GATK Best Practices for Variant Discovery

Somatic SNV and Indel Discovery

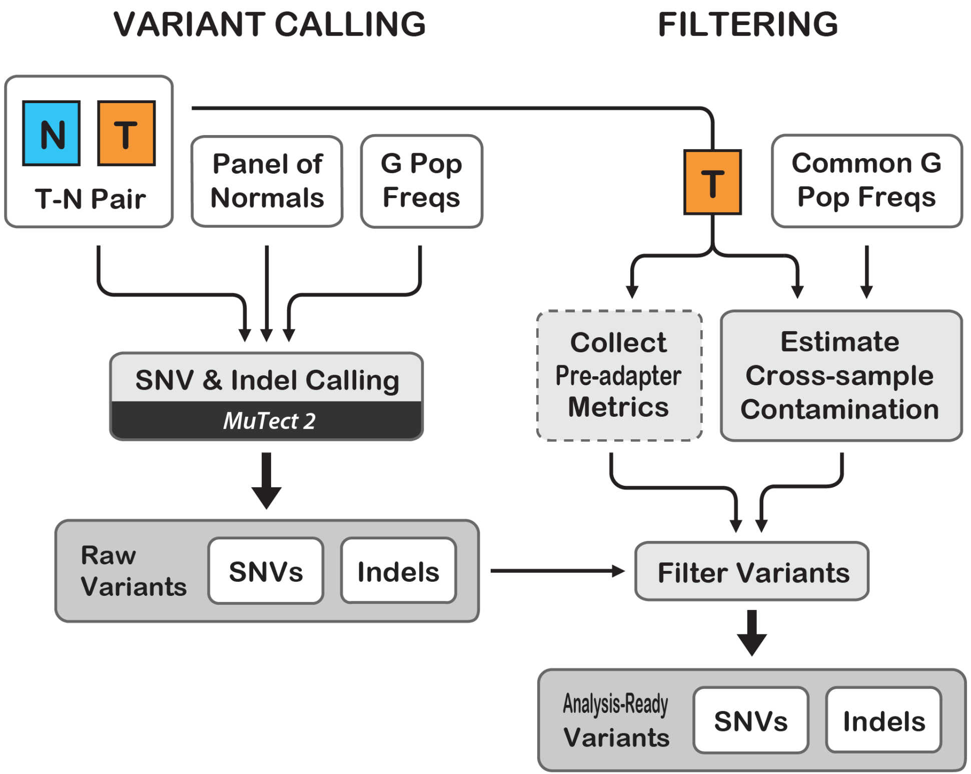
Basic operation and algorithm of
Mutect2 and related tools

BROAD INSTITUTE

https://software.broadinstitute.org/gatk/

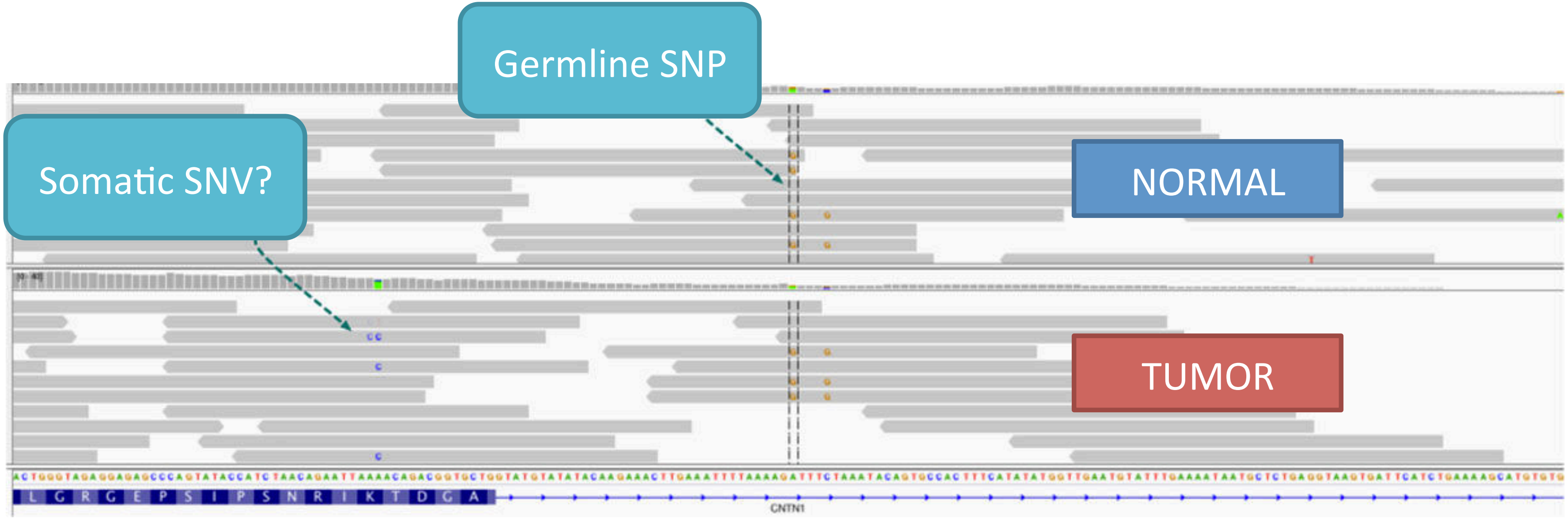
gatk

Somatic SNV & Indel discovery workflow



Logic of the Tumor-Normal workflow

Comparison to matched normal -> subtraction of germline background



Germline SNP

Somatic SNV?

NORMAL

TUMOR

Tumor-only analysis

- It is possible to run the workflow without a matched normal in “tumor-only mode” (normally used for PON creation)

- MUST have a good PON to eliminate common germline variation

- Will still require extra filtering (not described here)

Panel of Normals for SNVs & Indels

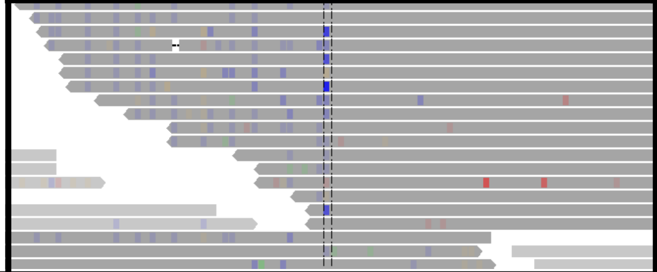
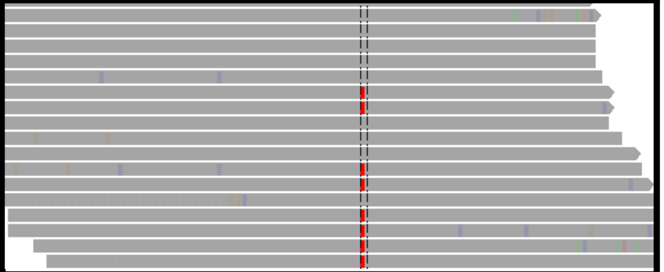
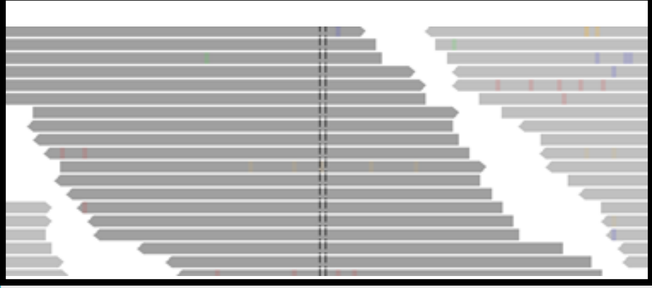
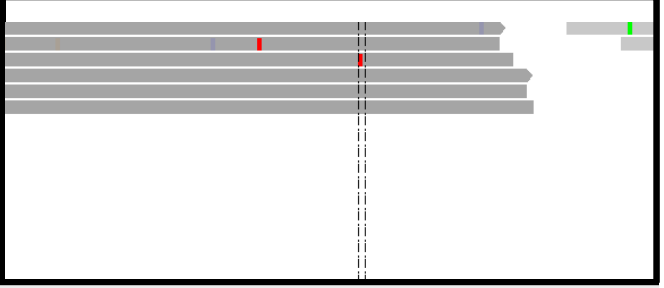
- VCF of calls made from a set of unrelated “normal” samples

- Main purpose:
  Eliminate common/recurring technical artifacts
  -> should use normals made using the same data generation techniques (eg same capture kit for exomes, same sequencing platform etc)

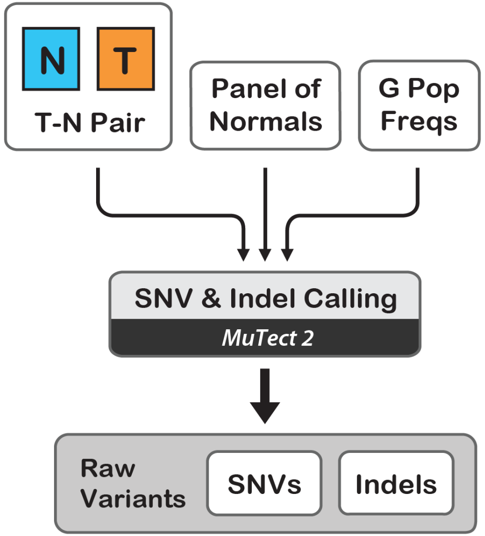
- Secondary purpose: also eliminates germline variants not called in the matched normal (or approximates the normal if none is available)

False positives from artifacts and germline variation

Somatic point mutations occur ~1 / Mbp

	ARTIFACT	GERMLINE EVENT
TUMOR		
NORMAL		
At risk	Every base	~1667 germline variants / Mbp
Source	• Misread bases • Sequencing artifacts • Misaligned reads	• Low coverage in NORMAL
Solutions	*filters, Panel of Normals (PoN)*	*dbSNP, ExAC, COSMIC, PoN* *gnomAD*

The 1000 Genomes Project Consortium, *Nature* 2015. *A global reference for human genetic variation.*
Exome Aggregation Consortium *<ExAC.broadinstitute.org>*,
Genome Aggregation Database <http://gnomad.broadinstitute.org/>



Step 1

CALL VARIANTS WITH MUTECT2

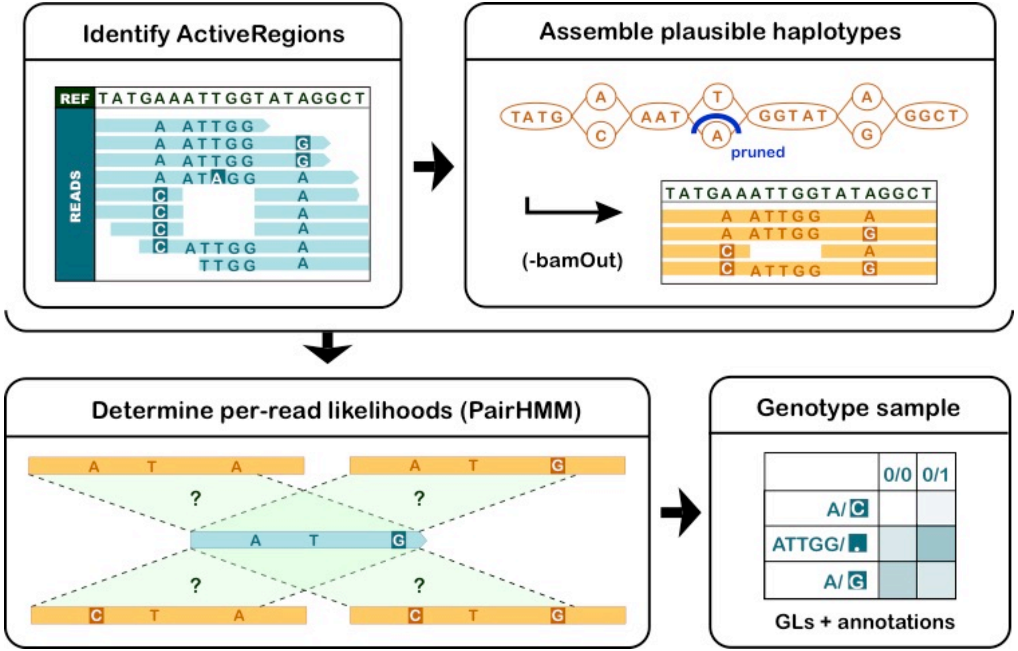
Mutect2 is based on HaplotypeCaller

Skip :

- Sites in *PoN*
- Sites with high fraction alt alleles in normal

Allele-specific calling:

- Distinguishes alleles in *the germline population frequency* resource and uses AF in calculating probability variant exists in normal *and* tumor



Somatic genotypes inferred from PairHMM likelihoods

$$\mathcal{L}(G_{\mathrm{ref}}|R) = \prod_j \mathcal{L}(G_{\mathrm{ref}}|R_j) \qquad \mathcal{L}(G_i|R) = \prod_j \left[\mathcal{L}(G_i|R_j) f_{\mathrm{alt}} + \mathcal{L}(G_{\mathrm{ref}}|R_j)(1 - f_{\mathrm{alt}})\right]$$

Likelihood of reference genotype given all reads

Likelihood of variant genotype *i* given all reads

Likelihoods of variant/reference alleles given read *j*

$$\mathrm{LR}_i = \log \mathcal{L}(G_i|R) - \log \mathcal{L}(G_{\mathrm{ref}}|R)$$

Log-likelihood ratio for genotype *i*

- No explicit ploidy assumptions (unlike HaplotypeCaller)
  - somatic genotype likelihoods weighted by variant allele fraction
- Statistical threshold for somatic call uses log-likelihood ratios
  - ≥ 5.3 in favor of the variant somatic genotype
  - Also filter based on the likelihood of the allele in the Normal

Converting likelihoods to probabilities uses AF (f)

- If variant is in gnomAD:

  Use the allele frequency f in gnomAD

- If variant is *not* in gnomAD:

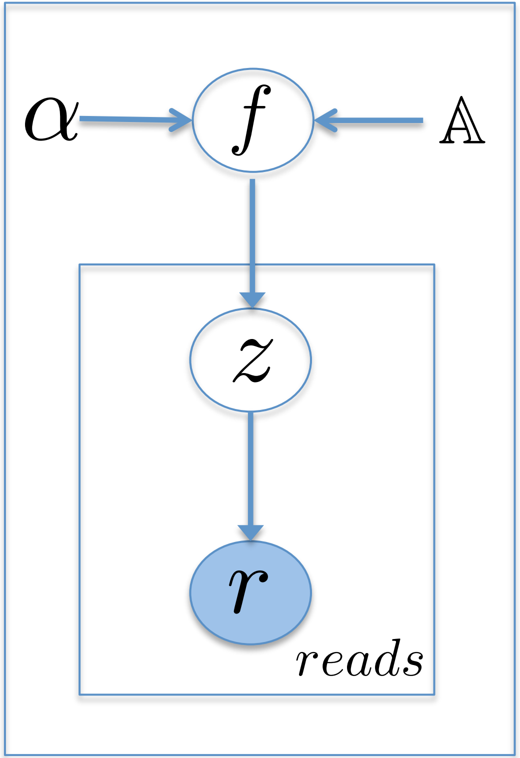
  Set f = 1/(2#samples + 2), which is a reasonable guess for the allele frequency of the variant*

- If we don't have gnomAD:

  Default to f = 0.001

Genotyping in GATK4 Mutect2

- GATK4 Mutect2 models the allele fractions and allele assignment to each read as latent variables f and z
- Choose the allele set A that maximizes model evidence
- If log odds > 3.0 (by default) then emit variant
- At low coverage sites, the Bayesian approach in GATK4 performs better than the prior frequentist approach in GATK3



$$f \sim \text{Dirichlet}(\alpha)$$

$$z|f \sim \text{Categorical}(f)$$

$$p(r|z_{ra}) = l_{ra}$$

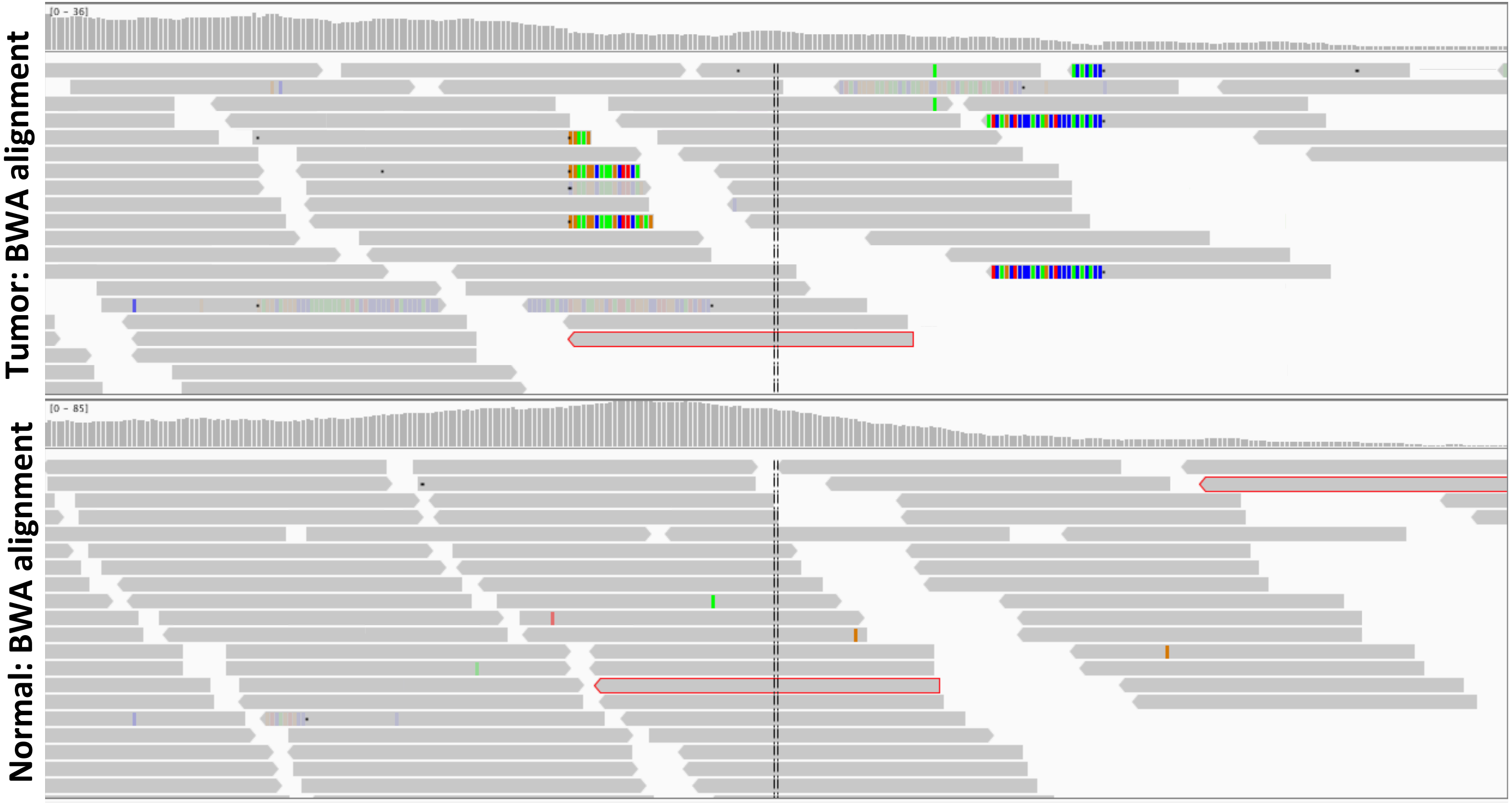
$$\ell_{ra} \equiv P(\text{read } r|\text{allele } a)$$

from PairHMM

$$\log \frac{p(\mathbb{R}|\mathbb{A}_{alt})}{p(\mathbb{R}|\mathbb{A}_{ref})} > \delta = 3.0$$

then emit variant

Case Study: 120 base deletion

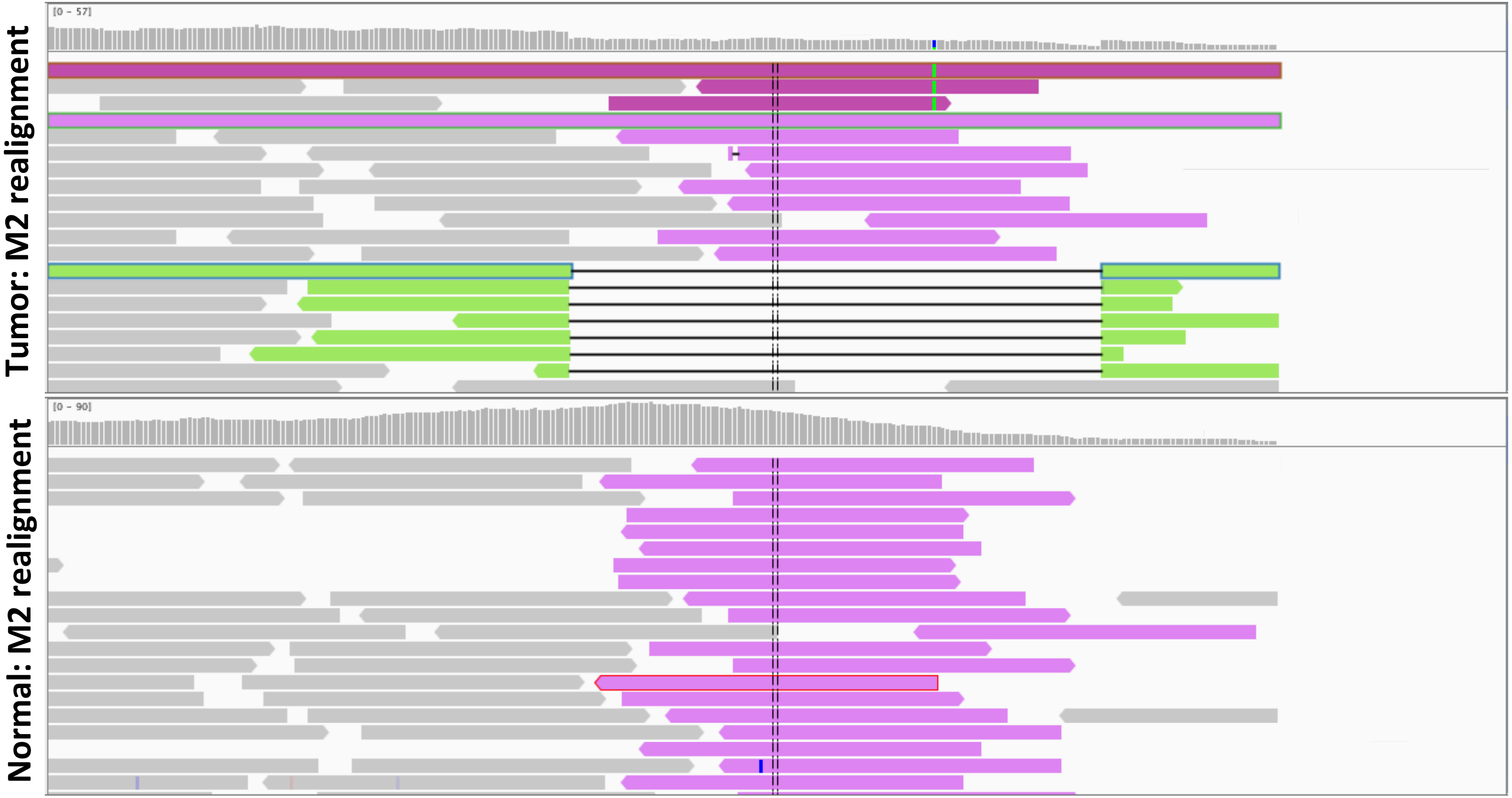


Clear evidence of some sort of event is present, but impossible for a traditional pileup-based caller to recover

Event would also not be caught with discordant read pair caller, since insert sizes of supporting reads are within normal range

Courtesy of J. Hess

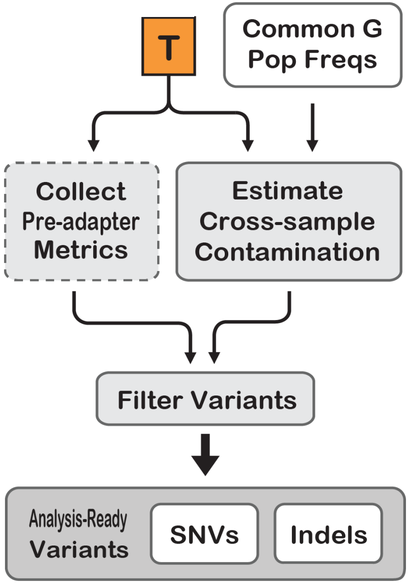
Case Study: 120 base deletion



MuTect2 reassembly recovers the 120 base deletion haplotype

It also discerns reads that are unambiguously phased into the WT haplotype, and a haplotype with insufficient likelihood.

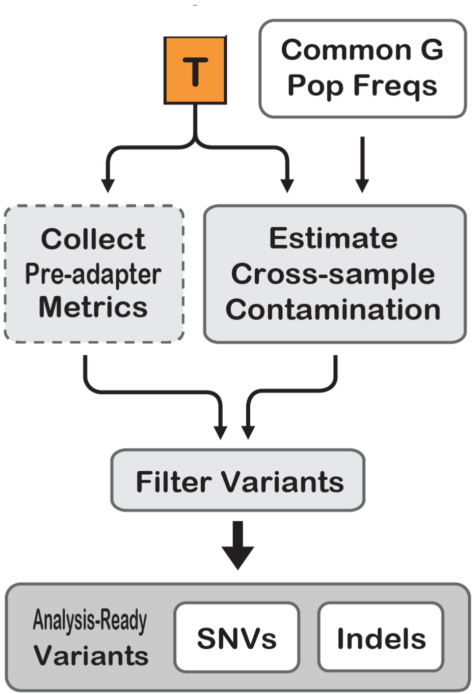
Courtesy of J. Hess



Step 2

FILTER RAW VARIANT CALLS

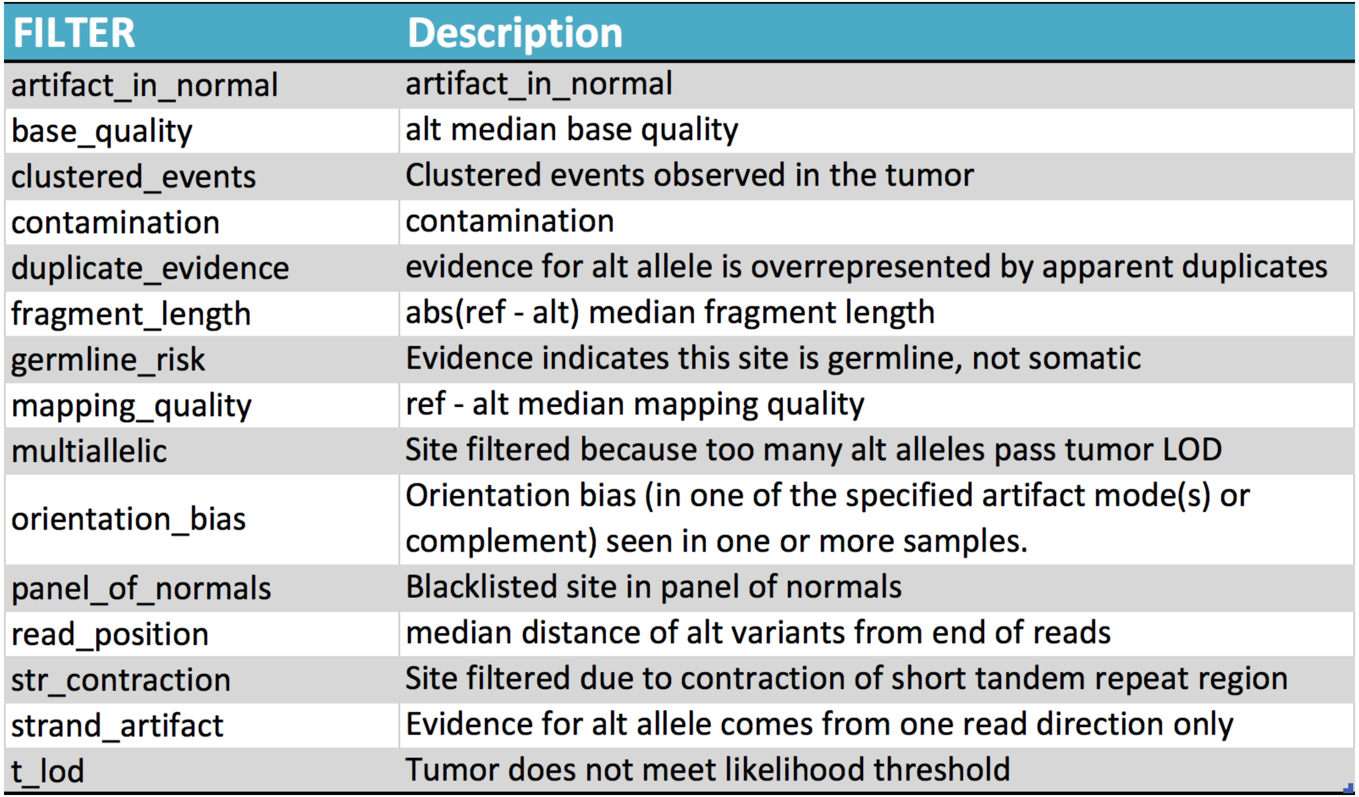
Filtering is based on annotations + contamination estimate



ANNOTATION	INFO field annotations
Coverage	**DP**
DepthPerAlleleBySample	**AD**
TandemRepeat	**STR**
OxoGReadCounts	**F1R2, F2R1**
ReadPosition	**MPOS**
BaseQuality	**MBQ**
MappingQuality	**MMQ**
FragmentLength	**MFRL**
StrandArtifact	**SA_POST_PROB, SA_MAP_AF**

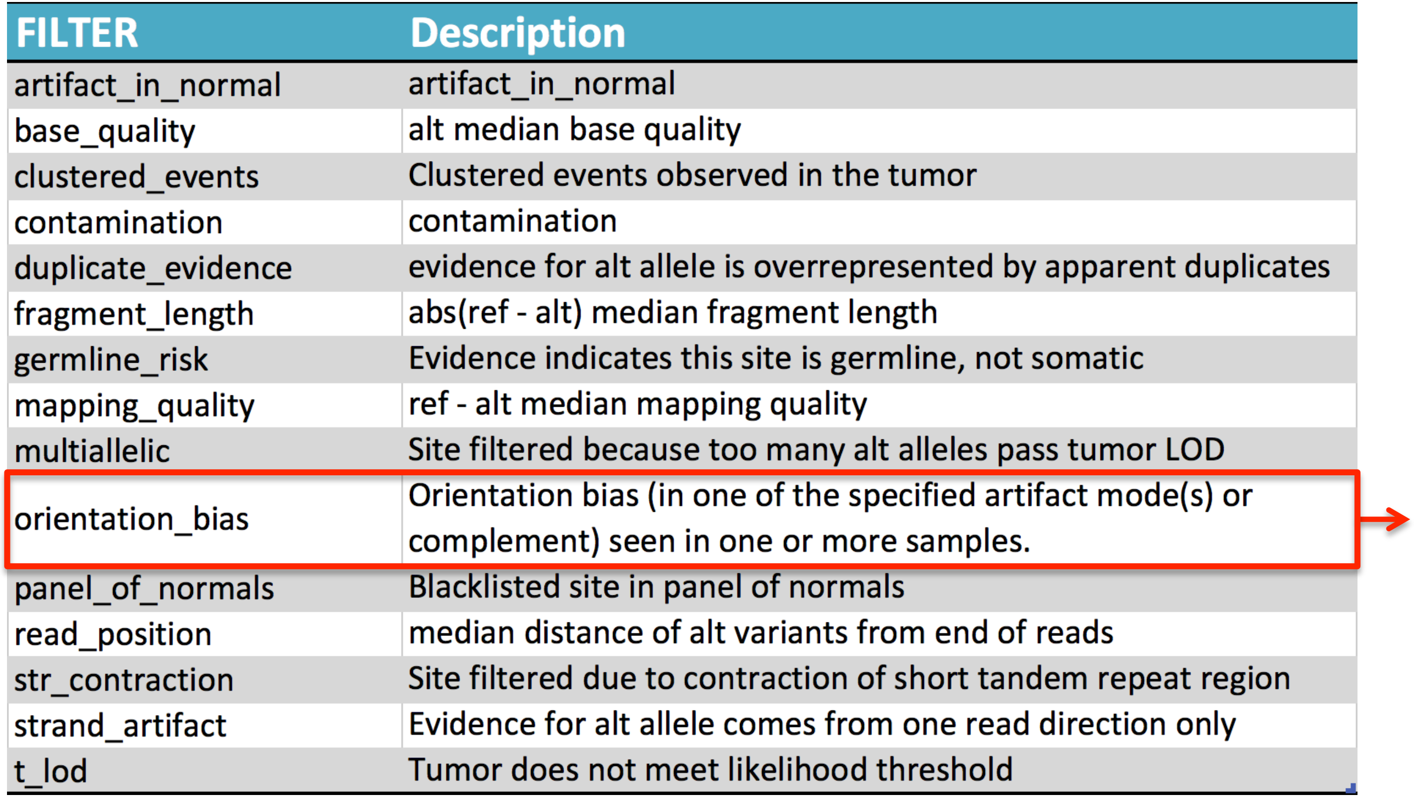
** Not a comprehensive list*

FilterMutectCalls filters for multiple criteria



FILTER	Description
artifact_in_normal	artifact_in_normal
base_quality	alt median base quality
clustered_events	Clustered events observed in the tumor
contamination	contamination
duplicate_evidence	evidence for alt allele is overrepresented by apparent duplicates
fragment_length	abs(ref - alt) median fragment length
germline_risk	Evidence indicates this site is germline, not somatic
mapping_quality	ref - alt median mapping quality
multiallelic	Site filtered because too many alt alleles pass tumor LOD
orientation_bias	Orientation bias (in one of the specified artifact mode(s) or complement) seen in one or more samples.
panel_of_normals	Blacklisted site in panel of normals
read_position	median distance of alt variants from end of reads
str_contraction	Site filtered due to contraction of short tandem repeat region
strand_artifact	Evidence for alt allele comes from one read direction only
t_lod	Tumor does not meet likelihood threshold

Additional filters for sequence context artifacts



FILTER	Description
artifact_in_normal	artifact_in_normal
base_quality	alt median base quality
clustered_events	Clustered events observed in the tumor
contamination	contamination
duplicate_evidence	evidence for alt allele is overrepresented by apparent duplicates
fragment_length	abs(ref - alt) median fragment length
germline_risk	Evidence indicates this site is germline, not somatic
mapping_quality	ref - alt median mapping quality
multiallelic	Site filtered because too many alt alleles pass tumor LOD
orientation_bias	Orientation bias (in one of the specified artifact mode(s) or complement) seen in one or more samples.
panel_of_normals	Blacklisted site in panel of normals
read_position	median distance of alt variants from end of reads
str_contraction	Site filtered due to contraction of short tandem repeat region
strand_artifact	Evidence for alt allele comes from one read direction only
t_lod	Tumor does not meet likelihood threshold

→ FilterByOrientationBias

E.g. likely OxoG G→T transversions

Mutect2 command and main options

Base command for PoN creation and tumor-only analysis:

```
gatk Mutect2 \
    -R ref_fasta.fa \
    -I sample.bam \
    -tumor sample_name \
    -L intervals.list \
    -O sample.vcf.gz
```

For matched-normal tumor calling add:

```
    -I normal.bam \
    -normal normal_sample_name \
    -bamout bamout.bam \
```

Reassembled BAM now recommended

To specify a germline AF resource:

```
    --germline_resource af-only-gnomad.vcf.gz \
    --af_of_alleles_not_in_resource 0.0000025 \
```

Germline resource must have allele-specific frequencies; af for not in gnomAD exomes

To specify a PoN:

```
    --normal_panel pon.vcf.gz \
```

Filtering commands and main options

Filter M2 calls for multiple contexts:

```
gatk FilterMutectCalls \
    -V tumor_matched_m2_snvs_indels.vcf.gz \
    -contaminationTable contamination.table \
    -O tumor_matched_m2_oncefiltered.vcf.gz
```

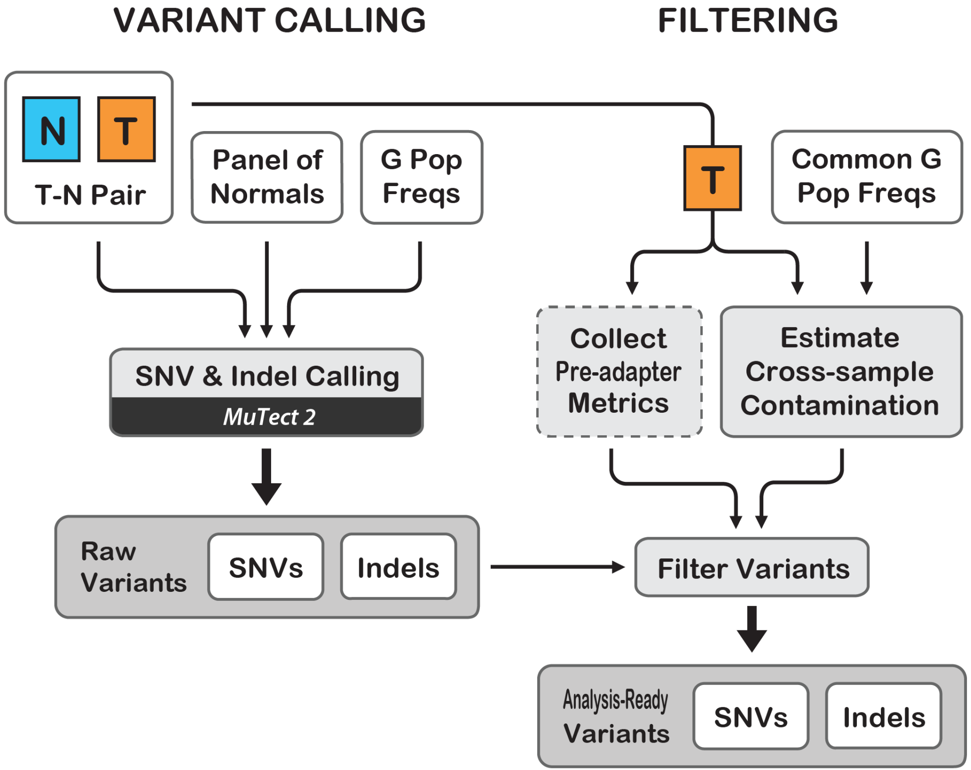
Output of CalculateContamination; FilterMutectCalls uses the first row listing BAM file-level contamination

Afterwards, optionally filter by orientation bias:

```
gatk FilterByOrientationBias \
    -V tumor_matched_m2_oncefiltered.vcf.gz \
    --artifactModes ‘G/T’ \
    -P tumor.preadapter_detail_metrics \
    -O tumor_oxog_twicefiltered.vcf.gz
```

Requires pre-adapter detailed metrics calculated by Picard CollectSequencingArtifactMetrics.

Somatic SNV & Indel discovery workflow


VARIANT CALLING
FILTERING
N
T
T-N Pair
Panel of Normals
G Pop Freqs
T
Common G Pop Freqs
SNV & Indel Calling
MuTect 2
Collect Pre-adapter Metrics
Estimate Cross-sample Contamination
Raw Variants
SNVs
Indels
Filter Variants
Analysis-Ready Variants
SNVs
Indels