Somatic SNV and Indel Discovery

New in GATK 3.5 – MuTect2 beta

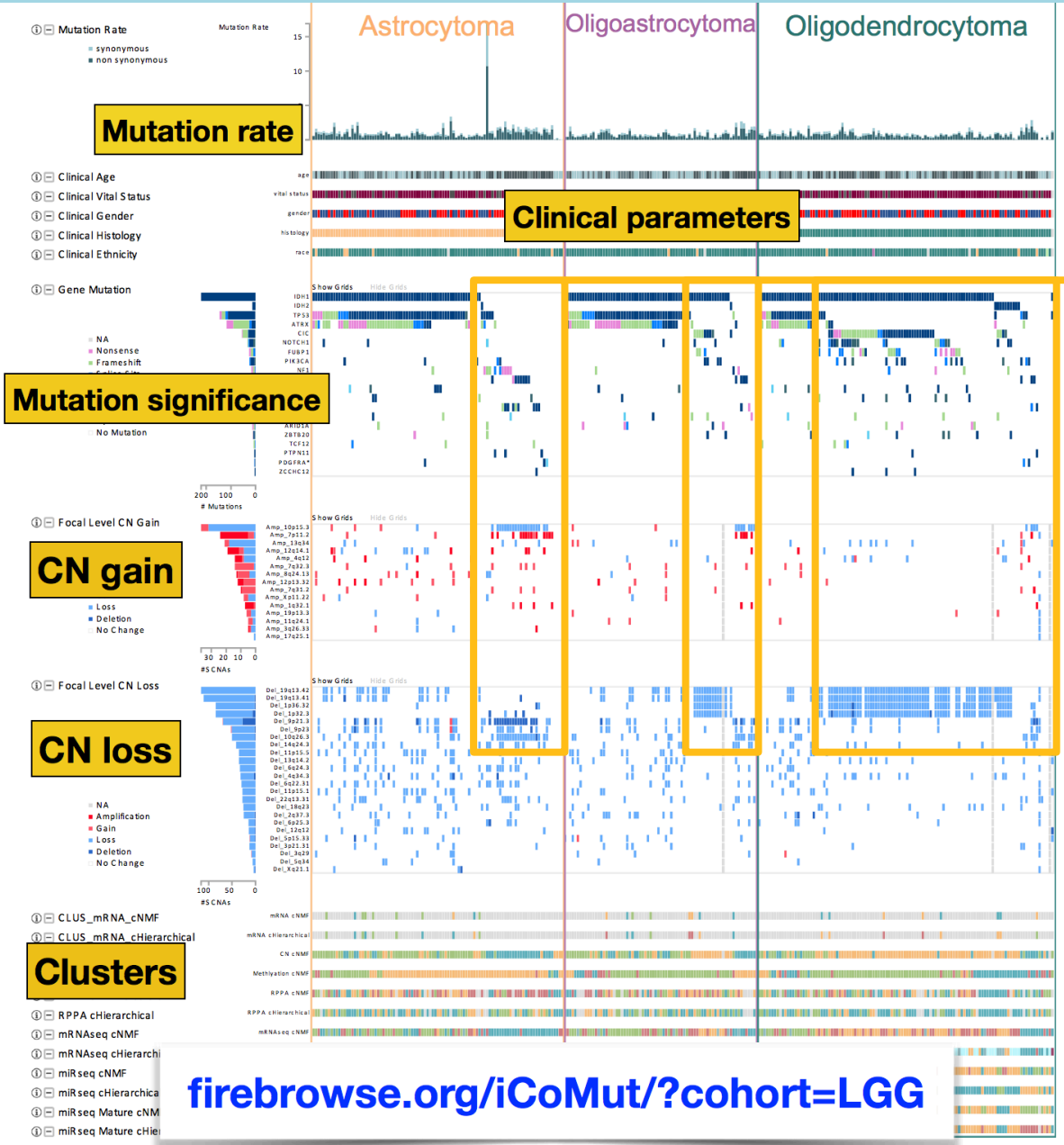
Somatic variant discovery with GATK

1. Cancer background refresher
2. Cross sample contamination estimation with ContEst
3. SNV and indel calling with MuTect2
4. [Copy number alteration calling with CNV] *next talk*

SOMATIC VARIANT DISCOVERY

A REFRESHER ON CANCER BACKGROUND

A cancer's genomic alterations are multilayered



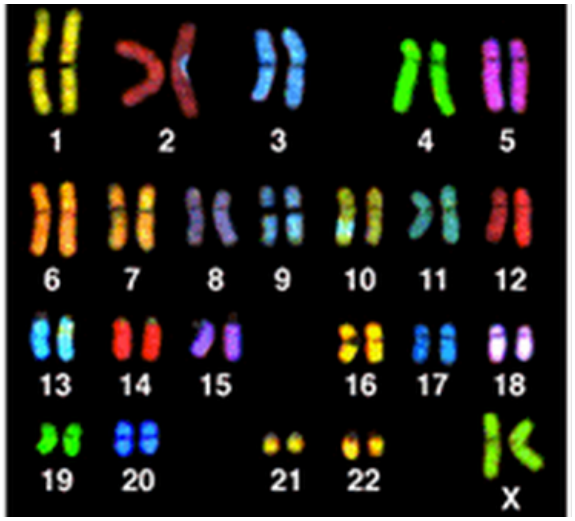
- iCoMut plot from Firebrowse for brain lower grade glioma
- Patients lacking characteristic mutations in IDH1/2, TP53 and ATRX have increased focal copy number alterations
- Cancer analysis must take into account a wide spectrum of alterations

Some ballpark metrics for cancer mutations

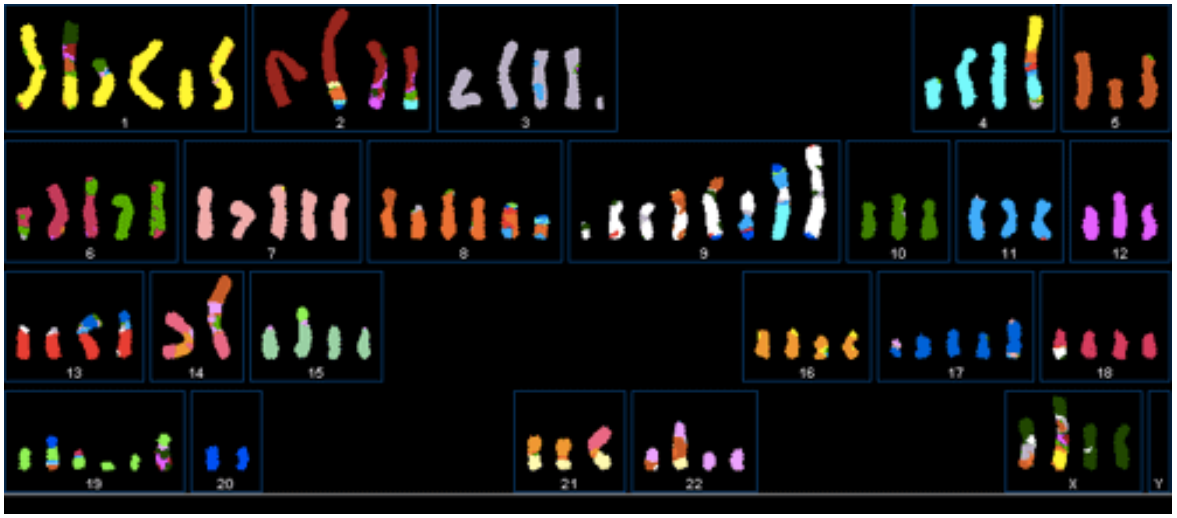
per genome (3 x 10^9 bp)	per Mbp	somatic alteration type	acronym
1000s to 10,000s	**1** (0.33 to >1000)	single nucleotide	SNV
100s to 1,000s	< 1	small insertions & deletions	Indel
100s to 1,000s	< 1	structural	SV
100s to 1,000s	< 1	copy number	CNA or CNV

Lawrence et al. Nature 2013

Normal cell



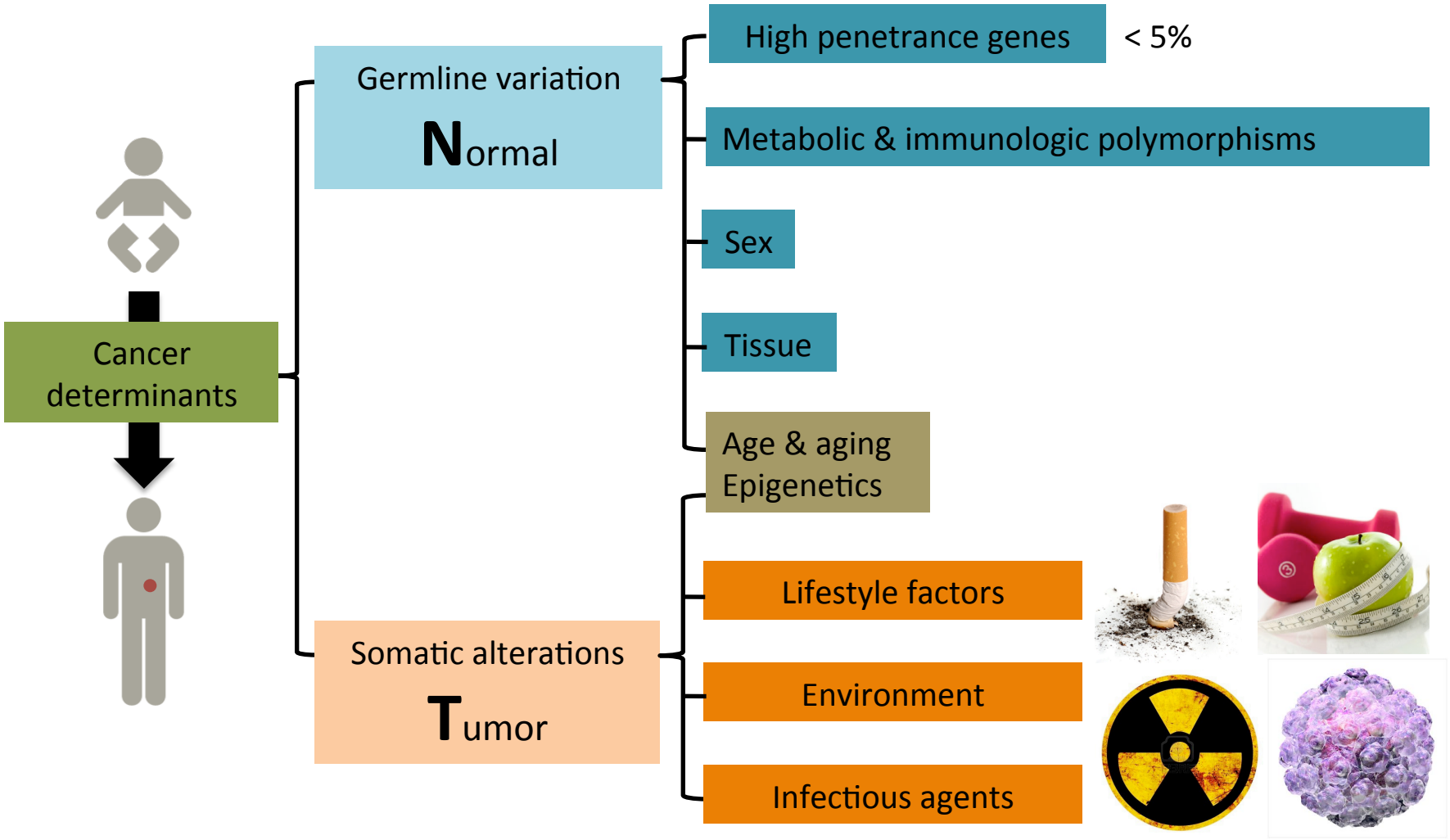
Breast cancer cell line HCC1954



Spectral karyotyping paints each chromosome pair with a color

http://old-www.path.cam.ac.uk/~pawefish/BreastCellLineDescriptions/HCC1954.html

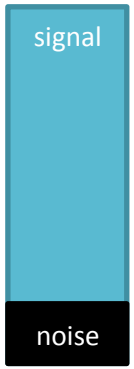
Cohorts of paired data to detect drivers of cancer



Having a cohort of matched sets empowers detection of driver mutations

Cancer-specific challenges confound analyses

N



T

sample prep artifacts sample swaps depth limitations

batch effects mapping artifacts

contamination sequencer artifacts sample swaps

X

A. Above plus

B. Purity: contamination of N/T, T/N

C. Tumor heterogeneity

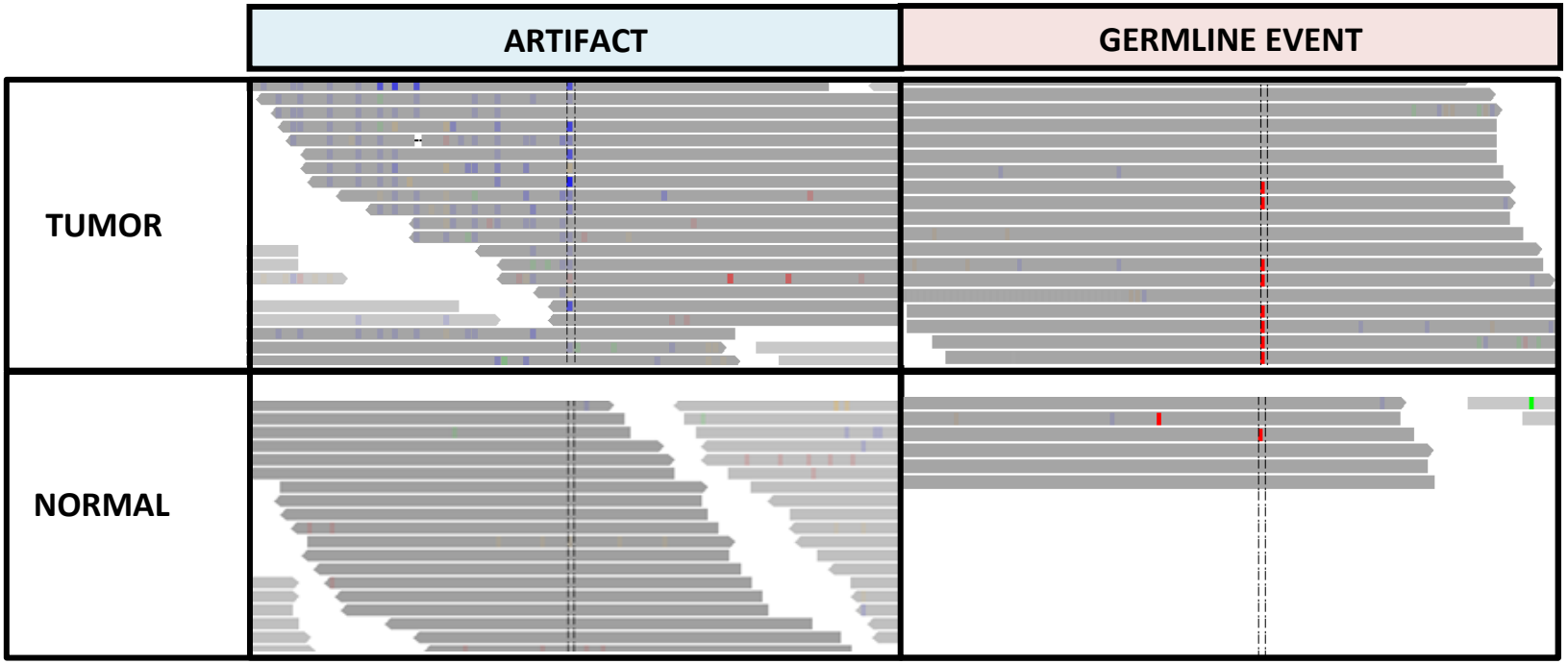
} **increases need for sensitivity**

D. Mutation spectrum: chromothripsis, aneuploidy, CNVs, SVs (>100bp), indels (<100bp), LOH (loss of heterozygosity), SNVs

→ **calls for varied tools**

False positives come from artifacts and germline variation

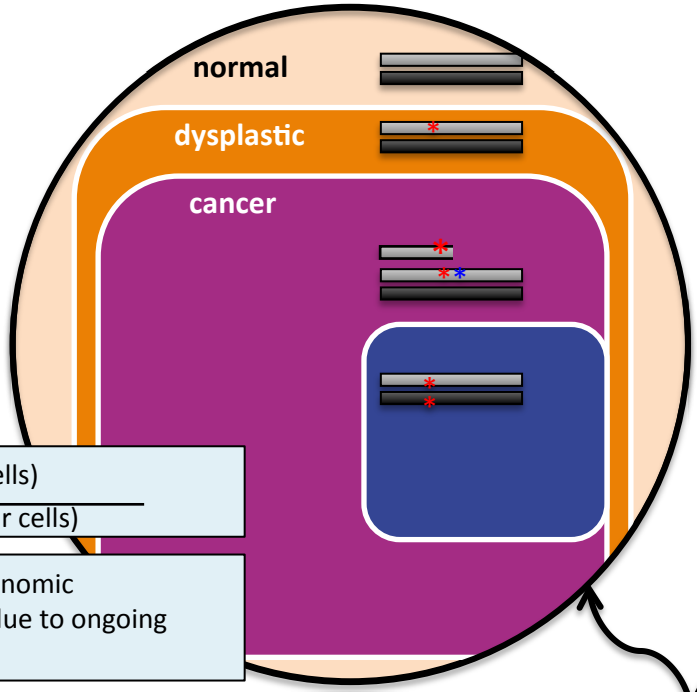
Somatic point mutations occur ~1 / Mbp

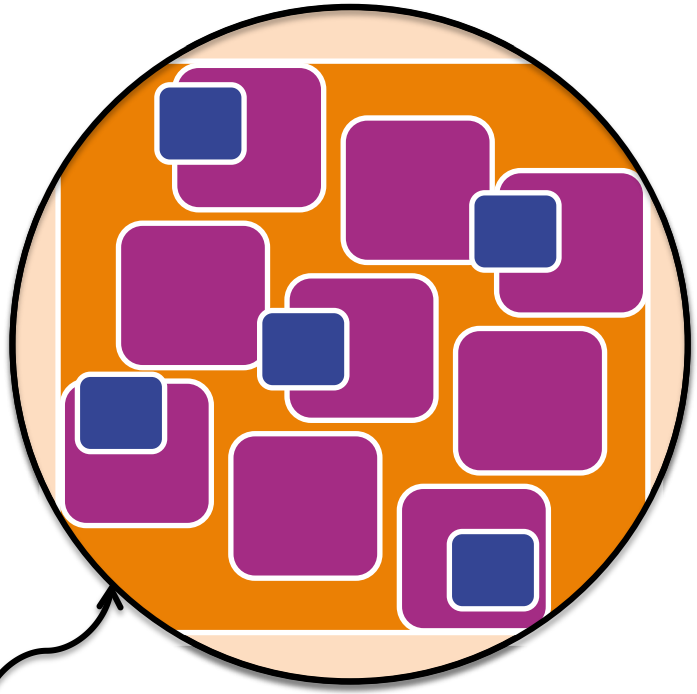


At risk	Every base	~1667 germline variants / Mbp
Source	• Misread bases • Sequencing artifacts • Misaligned reads	• Low coverage in NORMAL
Solutions	*filters, Panel of Normals (PoN)*	*dbSNP, ExAC, COSMIC, PoN*

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

Tumor samples are heterogeneous and often contain normal cells





Tumor **purity** = $\frac{\text{(tumor cells)}}{\text{(normal + tumor cells)}}$

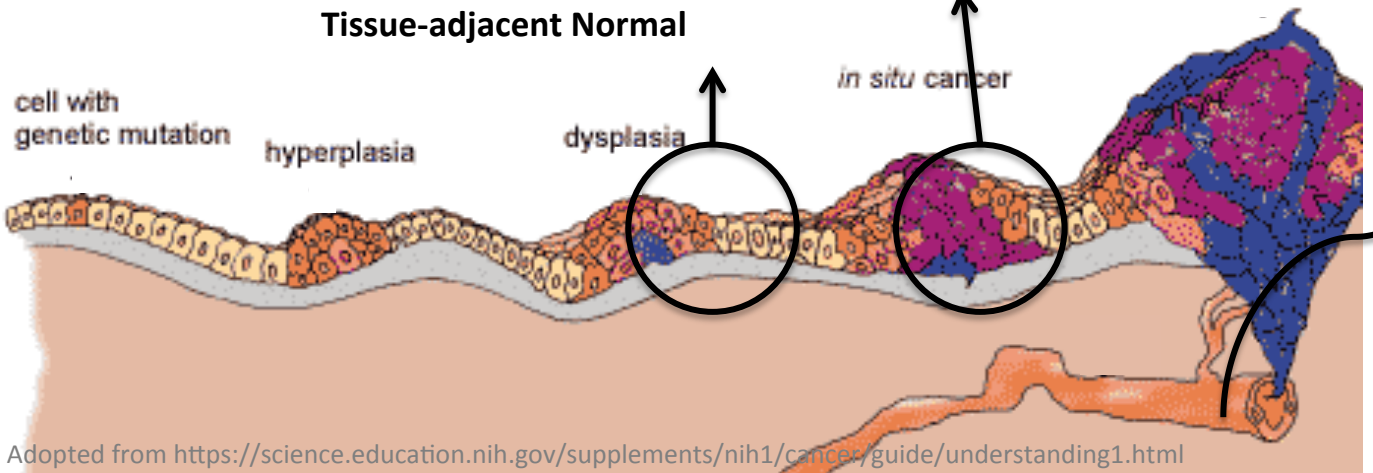
Tumor **heterogeneity** is based on polygenomic populations, segregated or intermixed, due to ongoing subclonal evolution.

Tumor Sample

invasive cancer

Tissue-adjacent Normal

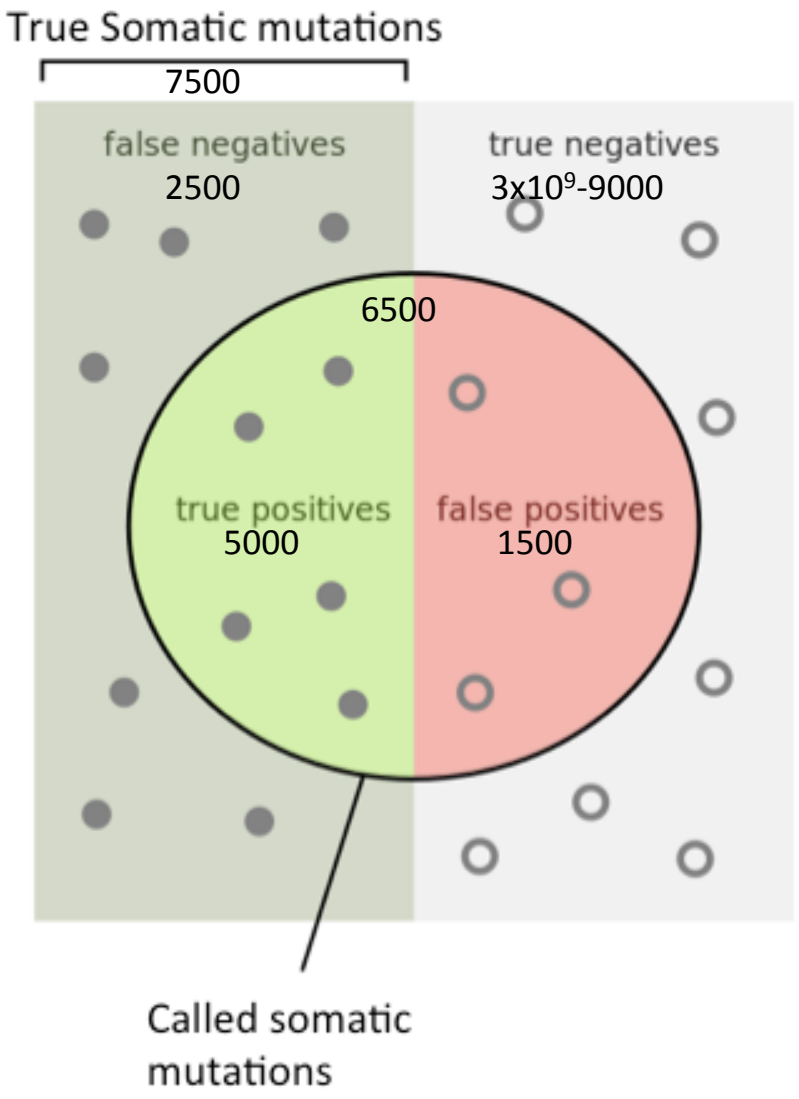
Blood Normal



Adopted from https://science.education.nih.gov/supplements/nih1/cancer/guide/understanding1.html

Multiple metrics measure callset quality

To detect somatic mutations, we need highly sensitive & specific mutation callers

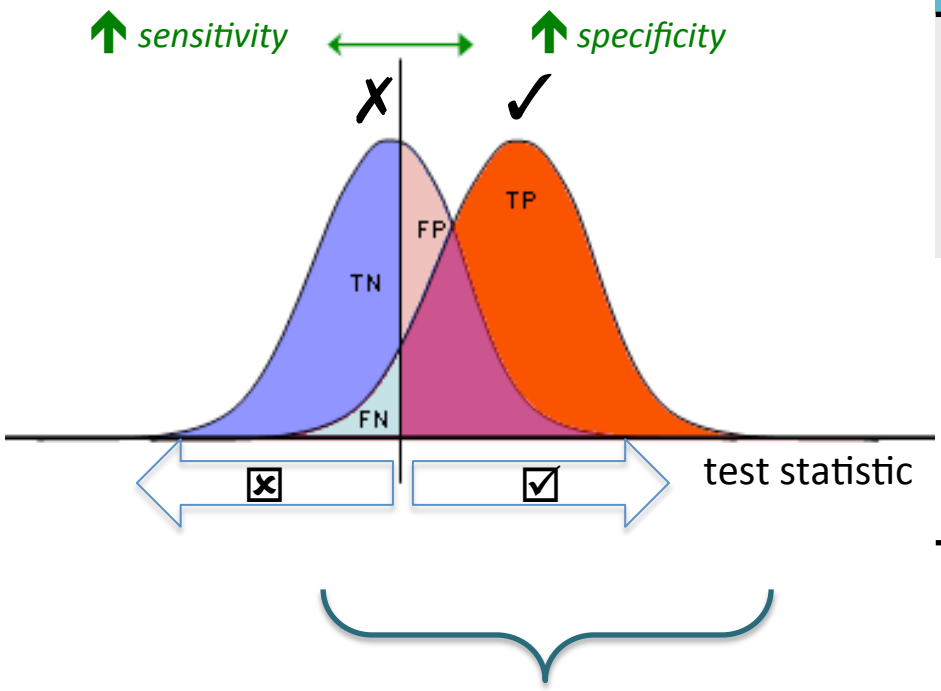


Metric	Definition	Example
Sensitivity	$\frac{TP}{FN+TP}$	$\frac{5000}{7500}$ = 67%
Specificity	$\frac{TN}{FP+TN}$	$\frac{3E9 - 9000}{3E9 - 7500}$ = 99.99995%
Precision	$\frac{TP}{TP+FP}$	$\frac{5000}{6500}$ = 77%
FPR	FP/[TN+FP]	0.5 FP mutations / Mbp
Balanced accuracy	[Sensitivity + Precision] / 2	[77% + 67%]/2 = 72%

Adapted from Wikipedia

Multiple metrics measure callset quality

To detect somatic mutations, we need highly sensitive & specific mutation callers

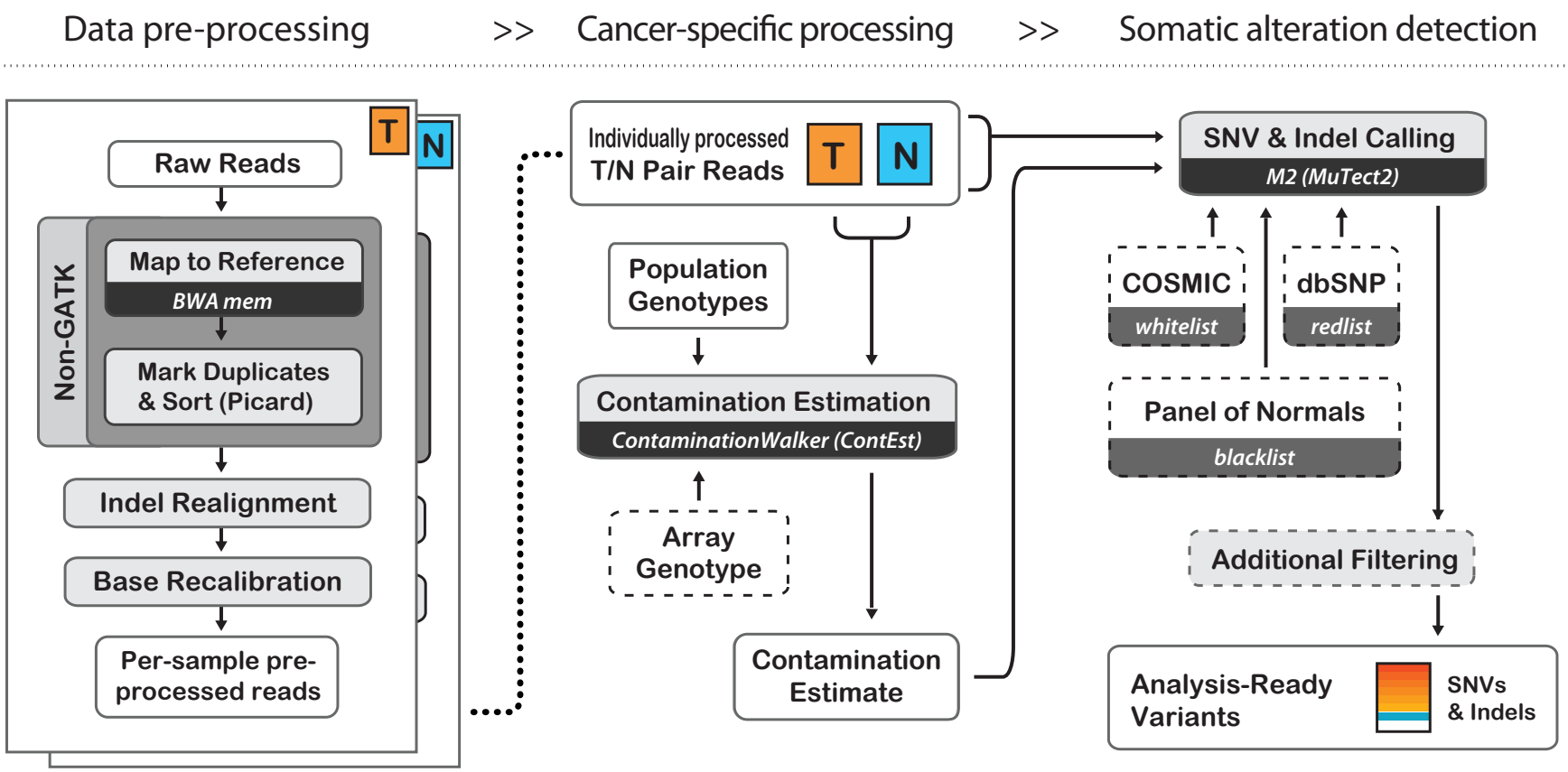


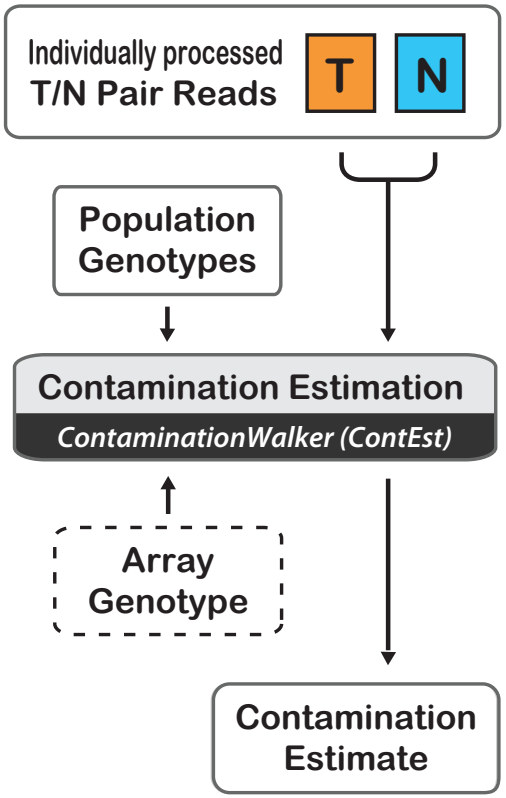
Metric	Definition	Impact
Sensitivity	$\frac{TP}{FN + TP}$	Greater sensitivity increases false positives
Specificity	$\frac{TN}{FP + TN}$	Greater specificity increases false negatives (decreases true positives)

Take these and apply filters to remove FPs

Adapted from Wikipedia <https://en.wikipedia.org/wiki/Receiver_operating_characteristic>

GATK workflow for somatic SNV and Indel detection

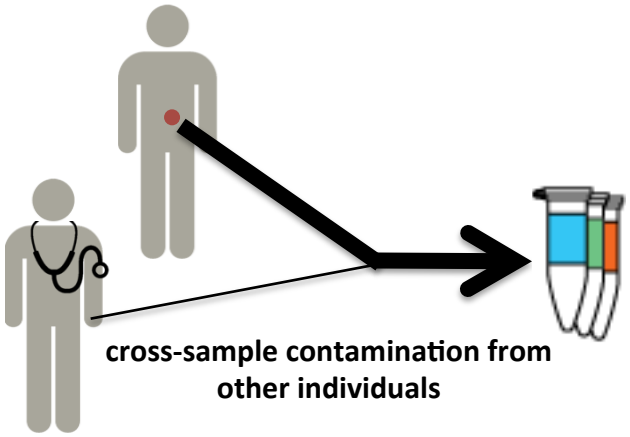




CONTAMINATION ESTIMATION

FOR CROSS-SAMPLE CONTAMINATION

Low levels of *cross-sample* contamination is common



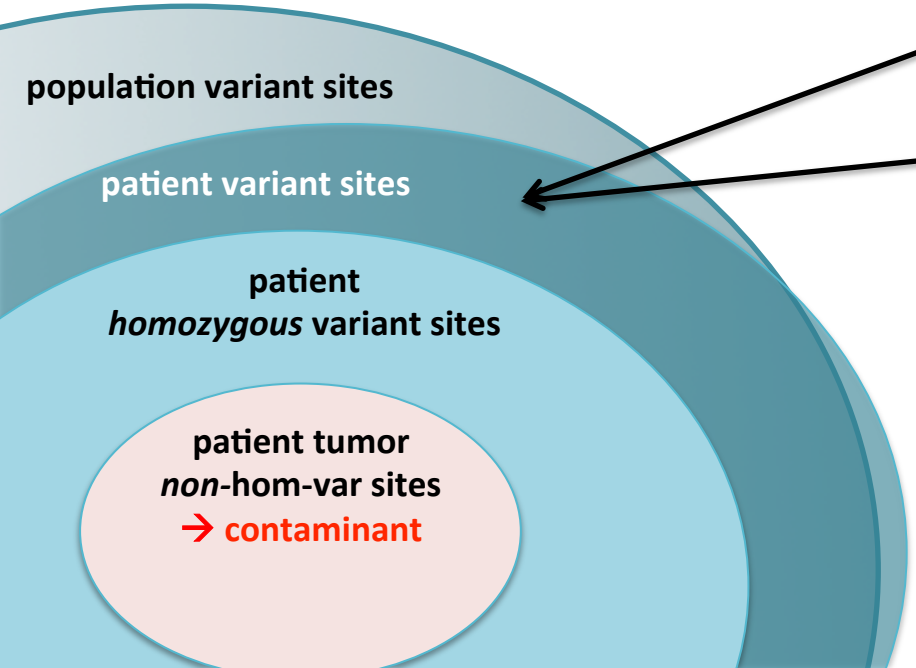
TCGA Ovarian cohort ~10% samples have > 1.5% contamination (N=206)	
Errors per sample	**~0.2 errors / Mbp**
Somatic point mutation rate = 1 / Mbp	
Fraction false positives	**~20%**

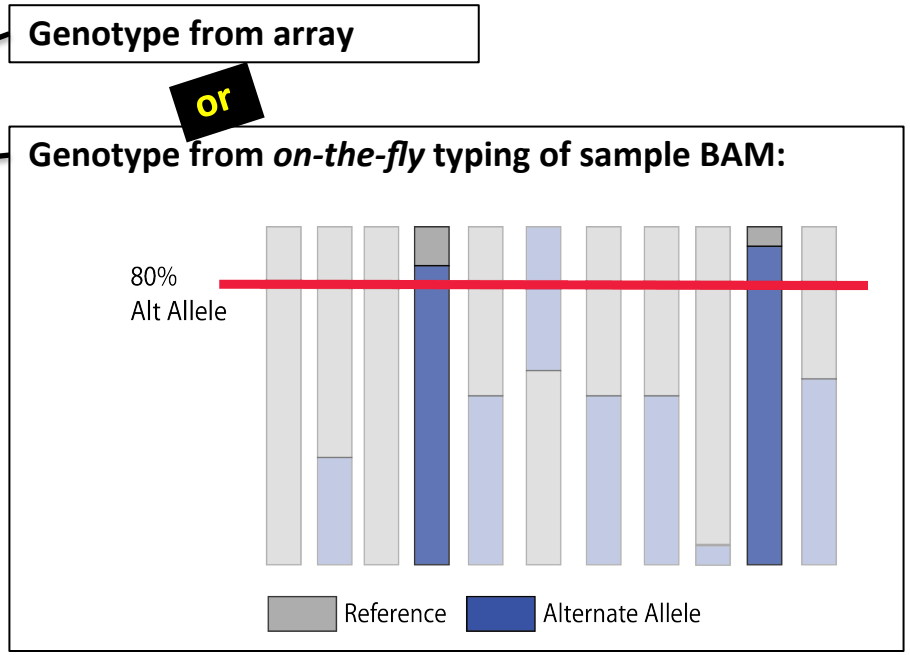
- ContEst performs *cross-sample* **cont**amination **est**imation
  - Not intended for stromal nor cross-species contamination
- Accurate estimates even with low average coverage (< 5x) for WES and WGS
- Determine % contamination of input bam by sample, by lane or in aggregate across all reads
- Output is text file of estimated percent contamination and 95% confidence interval

Cibulskis *et al. Bioinformatics 2011*

Contaminant sites vary from patient homozygous sites

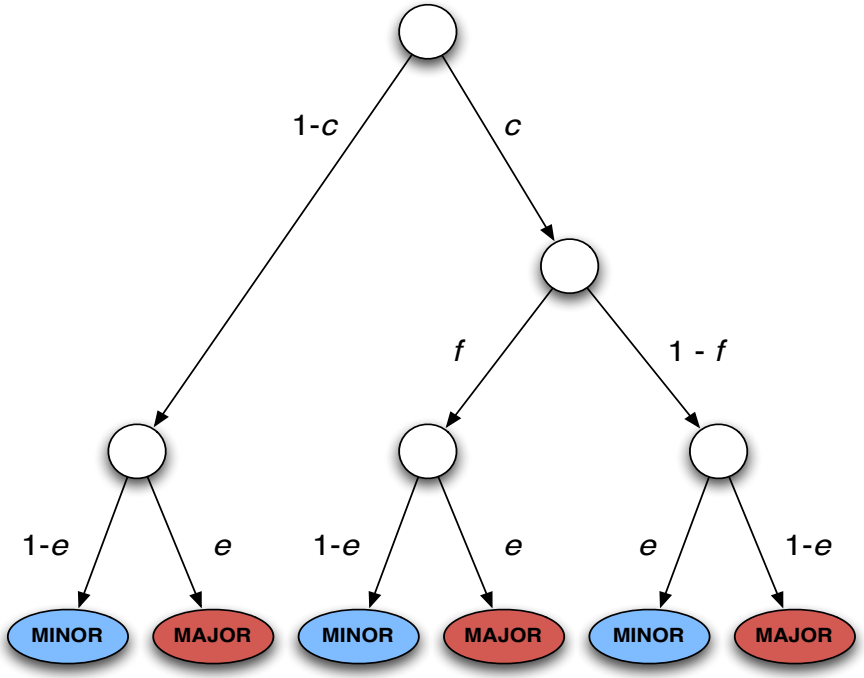
- Population file to pick a limited # of variant sites (uniform priors)
- Use subset homozygous variant in patient
  - Use genotype from matched normal over that from tumor as cancers have *loss of heterozygosity* (LOH) events
- Take lower bound of those not homozygous variant in tumor sample
  - Mismatching reads reflect part of true total contamination based on variation in population allele frequency
- Calculate probabilities of contamination from common germline SNPs





Underlying algorithm uses Bayesian approach

***c*: contamination**
***f*: minor allele frequency**
***e*: sequencing error rate**



P(MINOR | genotype) = (1-*c*)(1-*e*) + *cf*(1-*e*) + *c*(1-*f*)(*e*)

P(MAJOR | genotype) = (1-*c*)(*e*) + *cf*(*e*) + *c*(1-*f*)(1-*e*)

Bayesian approach to calculate the **posterior probability** of the contamination level and determine the **maximum a posteriori probability** (MAP) estimate of the contamination level

ContEst is GATK's ContaminationWalker

Example command to estimate cross-sample contamination:

```
java -jar GenomeAnalysisTK.jar \
    -T ContaminationWalker \
    -R reference.fasta \
    -I tumor.bam \
    --genotypes normalGenotypes.vcf \
    --popFile populationAlleleFrequencies.vcf \
    -L populationSites.interval_list
    [-L targets.interval_list] \
    -isr INTERSECTION \
    -o output.txt
```

genotype from VCF

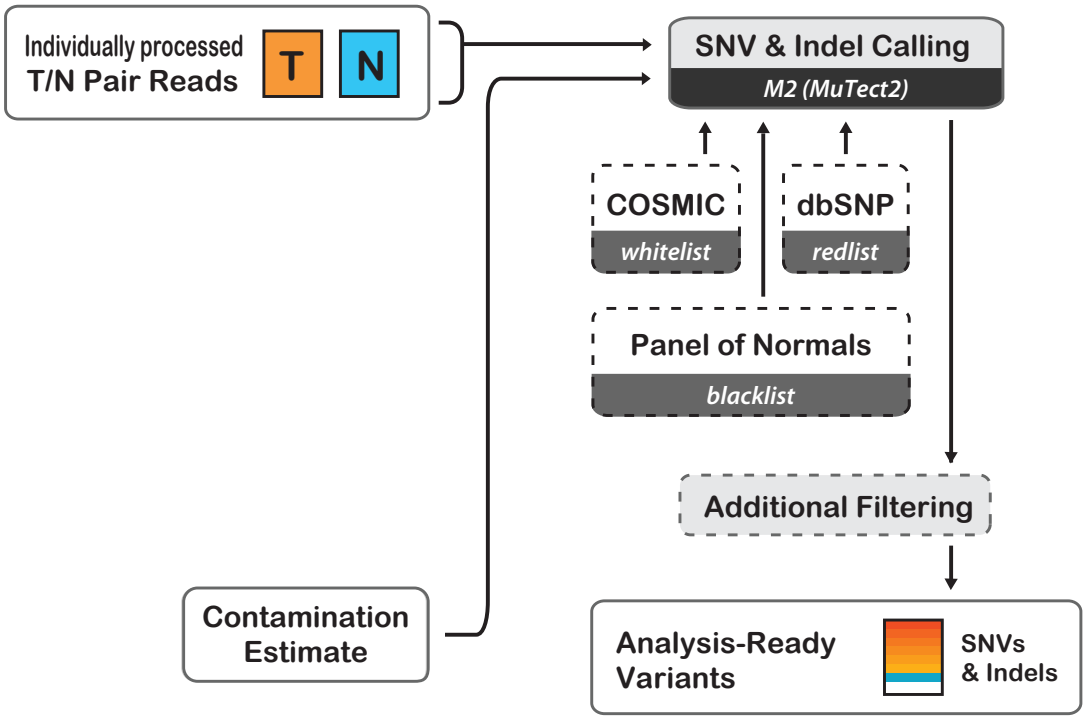
or

genotype on the fly

```
-I:genotype normal.bam \
```

Data interpretation guidelines:

Contamination	Label	Recommendation
0–2%	Fine	
2–5%	Slight contamination	Follow up if odd downstream results
5–15%	Moderate contamination	Salvageable; expect manual review
15–50%	Heavy contamination	Remove sample; follow up with project manager
>50%	Heavy contamination or sample swap	Consider sample swap as you approach 100%



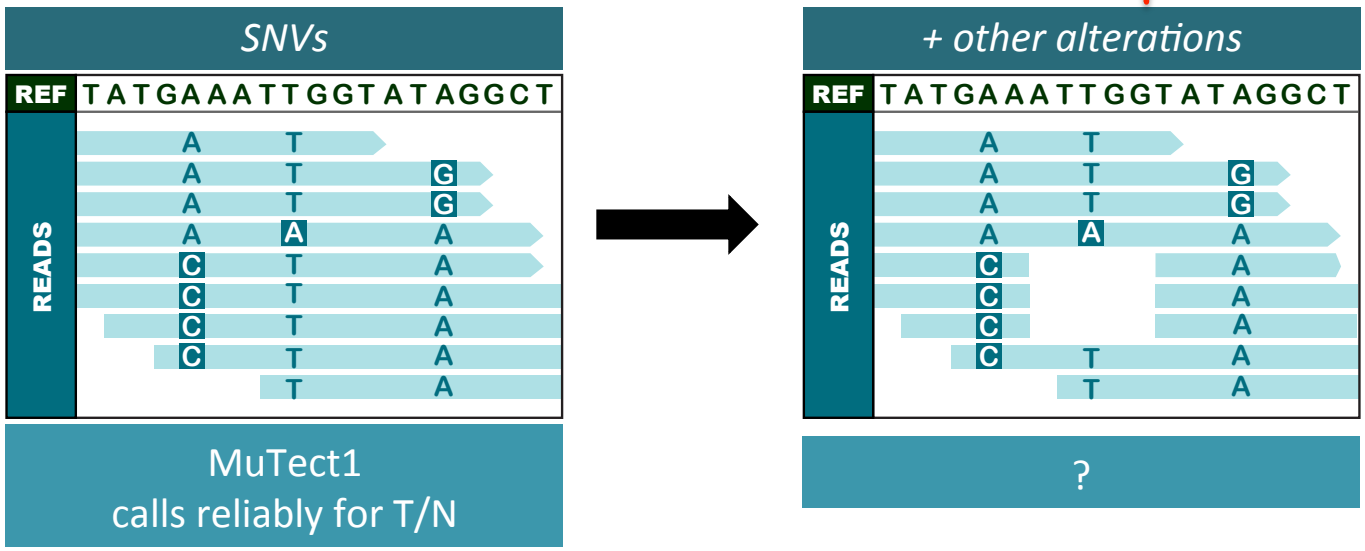
SNV AND INDEL DISCOVERY

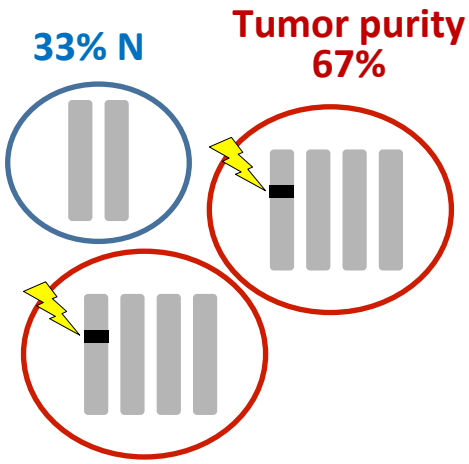
WITH MUTECT2 BETA

Tool is still in process of finalization

DO NOT USE IN PRODUCTION

Challenges of somatic calling: Indels, CNVs & SVs

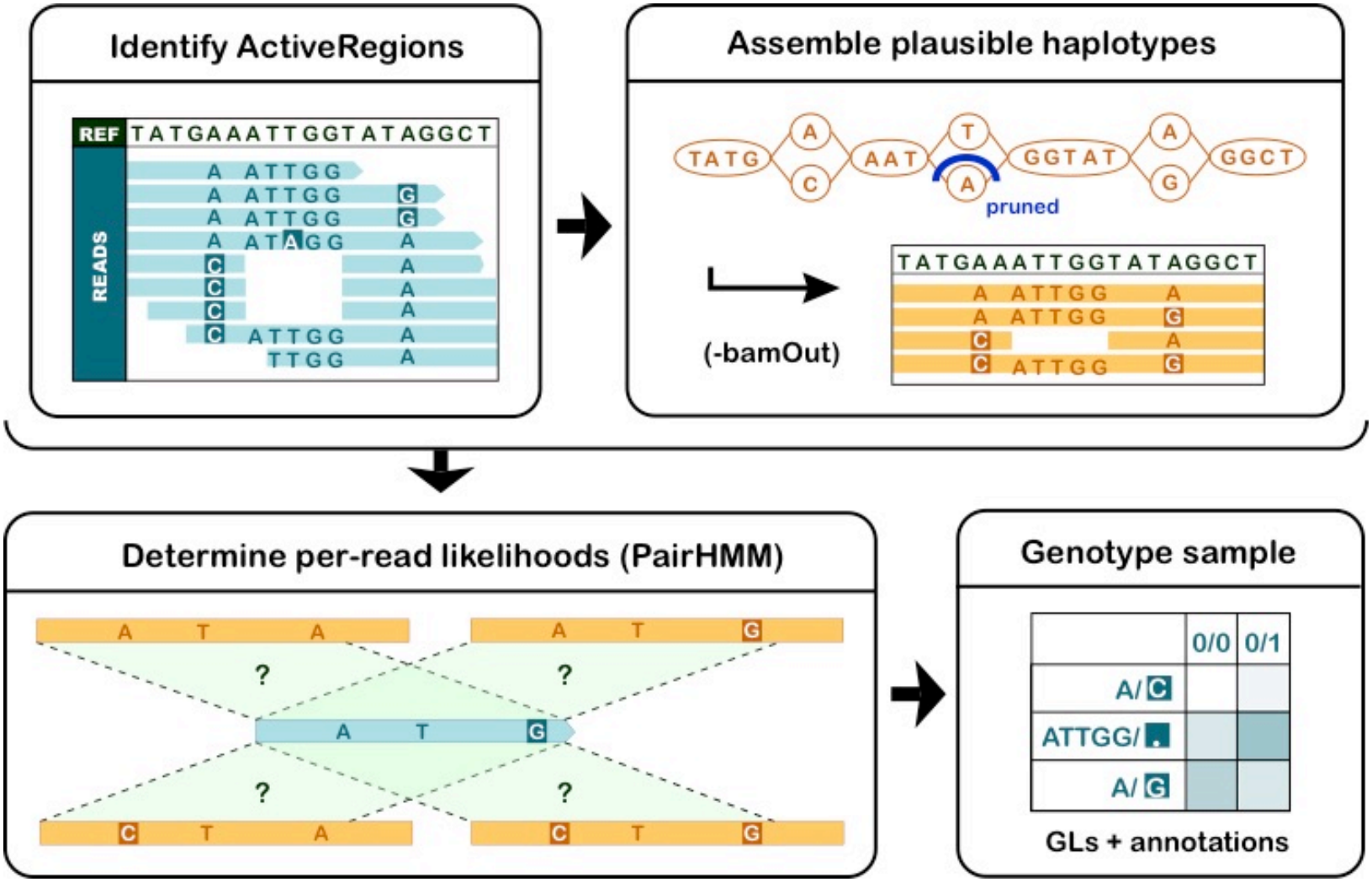




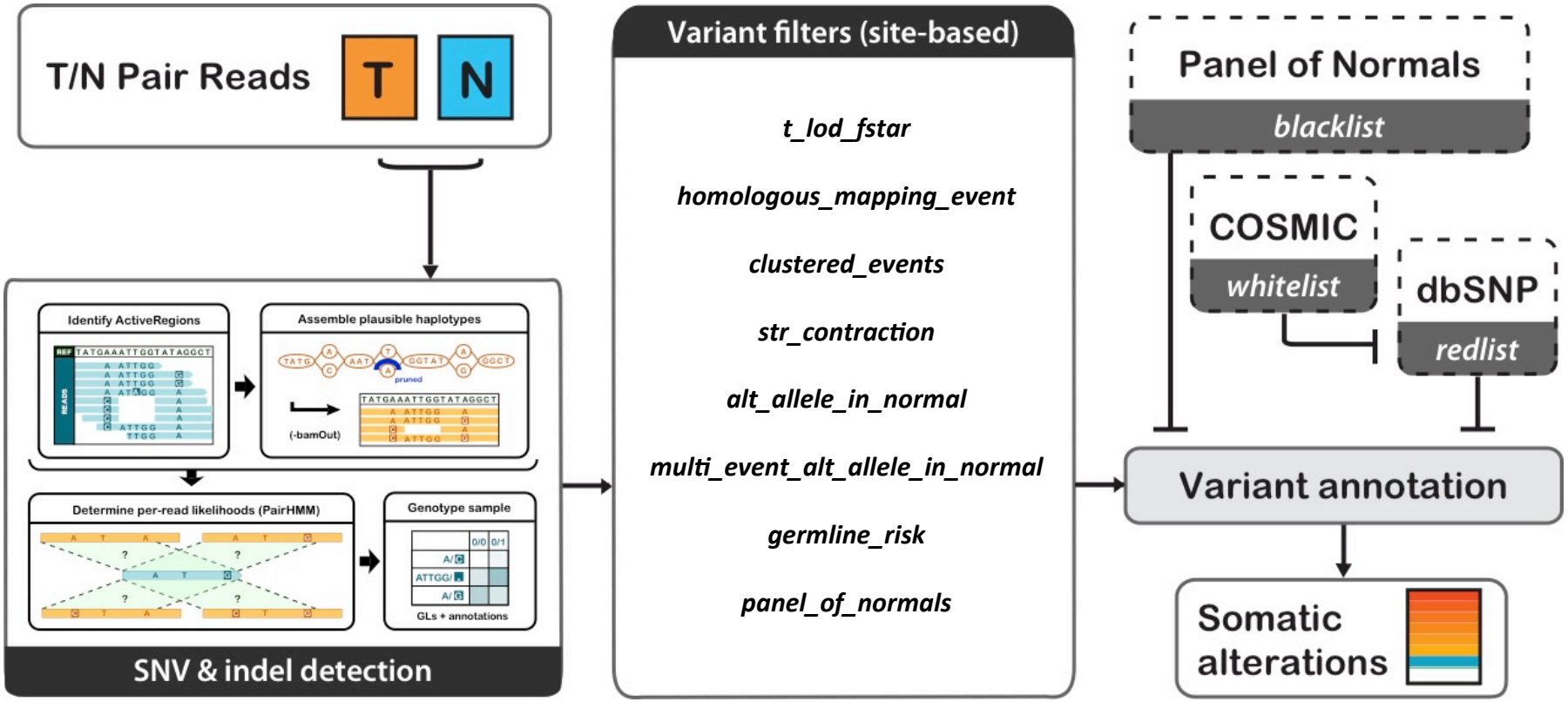
- Hard to call even in germline
- Rare events compared to SNPs
- Compounded by allele fraction
  - Tumor local ploidy = 4
  - Tumor # of mutated copies per cell = 1
  - ➔ Allelic fraction **AF**= 2/10 = 0.2
- We have a tool that calls indels

HaplotypeCaller

MuTect was hybridized with HaplotypeCaller to call somatic SNVs & indels

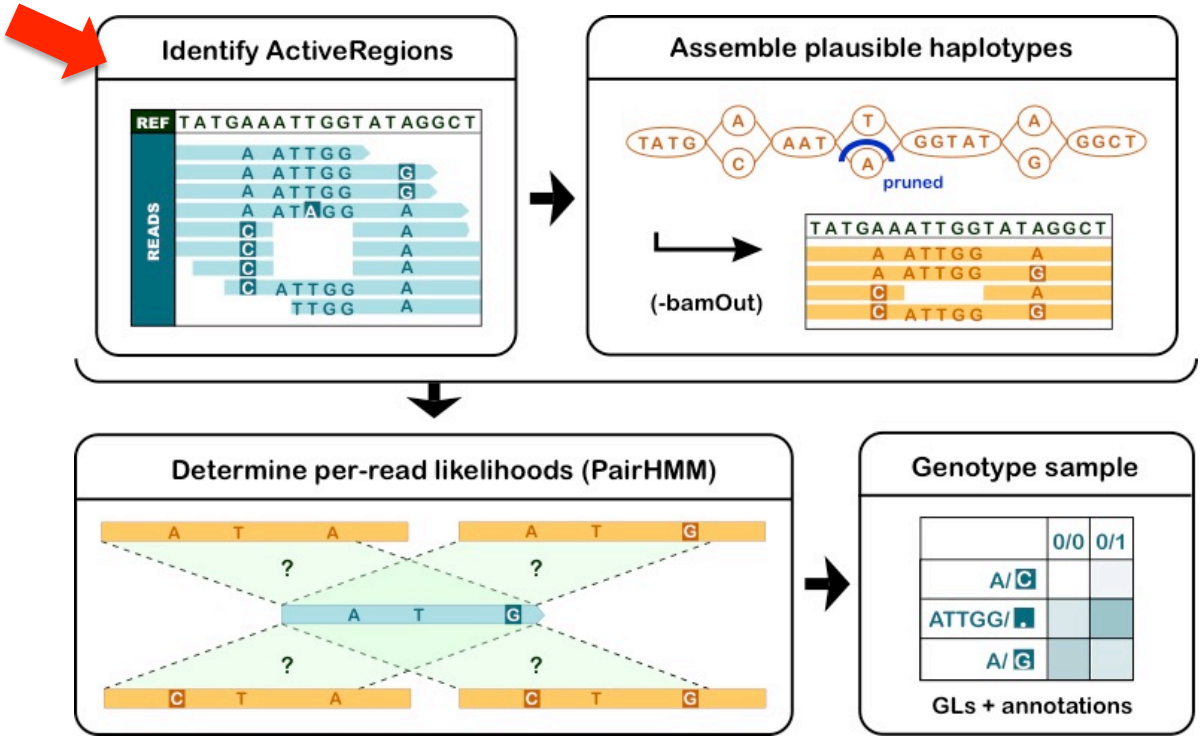


MuTect2 applies filters as VCF annotations



- HaplotypeCaller's reassembly method detects both SNVs and indels
- Built-in filters eliminate false positives, increasing specificity
- Optional filters screen for artifacts and germline events
- VCF annotations include new metrics

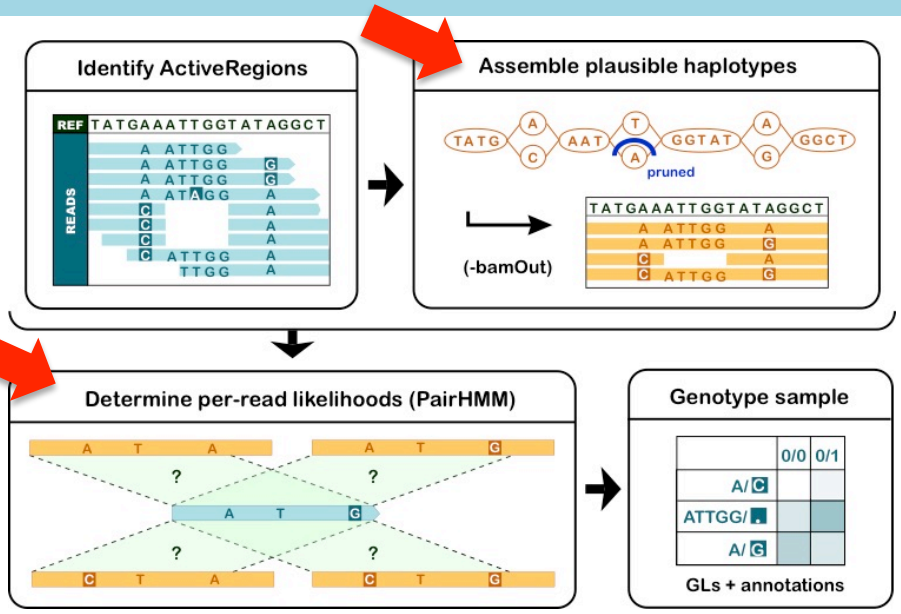
MuTect2 identifies **Active Regions**

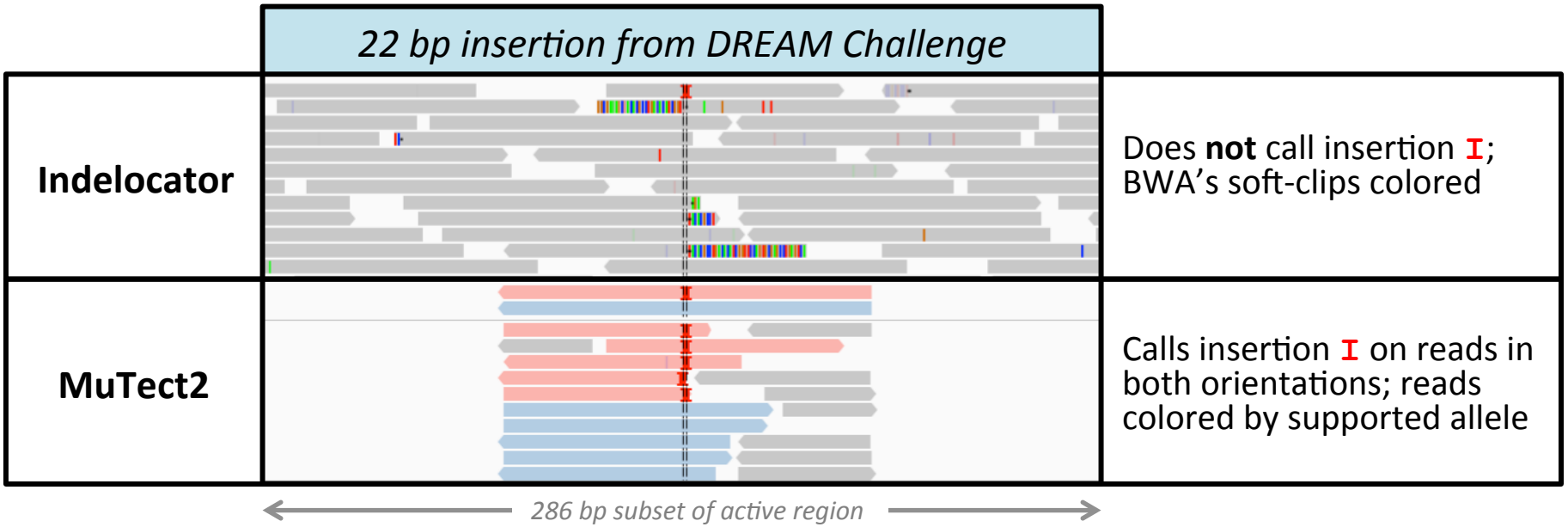


- Modified statistic threshold for somatic scenario uses LOD ≥ 4.0 in favor of the reference model
- Reads differentially filtered:
  - **Tumor is strict**: MAPQ ≥ Q20; discard discrepant overlapping reads
  - **Normal is permissive**: MAPQ ≥ Q0; keep alternate read from discrepant overlapping reads

MuTect2 **assembles haplotypes** and runs **PairHMM**

- Reassembles reads into possible haplotypes using only high quality reads
- Minor technical change lowers tolerance for discarding reads
  - For somatic events that are rare and at low allele fractions

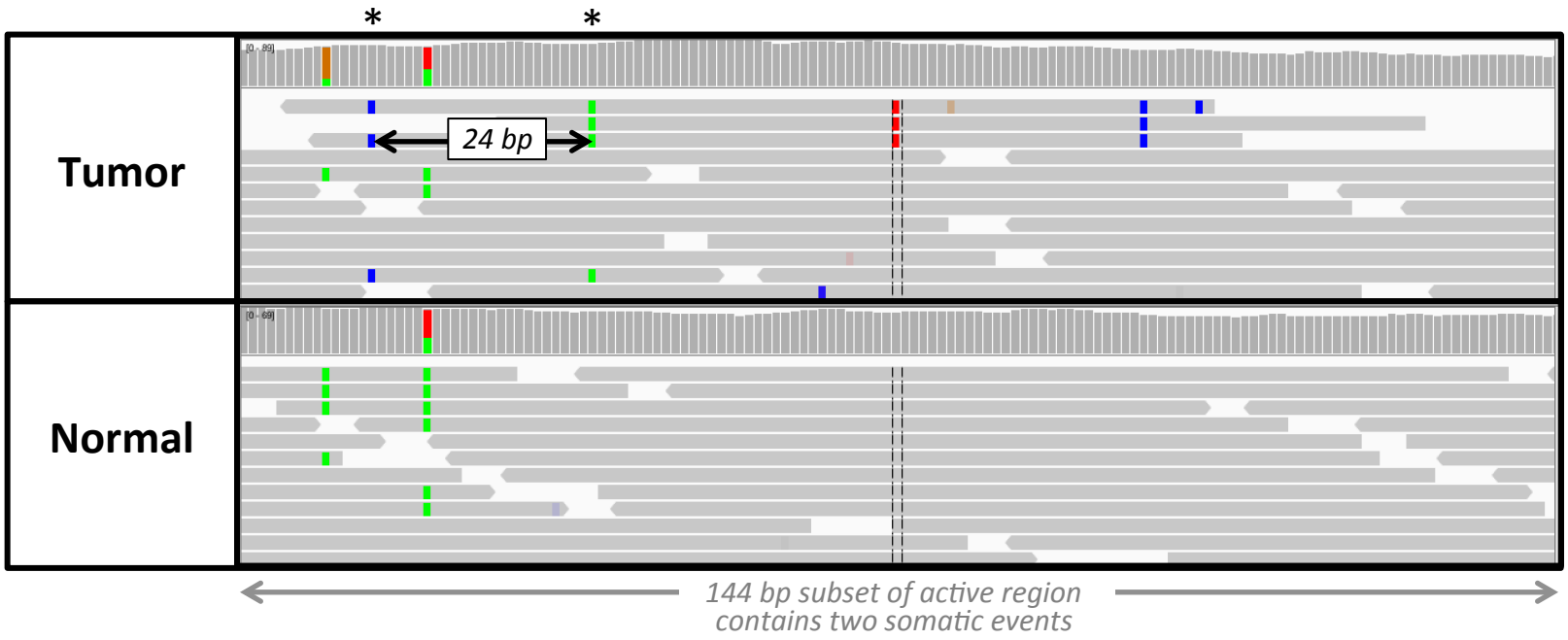




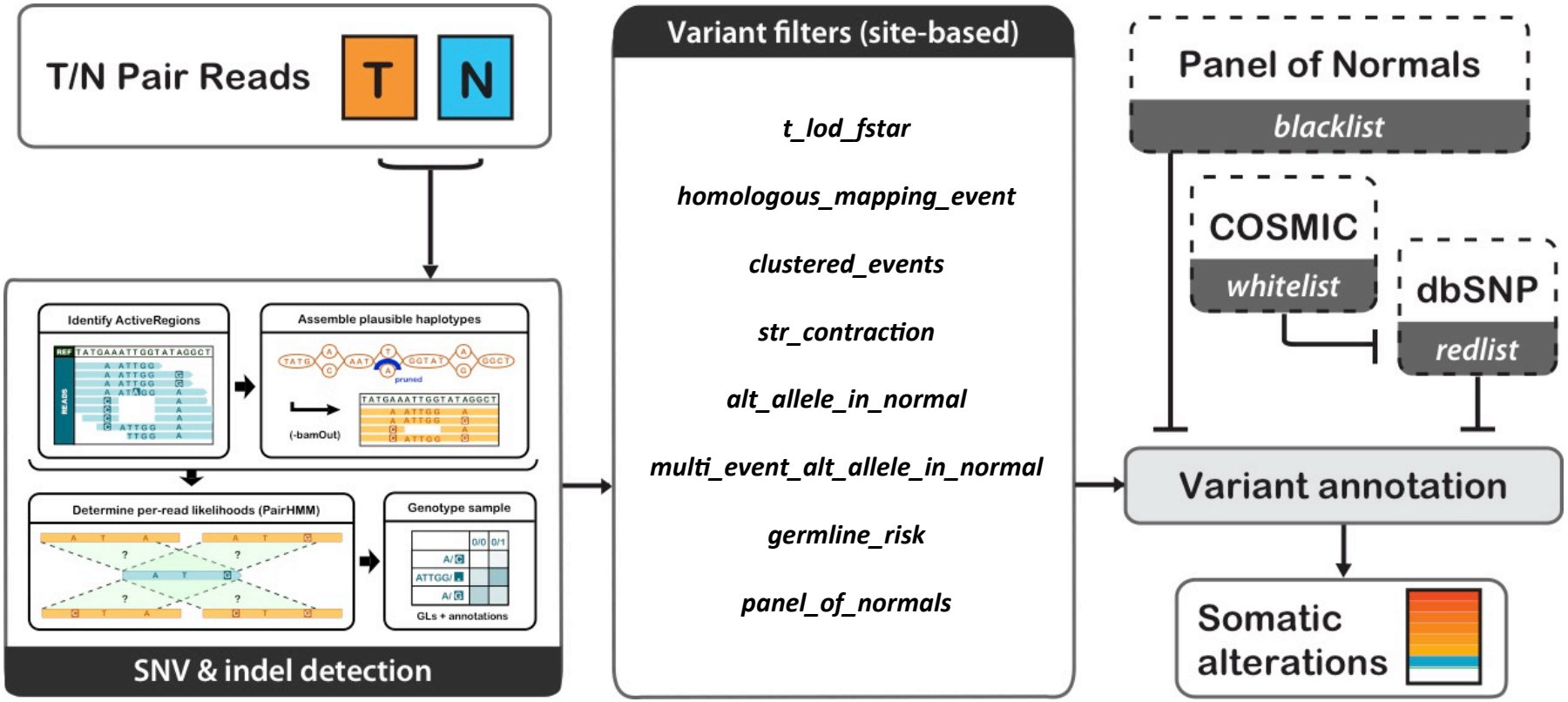
MuTect2 **calls genotypes** using M1 logic + PairHMM likelihoods

- Uses PairHMM likelihoods (like HC) instead of base quality scores (like M1)
- Also produces new metrics for each active region:

	New metric	Description	Example region
ECNT	Event Count	# of events in the haplotype	6
MIN_ED	Min Event Distance	Minimum distance between events	24 bp*
MAX_ED	Max Event Distance	Maximum distance between events	131 bp



MuTect2 applies filters as VCF annotations



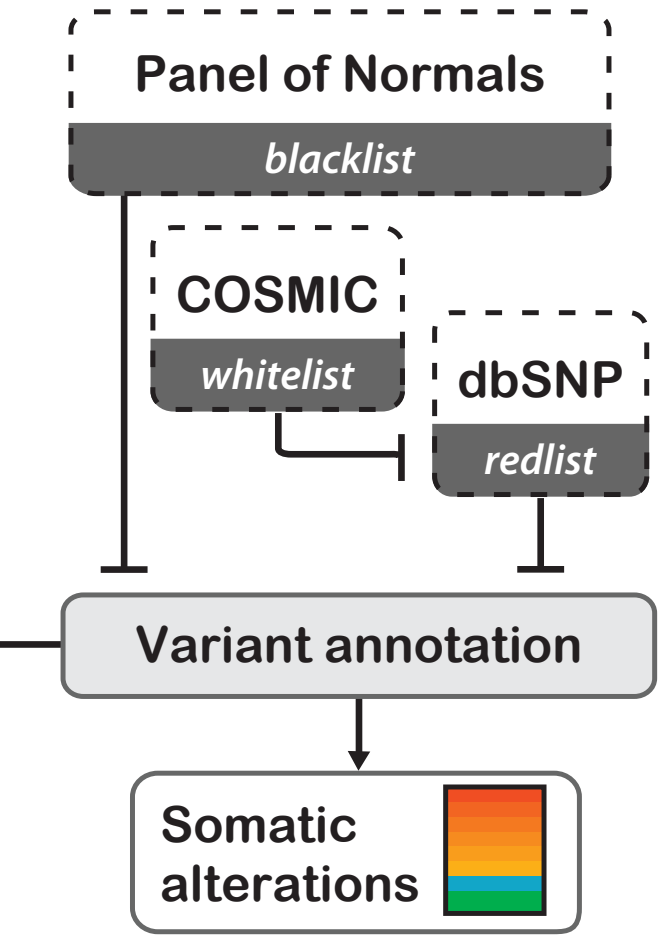
MuTect2's built-in filters mark false positives, increase specificity

Downstream hard-filters allow upstream sensitivity

Mutect2 filter field	Description of filter	Evidence source
t_lod_fstar	Tumor does not meet likelihood threshold	PairHMM likelihoods
homologous_mapping_event	More than 3 events were observed in the tumor	Tumor
clustered_events new	Clustered events (≥2) observed in the tumor (with distance between ≥3)	Tumor
str_contraction new	Site filtered due to contraction of short tandem repeat region.	Tumor
alt_allele_in_normal	Evidence seen in the normal sample	Normal
multi_event_alt_allele_in_normal	Multiple events observed in tumor and normal	Normal
germline_risk	Evidence indicates this site is germline, not somatic	Normal, dbSNP/COSMIC
panel_of_normals	Seen in at least 2 samples in the panel of normals	PoN

Optional filters screen for noise & germline events

M2 annotates filtered calls found in PoN or dbSNP



<u>If site is in PoN (≥ 2 samples), reject</u>

- VCF FILTER field: panel_of_normals
- VCF INFO field: PON = #, where # is the count from the PoN

<u>If a variant is in dbSNP but not in COSMIC, reject</u>

- VCF FILTER field: germline_risk
- Flag in VCF INFO field: DB
- M2 increases threshold for evidence of variant from the normal sample to prove not germline

MuTect2 is in GATK v3.5

Example command for T/N variant calling:

```
java -jar GenomeAnalysisTK.jar \
    -T MuTect2 \
    -R reference.fasta \
    -I:tumor tumor.bam \
    -I:normal normal.bam \
    [--dbsnp dbSNP.vcf] \
    [--cosmic COSMIC.vcf] \
    [-L targets.interval_list] \
    -o output.vcf
```

For a recalibrated BAM:

```
    -bamout bamout.bam
```

To specify a PON:

```
    -PON m2_pon.vcf
```

To specify cross-patient contamination file:

```
    -contaminationFile contamination.txt
```

Tab-separated file containing fraction of contamination in sequencing data (per sample) to remove. Format should be "<SampleID><TAB><Contamination>" per line.

Create your own *panel of normals* (PoN) in 2 steps

1. Create a sample-level PoN per sample using HaplotypeCaller:

```
java -jar GenomeAnalysisTK.jar \
    -T HaplotypeCaller \
    -R reference.fasta \
    -I:tumor normal1.bam \
    [--dbsnp dbSNP.vcf] \
    [--cosmic COSMIC.vcf] \
    --artifact_detection_mode \
    [-L targets.interval_list] \
    -o output.normal1.vcf
```

unusual nomenclature

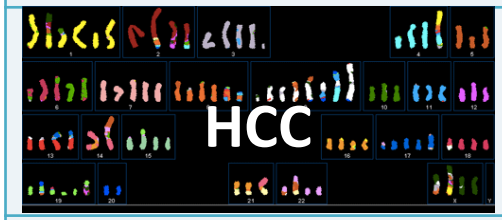
special mode

2. Combined PoN retains variants called in at least two samples:

```
java -jar GenomeAnalysisTK.jar \
    -T CombineVariants \
    -R reference.fasta \
    -V output.normal1.vcf -V output.normal2.vcf [-V ...] \
    -minN 2 \
    --setKey "null" \
    --filteredAreUncalled \
    --filteredrecordsmergetype KEEP_IF_ANY_UNFILTERED \
    [-L targets.interval_list] \
    -o M2_PON.vcf
```

event in at least two samples

Benchmark MuTect2 against various data & tools

Project	Data	Truth	Variant	Count (other tools)
ICGC-TCGA DREAM	**DREAM Synthetic Challenges 4 & 5** WG pre-aligned data of mixed purity & heterogeneity made with Bamsurgeon	known	SNV	24,218
			Indel	22,221
	Consensus data of 5 sequencing runs (2 Agilent, 1 ICE, 2 WGS) for breast cancer cell lines **HCC1143** & **HCC1954**	(partially) unknown	SNV	37,532
			Indel	594
The Cancer Genome Atlas	Large callsets (no indels) of real cancer samples from **TCGA** GBM, LUAD, STAD, & THCA cohorts	unknown	SNV	289,306
NA12878	WG sample used widely as the gold standard for germline tools	(mostly) known	SNV	36,587
			Indel	2,801

MuTect2 performance goals:

- As good as *MuTect1* (gold standard) on SNVs
- Better than *Indelocator* on indels

DREAM: MuTect1 ranked highly in SNV calling

- Broad's M1 workflow ranked highly in all 4 DREAM synthetic challenges
- Each challenge varies in purity and cancer cell fraction (CCF)
- 3 of 4 other top teams used M1 in workflow

simulation	Rank	Sensitivity	Precision	False Positive Rate (mutations per Mb)	Balanced Accuracy
100% purity	1st	0.967	0.984	0.021	0.975
80% purity	1st	0.961	0.992	0.010	0.977
100% purity 50%,33%,20% CCF	2nd	0.918	0.981	0.038	0.949
80% purity 50%,35%, CCF	1st	0.741	0.983	0.051	0.862

MuTect1 excels at accurately detecting low allele fraction mutations, hence uniquely suited for studying impure and heterogeneous tumors

HCC & TCGA: MuTect2 calls SNVs as well as M1

Differences in sensitivity rates range from –3.96 to +1.73%

MuTect1

Project	Dataset	%FN/TP (FN count)	False positives	True positives	Sensitivity	Precision	Balanced Accuracy
HCC	HCC1143	1 (154)	33	18690	0.99	0.998	0.995
	HCC1954	6 (1004)	49	17151	0.94	0.997	0.971
TCGA	GBM	24 (4943)	---	20849	0.81	---	---
	LUAD	4 (2949)	---	76989	0.96	---	---
	STAD	3 (2166)	---	62063	0.97	---	---
	THCA	22 (840)	---	3827	0.82	---	---

MuTect2

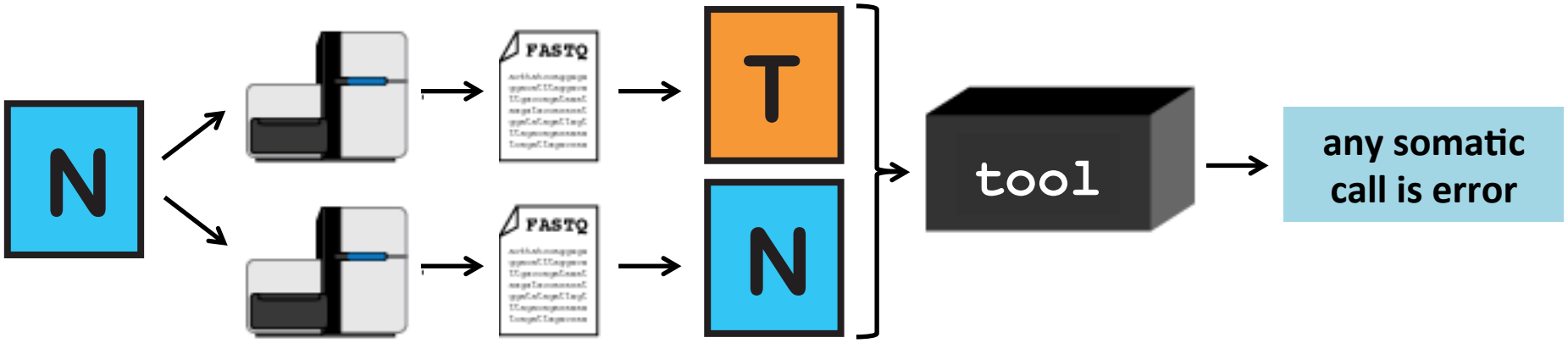
Project	Dataset	%FN/TP (FN count)	False positives	True positives	Sensitivity	Precision	Balanced Accuracy
HCC	HCC1143	4 (655)	11	18189	0.97	0.999	0.982
	HCC1954	8 (1319)	15	16836	0.93	0.999	0.963
TCGA	GBM	29 (5768)	---	20024	0.78	---	---
	LUAD	7 (5329)	---	74609	0.93	---	---
	STAD	6 (3700)	---	60529	0.94	---	---
	THCA	20 (774)	---	3893	0.83	---	---
					avg Δ –2%	**+0.16%**	**–1.03%**

MuTect2 specificity is similar to M1 for SNVs

False Positive Rate (errors/Mbp) @ ~0.24 (M1) & ~0.25 (M2)

	SNVs errors/Mbp	Indels errors/Mbp
MuTect1	0.238 ± 0.002	N/A
MuTect2	0.247 ± 0.001	0.071± 0.002

To measure specificity, need to know the truth. We approximate:



12 N:N combinations from 4 clinical sequencing (CRSP) lab runs of NA12878

DREAM: M2 calls SNVs as well as M1, Indels better than Indelocator

DREAM_Synthetic_4: BWA MEM, 80% purity, 50 & 35% subclones

Mode	ID	Submission Name	Team	# of Calls	Sensitivity	Precision	Balanced Accuracy *
SNV	2559951	MuTect-RSp2	Broad SMC	12272	0.741	0.983	0.862
	2549704	MuTect-RSp	Broad SMC	12282	0.741	0.982	0.862
	2528605	MuTect-R	Broad SMC	12288	0.741	0.981	0.861
	2510567	**MuTectStock**	Broad SMC	12879	0.754	0.952	**0.853**
	2550741	**EmTooS**	Broad SMC	14972	0.740	0.804	**0.772**
Indel	2559911	novoBreak_indel	Ken_Chen_Lab	12059	0.788	0.928	0.858
	2556244	EmTooNVDBS	Broad SMC	11907	0.772	0.921	0.846
	2559878	novoBreak_indel	Ken_Chen_Lab	11341	0.751	0.940	0.845
	2550739	**EmTooS**	Broad SMC	12168	0.776	0.906	**0.841**
	2550981	EmTooS	Broad SMC	12127	0.774	0.906	0.840
	2551191	EmTooSH2	Broad SMC	12068	0.772	0.908	0.840
	2519991	IndelStock	Broad SMC	4805	0.186	0.549	0.368

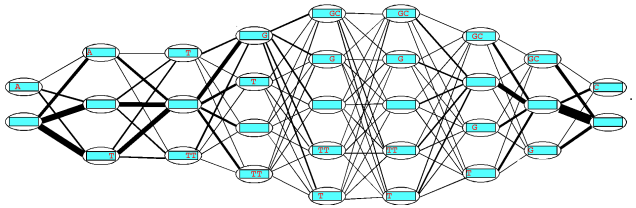
DREAM_Synthetic_5_leaderboard: Novoalign v3.02.05, 80% purity, 50% subclone

Mode	ID	Submission Name	Team	# of Calls	TPs	FNs	Sensitivity	Precision	Balanced Accuracy *
SNV	4869621	CASpoint.BIG-Cancerevo	Cancerevo	45356	45269	114	0.998	0.998	0.998
	4868623	EMILY_0.0001	BinaTeam	45374	45228	155	0.997	0.997	0.997
	3154033	ADA*0.3*iFilter_M0.35	BinaTeam	42707	42705	2678	0.941	1.000	0.971
	3160884	**M2-stock**	Broad SMC	42337	42324	3059	0.933	1.000	**0.966**
	3160968	**MuTect-stock**	Broad SMC	42106	42100	3283	0.928	1.000	**0.964**
	2924453	mutect*default*	SMCAdmins	41820	41810	3573	0.921	1.000	0.961
Indel	4868651	ETHAN_0.8	BinaTeam	15915	11576	4852	0.705	0.727	0.716
	4771022	ADRIAN_0.5	BinaTeam	16008	11606	4822	0.707	0.725	0.716
	3160888	**M2-stock**	Broad SMC	14999	11005	5423	0.670	0.734	**0.702**
	812846	LoFreq Somatic indel: v2.1.2 default	LoFreq Somatic - GIS	10436	8426	8002	0.513	0.807	0.660
	3160993	Indl-stock	Broad SMC	5853	0	16428	0.000	0.000	0.000

**Sorted on balanced accuracy then rounded. Selected submissions shown with #1 submission at top (pink).*

ICGC-DREAM Challenge website: <https://www.synapse.org/#!Synapse:syn312572/wiki/58893>

Cheat sheet of MuTect1 vs. MuTect2 features

	MuTect1	MuTect2
Names	Originally *μTector* (2009)	GATK v3.5 jar's *MuTect2*
Reference	*Cibulskis et al. Nature Biotech*, 2013	β-release
Detects	SNV only	SNVs and **indels**
Core statistical model	Locus-based and reliant on original alignment. Bayesian classifier, for both detection & classification of SNV $P(b_i \mid e_i, r, m, f) = \begin{cases} f^{e_i}/3 + (1-f)(1-e_i) & \text{if } b_i = r \\ f(1-e_i) + (1-f)^{e_i}/3 & \text{if } b_i = m \\ {}^{e_i}/3 & \text{otherwise} \end{cases}$	**HaplotypeCaller's local reassembly engine** in active region detects SNVs and indels and uses **PairHMM per-read likelihoods** 
Genotyping	By evaluating likelihood based on base quality scores of candidate SNVs	M1's extended to handle indels
Filtering	Pragmatic, hard-filters (6) to control FP rate: *Proximal gap, poor mapping, triallelic site, observed in control, strand bias, clustered position*	No VQSR *yet*. Hard-filters: *Alt_allele_in_normal, multi_event_alt_allele_in_normal, germline_risk, clustered_events, homologous_mapping_event, panel_of_normals, str_contraction, t_lod_fstar*
Performance	The benchmark for somatic SNV calling	• As good as *MuTect1* on SNVs • Better than *Indelocator* on indels
Formats	Callstats file	**VCF** with **new annotations**

Further reading

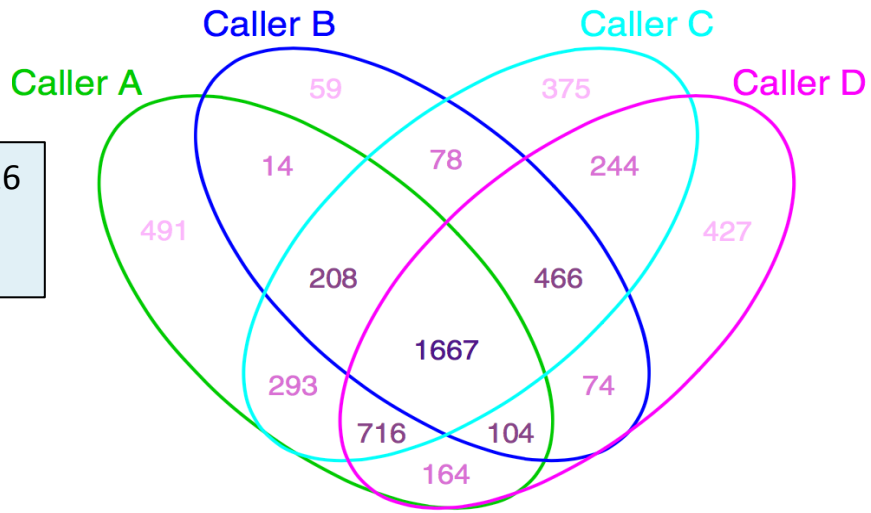
http://www.broadinstitute.org/gatk/best-practices

New documentation will be posted for MuTect2 and the somatic workflow in the near future

Bonus slides follow

Before examining M2's performance, review benchmarking

Counts of overlapping mutations detected by four callers in 16 TCGA lung cancer T:N exome-seq pairs. Only 31.0% (1,667/5,380) of SNVs were identified by all four callers.



Benchmarking itself is challenging:

a. Variant-calling algorithms are highly parameterized
b. Sequencing error profiles can vary between & within centers
c. Validation data from independent tech or higher-depth sequencing undergo routines prone to same errors, e.g. alignment artifacts
d. Most research has focused on coding aberrations, restricting validation to <2% of genome

Field moves forward by sharing and iteration

Kim & Speed, BMC Bioinformatics 2013. "Comparing somatic mutation-callers: beyond Venn diagrams"
Ewing et al, Nature 2015. "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection"

Benchmarking approaches are self-limiting

Cannot model full diversity of nonrandom sequencing errors

Unknown ground truth

Experimental validation expensive & impractical for general application. *E.g.* To confirm clonal events:

- study several metastases from same patient
- ultra-deep sequencing
- sequence very small numbers of single cells

E.g. Virtual tumors method

- We know all true mutations with certainty
- Simulate somatic mutations at controlled allelic fractions
- *Limitations*: simulates somatic mutations using germline events, which differ from somatic mutations in nucleotide substitution frequencies and context

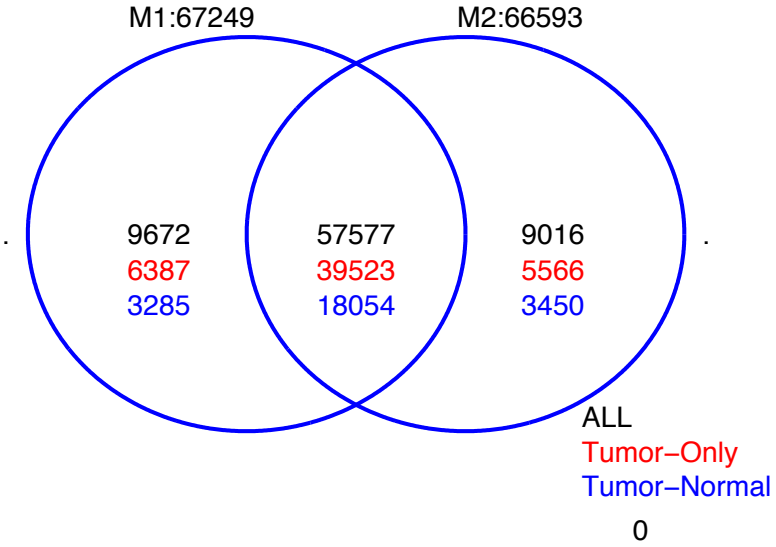
E.g. Downsampling

- Subset reads of validated somatic mutations to measure sensitivity of caller on known mutations at various *simulated* depths of coverage
- *Limitations*: larger error for sensitivity estimate, only previously validated allele fractions considered, excludes mutations not originally detected thereby overestimating true sensitivity, specificity cannot be measured

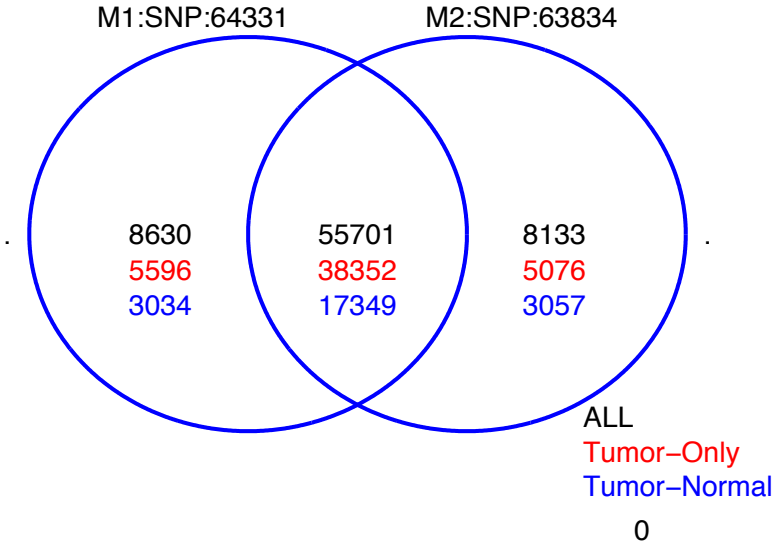
Conclusion: use multiple benchmarking approaches

Cibulskis et al. Nature Biotech, 2013

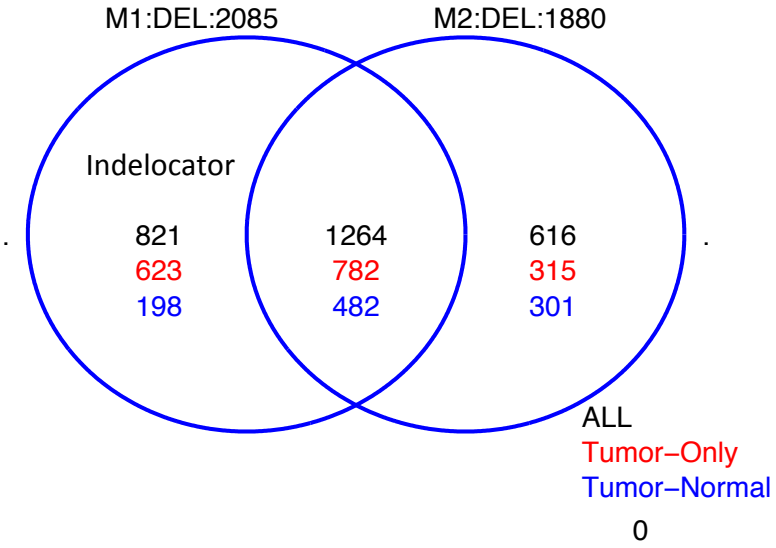
DLBCL Venns
(PoN + alt 2 indelocator filter)



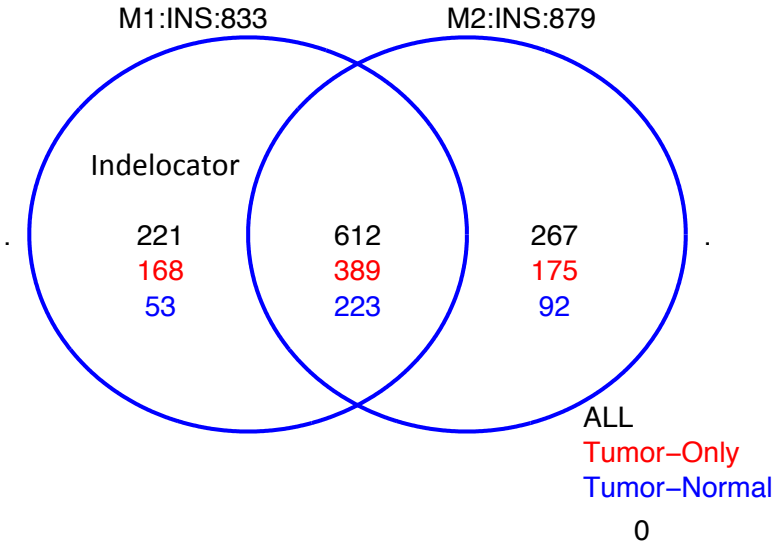
0



0



0



0

M2 VCF output

Header:

a. File format specified
b. Filters described
c. FORMAT field tags described
d. INFO field items described
e. Input sample files
f. Contigs listed per row
g. Reference fasta

```
1   ##fileformat=VCFv4.1
2   ##FILTER=<ID=alt_allele_in_normal,Description="Evidence seen in the normal sample">
3   ##FILTER=<ID=clustered_events,Description="Clustered events observed in the tumor">
4   ##FILTER=<ID=germline_risk,Description="Evidence indicates this site is germline, not somatic">
5   ##FILTER=<ID=homologous_mapping_event,Description="More than three events were observed in the tumor">
6   ##FILTER=<ID=multi_event_alt_allele_in_normal,Description="Multiple events observed in tumor and normal">
7   ##FILTER=<ID=panel_of_normals,Description="Seen in at least 2 samples in the panel of normals">
8   ##FILTER=<ID=str_contraction,Description="Site filtered due to contraction of short tandem repeat region">
9   ##FILTER=<ID=t_lod_fstar,Description="Tumor does not meet likelihood threshold">
10  ##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
11  ##FORMAT=<ID=AF,Number=1,Type=Float,Description="Allele fraction of the event in the tumor">
12  ##FORMAT=<ID=ALT_F1R2,Number=1,Type=Integer,Description="Count of reads in F1R2 pair orientation supporting the
13  ##FORMAT=<ID=ALT_F2R1,Number=1,Type=Integer,Description="Count of reads in F2R1 pair orientation supporting the
14  ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad ma
15  ##FORMAT=<ID=FOXOG,Number=1,Type=Float,Description="Fraction of alt reads indicating OxoG error">
16  ##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
17  ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
18  ##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the al
19  ##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within
20  ##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as define
21  ##FORMAT=<ID=QSS,Number=A,Type=Integer,Description="Sum of base quality scores for each allele">
22  ##FORMAT=<ID=REF_F1R2,Number=1,Type=Integer,Description="Count of reads in F1R2 pair orientation supporting the
23  ##FORMAT=<ID=REF_F2R1,Number=1,Type=Integer,Description="Count of reads in F2R1 pair orientation supporting the
24  ##GATKCommandLine.M2=<ID=M2,Version=3.4-33-g9cd2cd4,Date="Tue Jul 28 20:52:25 EDT 2015",Epoch=1438131145
25  ##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
26  ##INFO=<ID=ECNT,Number=1,Type=String,Description="Number of events in this haplotype">
27  ##INFO=<ID=HCNT,Number=1,Type=String,Description="Number of haplotypes that support this variant">
28  ##INFO=<ID=MAX_ED,Number=1,Type=Integer,Description="Maximum distance between events in this active region">
29  ##INFO=<ID=MIN_ED,Number=1,Type=Integer,Description="Minimum distance between events in this active region">
30  ##INFO=<ID=NLOD,Number=1,Type=String,Description="Normal LOD score">
31  ##INFO=<ID=PON,Number=1,Type=String,Description="Count from Panel of Normals">
32  ##INFO=<ID=RPA,Number=.,Type=Integer,Description="Number of times tandem repeat unit is repeated, for each allele
33  ##INFO=<ID=RU,Number=1,Type=String,Description="Tandem repeat unit (bases)">
34  ##INFO=<ID=STR,Number=0,Type=Flag,Description="Variant is a short tandem repeat">
35  ##INFO=<ID=TLOD,Number=1,Type=String,Description="Tumor LOD score">
36  ##SAMPLE=<ID=NORMAL,File=/seq/picard_aggregation/C500/TCGA-CF-A8HY-10A-01D-A362-08/v1/TCGA-CF-A8HY-10A-
37  ##SAMPLE=<ID=TUMOR,File=/seq/picard_aggregation/C500/TCGA-CF-A8HY-01A-11D-A364-08/v1/TCGA-CF-A8HY-01A-1
38  ##contig=<ID=1,length=249250621>
122 ##contig=<ID=NC_007605,length=171823>
123 ##reference=file:///seq/references/Homo_sapiens_assembly19/v1/Homo_sapiens_assembly19.fasta
```

11 Data columns:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	TUMOR	NORMAL
2	25469291	.	T	C	.	t_lod_fstar	ECNT=1;HCNT	GT:AD:AF:ALT	0/1:32,2:0.05	0/0:31,0:0.00:
6	32816998	rs12527715	T	C	.	germline_risk	DB;ECNT=1;H	GT:AD:AF:ALT	0/1:1,3:0.750	0/0:4,0:0.00:0