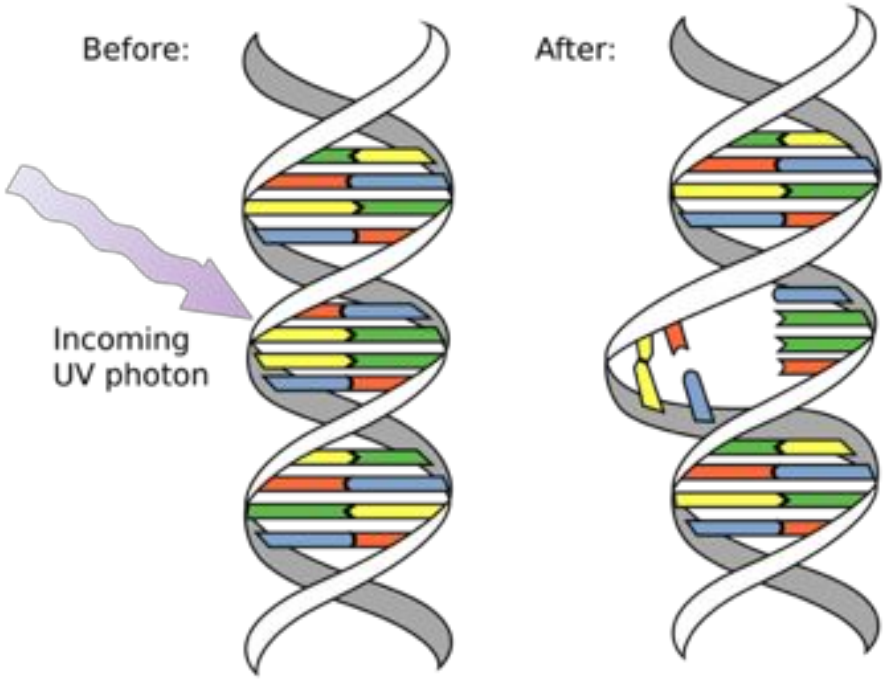
Somatic Mutation Calling

BFX Workshop Week 08

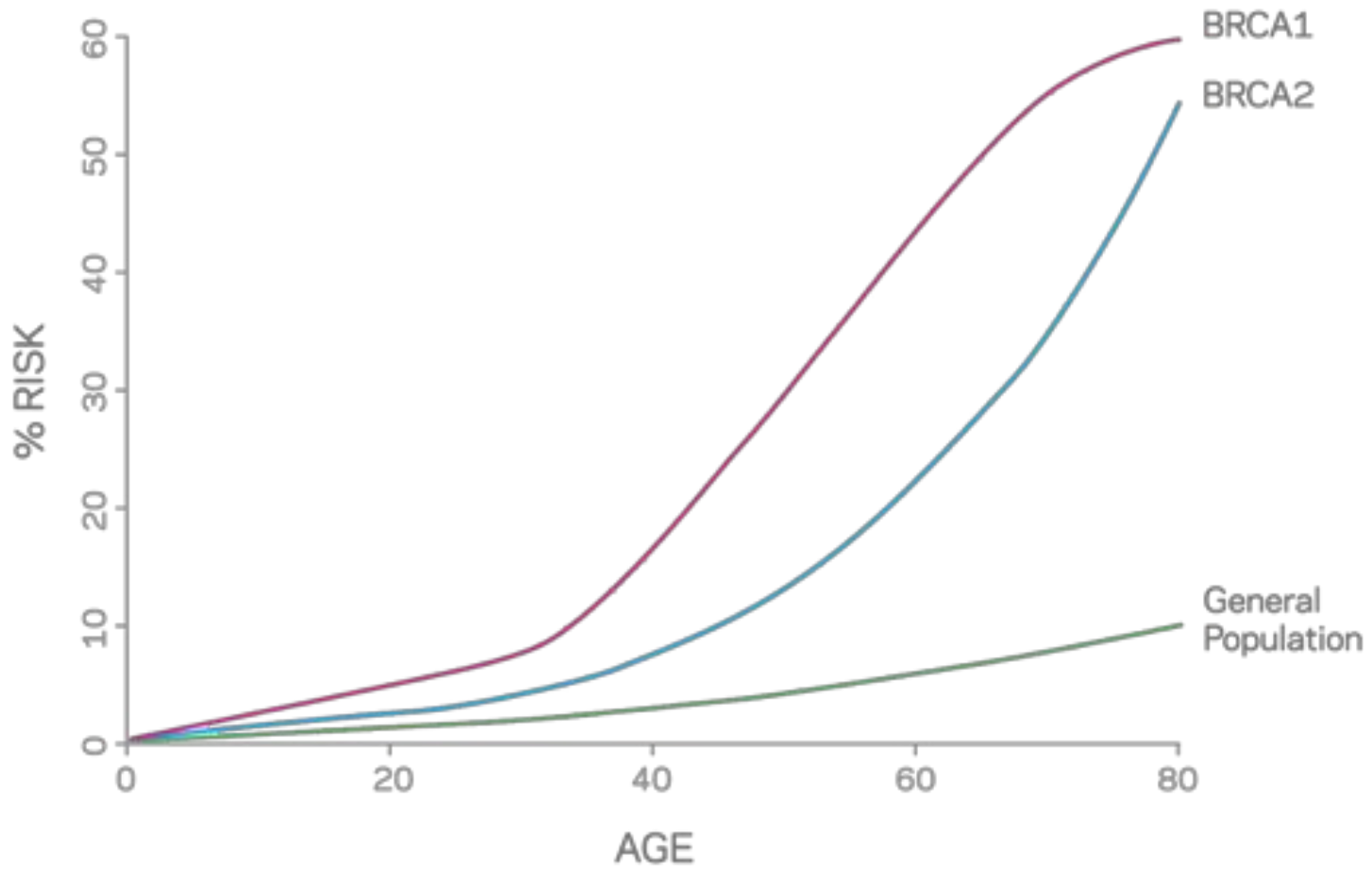
Chris Miller

Cancer is a disease of the genome

- Cancer is caused by **somatic** mutations
- These mutations are introduced into the genome of a cell (errors in DNA copying, UV light, chemicals)
- Most cancers require around 3 driver mutations



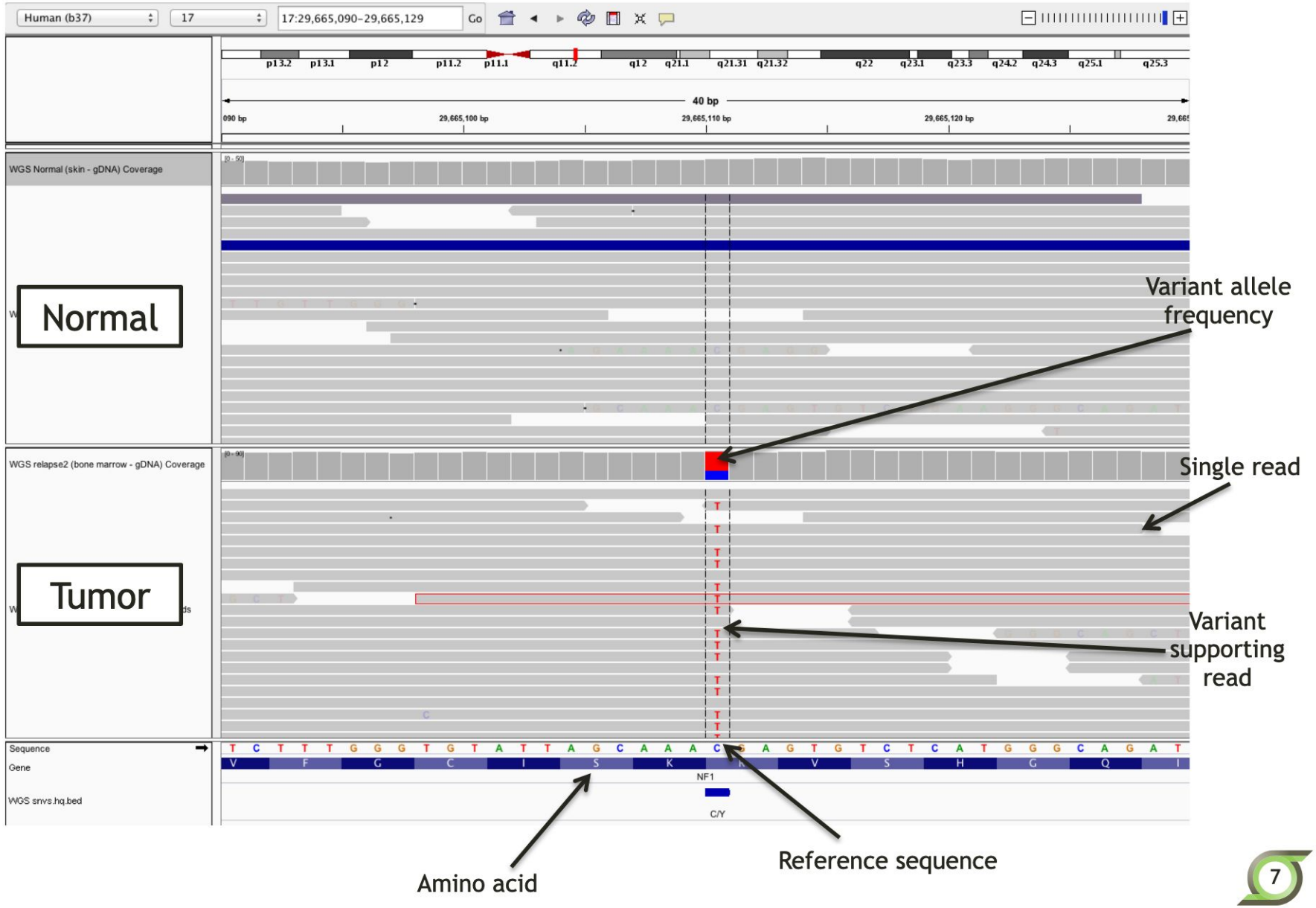
Germline Predisposition



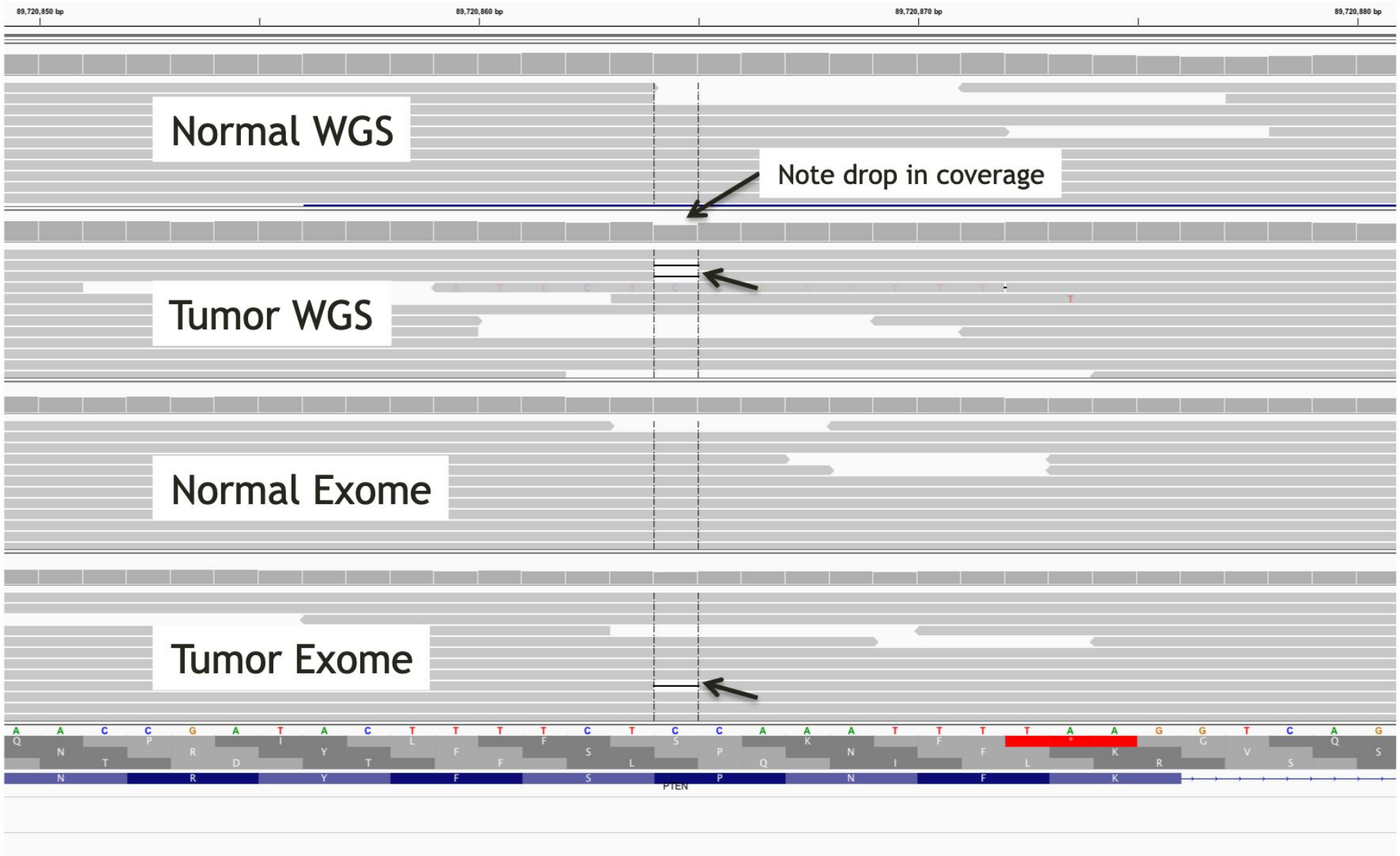
Cancer Sequencing

- In cancer, we have to (at least) double sequencing costs
- Uses both a tumor sample and a matched normal
- We compare them to find somatic mutations

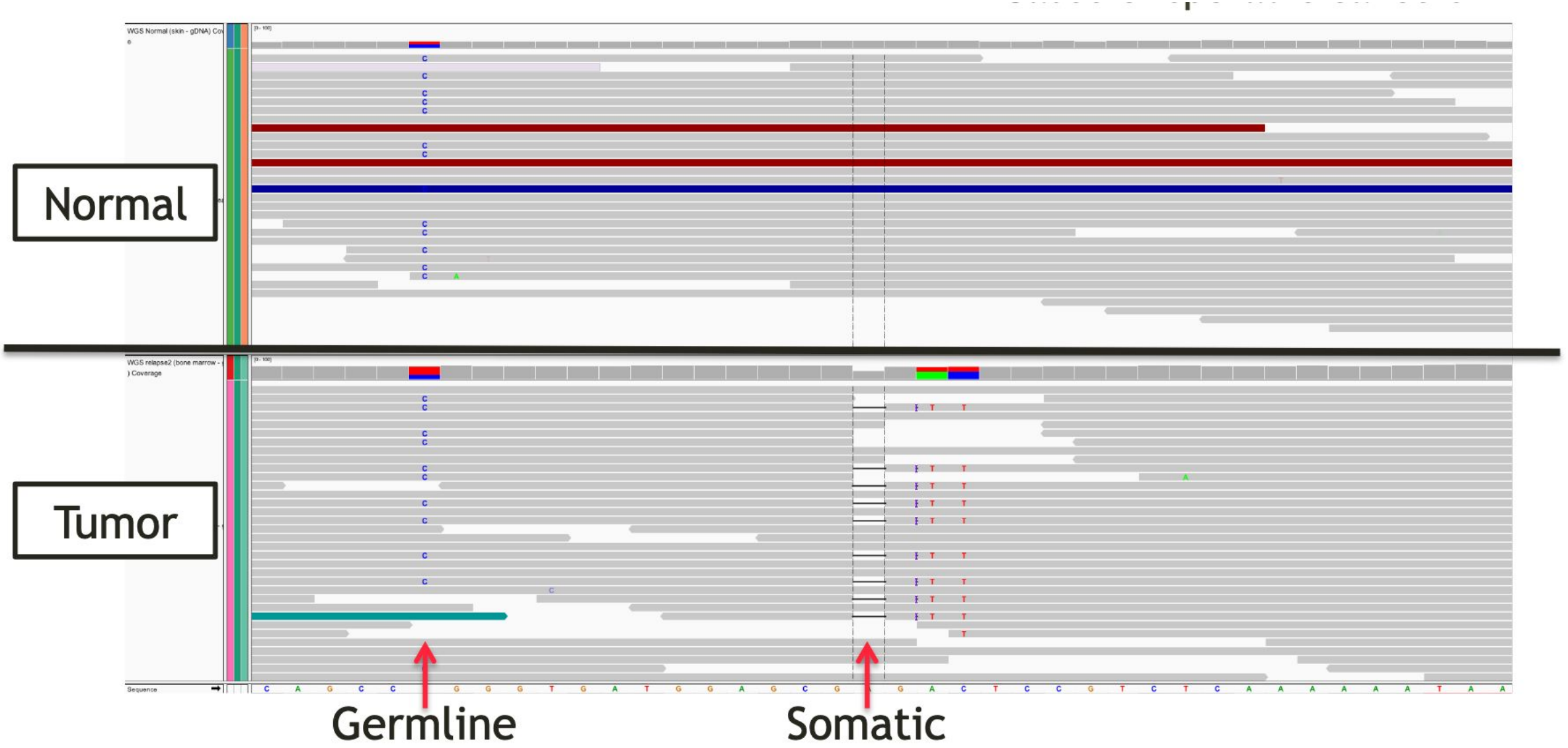
What do somatic variants look like?



Indels


89,720,850 bp
89,720,860 bp
89,720,870 bp
89,720,880 bp
Normal WGS
Note drop in coverage
Tumor WGS
Normal Exome
Tumor Exome
PTEN

Germline vs Somatic

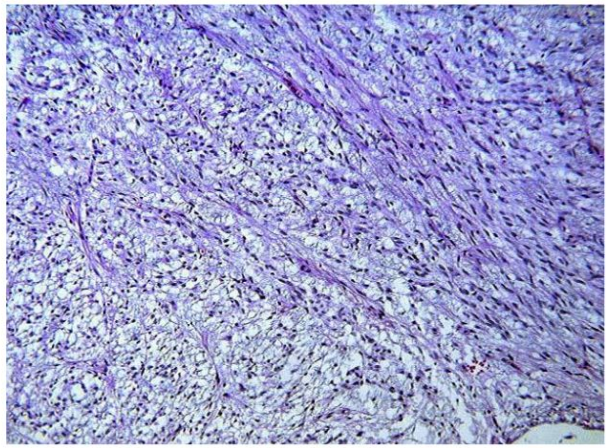


VAF = Variant reads / Total reads

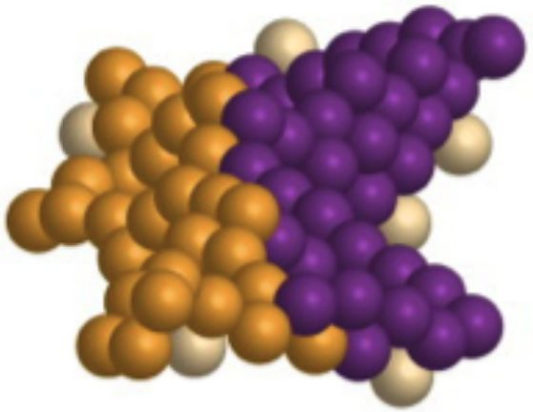
VAF = 0/20 = 0%

VAF = 14/20 = 70%

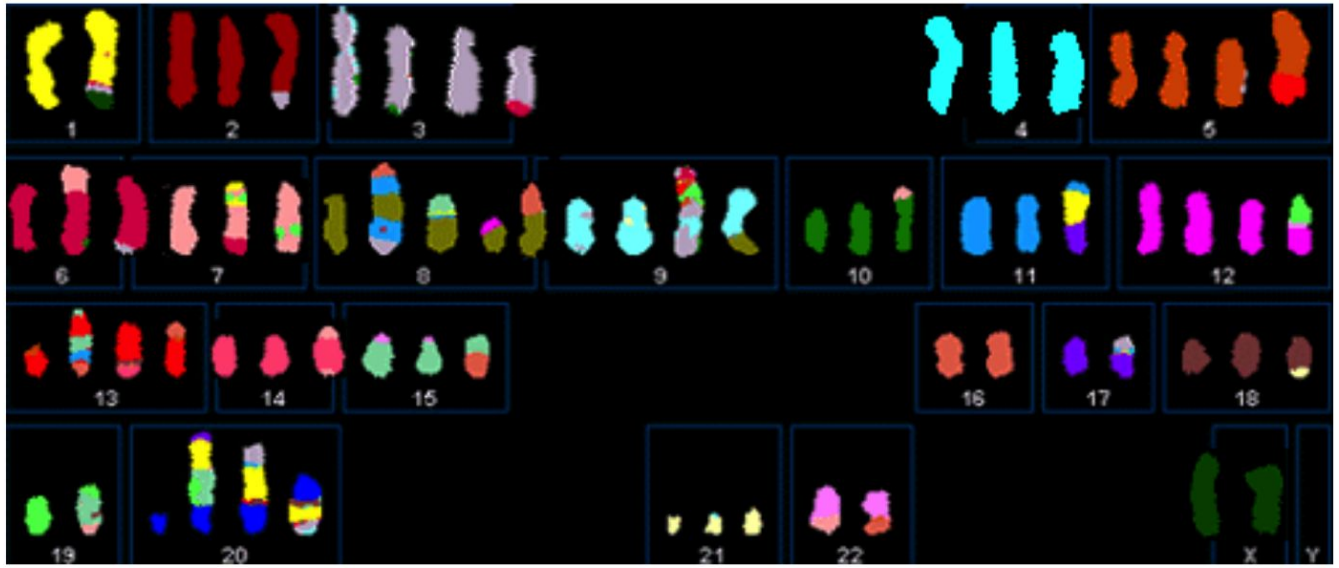
Tumors are often impure, heterogeneous, and aneuploid



Tumors are often impure
(contain normal cells)



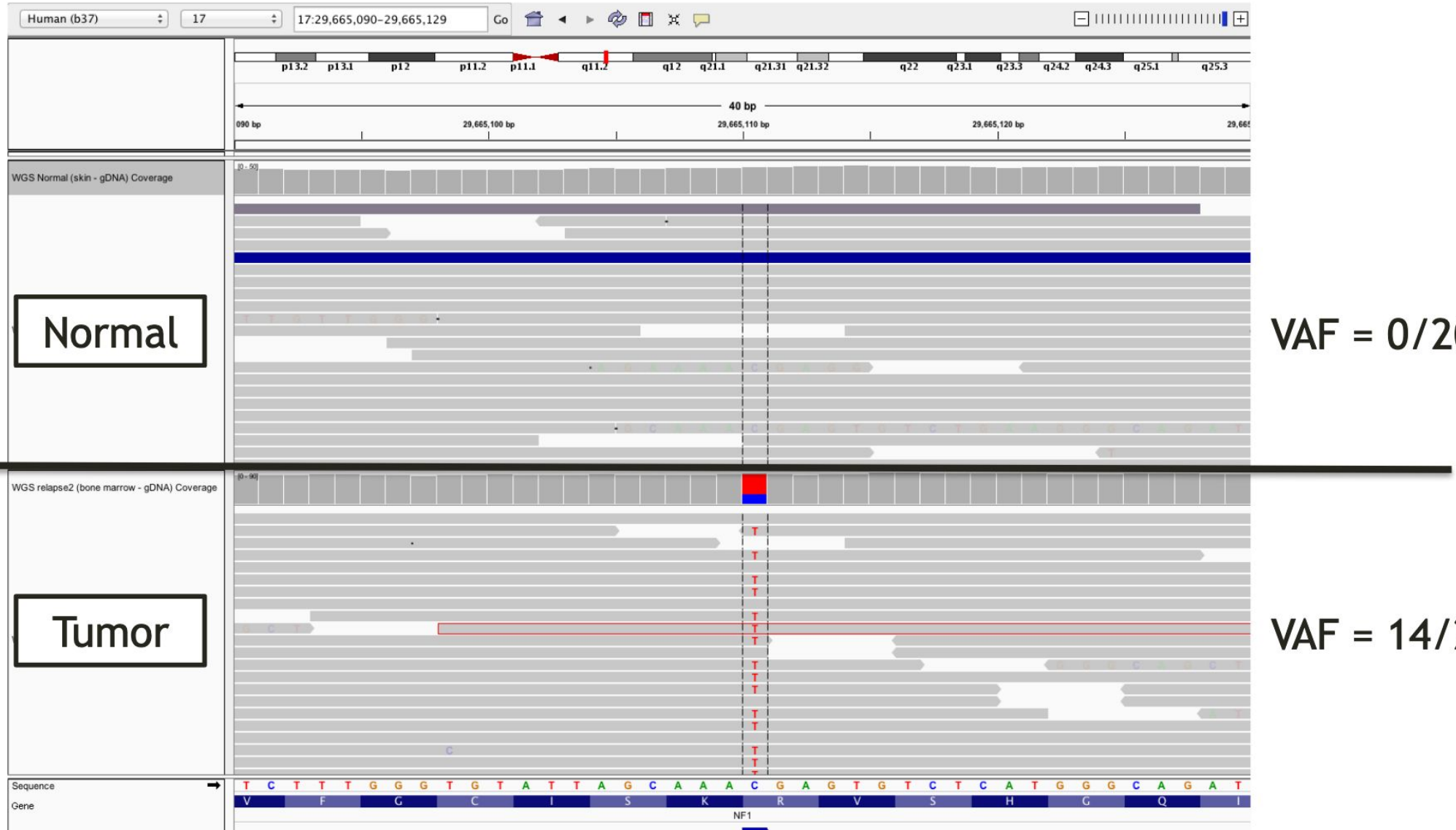
Tumors are often genetically
diverse collections of cells



Tumors may be aneuploid

How does purity influence VAF?

VAF = Variant reads / Total reads

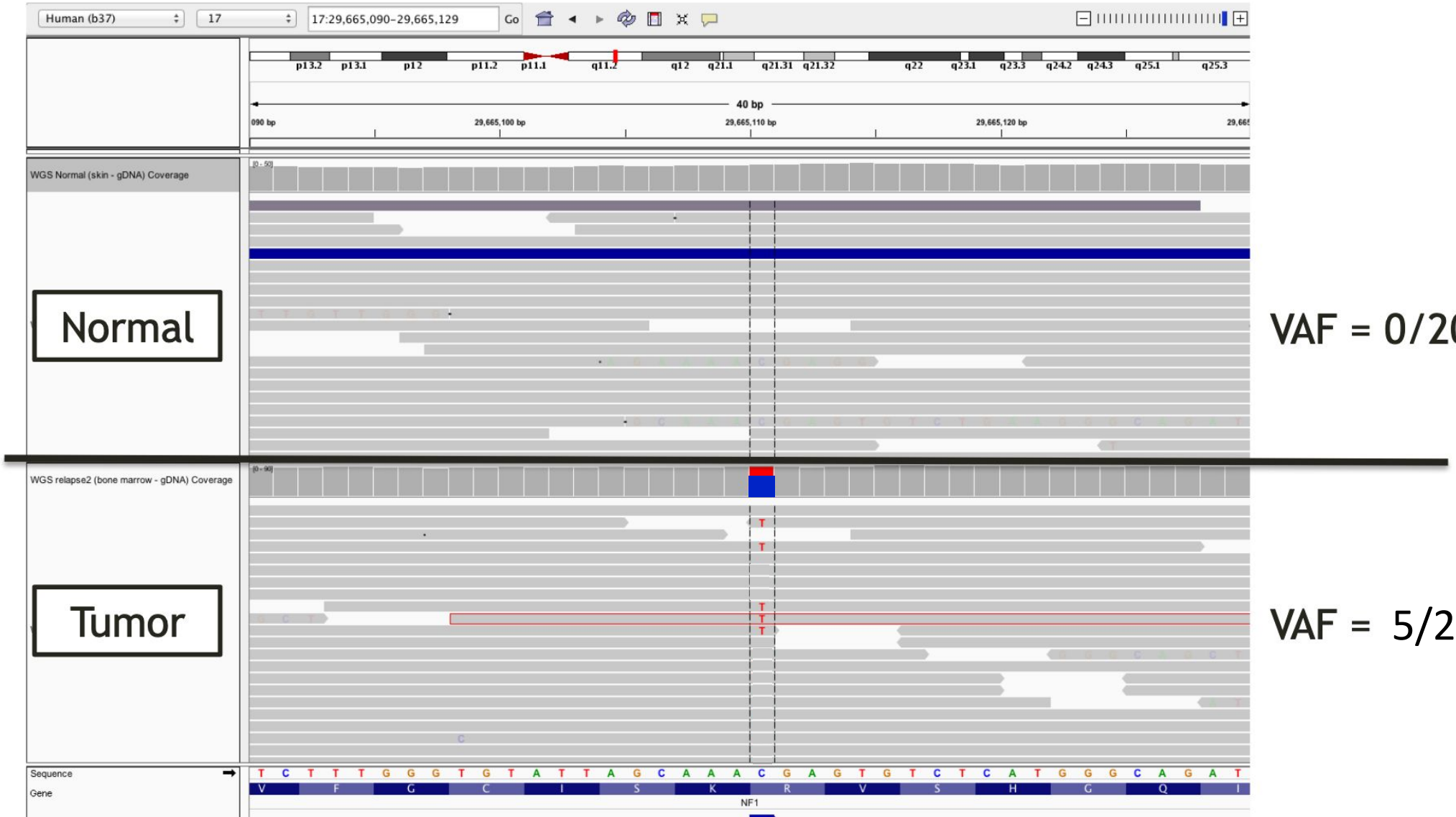


VAF = 0/20 = 0%

VAF = 14/20 = 70%

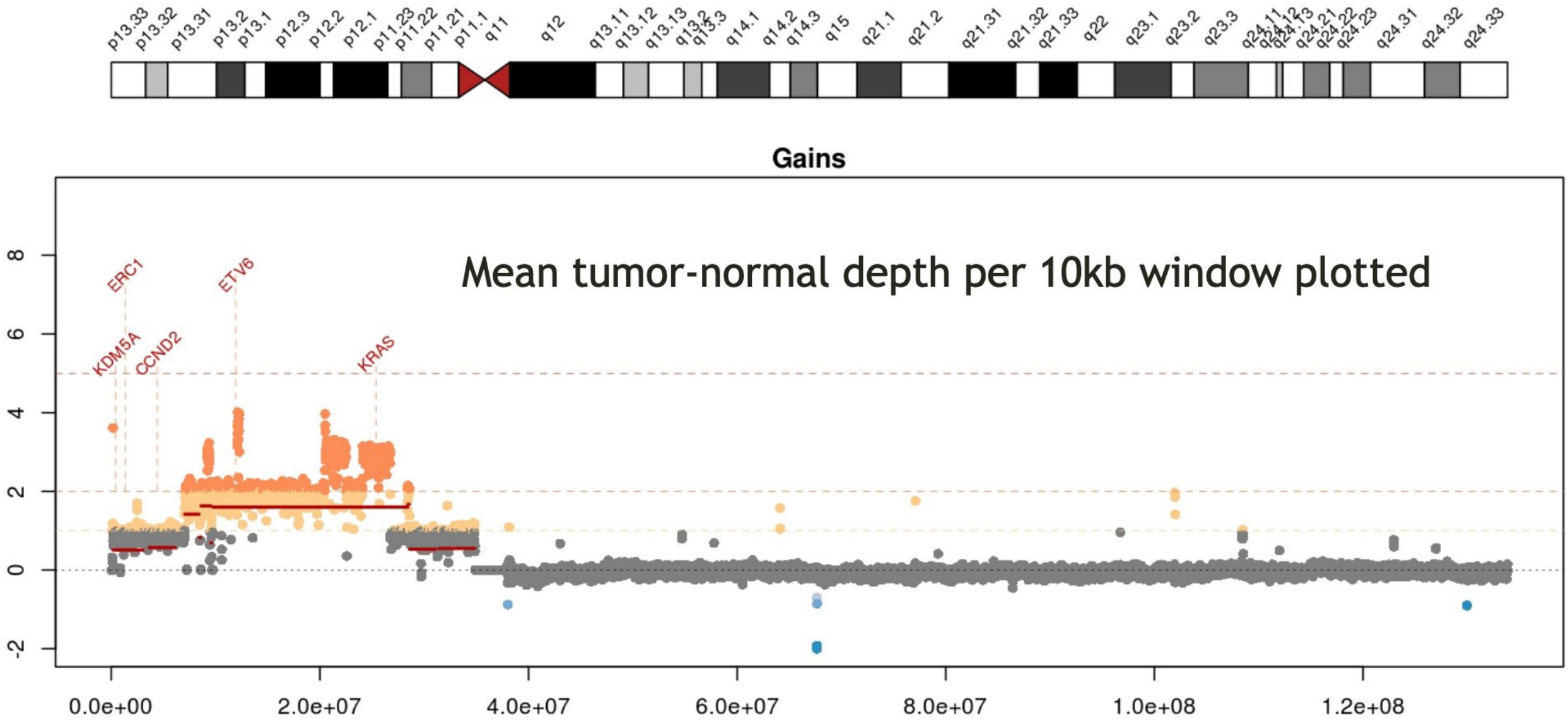
How does purity influence VAF?

VAF = Variant reads / Total reads



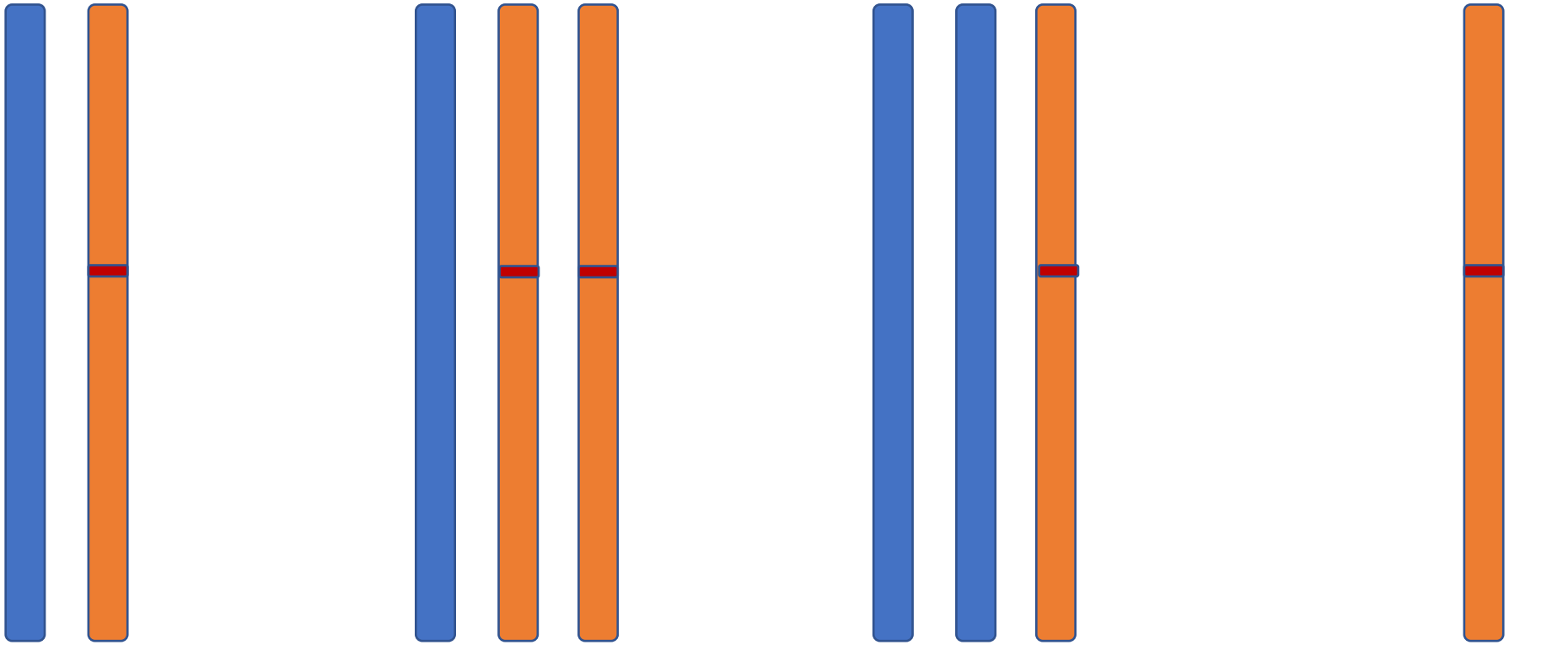
VAF = 0/20 = 0%

VAF = 5/20 = 25%



KRAS amplification in a metastatic breast cancer

How does copy number influence VAF?



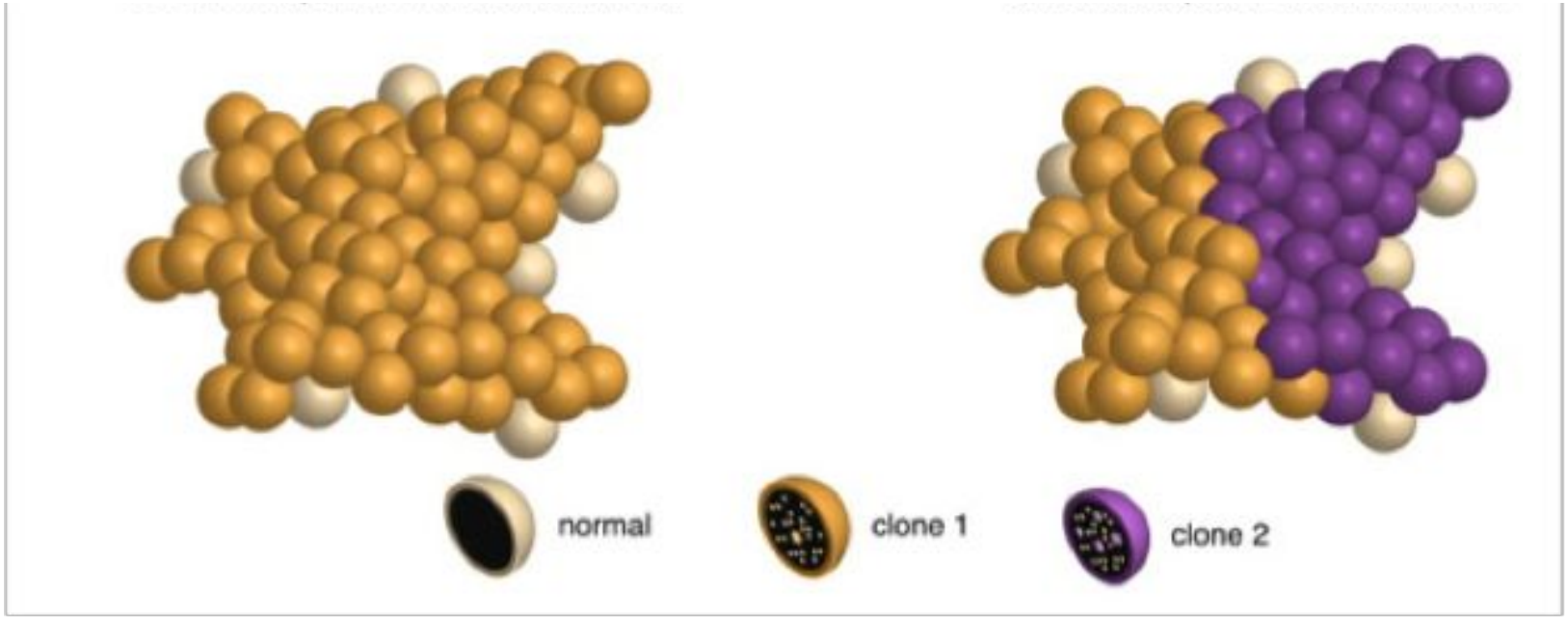
2 copies
50% VAF

3 copies
66% VAF

3 copies
33% VAF

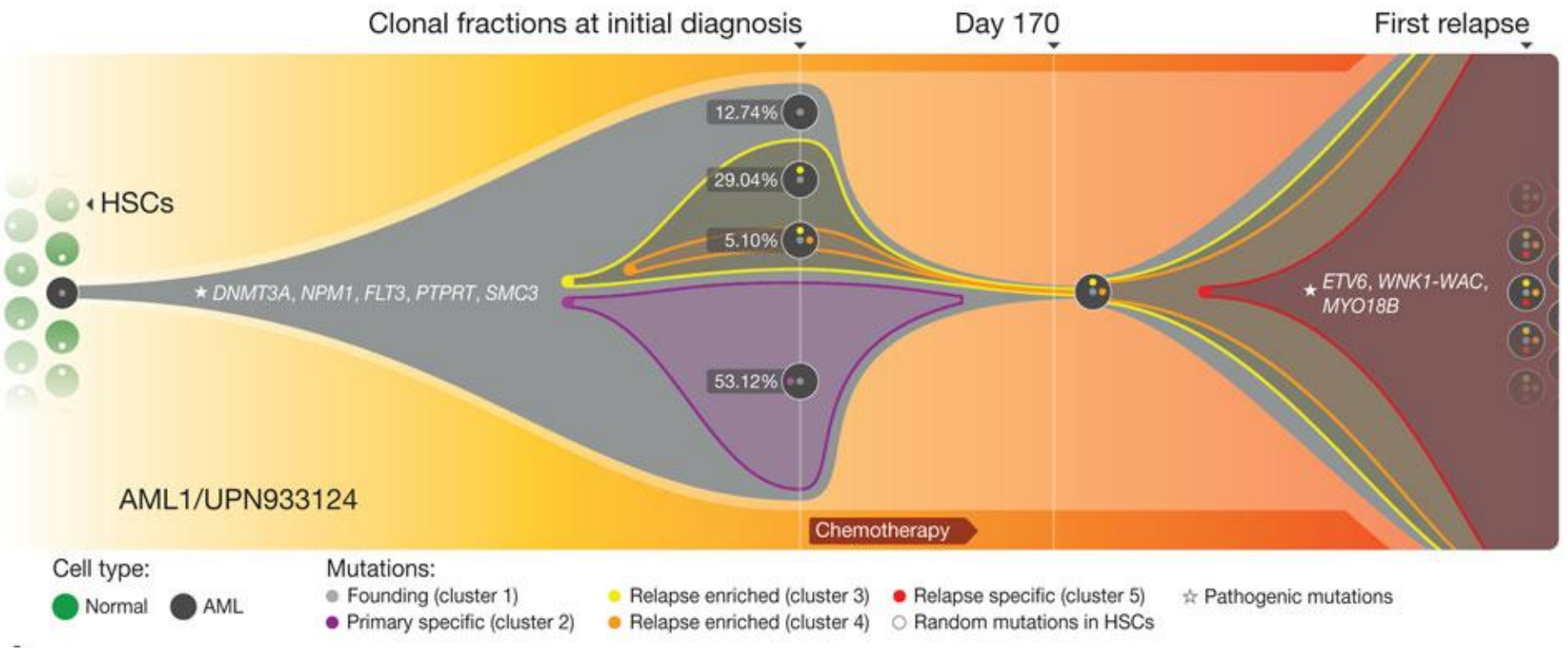
1 copy
100% VAF

How does clonality influence VAF?



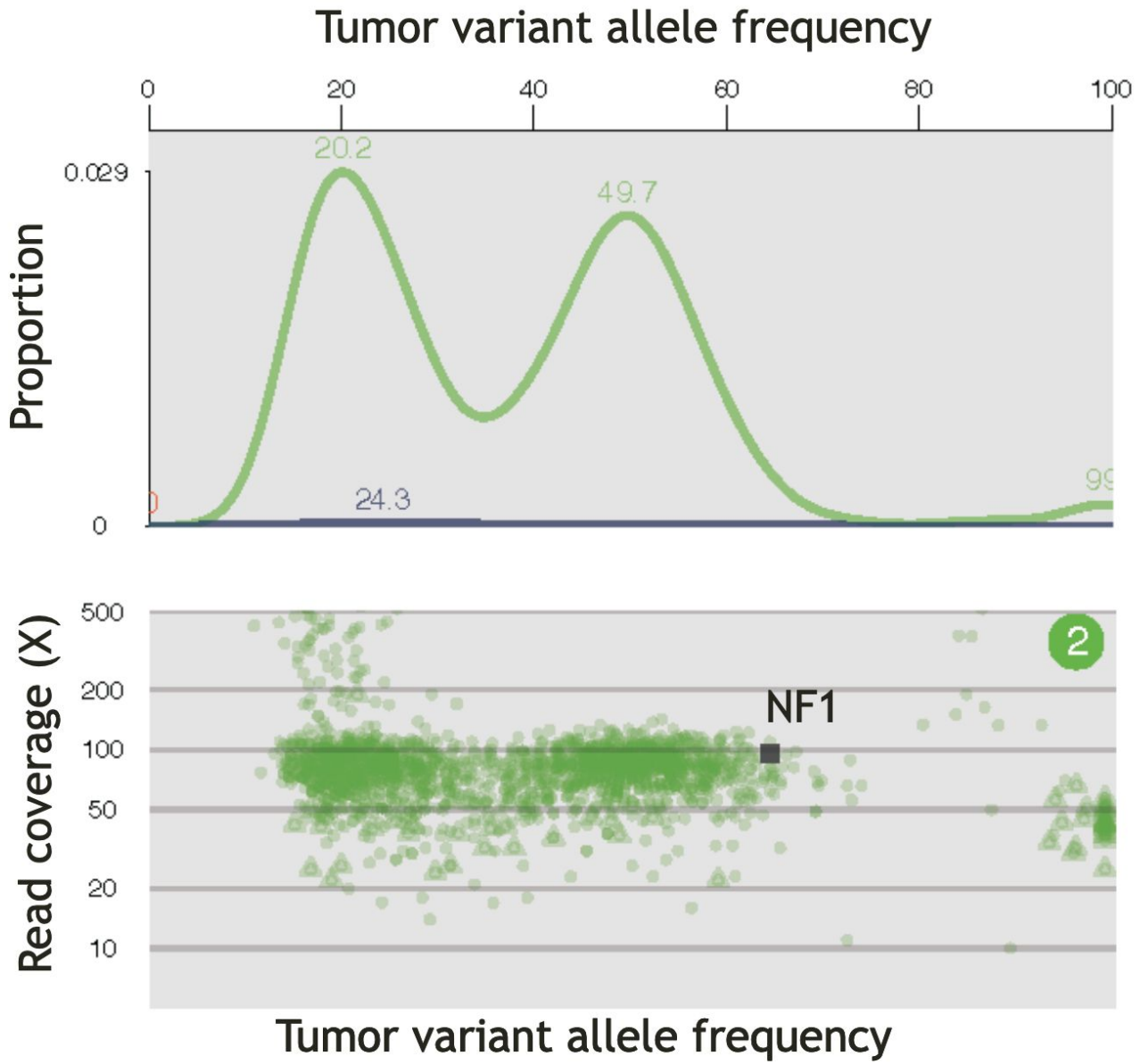
- Subclones contain genetically diverse populations of cells
- Evolution occurs at the molecular and cellular levels
- The growth rates for subclones are often different

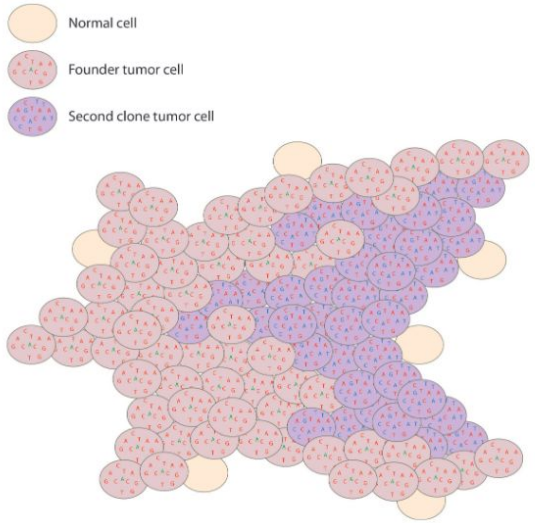
Clonal evolution in relapsed AML



Nature 481, 506–510 doi:10.1038/nature10738

Dominant clone vs. sub-clonal (and driver vs. passenger)





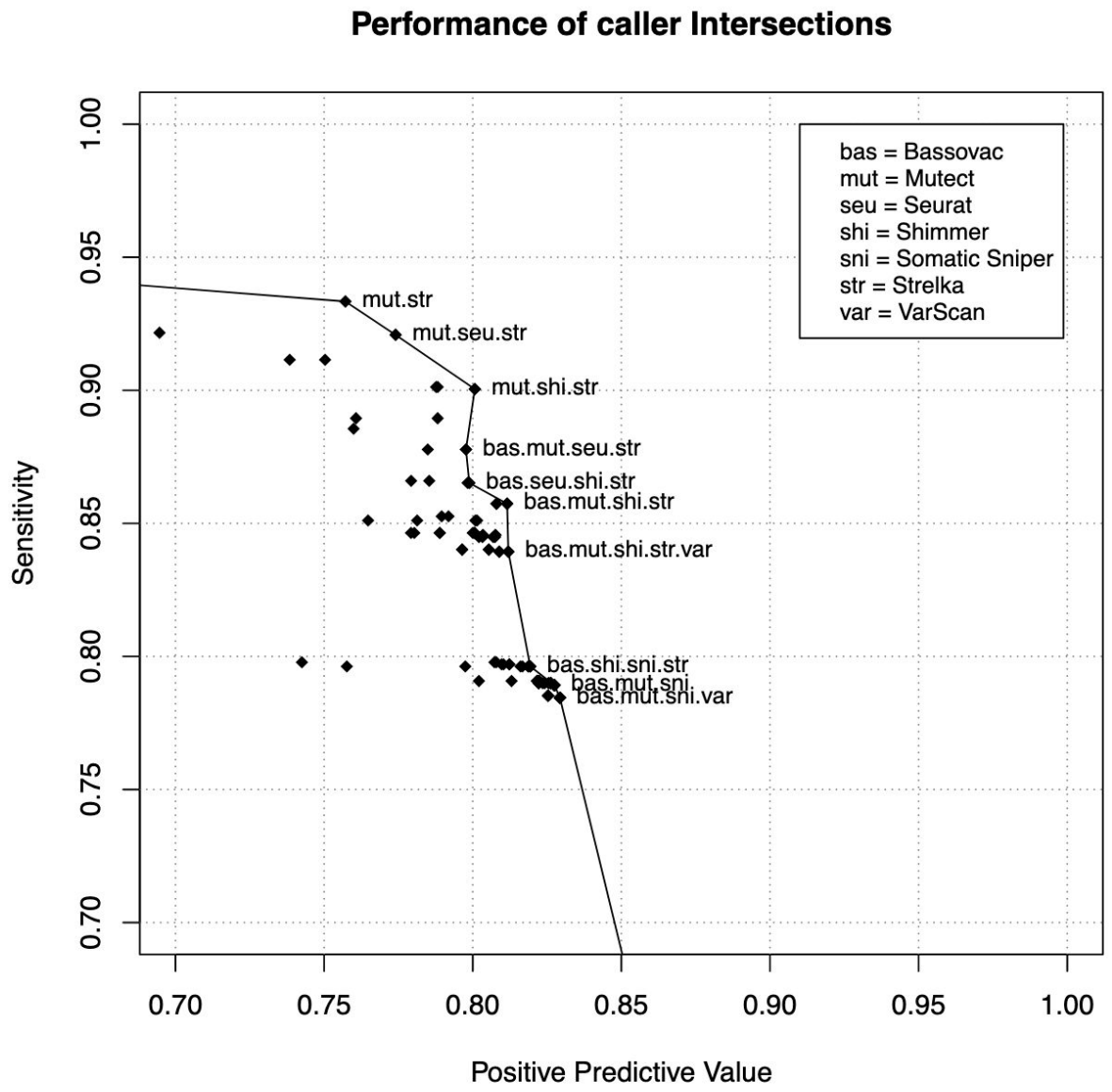
WGS data

Somatic variant calling is harder

- There are more factors to consider, a wider range of possibilities, and often, more sketchy samples

- One class of callers tries to address this problem with Bayesian statistics

Use of multiple variant callers can improve sensitivity and accuracy



Somatic Variant detection workflow

- Step 1 – run the callers
  - Mutect, Strelka, Varscan, Pindel

Somatic Variant detection workflow

- Varscan
  - `java -jar VarScan.v2.3.6.jar somatic <(samtools mpileup … normal.bam) <(samtools mpileup … tumor.bam)`
- Mutect
  - `/gatk/gatk Mutect2 -O output.vcf -R ref.fa -I tumor.bam -tumor TUMORNAME -I nomal.bam -normal NORMALNAME -L interval.list`

Somatic Variant detection workflow

- Step 1 – run the callers
  - Mutect, Strelka, Varscan, Pindel

- Step 2 – Merge the VCFs
  - GATK - CombineVariants

Somatic Variant detection workflow

- Combine Variants

```
/usr/bin/java -jar /opt/GenomeAnalysisTK.jar -T
CombineVariants -genotypeMergeOptions PRIORITIZE
--rod_priority_list mutect,varscan -o combined.vcf.gz
--variant:mutect mutect_output.vcf --variant:varscan
varscan_output.vcf
```

Somatic Variant detection workflow

- Step 1 – run the callers
  - Mutect, Strelka, Varscan, Pindel
- Step 2 – Merge the VCFs
  - GATK - CombineVariants
- Step 3: Annotate and filter

Homework

- Take a tumor and normal bam file and identify somatic variants using multiple callers
- Produce a merged VCF with your results