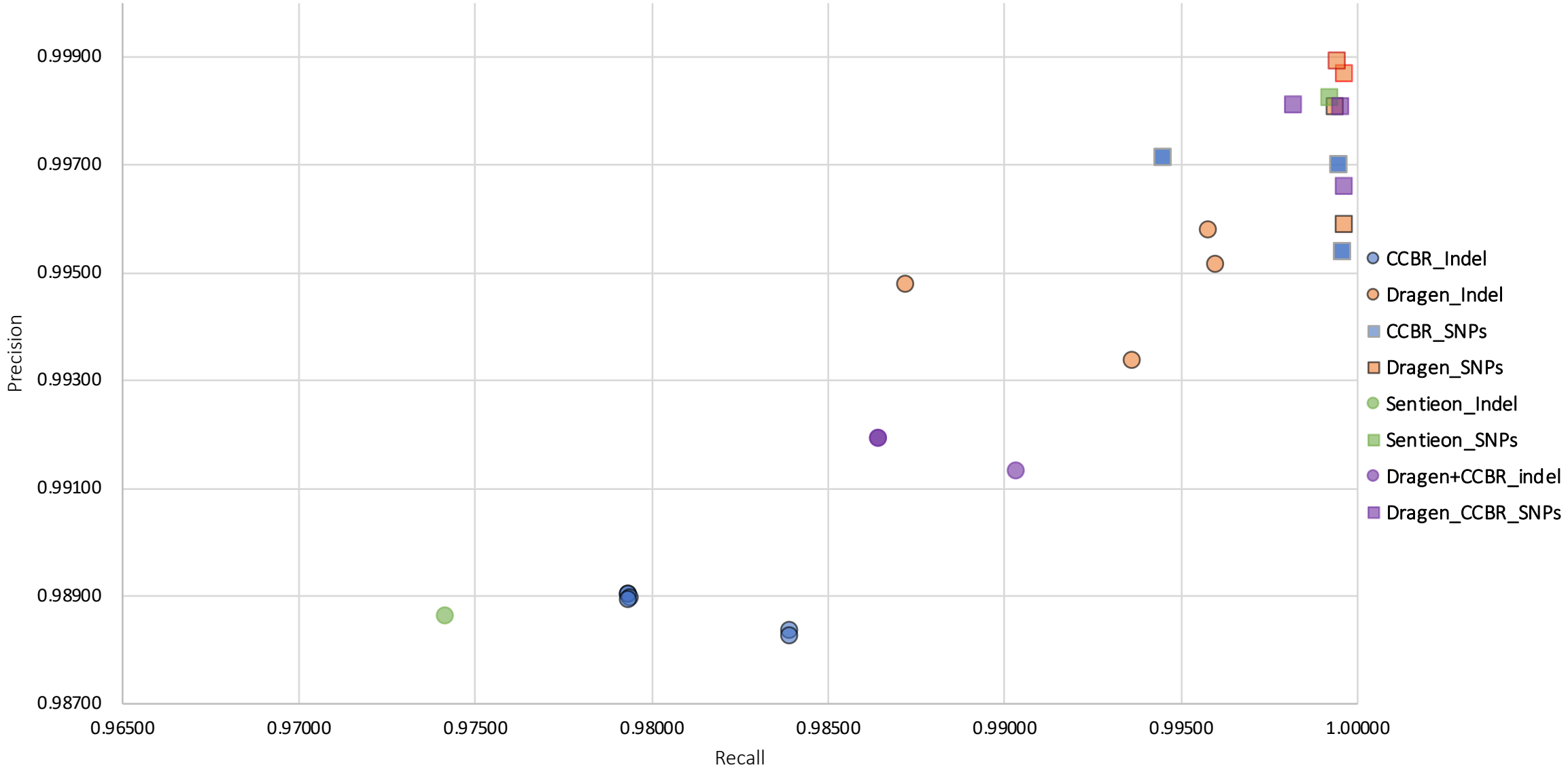
FDA Consistency/Truth Challenges

- Performance of most recent pipeline version



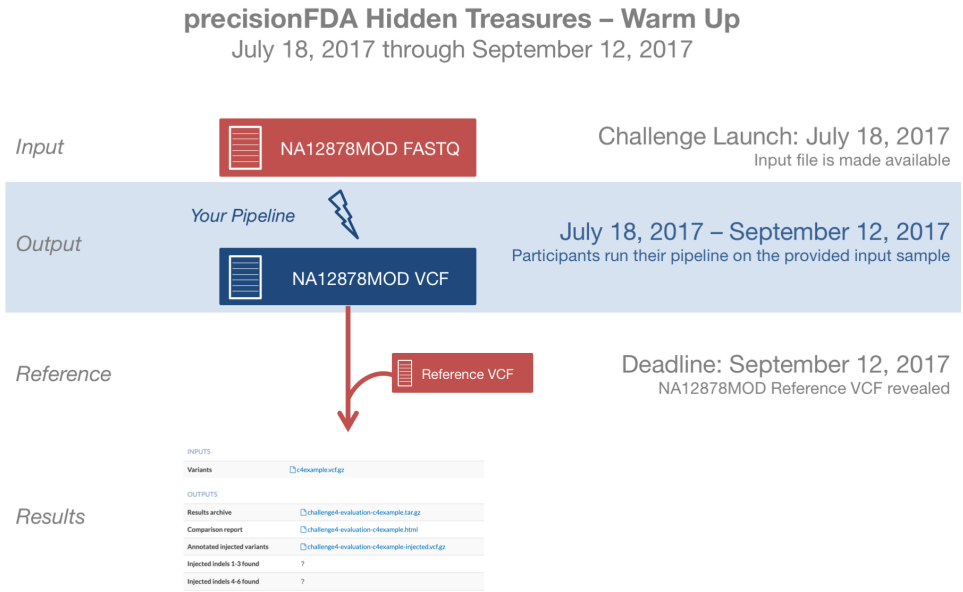


Pipeline performance vs Sentieon and Dragen - Recall vs Precision (with Keyur Talsania, CCR Sequencing Facility)



FDA Hidden Treasure Challenge

- NA12878MOD generated by *in silico* modification of NA12878
- 50 Spike-in variants >=0.2 frequency
- INDELs <= 40bp
- Evaluate ability to detect *in silico* variants
- FP/FN balance for SNPs and INDELs
- Ability to accurately call allele frequencies

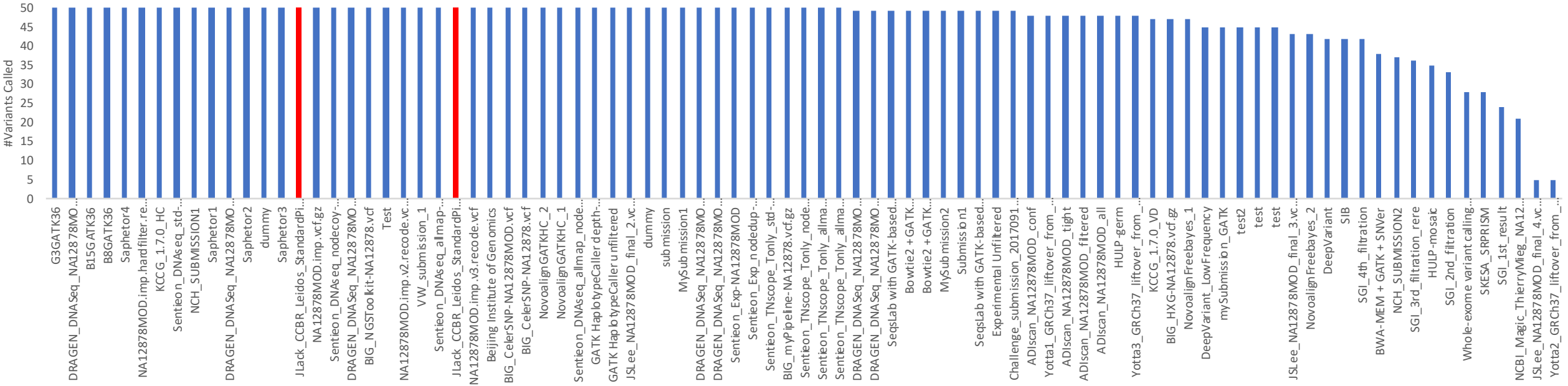


FDA Hidden Treasure Challenge

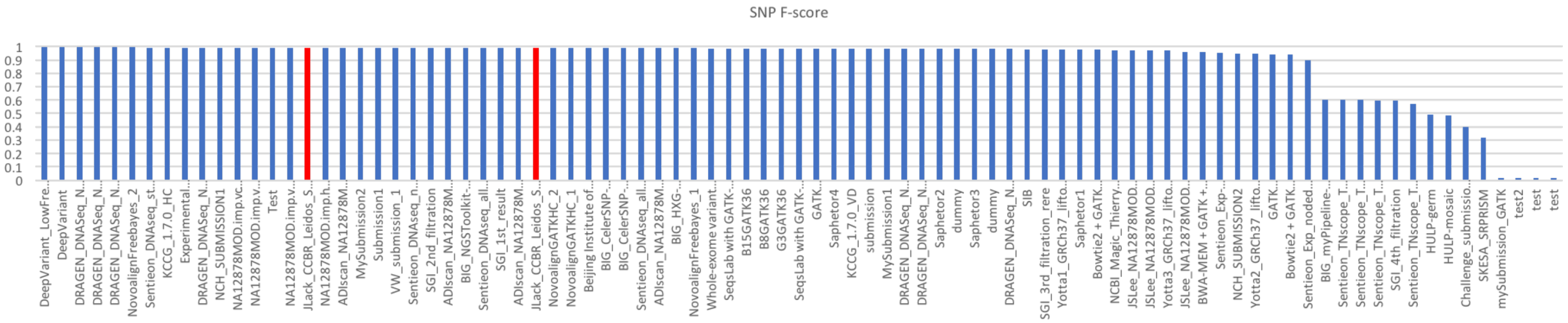
- 86 entries
- Ran the full somatic pipeline with both strict and relaxed filtering criteria

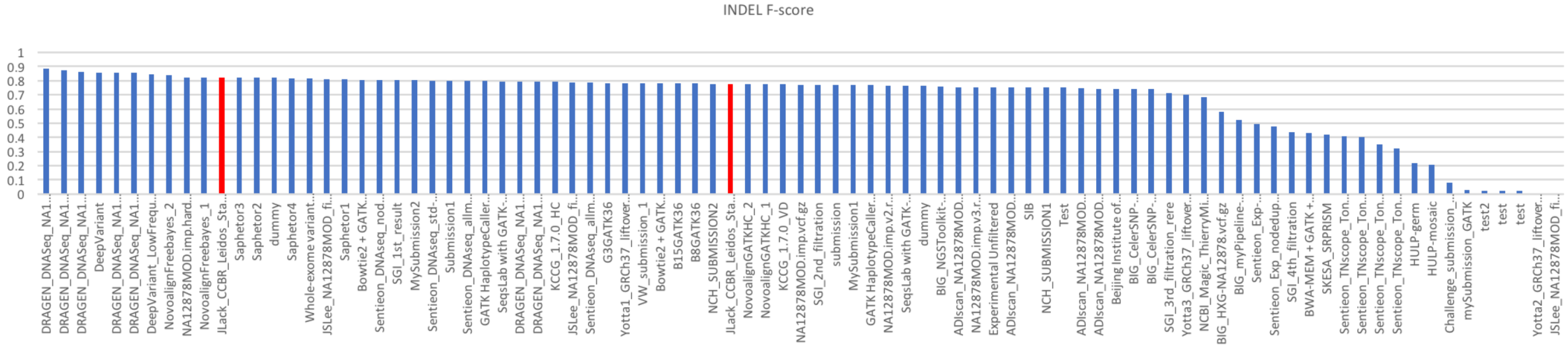
FDA Hidden Treasure Challenge

- 86 entries
- Ran the full somatic pipeline with both strict and relaxed filtering criteria
- Successfully recovered all 50 spike-in variants



Performance - F-score





A Pipeliner Demo…

Questions?

GATK - Variant Quality Score Recalibration (VQSR)