



# Genomics in the Cloud

Book Club - Week 8

January 18, 2021

# Agenda

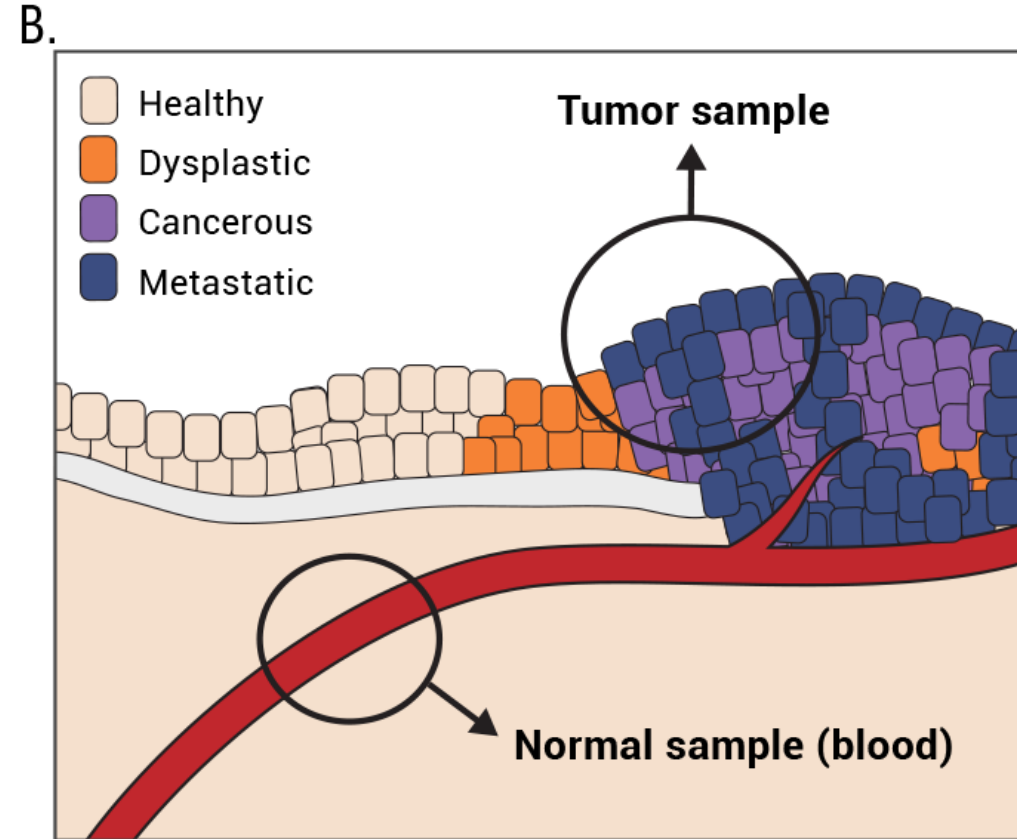
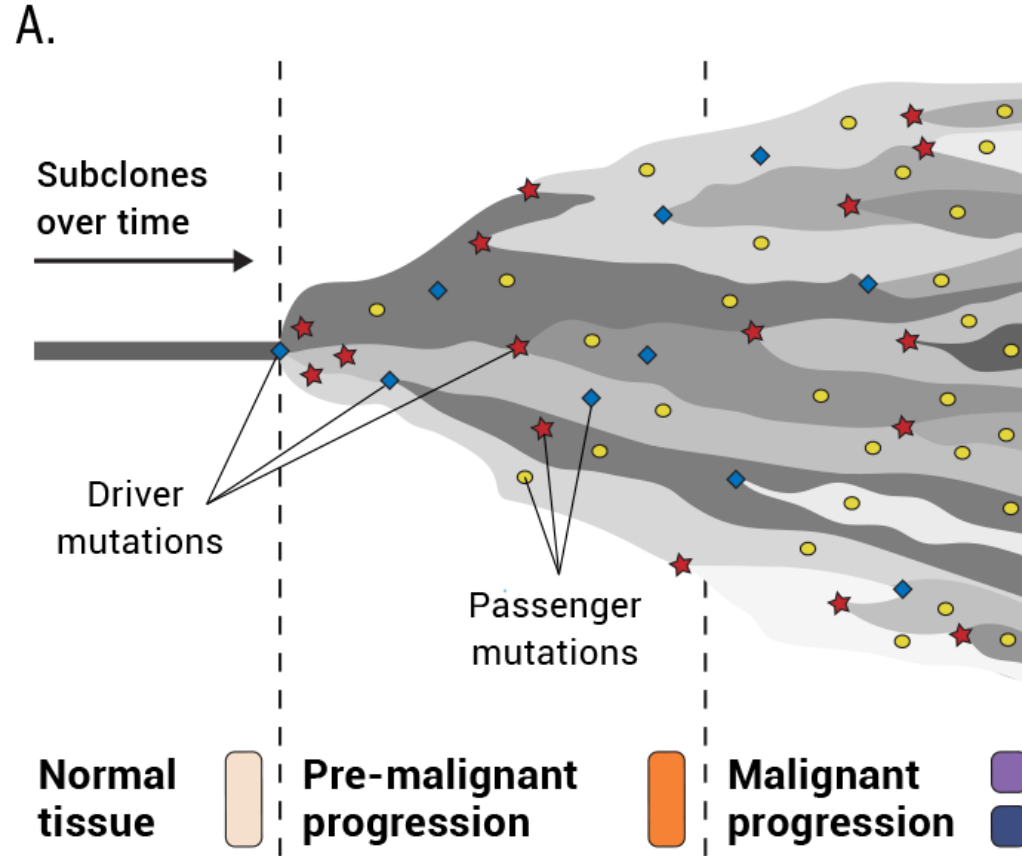
- Chapter 7: GATK Best Practices for Somatic Variant Discovery
- Additional resources
- Open discussion



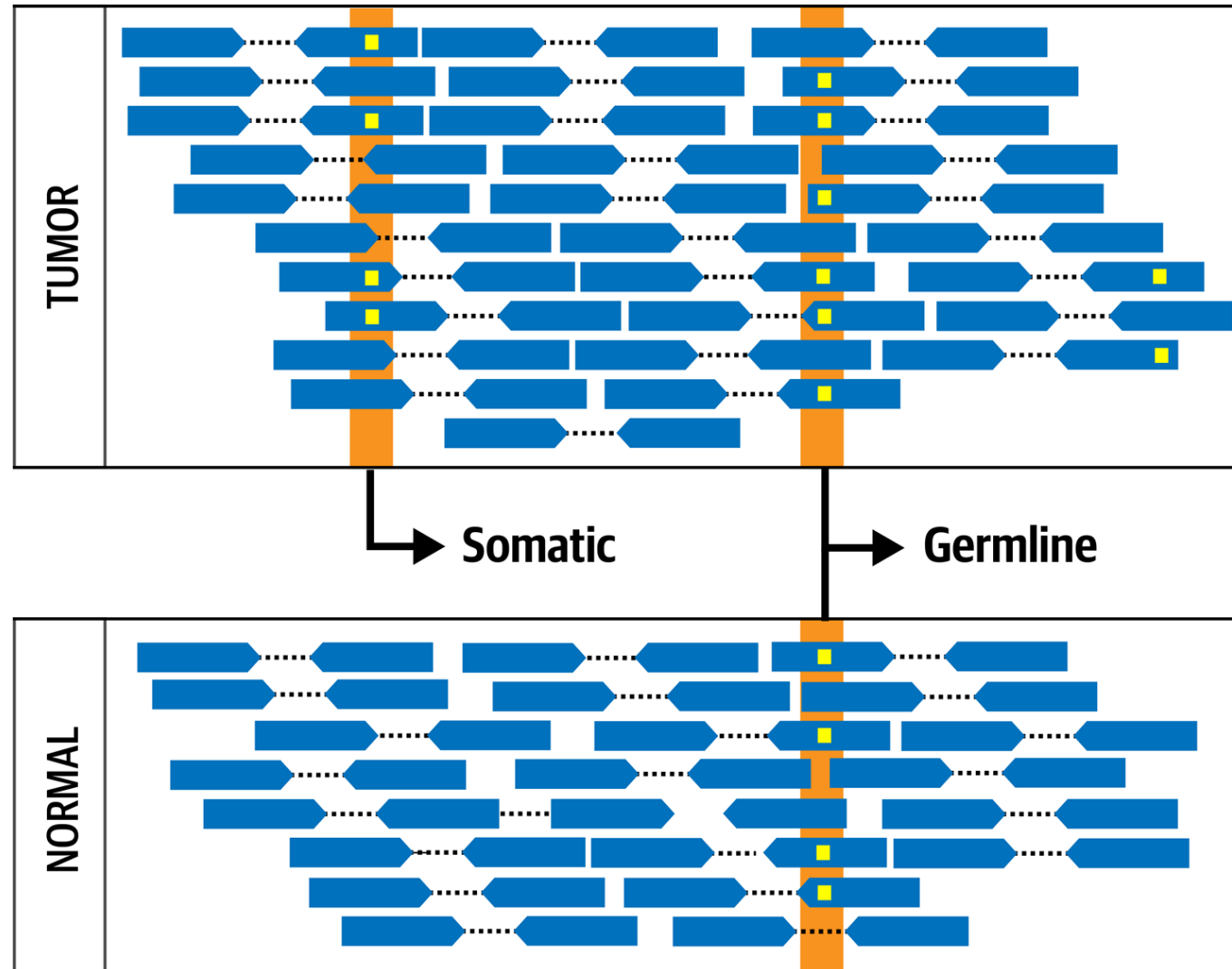
# Chapter 7: GATK Best Practices for Somatic Variant Discovery

*Genomics in the Cloud* by Geraldine A. Van der Auwera and Brian D. O'Connor (O'Reilly). Copyright 2020 The Broad Institute, Inc. and Brian O'Connor, 978-1-491-97519-0.

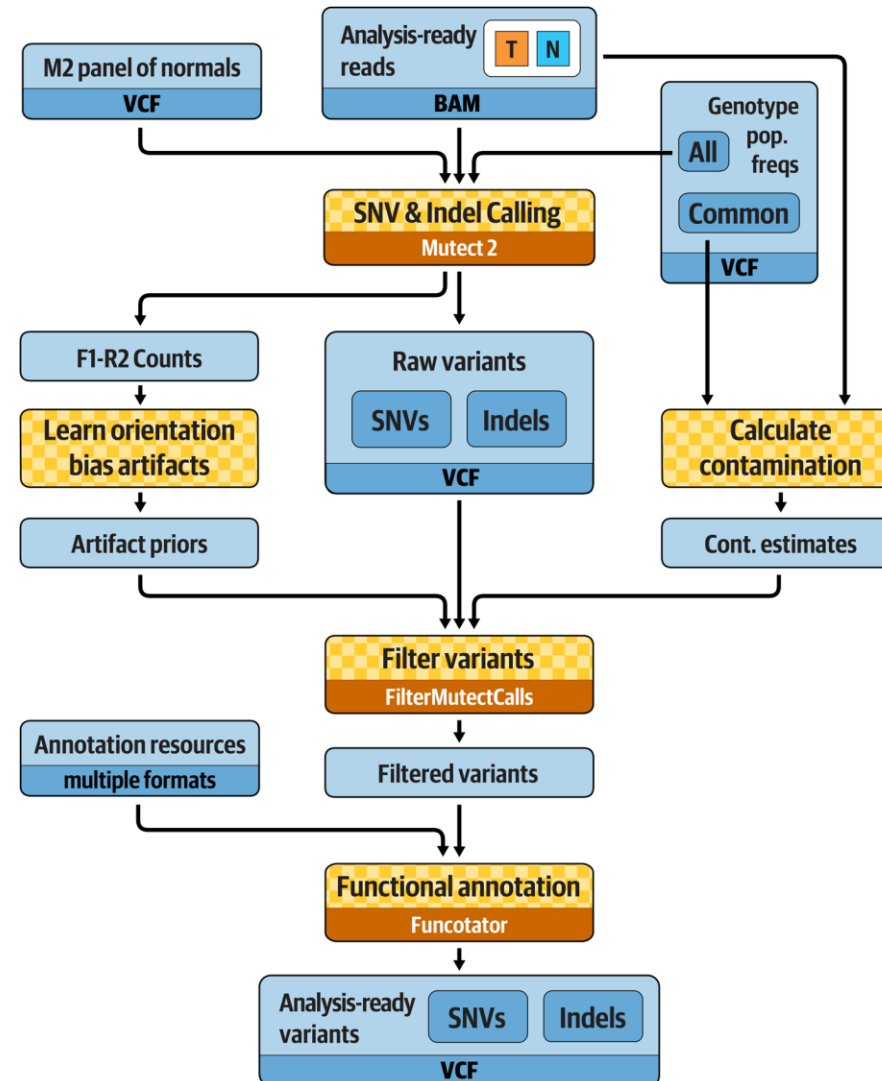
# Tumor progression



# Tumor-Normal comparison



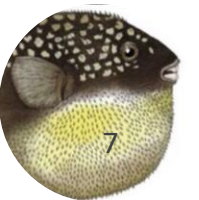
# Best practices for somatic short variant discovery



# 1. Creating a Mutect2 PoN

```
# gatk Mutect2 \  
  -R reference.fasta \  
  -I normal_1.bam \  
  -O normal_1.vcf.gz \  
  --max-mnp-distance 0
```

```
# gatk GenomicsDBImport \  
  -R reference.fasta \  
  -L intervals.interval_list \  
  -V normal_1.vcf.gz \  
  -V normal_2.vcf.gz \  
  -V normal_3.vcf.gz \  
  --genomicsdb-workspace-path pon_db
```



# 1. Creating a Mutect2 PoN (continued)

```
# gatk CreateSomaticPanelofNormals \  
  -R reference.fasta \  
  -V gendb://pon_db \  
  --germline-resource af-only-gnomad.vcf.gz \  
  -O pon.vcf.gz
```

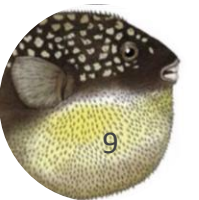
```
# zcat resources/chr17_m2pon.vcf.gz | grep -v '##' | head -3  
#CHROM POS ID REF ALT QUAL FILTER INFO  
chr6 29941027 . G A . . .  
chr6 29941061 . G C . . .
```





## 2. Running Mutect2 on the T-N pair

```
# gatk Mutect2 \  
  -R ref/Homo_sapiens_assembly38.fasta \  
  -I bams/tumor.bam \  
  -I bams/normal.bam \  
  -normal HCC1143_normal \  
  -L resources/chr17plus.interval_list \  
  -pon resources/chr17_m2pon.vcf.gz \  
  --germline-resource resources/chr17_af-only-gnomad_grch38.vcf.gz \  
  -bamout sandbox/m2_tumor_normal.bam \  
  -O sandbox/m2_somatic_calls.vcf.gz
```



### 3. Estimating cross-sample contamination

```
# gatk GetPileupSummaries \  
  -I bams/normal.bam \  
  -V resources/chr17_small_exac_common_3_grch38.vcf.gz \  
  -L resources/chr17_small_exac_common_3_grch38.vcf.gz \  
  -O sandbox/normal_getpileupsummaries.table
```

```
# gatk GetPileupSummaries \  
  -I bams/tumor.bam \  
  -V resources/chr17_small_exac_common_3_grch38.vcf.gz \  
  -L resources/chr17_small_exac_common_3_grch38.vcf.gz \  
  -O sandbox/tumor_getpileupsummaries.table
```



### 3. Pileup summary tables

```
# head -5 sandbox/normal_getpileupsummaries.table
#<METADATA>SAMPLE=HCC1143_normal
contig position ref_count alt_count other_alt_count allele_frequency
chr6 29942512 7          4          0          0.063
chr6 29942517 12          4          0          0.062
chr6 29942525 13          7          0          0.063
```

```
# head -5 sandbox/tumor_getpileupsummaries.table
#<METADATA>SAMPLE=HCC1143_tumor
contig position ref_count alt_count other_alt_count allele_frequency
chr6 29942512 9           0          0          0.063
chr6 29942517 13           1          0          0.062
chr6 29942525 13           7          0          0.063
```



## 4. Calculate contamination estimate

```
# gatk CalculateContamination \  
  -I sandbox/tumor_getpileupsummaries.table \  
  -matched sandbox/normal_getpileupsummaries.table \  
  -tumor-segmentation sandbox/segments.table \  
  -O sandbox/pair_calculatecontamination.table
```

```
# cat sandbox/pair_calculatecontamination.table  
sample contamination error  
HCC1143_tumor 0.0114853 0.0019180
```

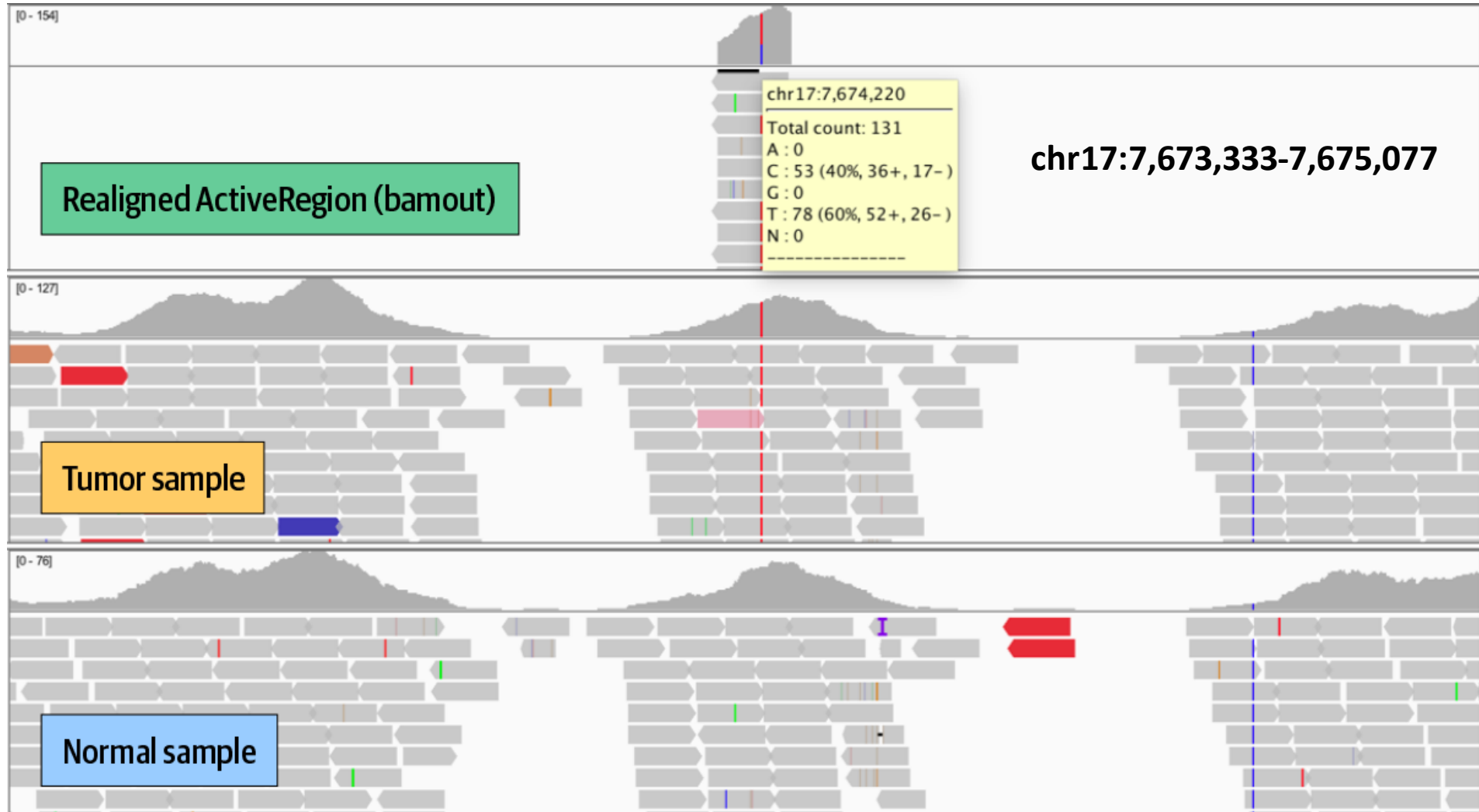


## 5. Filtering Mutect2 calls

```
# gatk FilterMutectCalls \  
  -R ref/Homo_sapiens_assembly38.fasta \  
  -V sandbox/m2_somatic_calls.vcf.gz \  
  --contamination-table sandbox/pair_calculatecontamination.table \  
  -O sandbox/m2_somatic_calls.filtered.vcf.gz \  
  --stats sandbox/m2_somatic_calls.vcf.gz.stats \  
  --tumor-segmentation sandbox/segments.table
```



# Visualizing TP53 in IGV



## 6. Annotating with Funcotator

```
# gatk Funcotator \  
  --data-sources-path  
  resources/funcotator_dataSources_GATK_Workshop_20181205 \  
  --ref-version hg38 \  
  -R ref/Homo_sapiens_assembly38.fasta \  
  -V sandbox/m2_somatic_calls.filtered.vcf.gz \  
  -O sandbox/m2_somatic_calls.funcotated.vcf.gz \  
  --output-file-format VCF
```



## 6. Funcotator output for TP53 mutation

```
# zcat sandbox/m2_somatic_calls.funcotated.vcf.gz | grep 7674220
chr17 7674220 . C T . PASS
CONTQ=93;DP=134;ECNT=1;FUNCOTATION=[TP53|hg38|chr17|7674220|7674220|MISSENSE||SNP|C|C|
T|g.chr17:7674220C>T|ENST00000269305.8|-|7|933|c.743G>A|c.(742-
744)cGg>cAg|p.R248Q|0.5660847880299252|GATGGGCCTCCGGTTCATGCC|TP53_ENST00000445888.6_M
ISSENSE_p.R248Q/TP53_ENST
00000420246.6_MISSENSE_p.R248Q/TP53_ENST00000622645.4_MISSENSE_p.R209Q/TP53_ENST000006
10292.4_MISSENSE_p.R209Q/TP53_ENST00000455263.6_MISSENSE_p.R248Q/TP53_ENST00000610538.4
_MISSENSE_p.R209Q/TP53_ENST00000620739.4_MISSENSE_p.R209Q/TP53_ENST00000619485.4_MISSE
NSE_p.R209Q/TP53_ENST00000510385.5_MISSENSE_p.R116Q/TP53_ENST00000618944.4_MISSENSE_p.
R89Q/TP53_ENST000005
04290.5_MISSENSE_p.R116Q/TP53_ENST00000610623.4_MISSENSE_p.R89Q/TP53_ENST00000504937.5_
MISSENSE_p.R116Q/TP53_ENST00000619186.4_MISSENSE_p.R89Q/TP53_ENST00000359597.8_MISSENS
E_p.R248Q/TP53_ENST00000413465.6_MISSENSE_p.R248Q/TP53_ENST00000615910.4_MISSENSE_p.R23
7Q/TP53_ENST00000617185.4_MISSENSE_p.R248Q];GERMQ=93;MBQ=31,32;MFRL=146,140;MMQ=60,60
;MPOS=21;NALOD=1.73;NLOD=15.33;POPAF=6.00;SEQQ=93;STRANDQ=93;TLOD=264.54
GT:AD:AF:DP:F1R2:F2R1:SB 0/0:51,0:0.018:51:22,0:29,0:36,15,0,0 0/1:0,76:0.987:76:0,38:0,38:0,0,52,24
```

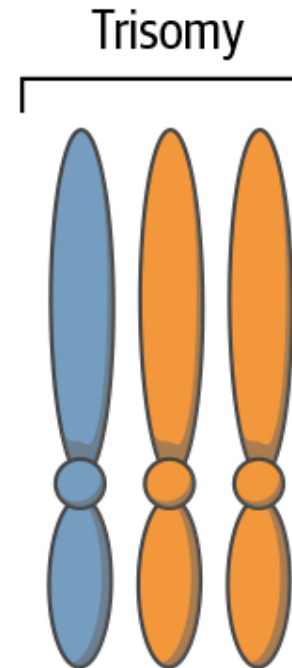
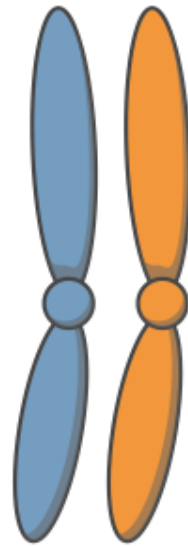




# Somatic Copy-Number Alterations

# Difference between copy number and copy ratio

Chromosomes  
(chromatid form)  
of a **diploid**  
organism



*Duplicated*

**Copy Number**

**2**

**2**

**3**

**2**

**Copy Ratio**

**1**

**1**

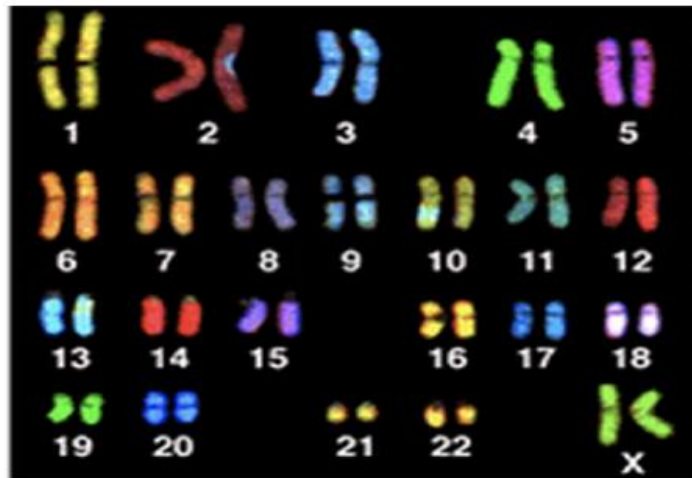
**1.5**

**1**



# Karyotyping or normal vs tumor cells

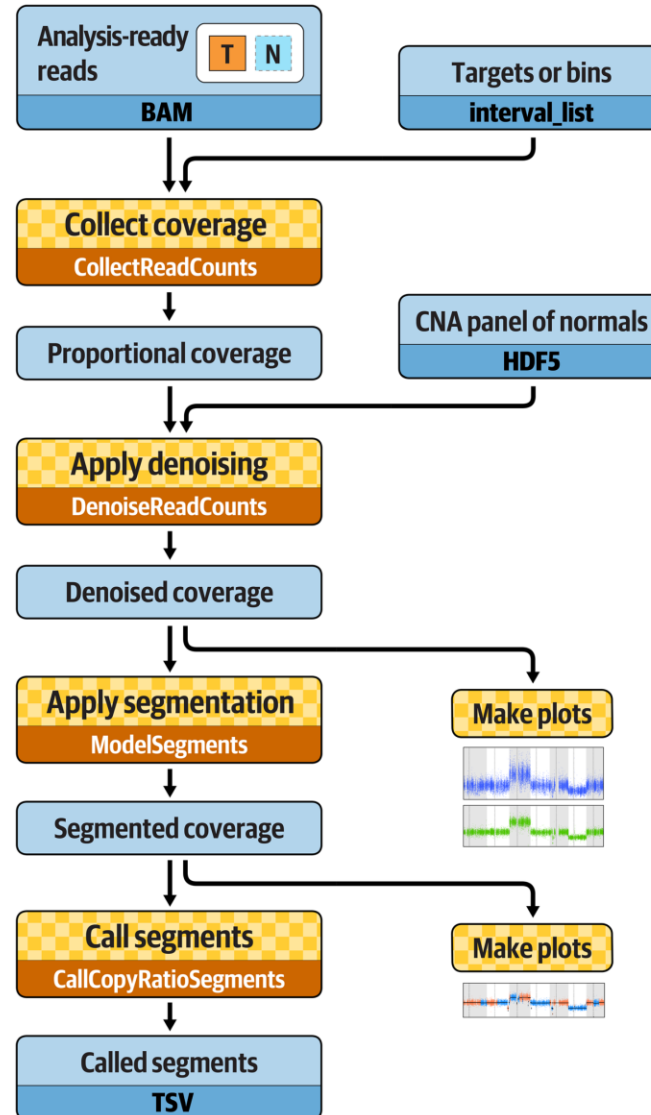
**Normal cell line**



**Cancer cell line HCC1954**



# Best practices workflow for somatic CNA discovery



# 1. Calculating coverage counts

```
# gatk PreprocessIntervals \  
  -R ref/Homo_sapiens_assembly38.fasta \  
  -L resources/targets_chr17.interval_list \  
  -O sandbox/targets_chr17.preprocessed.interval_list \  
  --padding 250 \  
  --bin-length 0 \  
  --interval-merging-rule OVERLAPPING_ONLY
```



## 2. Collect read counts

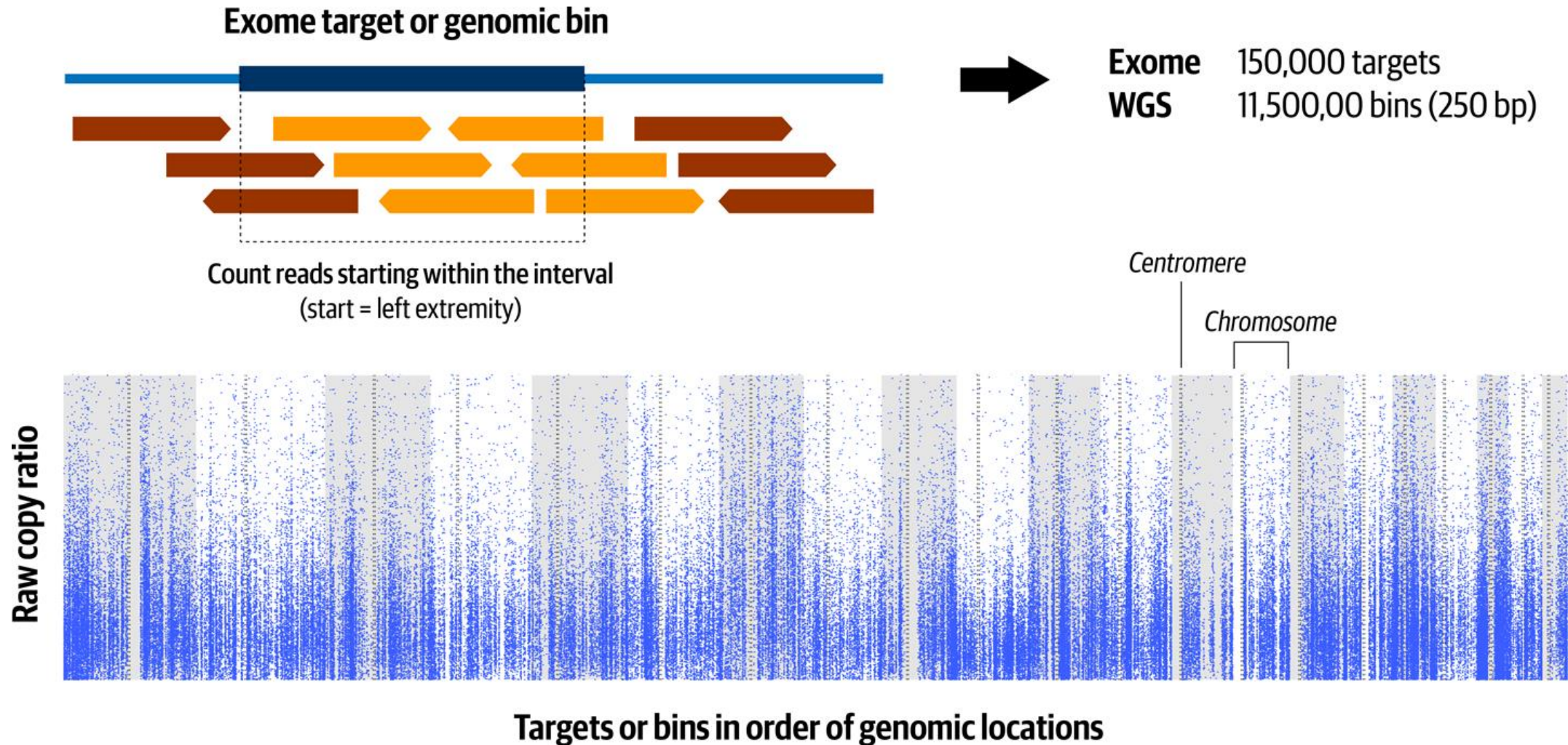
```
# gatk CollectReadCounts \  
  -I bams/tumor.bam \  
  -L sandbox/targets_chr17.preprocessed.interval_list \  
  -R ref/Homo_sapiens_assembly38.fasta \  
  -O sandbox/tumor.counts.tsv \  
  --format TSV \  
  --imr OVERLAPPING_ONLY
```

```
# tail -5 sandbox/tumor.counts.tsv  
chr17 83051485 83052048 1  
chr17 83079564 83080237 0  
chr17 83084686 83085575 1010  
chr17 83092915 83093478 118  
chr17 83094004 83094827 484
```





# Read counts for estimating segmented copy ratio



### 3. Applying denoising

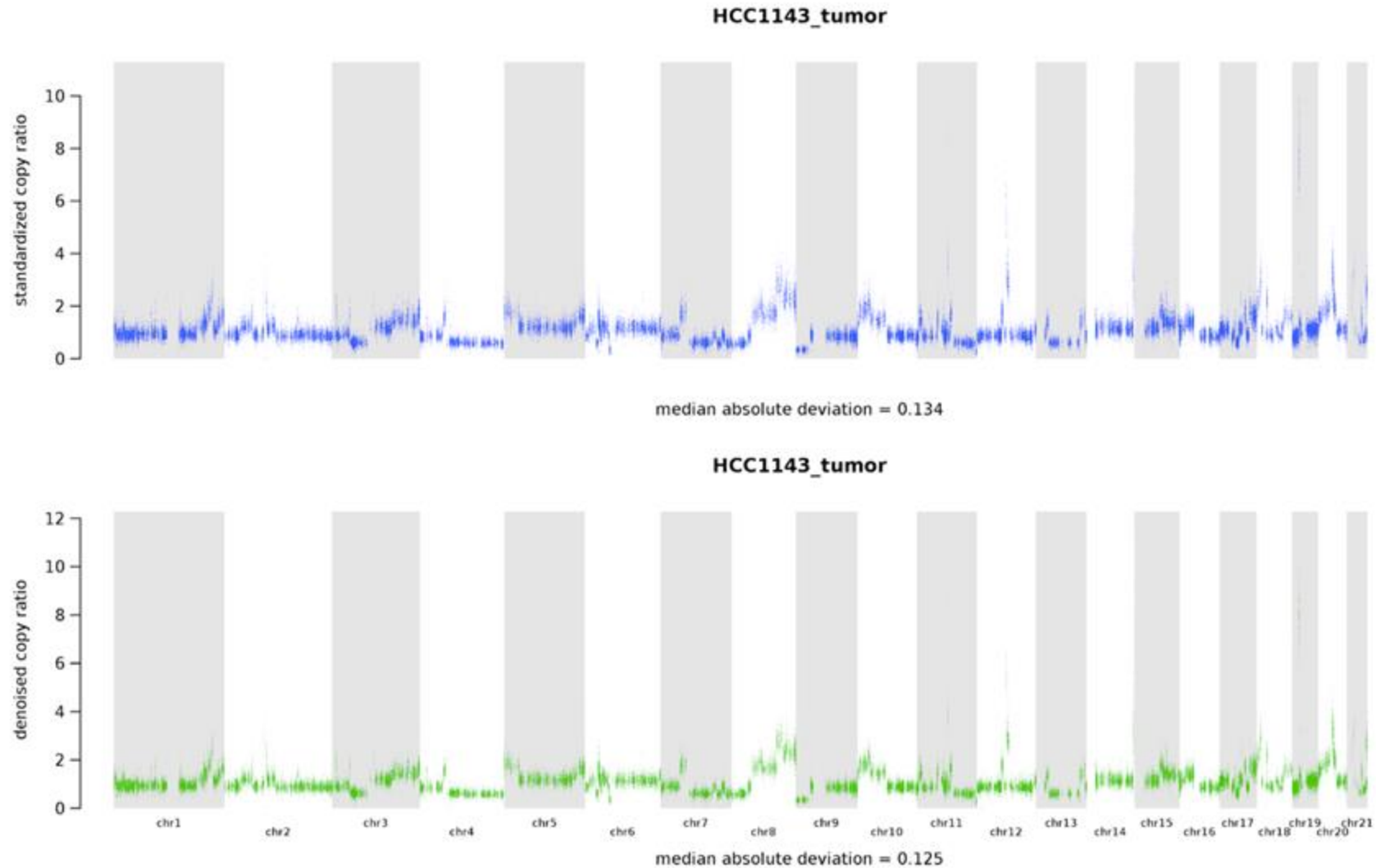
```
# gatk DenoiseReadCounts \  
  -I cna_inputs/hcc1143_T_clean.counts.hdf5 \  
  --count-panel-of-normals cna_inputs/cnaponC.pon.hdf5 \  
  --standardized-copy-ratios sandbox/hcc1143_T_clean.standardizedCR.tsv \  
  --denoised-copy-ratios sandbox/hcc1143_T_clean.denoisedCR.tsv
```

```
# gatk PlotDenoisedCopyRatios \  
  --sequence-dictionary ref/Homo_sapiens_assembly38.dict \  
  --standardized-copy-ratios  
  sandbox/hcc1143_T_clean.standardizedCR.tsv \  
  --denoised-copy-ratios sandbox/hcc1143_T_clean.denoisedCR.tsv \  
  --minimum-contig-length 46709983 \  
  --output sandbox/cna_plots \  
  --output-prefix hcc1143_T_clean
```





# CNA plots with 2-step denoising



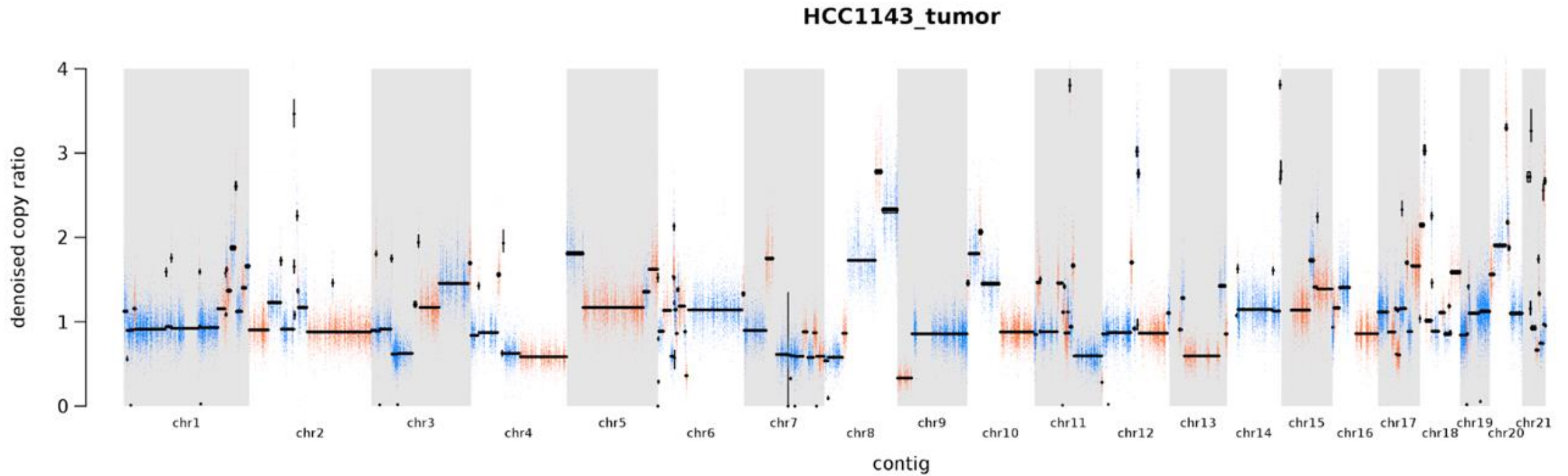
## 4. Performing segmentation and call CNAs

```
# gatk ModelSegments \  
  --denoised-copy-ratios sandbox/hcc1143_T_clean.denoisedCR.tsv \  
  --output sandbox \  
  --output-prefix hcc1143_T_clean
```

```
# gatk PlotModeledSegments \  
  --denoised-copy-ratios sandbox/hcc1143_T_clean.denoisedCR.tsv \  
  --segments sandbox/hcc1143_T_clean.modelFinal.seg \  
  --sequence-dictionary ref/Homo_sapiens_assembly38.dict \  
  --minimum-contig-length 46709983 \  
  --output sandbox/cna_plots \  
  --output-prefix hcc1143_T_clean
```



# Plot of segments based on denoised copy ratios



## 5. Get final CNA calls

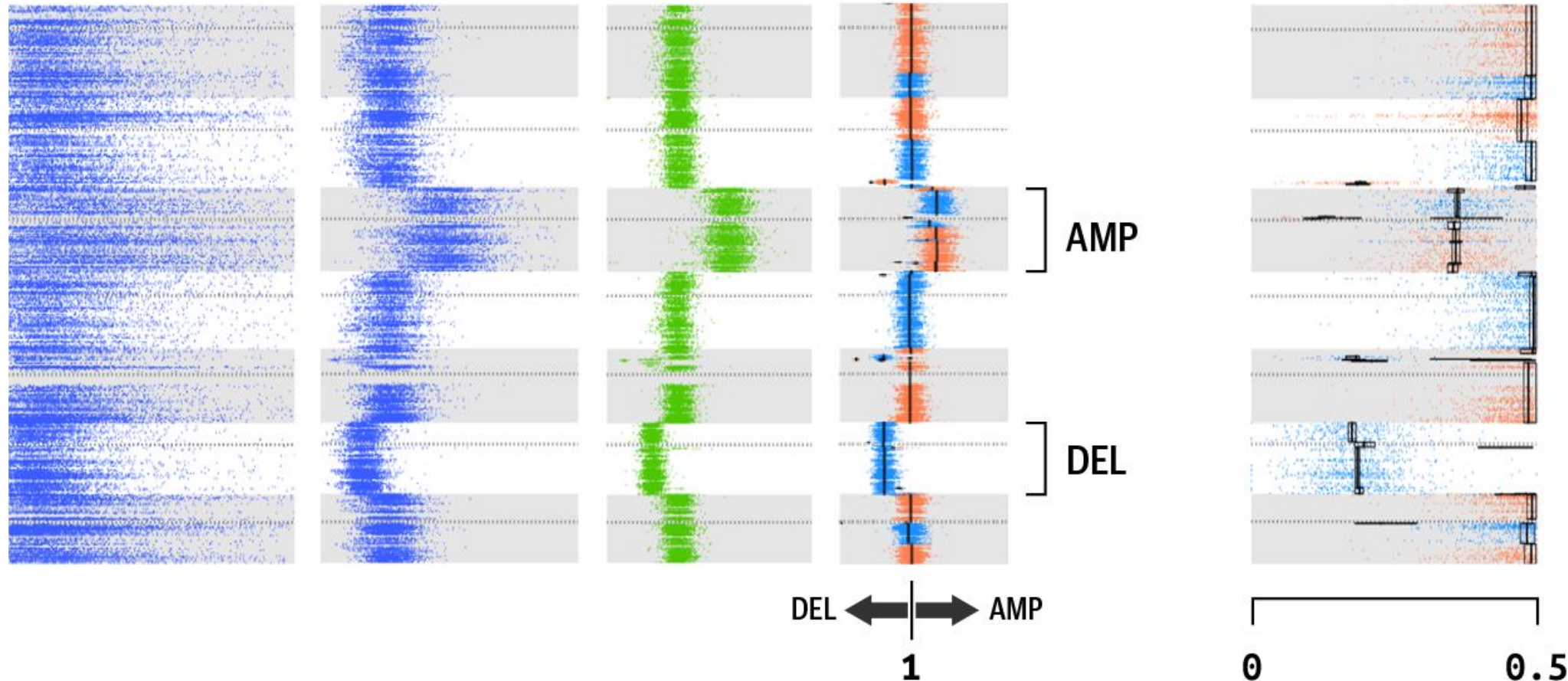
```
# gatk CallCopyRatioSegments \  
  -I sandbox/hcc1143_T_clean.cr.seg \  
  -O sandbox/hcc1143_T_clean.called.seg
```

```
# tail -5 sandbox/hcc1143_T_clean.called.seg  
chrX 118974529 139746109 864 0.475183 +  
chrX 139749773 139965748 28 -0.925385 -  
chrX 140503468 153058699 277 -0.366860 -  
chrX 153182138 153580550 47 0.658197 +  
chrX 153588113 156010661 544 0.075279 0
```



# Full progression from raw data to results

Raw copy ratio to called segments



# Additional resources

- GATK documentation: Somatic variant discovery (SNVs + Indels)
  - <https://gatk.broadinstitute.org/hc/en-us/articles/360035894731>
- About Mutect2
  - <https://gatk.broadinstitute.org/hc/en-us/articles/360051306691-Mutect2>
- About Funcotator
  - <https://gatk.broadinstitute.org/hc/en-us/articles/360051304411-Funcotator>
- GATK documentation: Functional annotations
  - <https://gatk.broadinstitute.org/hc/en-us/articles/360035531732-Funcotator-Annotation-Specifications>



# Additional resources (continued)

- About HDF5 format
  - <https://gatk.broadinstitute.org/hc/en-us/articles/360035531712?id=11508>
- About somatic copy number variant discovery (CNV)
  - <https://gatk.broadinstitute.org/hc/en-us/articles/360035535892>
- BroadE: GATK - Introduction to Somatic Variant Discovery
  - [https://youtu.be/0q5\\_e2Nfph4](https://youtu.be/0q5_e2Nfph4)





A pufferfish is shown in profile, facing right. Its body is divided into two distinct horizontal sections. The upper section is dark grey or black, covered in numerous small, irregular white spots. The lower section is a lighter, yellowish-tan color with a fine, grid-like or woven texture. The fish's head is slightly inflated, and its mouth is small and closed. The background is a plain, light grey.

# Thank you for joining us today!

Next week: Chapter 8

Next meeting: January 25, 2021