Somatic Copy Number Variant Discovery

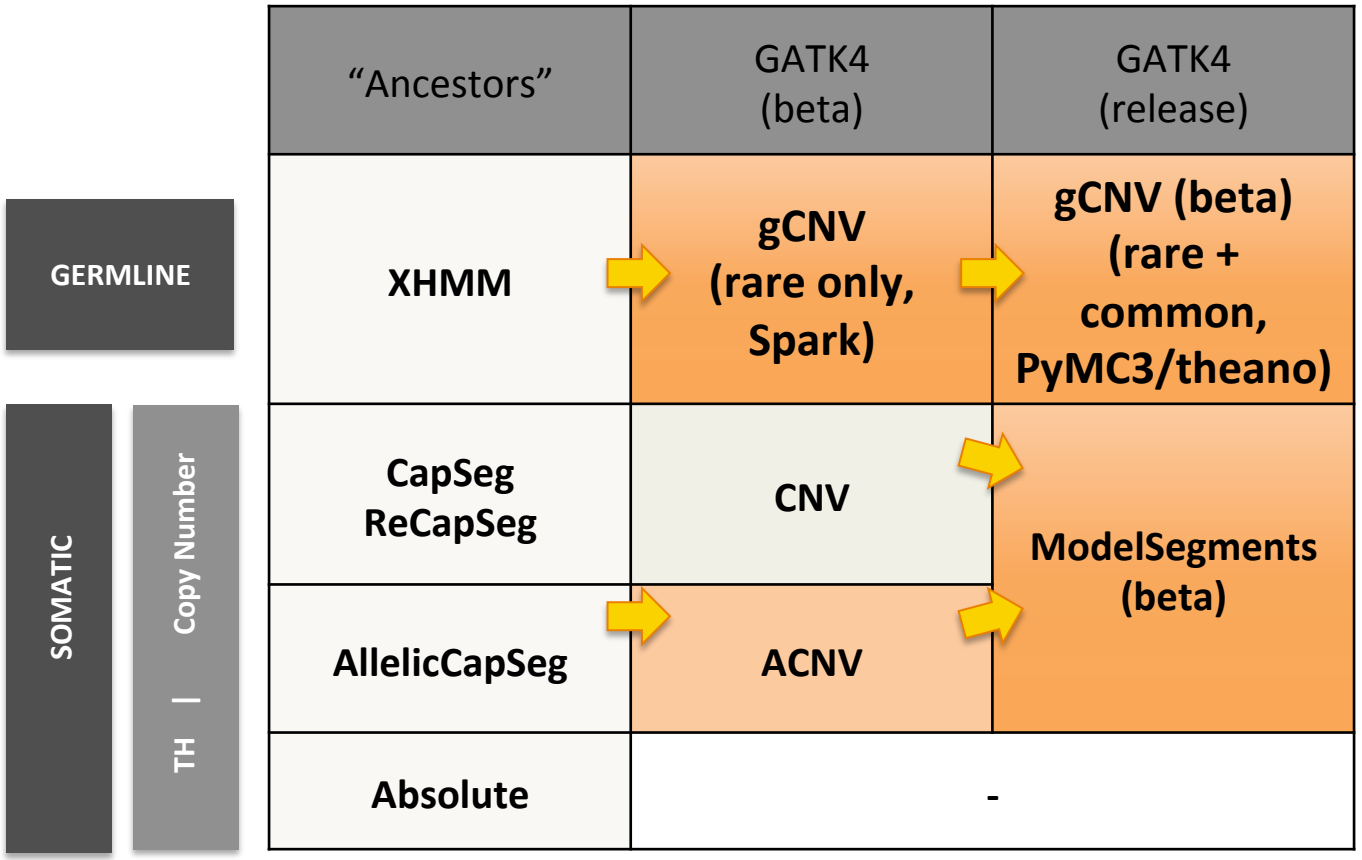
Somatic CNV workflow
in GATK4

BROAD INSTITUTE

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

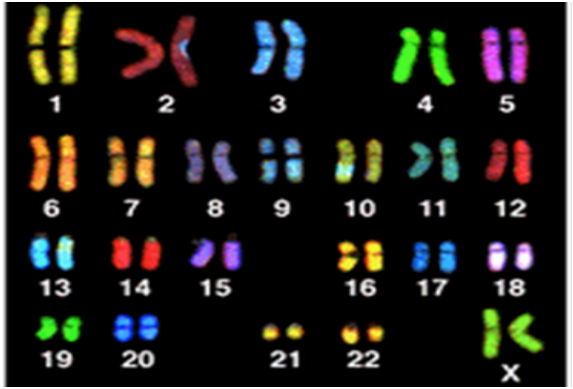
gatk

History of GATK CNV tool development



Somatic copy-number variation can be dramatic

Normal Cell



Cancer Cell Line HCC1954

- Spectral karyotyping paints each chromosome pair with a color
- Alterations can vary dramatically between cancers and within cancers

Why do we care about copy-number variants?

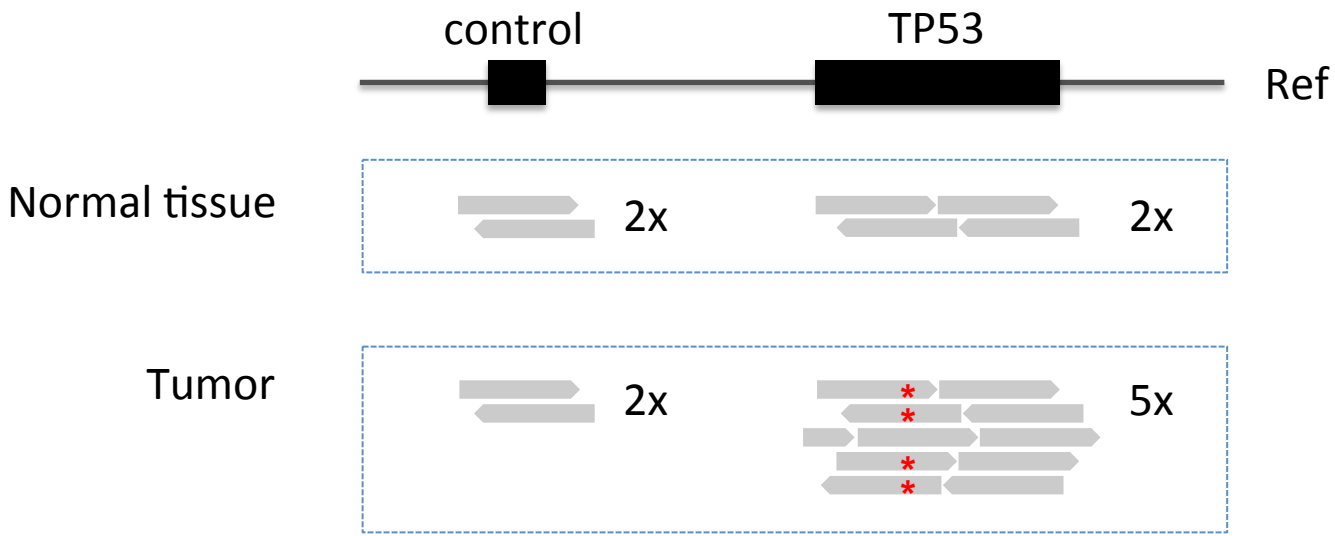
Somatic (cancer):

- HER2, EGFR, ERBB2, KIT, KRAS, Wnt, Notch, Hedgehog :amplified
- APC, BRCA1, BRCA2, PTEN, p53, NF1, NF2, PTC, DPC4 :deleted
- Infer copy-number variation → tumor purity, ploidy, phylogeny, etc.

Germline – e.g., 22q11.2 deletion:

- Cardiac abnormality
- Abnormal facies
- Thymic aplasia
- Cleft palate
- Hypocalcemia/Hypoparathyroidism
- 20 to 30-fold increased risk of schizophrenia

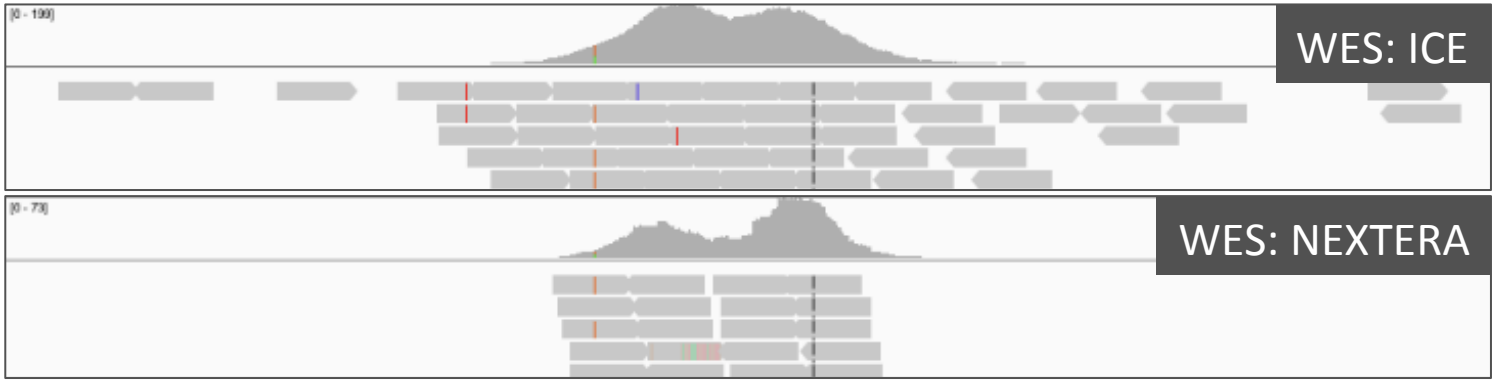
How do we identify CNVs?



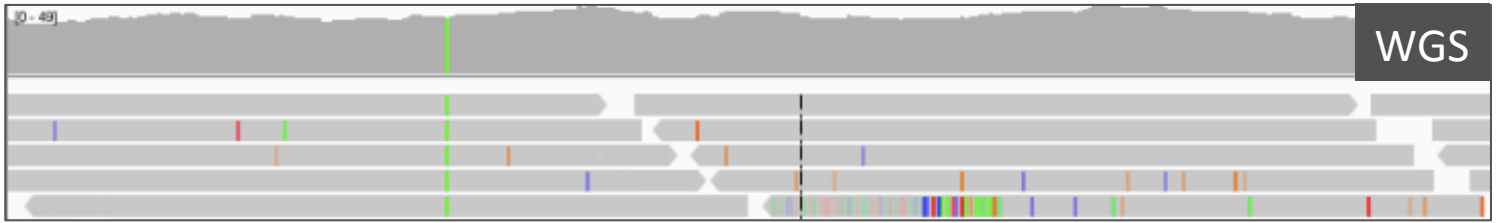
Copy number variants cause coverage imbalance

Coverage is variable across WES *targets* and *kits*

WES bait-capture and library amplification add to variability.



In comparison, WGS gives even coverage.



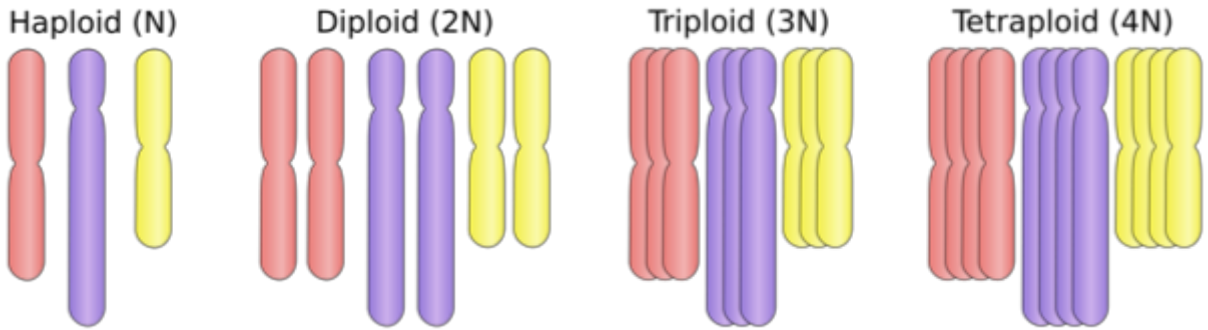
Copy **number** vs. copy **ratio**

Copy-number profile

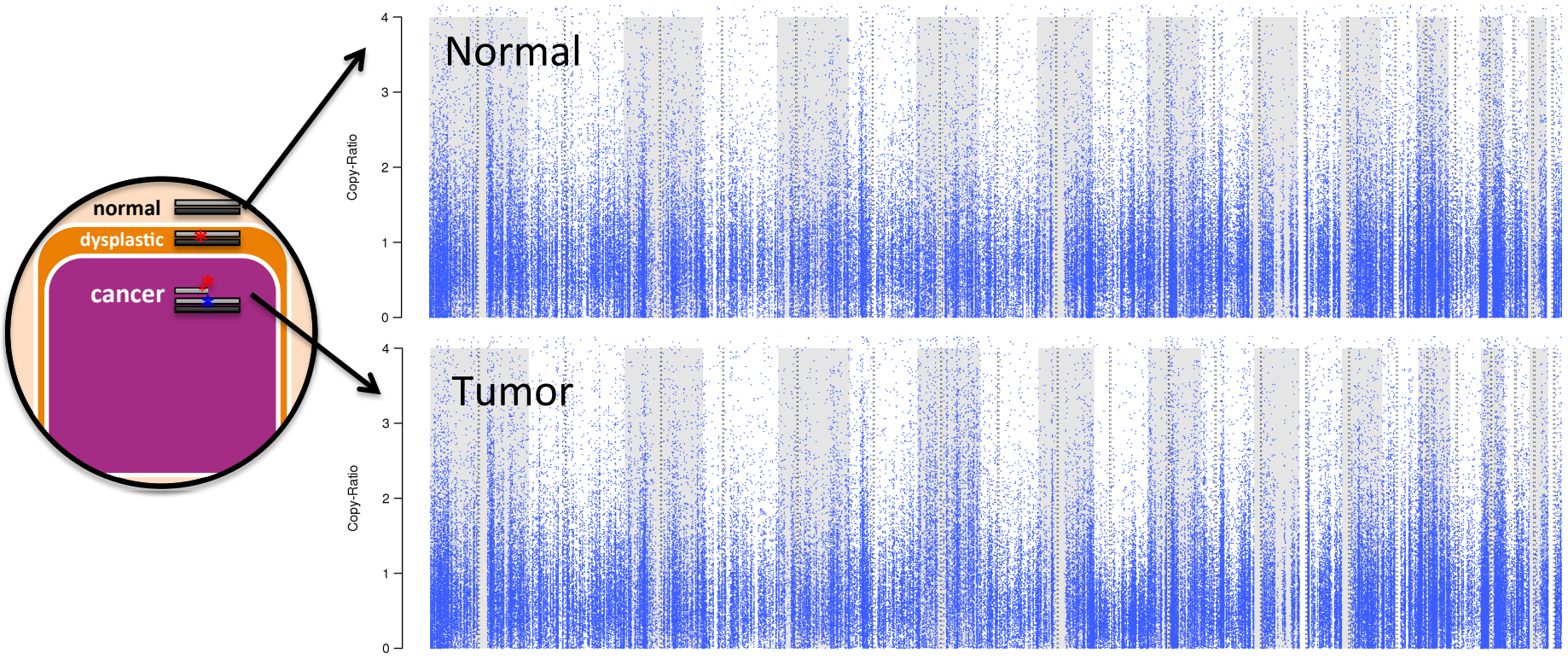
Absolute, integer-valued
number of copies
of each locus

Copy-ratio profile

Relative, real-valued
ratio of **number of copies** of each locus ***to*** **average ploidy**

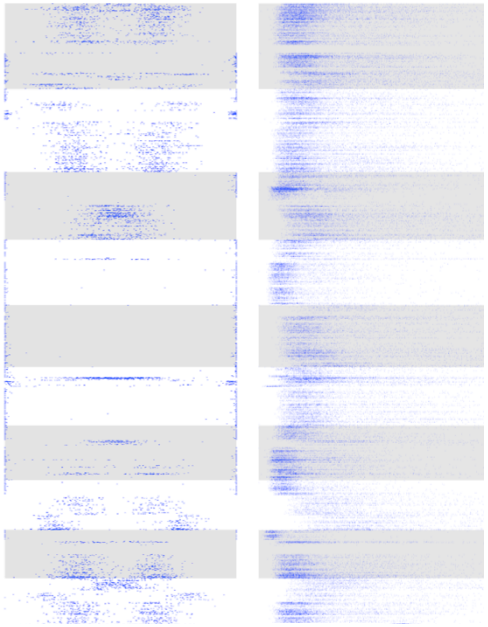


Raw copy-ratio profiles from exomes are noisy

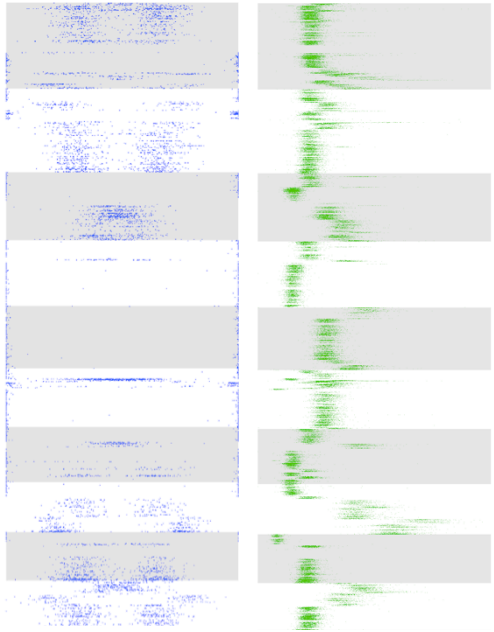


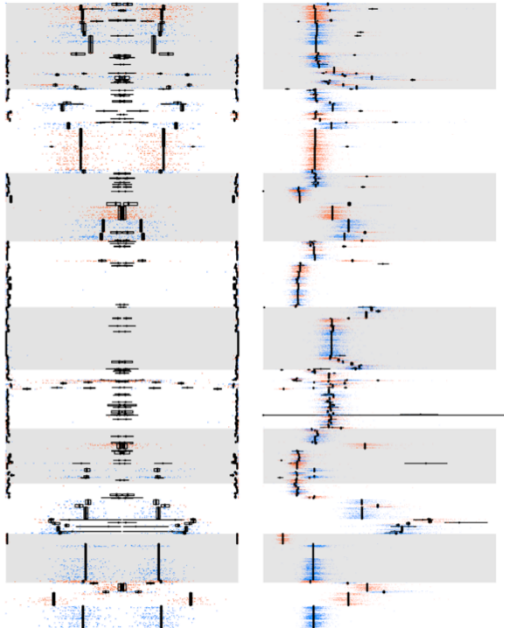
Copy-number variation: teasing out signals from the noise

data collection → *copy-ratio denoising* → *segmentation + modeling*



allele fractions fragment counts

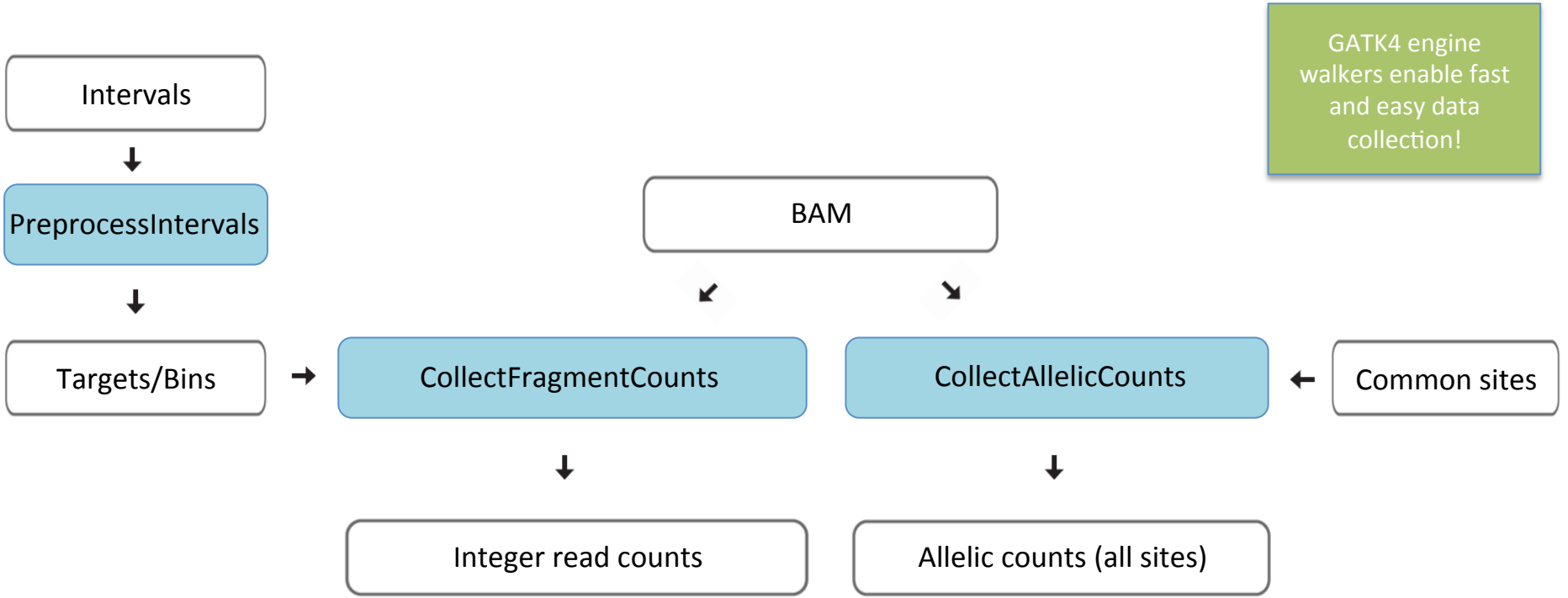




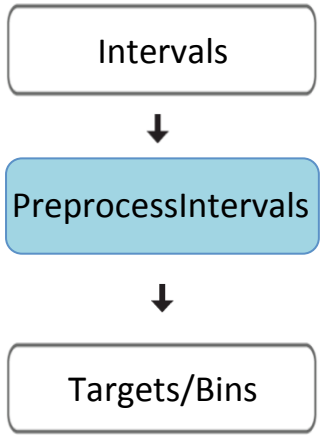
DEL ← | → AMP

event calling 1

ModelSegments pipeline: data collection



Step 1: Preprocess intervals



Exome

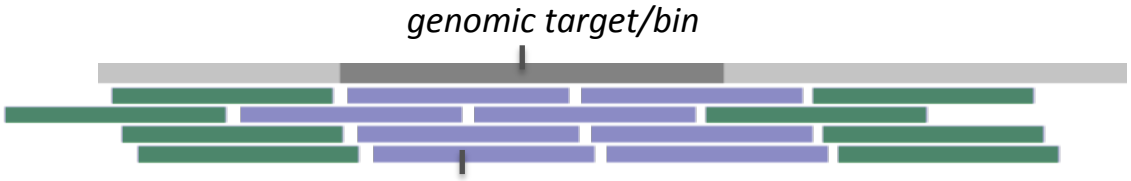
```
gatk PreprocessIntervals \
   -R reference.fasta \
   -L intervals.interval_list \
   --bin-length 0 \
   --padding 250 \
   -O preprocessed_intervals.interval_list
```

Genome

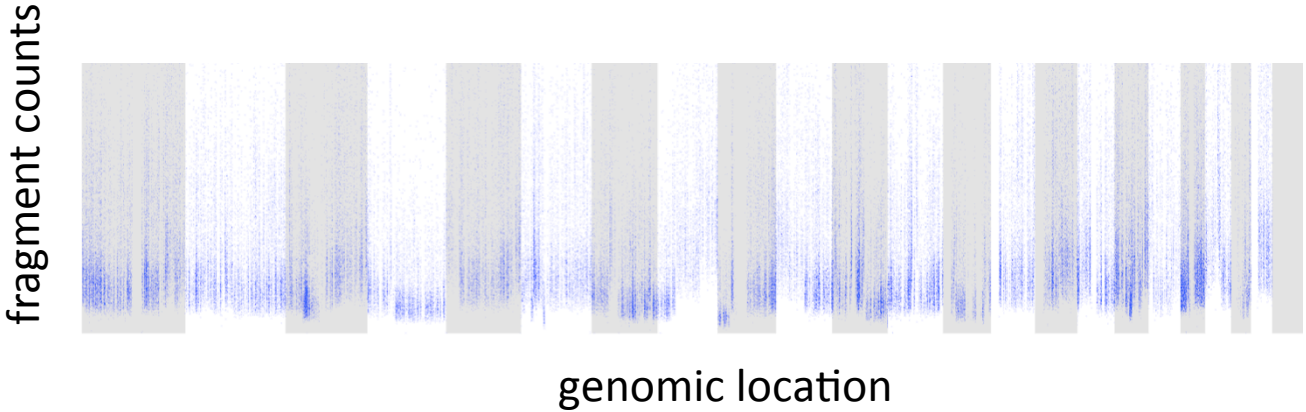
```
gatk PreprocessIntervals \
   -R reference.fasta \
   --bin-length 1000 \
   --padding 0 \
   -O preprocessed_intervals.interval_list
```

Step 2: Collect fragment counts

Fragment counts in each genomic **target/bin** allow estimates of **segmented copy ratio**

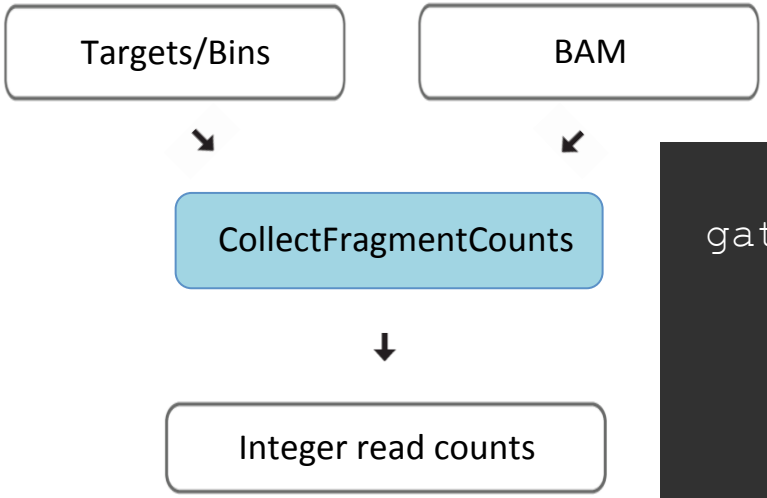


fragments w/ centers within target/ bin are counted



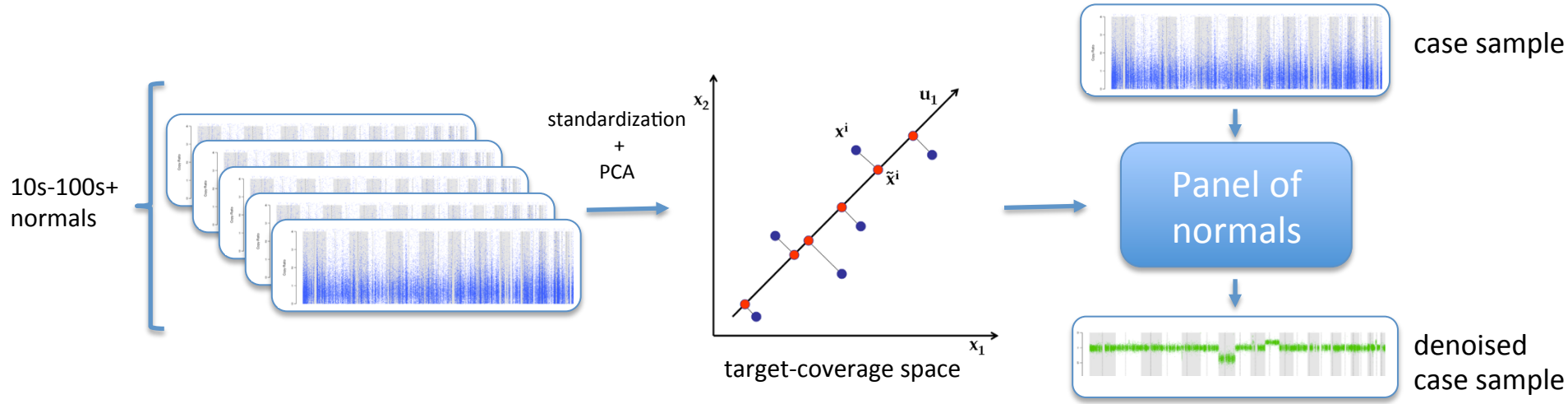
WES: ~150k targets
WGS: ~11.5M 250bp bins

Step 2: Collect fragment counts




```
gatk CollectFragmentCounts \
   -I sample.bam \
   -L intervals.interval_list \
   --interval-merging-rule OVERLAPPING_ONLY \
   -O sample.counts.hdf5
```

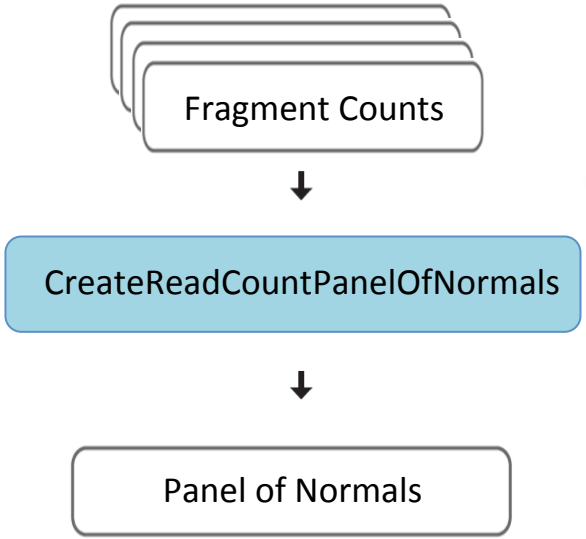
PCA denoising via a panel of normals



- Learn a hyperplane (eigenbasis) representing systematic bias from a panel of N normals
- Denoise case sample by subtracting its projection onto the first K < N eigenvectors

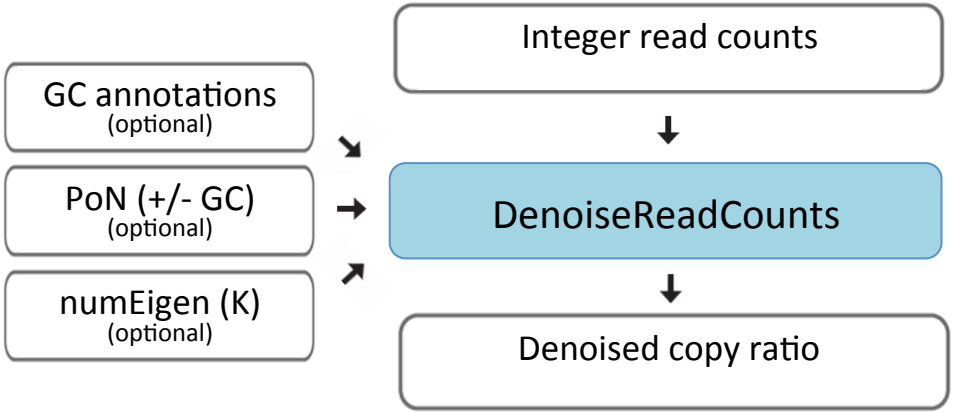
Caveat: PCA denoising cannot handle mixed-sex PoNs or common CNVs in the PoN!

Step 3: Create panel of normals



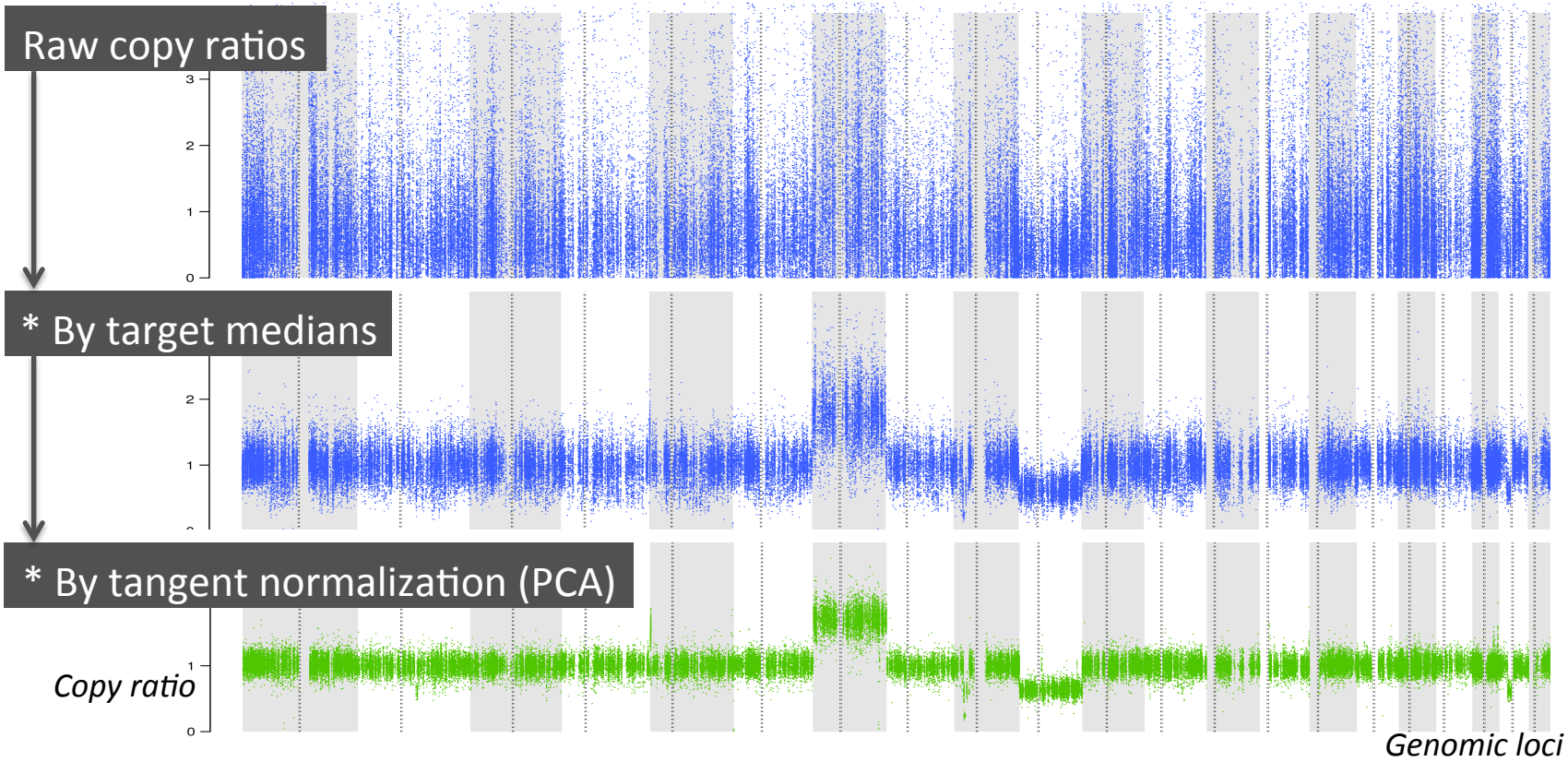

```
gatk CreateReadCountPanelOfNormals \
    -I sample_1.counts.hdf5 \
    -I sample_2.counts.tsv \
    ... \
    --annotated-intervals annotated_intervals.tsv \
    -O cnv.pon.hdf5
```

Step 4: Denoise coverage data

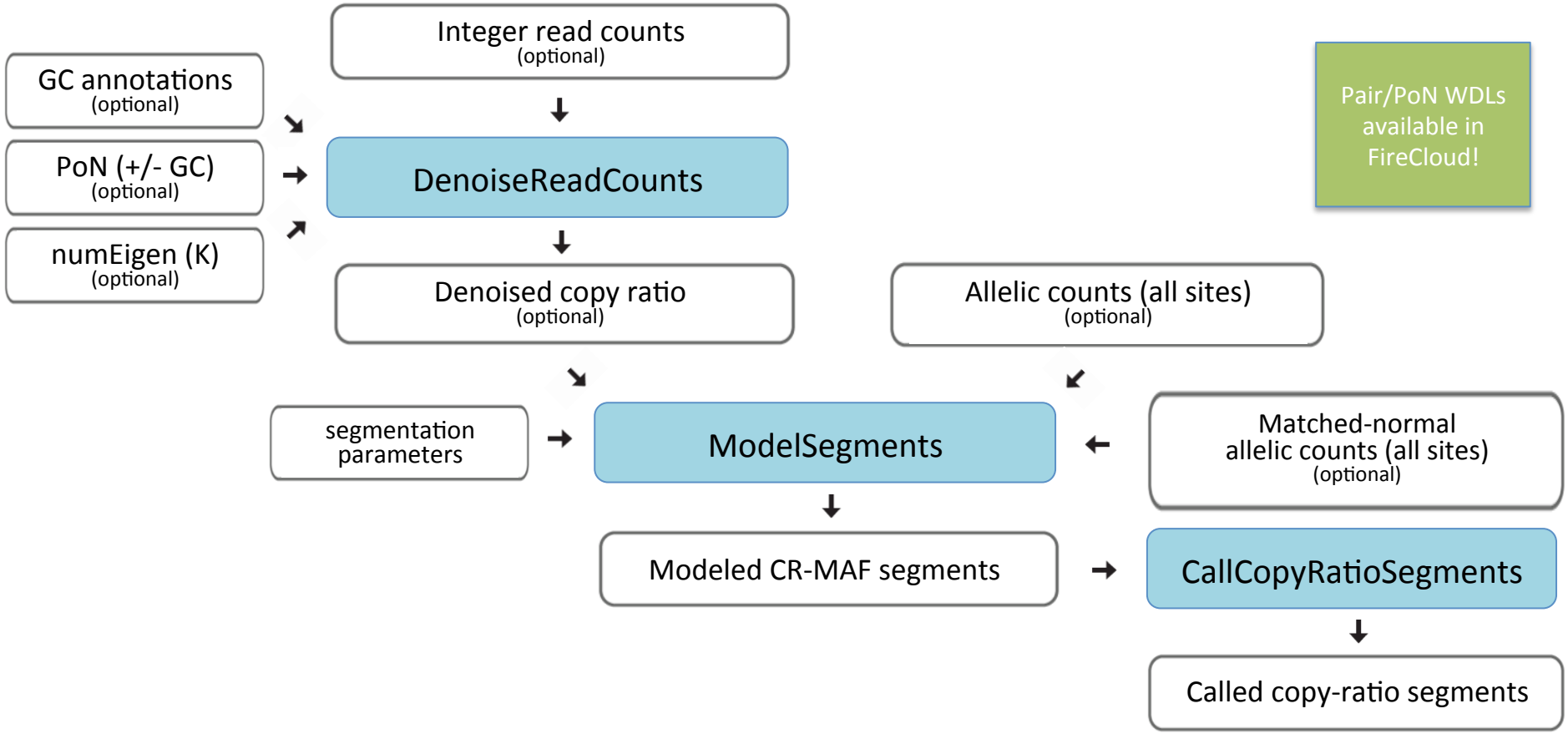



```
gatk DenoiseReadCounts \
   -I sample.counts.hdf5 \
   --count-panel-of-normals panel_of_normals.pon.hdf5 \
   --standardized-copy-ratios sample.standardizedCR.tsv \
   --denoised-copy-ratios sample.denoisedCR.tsv
```

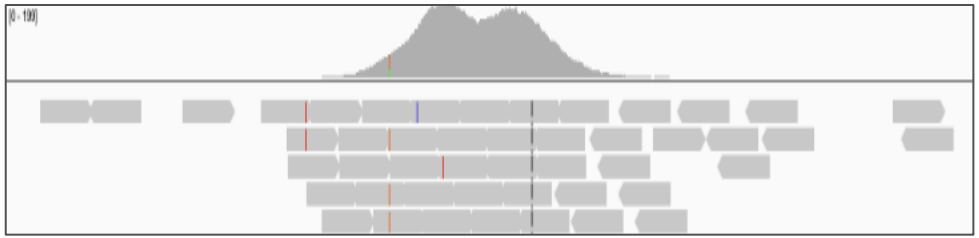
Remove noise* to reveal copy-number variation



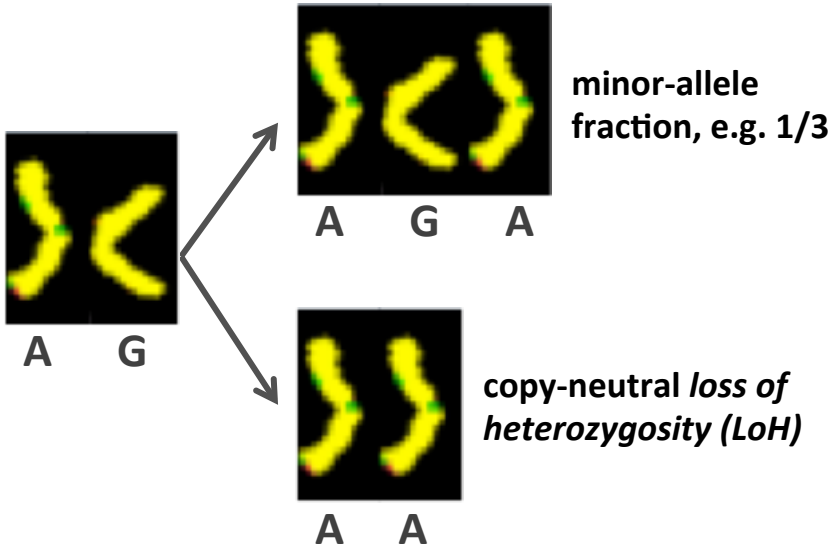
ModelSegments pipeline: case workflow



Segmentation: two sources of data



Fragment Counts



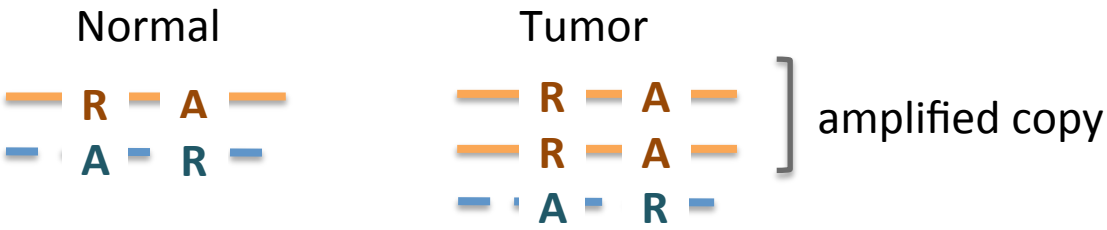
minor-allele fraction, e.g. 1/3

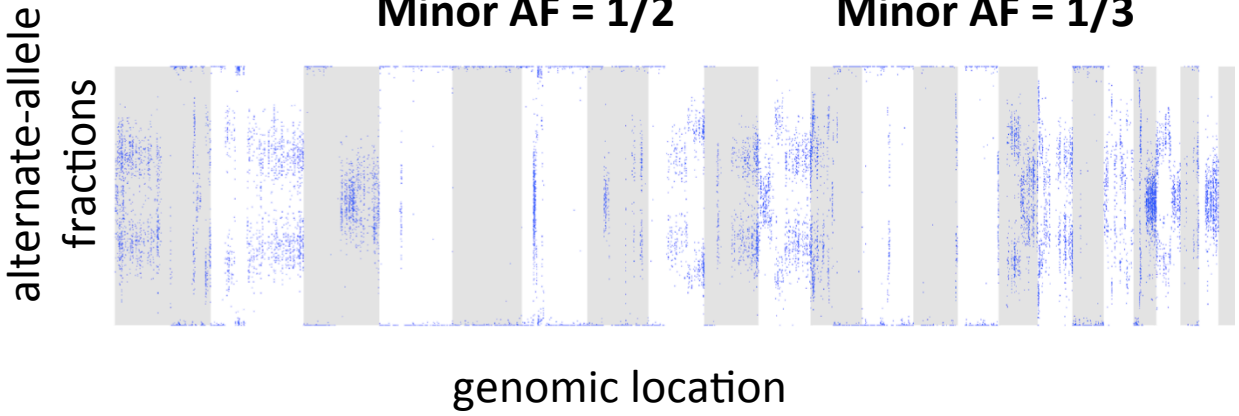
copy-neutral *loss of heterozygosity (LoH)*

Allelic Counts

Allele counts provide a secondary data source

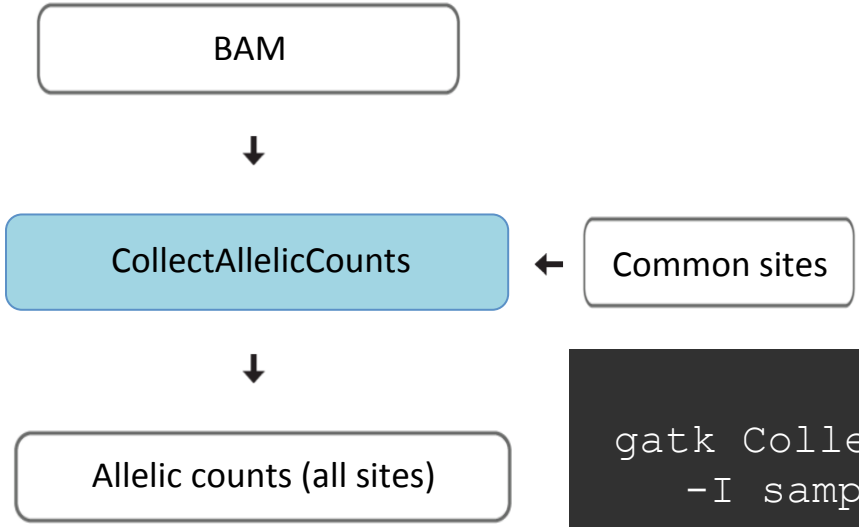
Allele counts at **germline heterozygous sites** allow estimates of **segmented minor allele fraction**





WES: ~30k germline hets
WGS: ~1.5m germline hets

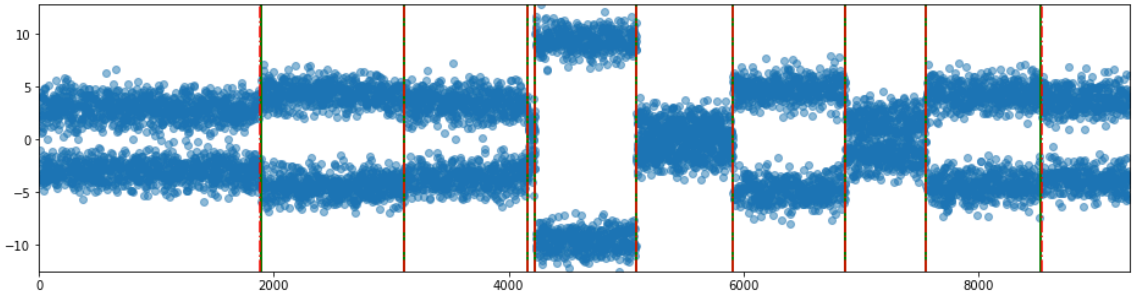
Step 5: Collect allelic counts



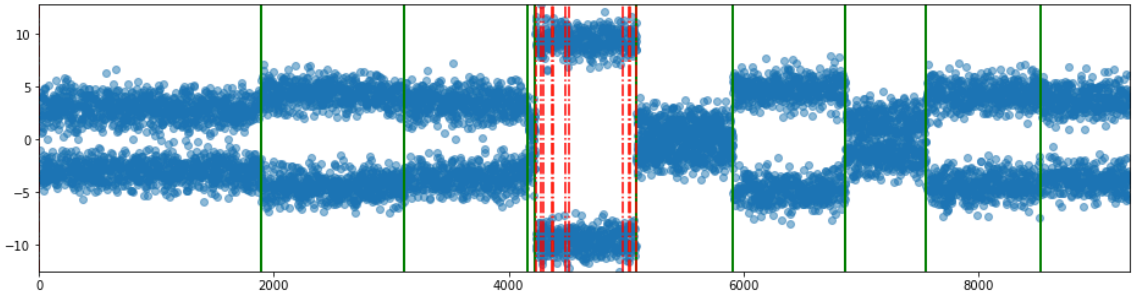

```
gatk CollectAllelicCounts \
   -I sample.bam \
   -R reference.fa \
   -L sites.interval_list \
   -O sample.allelicCounts.tsv
```

Kernel segmentation: multidimensional data

A *nonlinear* kernel allows sensitivity to changes in all moments of the distribution, so we can segment multimodal data (e.g., alternate-allele fractions):



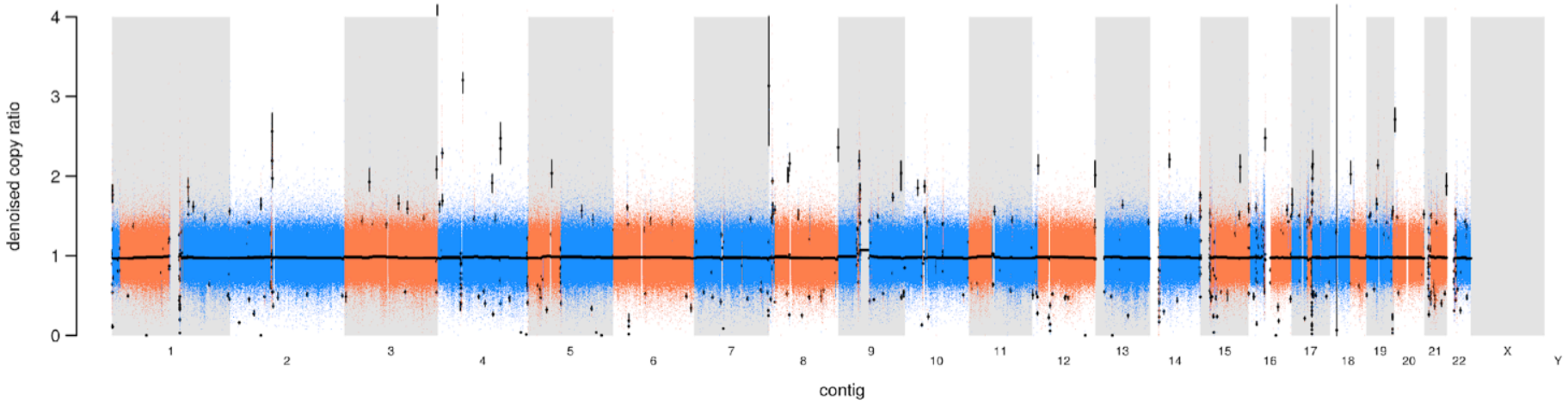
In contrast, the CBS test statistic is only sensitive to changes in the mean:



We can also segment multidimensional data by simply summing multiple kernels, allowing for joint segmentation of copy ratio and allele fraction.

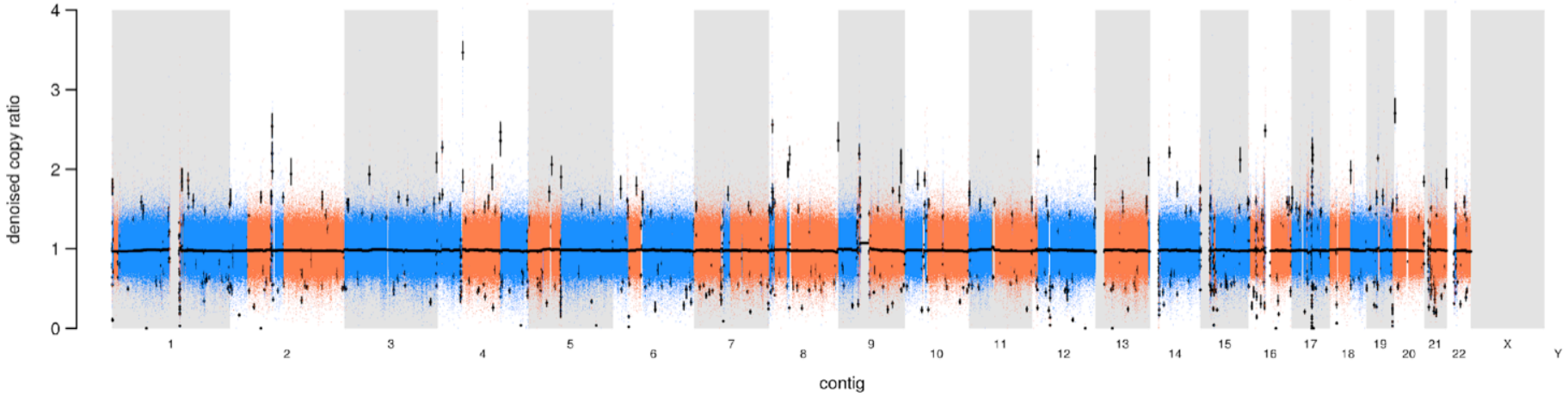
Kernel segmentation: adjustable sensitivity

changepoint penalty A = 0.5



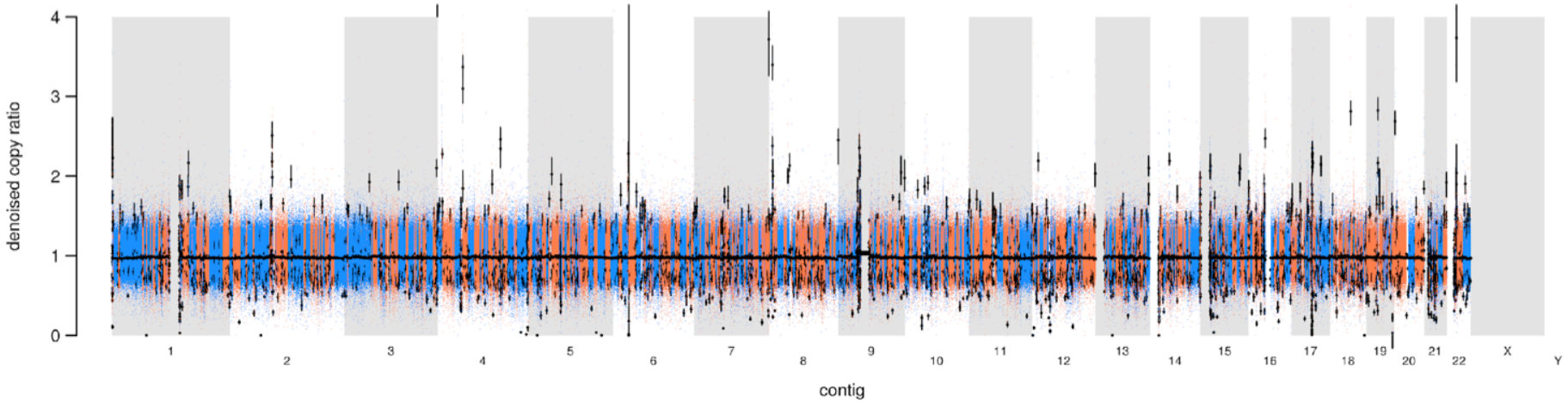
Kernel segmentation: adjustable sensitivity

changepoint penalty A = 0.3

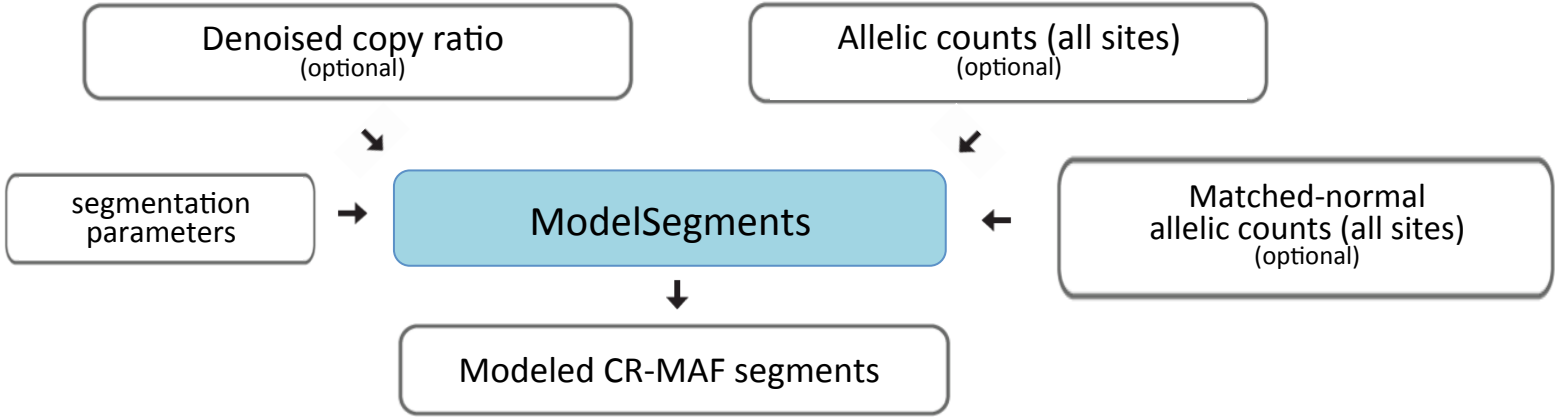


Kernel segmentation: adjustable sensitivity

changepoint penalty A = 0.1

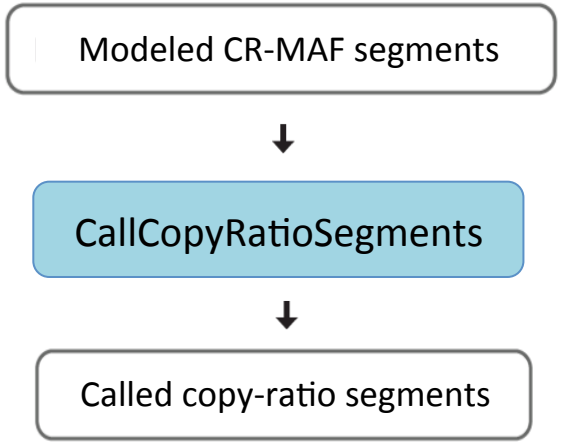


Step 6: Segment coverage



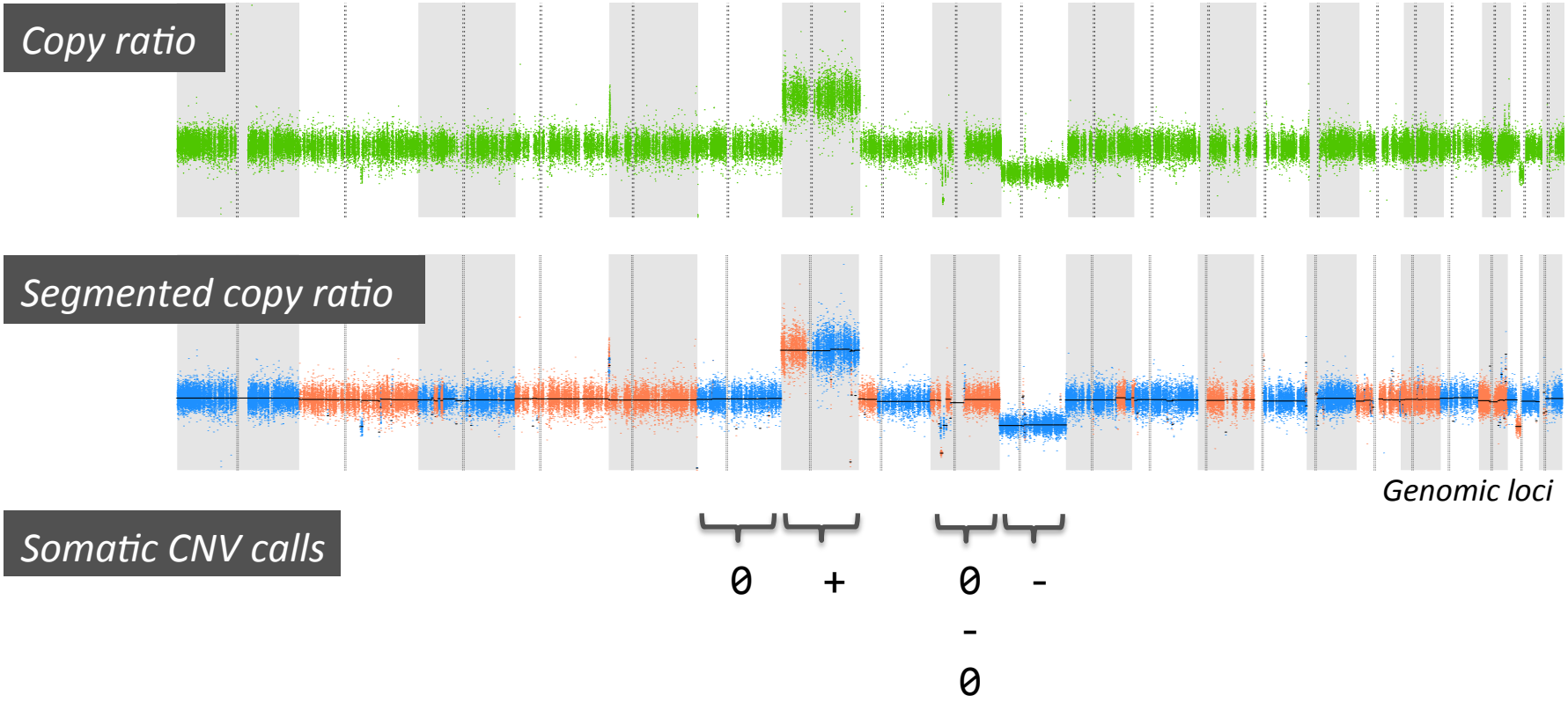

```
gatk ModelSegments \
   --denoised-copy-ratios tumor.denoisedCR.tsv \
   --allelic-counts tumor.allelicCounts.tsv \
   --normal-allelic-counts normal.allelicCounts.tsv \
   --output-prefix tumor \
   -O output_dir
```

Step 7: Call copy number events

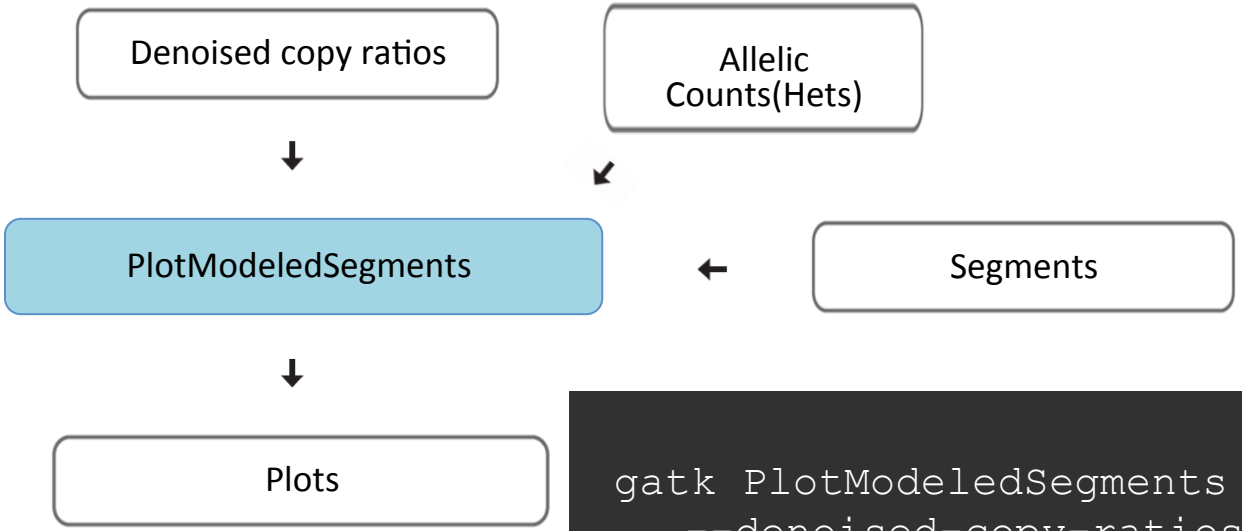



```
gatk CallCopyRatioSegments \
   -I tumor.cr.seg \
   -O tumor.called.seg
```

Segment and call copy-number events

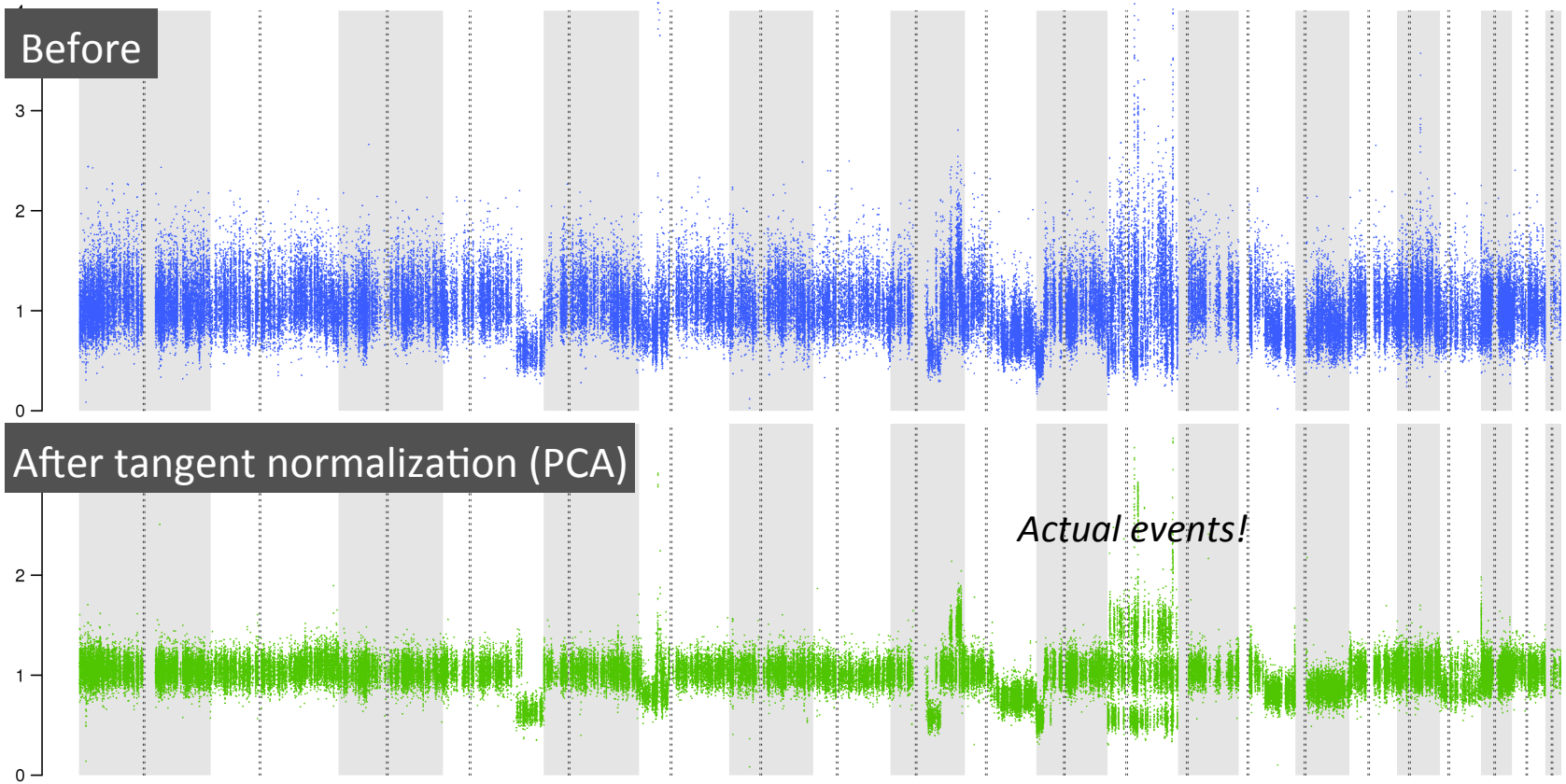


Step 8 (optional): Plot segmented coverage profile




```
gatk PlotModeledSegments \
   --denoised-copy-ratios tumor.denoisedCR.tsv \
   --allelic-counts tumor.hets.tsv \
   --segments tumor.modelFinal.seg \
   --sequence-dictionary contigs_to_plot.dict \
   --output-prefix tumor \
   -O output_dir
```

Here's one more CNV plot for the road



Further reading

- Technical whitepaper on CNV methods:
- https://github.com/broadinstitute/gatk/blob/master/docs/CNVs/CNV-methods.pdf

- GATK4 repository, releases and simple install instructions:
- https://github.com/broadinstitute/gatk
- https://github.com/broadinstitute/gatk/releases