When Good Experiments Go Bad

Chris Miller
Applied Computational Genomics I
BFX Workshop - Week 12

However improbable we regard this event, or any of the steps which it involves, given enough time it will almost certainly happen at least once.

--George Wald

Anything that can go wrong, will go wrong

--Murphy

Shit happens.

--Forrest Gump

Case #1

- Exome sequencing – Glioblastoma Tumor/Normal pairs
- Alignment, somatic variant calling, filtering

```
$ wc -l H_RL-01-0*/snvs.indels.annotated

  159 H_RL-01-0203-1412449/snvs.indels.annotated
  219 H_RL-01-0216-1412454/snvs.indels.annotated
10583 H_RL-01-0334-1412447/snvs.indels.annotated
```

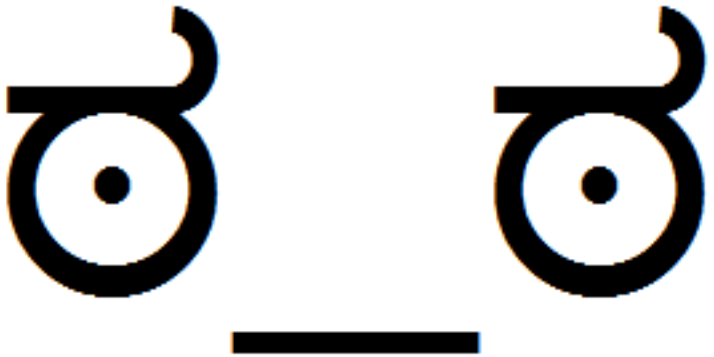
```
$ wc -l H_RL-01-0*/snvs.indels.annotated

  159 H_RL-01-0203-1412449/snvs.indels.annotated
  219 H_RL-01-0216-1412454/snvs.indels.annotated
10583 H_RL-01-0334-1412447/snvs.indels.annotated
```

```
$ wc -l H_RL-01-0*/snvs.indels.annotated

  159 H_RL-01-0203-1412449/snvs.indels.annotated
  219 H_RL-01-0216-1412454/snvs.indels.annotated
34-1412447/snvs.indels.annotated
```

10583

```
$ wc -l H_RL-01-0*/snvs.indels.annotated

  159 H_RL-01-0203-1412449/snvs.indels.annotated
  219 H_RL-01-0216-1412454/snvs.indels.annotated
34-1412447/snvs.indels.annotated
```

10583



```
$ wc -l H_RL-01-0*/snvs.indels.annotated

  159 H_RL-01-0203-1412449/snvs.indels.annotated
  219 H_RL-01-0216-1412454/snvs.indels.annotated
10583 H_RL-01-0334-1412447/snvs.indels.annotated
```

- How many of these occur at known dbSNP sites?

```
$ wc -l H_RL-01-0*/snvs.indels.annotated

  159 H_RL-01-0203-1412449/snvs.indels.annotated
  219 H_RL-01-0216-1412454/snvs.indels.annotated
10583 H_RL-01-0334-1412447/snvs.indels.annotated
```

- How many of these occur at known dbSNP sites?
  ~85%

```
$ wc -l H_RL-01-0*/snvs.indels.annotated

  159 H_RL-01-0203-1412449/snvs.indels.annotated
  219 H_RL-01-0216-1412454/snvs.indels.annotated
10583 H_RL-01-0334-1412447/snvs.indels.annotated
```

- How many of these occur at known dbSNP sites?
  ~85%

- What is their VAF?

```
$ wc -l H_RL-01-0*/snvs.indels.annotated

  159 H_RL-01-0203-1412449/snvs.indels.annotated
  219 H_RL-01-0216-1412454/snvs.indels.annotated
10583 H_RL-01-0334-1412447/snvs.indels.annotated
```

- How many of these occur at known dbSNP sites?
  ~85%

- What is their VAF?
  ~50% or 100%

```
$ wc -l H_RL-01-0*/snvs.indels.annotated

  159 H_RL-01-0203-1412449/snvs.indels.annotated
  219 H_RL-01-0216-1412454/snvs.indels.annotated
10583 H_RL-01-0334-1412447/snvs.indels.annotated
```

- How many of these occur at known dbSNP sites?
  ~85%

- What is their VAF?
  ~50% or 100%

dx: SAMPLE SWAP

Damage Control

- Check other samples in the cohort
  - May not be resolvable!

    S1 Tumor vs S1 Normal ######
    S2 Tumor vs S2 Normal ######

    S1 Tumor vs S2 Normal ###
    S2 Tumor vs S1 Normal ###

    are the tumors or normals swapped?

- Often need more information to resolve (RNAseq? Cytogenetics?)
- Check other lanes/indices on the same machine/batch

- Often, the resolution is to drop the samples

Case #2

- 2 projects, 3 patients
  - Patient 1 – Normal, CML, AML
  - Patient 2 – Normal, CML, AML
  - Patient 3 – Normal, Tumor

- Exome Sequencing
  - 1 lane of HiSeq2500 - 79-99X mean depth

Variant	Patient 1 CML	Patient 1 AML	Patient 2 CML	Patient 2 AML
Tier 1 SNVs	14430	669	5276	87
Tier 1 indels	255	115	55	17

How many SNVs are population variants?

Patient 1 AML - < 5% dbSNP

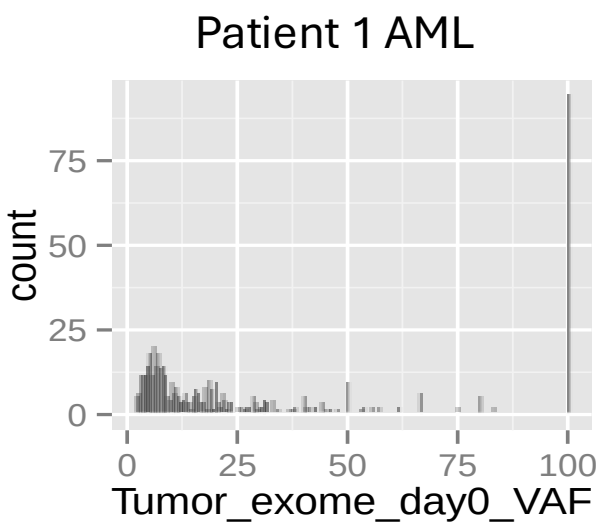
Patient 1 CML - > 90% dbSNP

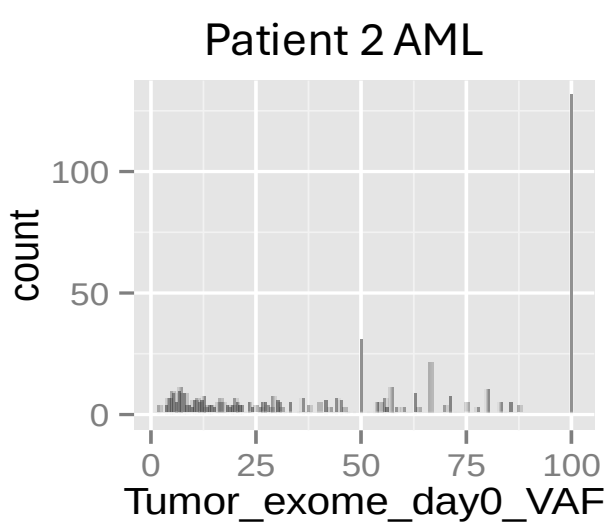
Patient 2 AML - < 5% dbSNP

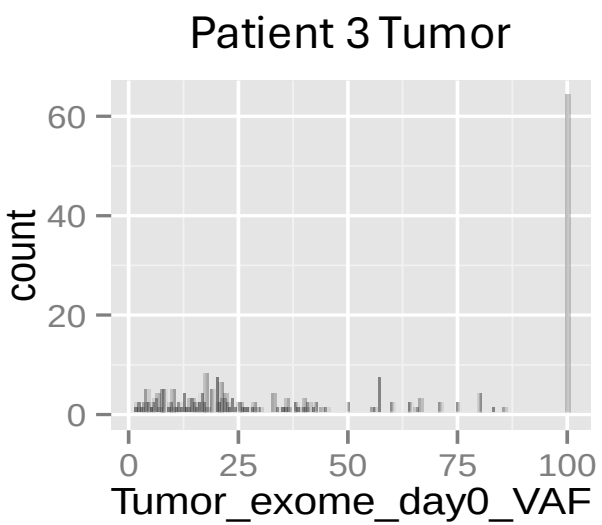
Patient 2 CML - > 90% dbSNP

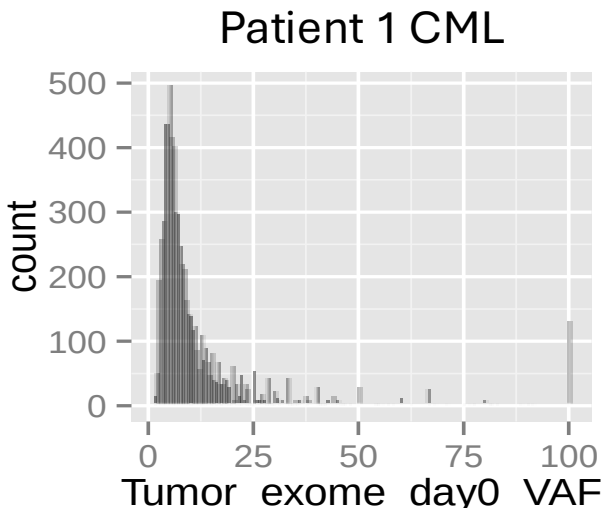
Patient 3 AML - < 5% dbSNP

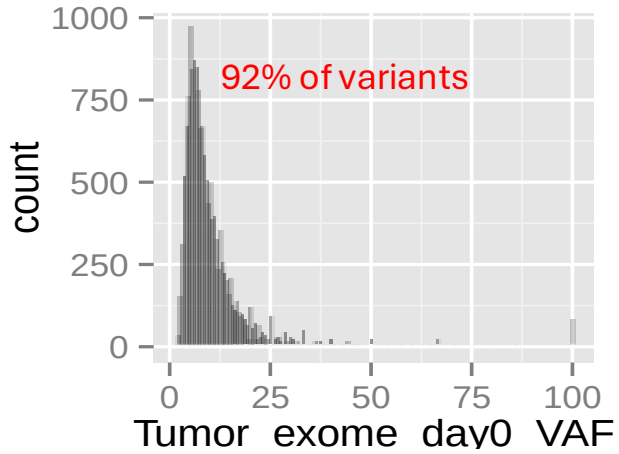
Coding variants with an rsID (dbSNP)



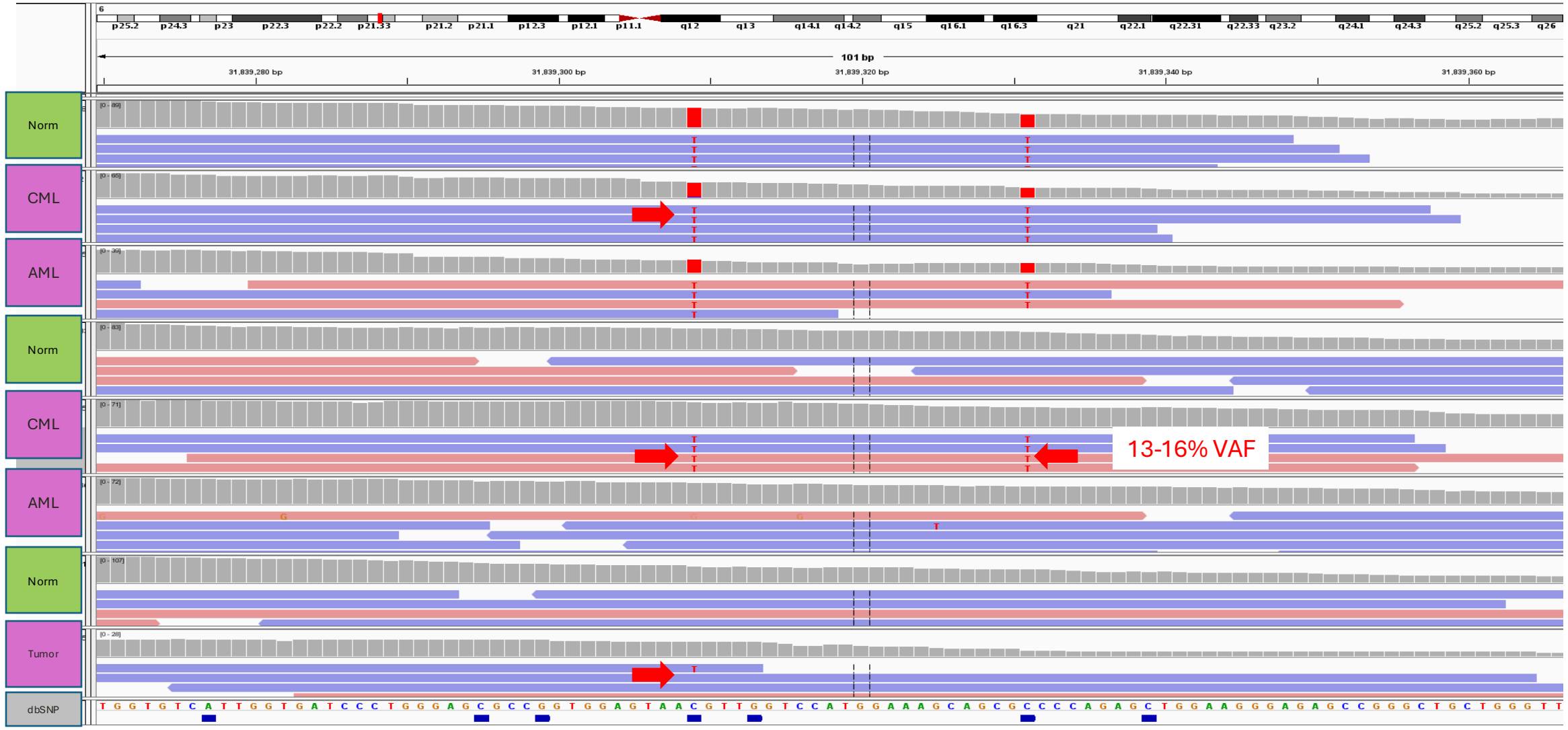




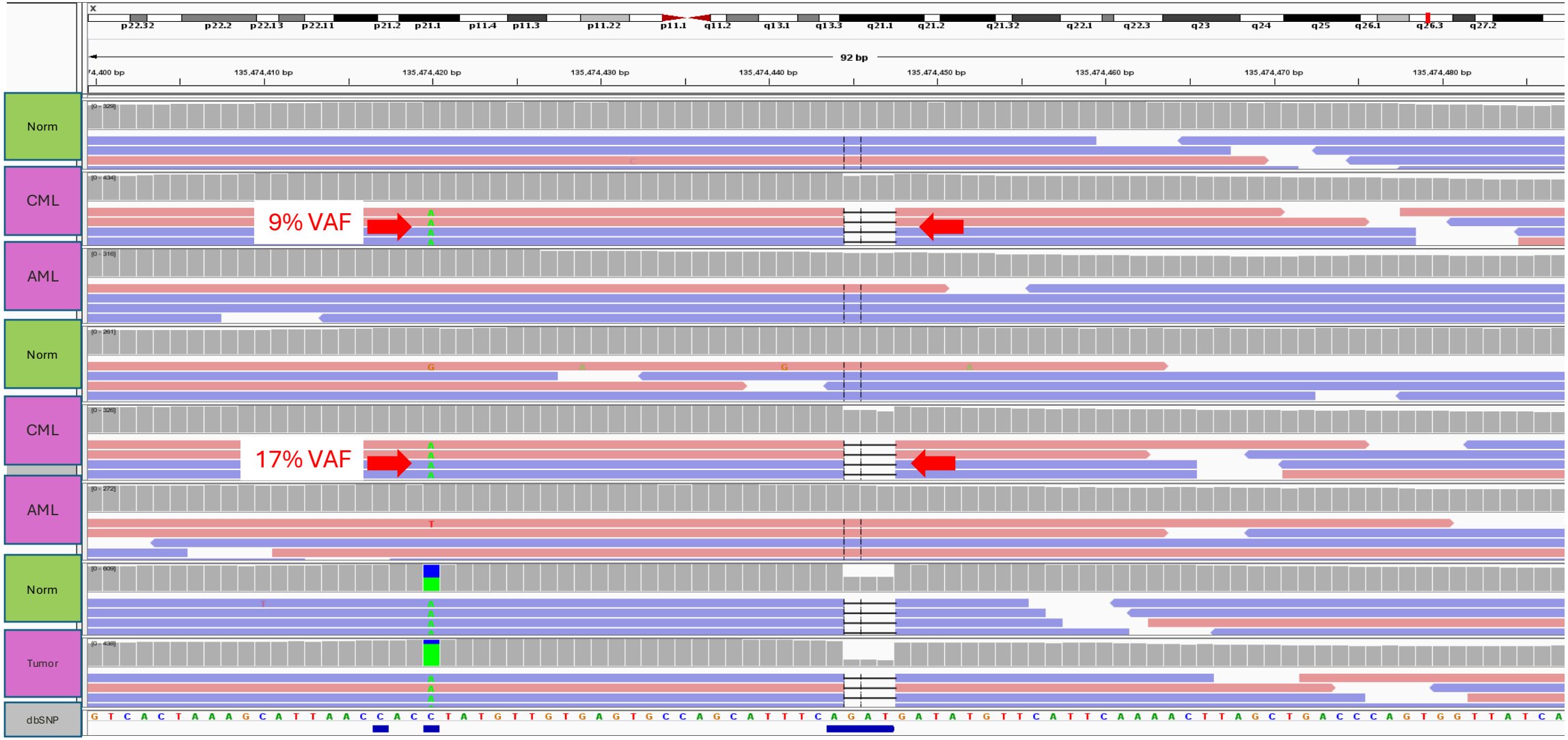




IGV inspection of variants/reads



IGV inspection of variants/reads



- How many of these occur at known dbSNP sites?
    most

- What is their VAF?
        NOT ~50% or 100%

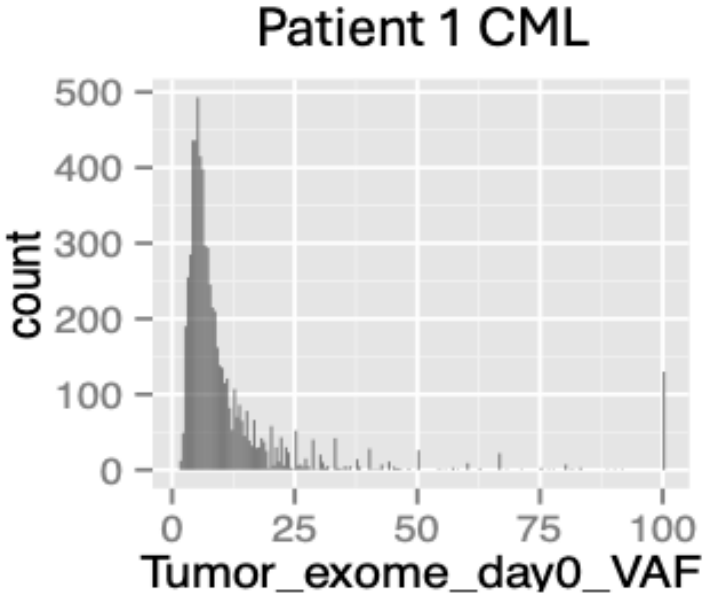
dx: SAMPLE CONTAMINATION

Damage Control

- Check other samples in the cohort to figure out source
  - May not be resolvable! (what if it's from someone else's samples?)

- If it's low-level enough, could apply filters
  - Only keep VAFs >30%
  - Downside: you may miss real events!

- Best solution is to make new libraries from the original source tissue



Case #3

- Single-cell RNA sequencing data
- Transcriptome alignment
  – we expect high level: 90%+
- Our data had ~10% alignment

This example is
good, ours was not!

Mapping

Reads Mapped to Genome	100.0%
Reads Mapped Confidently to Genome	21.4%
Reads Mapped Confidently to Intergenic Regions	2.6%
Reads Mapped Confidently to Intronic Regions	12.5%
Reads Mapped Confidently to Exonic Regions	6.3%
Reads Mapped Confidently to Transcriptome	16.3%
Reads Mapped Antisense to Gene	2.0%

Case #3

- Checked the kit - 3’ vs 5’ (matched)
- Checked the data – blatted a read at random

https://genome.ucsc.edu/cgi-bin/hgBlat

Human BLAT Search

BLAT Search Genome

Genome: ☐ Search all genomes — Human

Assembly: Dec. 2013 (GRCh38/hg38)

Query type: BLAT's guess

Sort output: query,score

Output type: hyperlink

☐ All Results (no minimum matches)

Submit | I'm feeling lucky | Clear

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

Human (hg38) BLAT Results

BLAT Search Results

Custom track name: blat YourSeq

Custom track description: blat on YourSeq

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	YourSeq	146	1	146	146	100.0%	chr2	-	25229049	25229194	146
browser details	YourSeq	21	81	102	146	100.0%	chr12	-	18672011	18672033	23

Case #3

- Checked the kit - 3’ vs 5’ (matched)
- Checked the data – blatted a read at random (matched to human)
- Checked 10x indices – do they appear in the whitelist (yes)
- Pulled our hair out, contacted production

Case #3

- Retraced our steps double checking all of our work

- Blatted a few more reads
  - They all matched poorly to the human genome
  - They all matched well to the mouse genome

Case #3

- Retraced our steps double checking all of our work

- Blatted a few more reads
  - They all matched poorly to the human genome
  - They all matched well to the mouse genome

dx: SPECIES MIXUP

- Just by chance, the first read we checked was from a very highly conserved gene!

Damage Control

- Realign to the correct species
  - Gave expected high alignment rate

- Still have to resolve what happened with the sample naming
  - Was the species designation just wrong?
  - Is the entire sample named wrong? (swap)

Xenograft contamination

- Related topic is dealing with Xenograft data
  - e.g. human tumors implanted in a mouse
- Mouse reads with homology to human genome
- One solution is Xenosplit – alignment-based read filtering
  - Human, mouse, ambiguous

Global alignment/mismatch issues

- Sample swaps
  - check SNP concordance
  - Somalier is a tool for rapid sample identity checking
- Contamination
  - VAFs, IGV inspection are your friends
- Species swaps
  - Check a few reads, some concordance is expected!

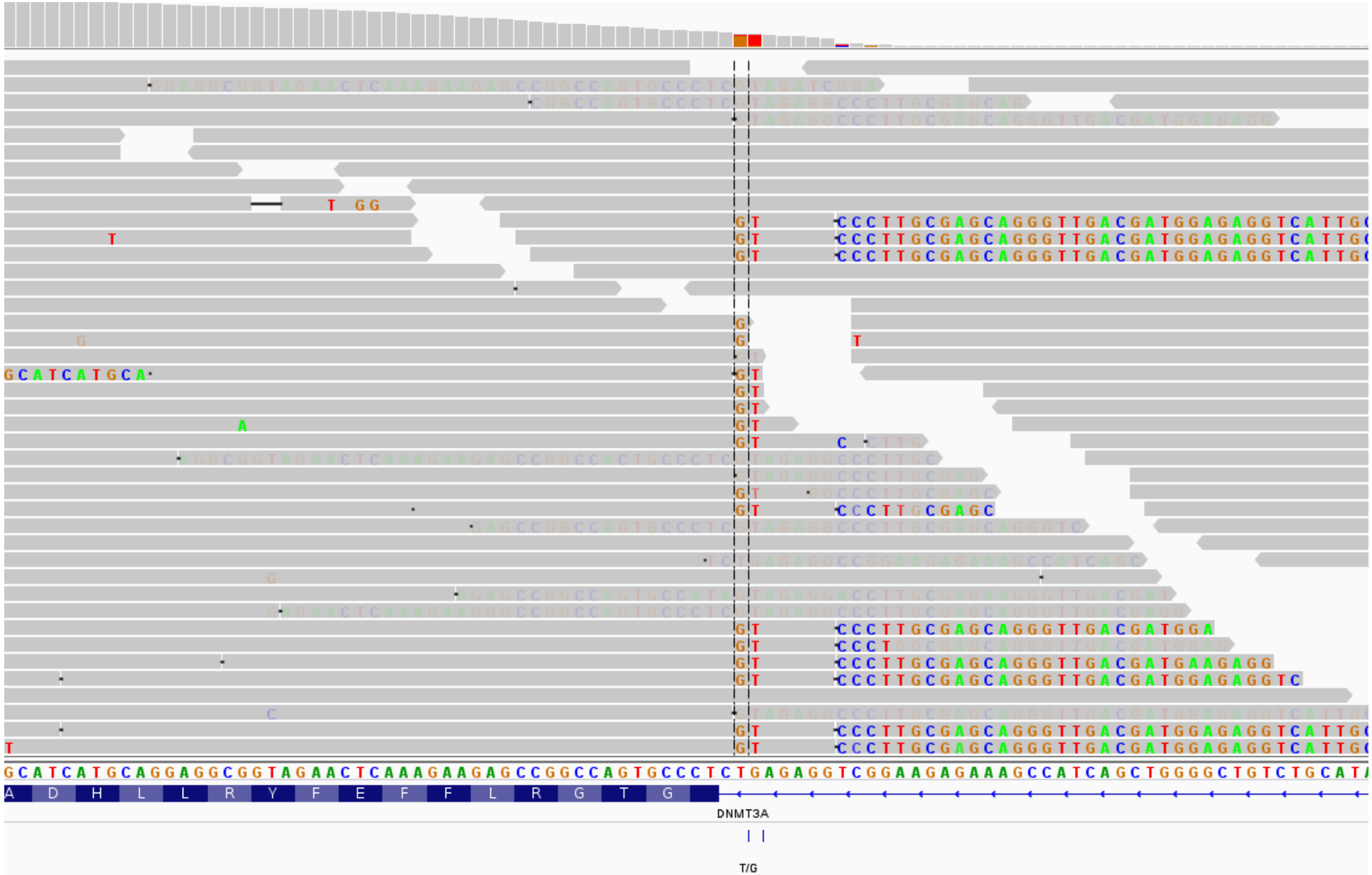
Case #4

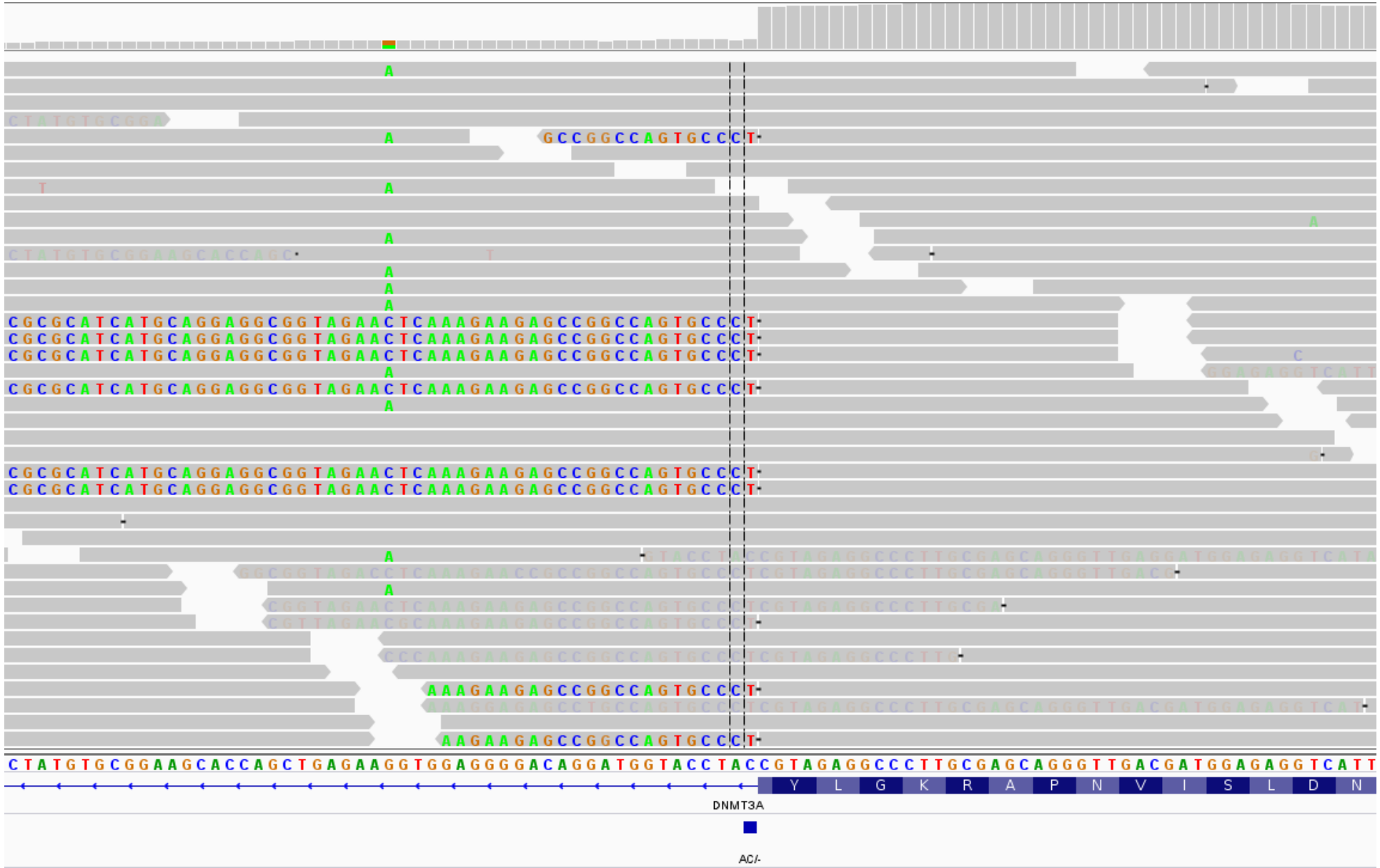
Mutations from a single AML sample, sorted by the number of times a single gene is hit

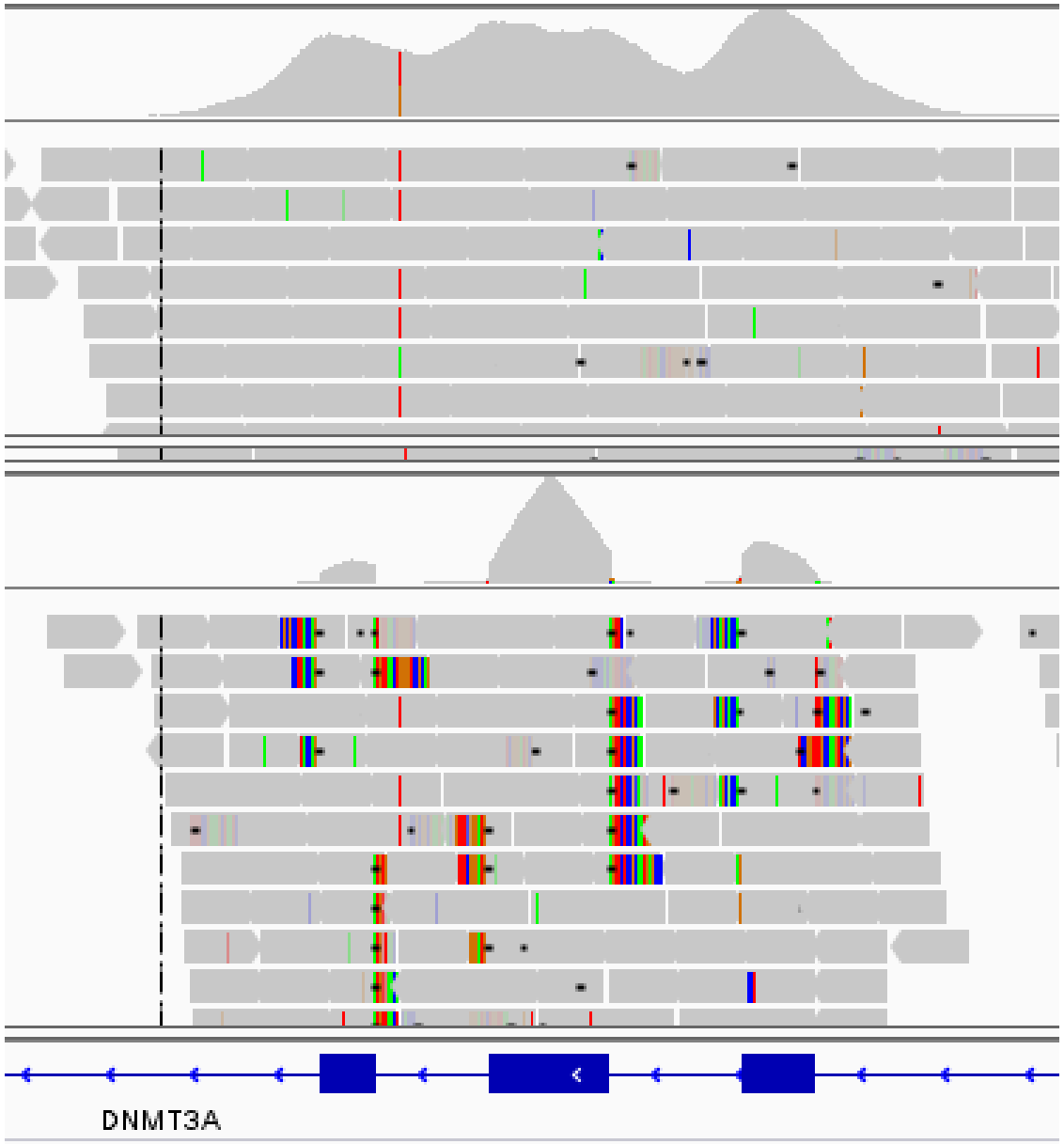
```
43 DNMT3A
 2 WT1
 2 SLC35F3
 1 UNC93B5
 1 TSLP
 1 TRPS1
 1 TARDBP
 1 SUN3
 1 SREBF1
 1 SPTBN2
 1 SPAST
 1 SNX1
 1 SNRNP40
 1 SLC17A3
 1 SELK
 1 RUNX1
 1 RCC1
 1 PTPN11
```

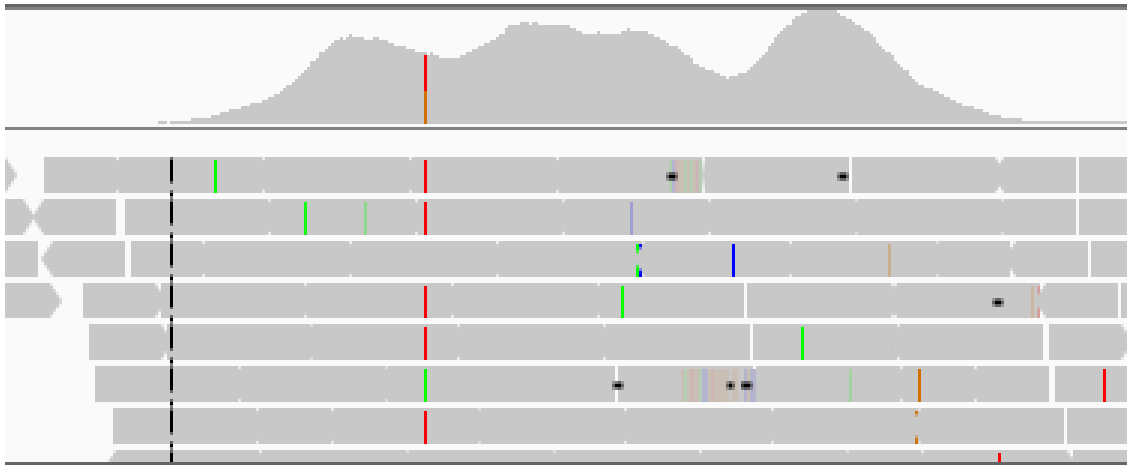
Variants called in DNMT3A

2	25457242	C	T	DNMT3A	missense	p.R882H
2	25458572	G	A	DNMT3A	splice_region	e21+4
2	25463169	A	G	DNMT3A	splice_site	e18+2
2	25463170	C	A	DNMT3A	splice_site	e18+1
2	25463321	T	G	DNMT3A	splice_site	e18-2
2	25463322	G	T	DNMT3A	splice_region	e18-3
2	25463507	AC	-	DNMT3A	splice_site_del	e17+1
2	25463600	C	A	DNMT3A	splice_site	e17-1
2	25464428	T	G	DNMT3A	splice_region	e16+3
2	25464430	C	T	DNMT3A	splice_site	e16+1
2	25467021	C	G	DNMT3A	splice_region	e14+3
2	25467022	A	T	DNMT3A	splice_site	e14+2
2	25467211	G	C	DNMT3A	splice_region	e14-4
2	25467213	A	G	DNMT3A	splice_region	e14-6
2	25467403	CCT	-	DNMT3A	splice_region_del	e13+4
2	25467524	A	T	DNMT3A	splice_region	e13-3
2	25467526	A	C	DNMT3A	splice_region	e13-5
2	25469028	C	T	DNMT3A	splice_site	e10+1
2	25469181	G	T	DNMT3A	splice_region	e10-3

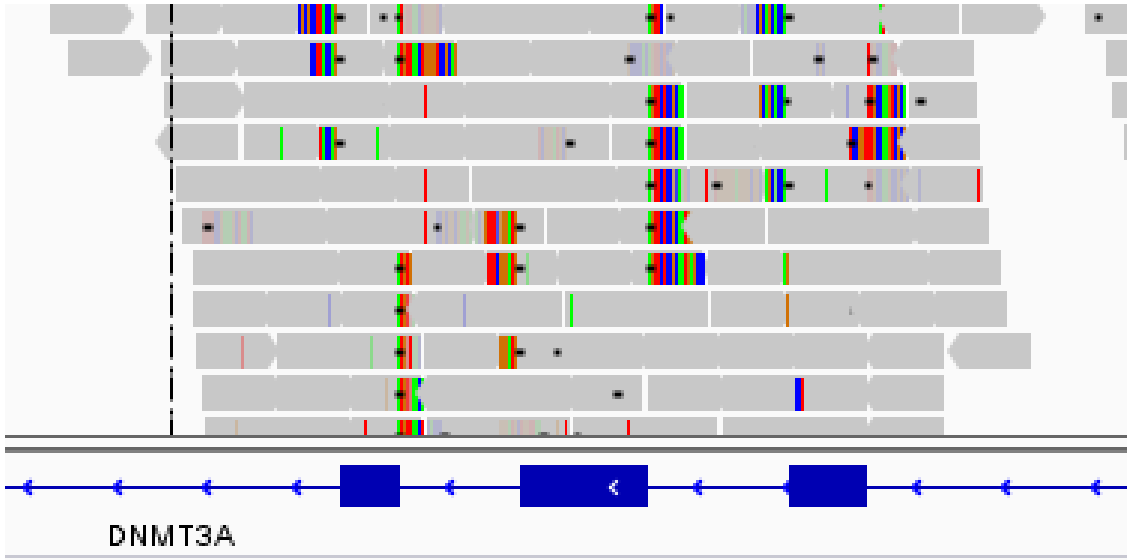

DNMT3A
T/G


CTATGTGCGGAAGCACCAGCTGAGAAGGTGGAGGGGACAGGATGGTACCTACCGTAGAGGCCCTTGCGAGCAGGGTTGACGATGGAGAGGTCATT
Y L G K R A P N V I S L D N
DNMT3A
AC/-


DNMT3A



dx: cDNA contamination



Damage Control

- If it is limited to a single gene (as in this case), could remove all splice-site adjacent mutations in that gene
  - If it's many genes/widespread, you might miss a lot of real events!

- Remake the libraries, resequence the sample

Case #5 – WGS serial samples

```
$ wc -l AML30_final_filtered_clean_b20_q10.hq.txt

  10114  AML30_final_filtered_clean_b20_q10.hq.txt
```

Case #5 – WGS serial samples

```
$ wc -l AML30_final_filtered_clean_b20_q10.hq.txt

  10114  AML30_final_filtered_clean_b20_q10.hq.txt
```

Max from TCGA AML cohort:
1298

Case #5 – WGS serial samples

```
$ wc -l AML30_final_filtered_clean_b20_q10.hq.txt

  10114  AML30_final_filtered_clean_b20_q10.hq.txt
```

Max from TCGA AML cohort:
1298

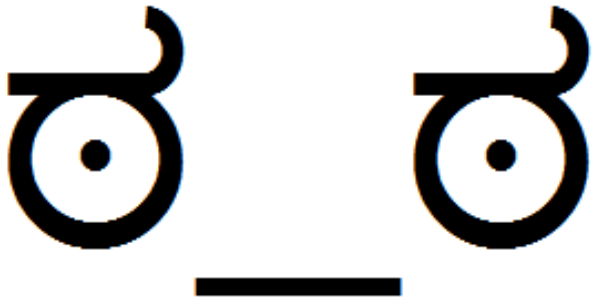
Primary tumor from this sample: 573

Case #5 – WGS serial samples

```
$ wc -l AML30_final_filtered_clean_b20_q10.hq.txt

  10114  AML30_final_filtered_clean_b20_q10.hq.txt
```

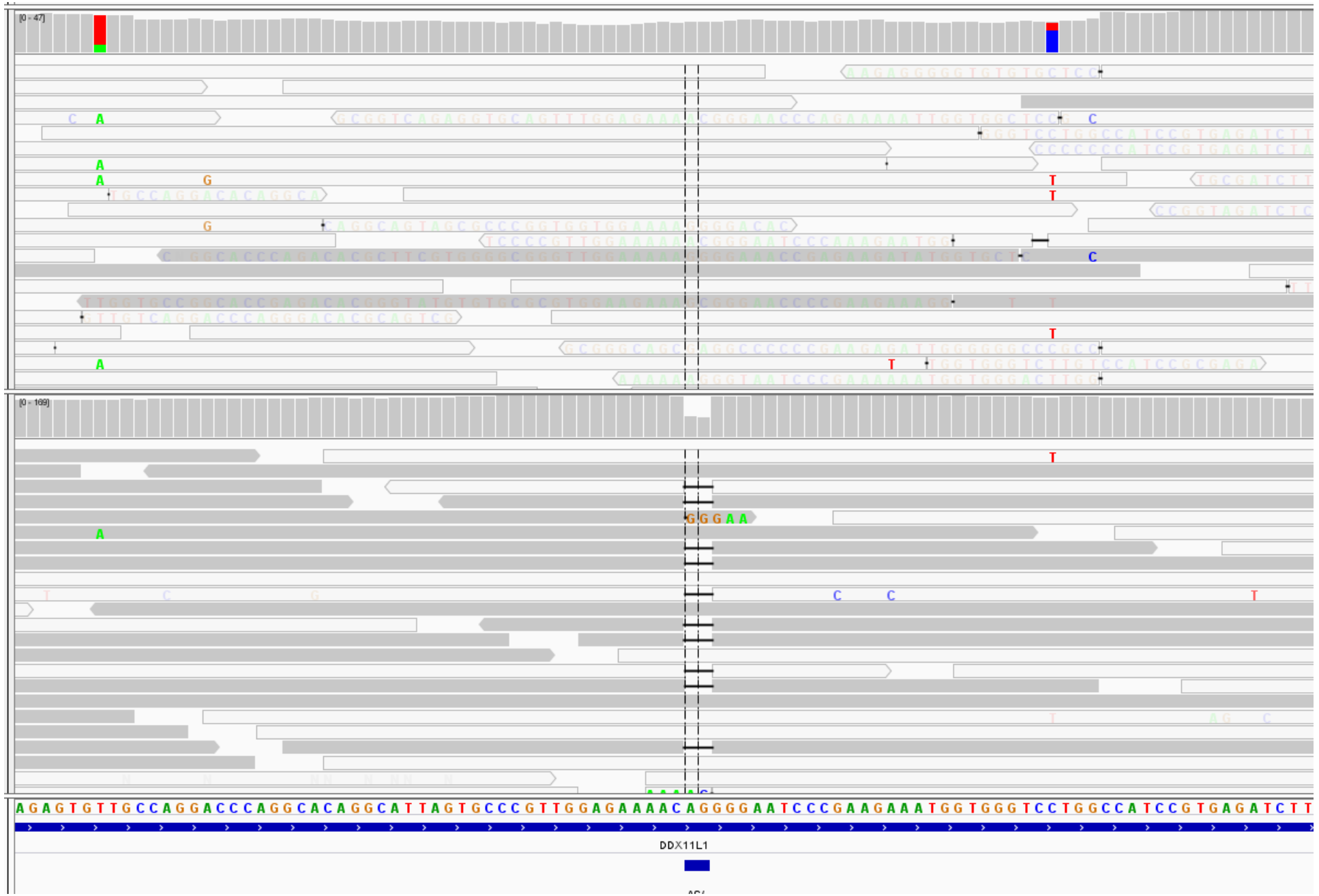
Max from TCGA AML cohort:
1298

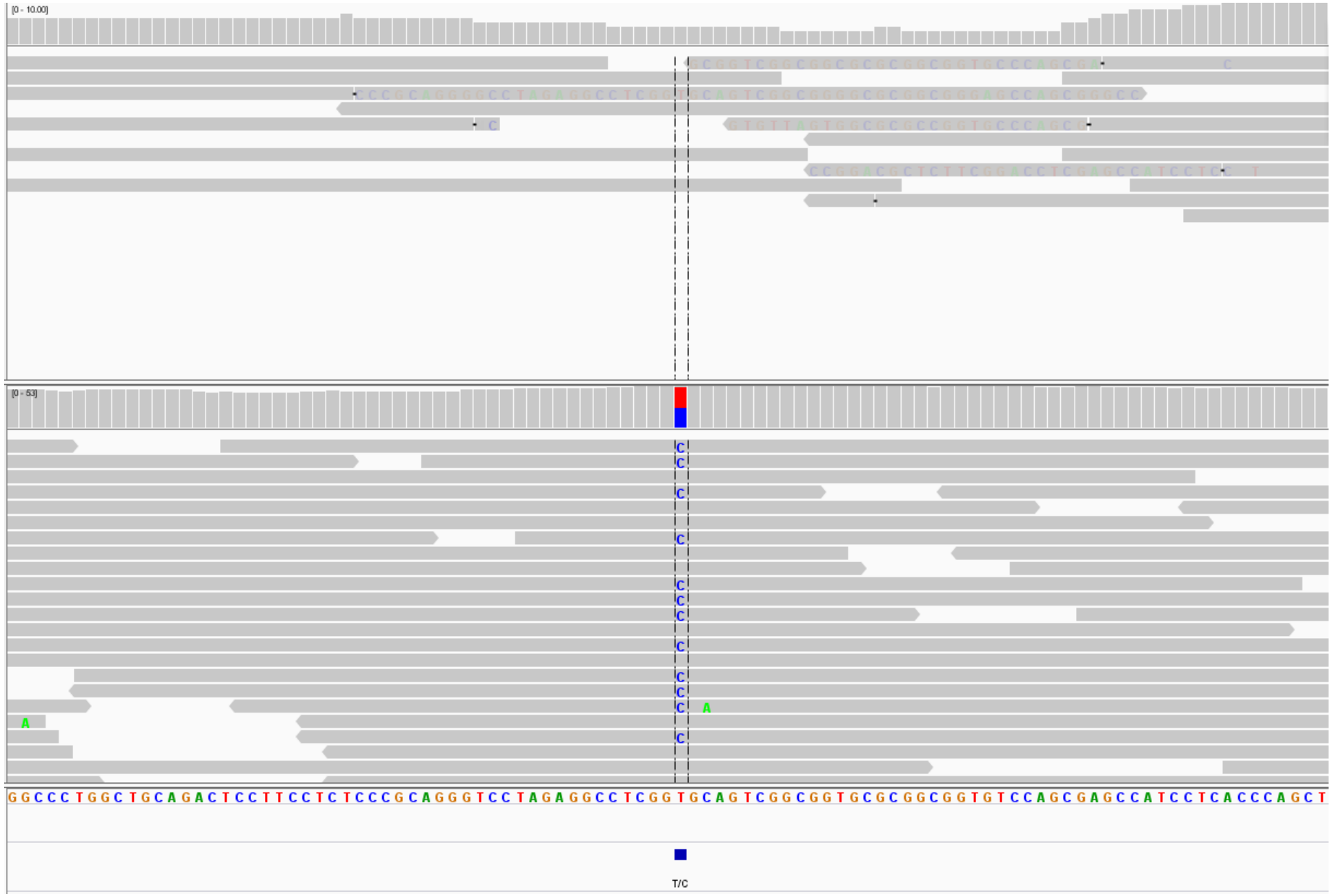
Primary tumor from this sample: 573



Case #5 – WGS serial samples

- Sample swap?
- Mismatch repair?
- Mutagenic therapy?

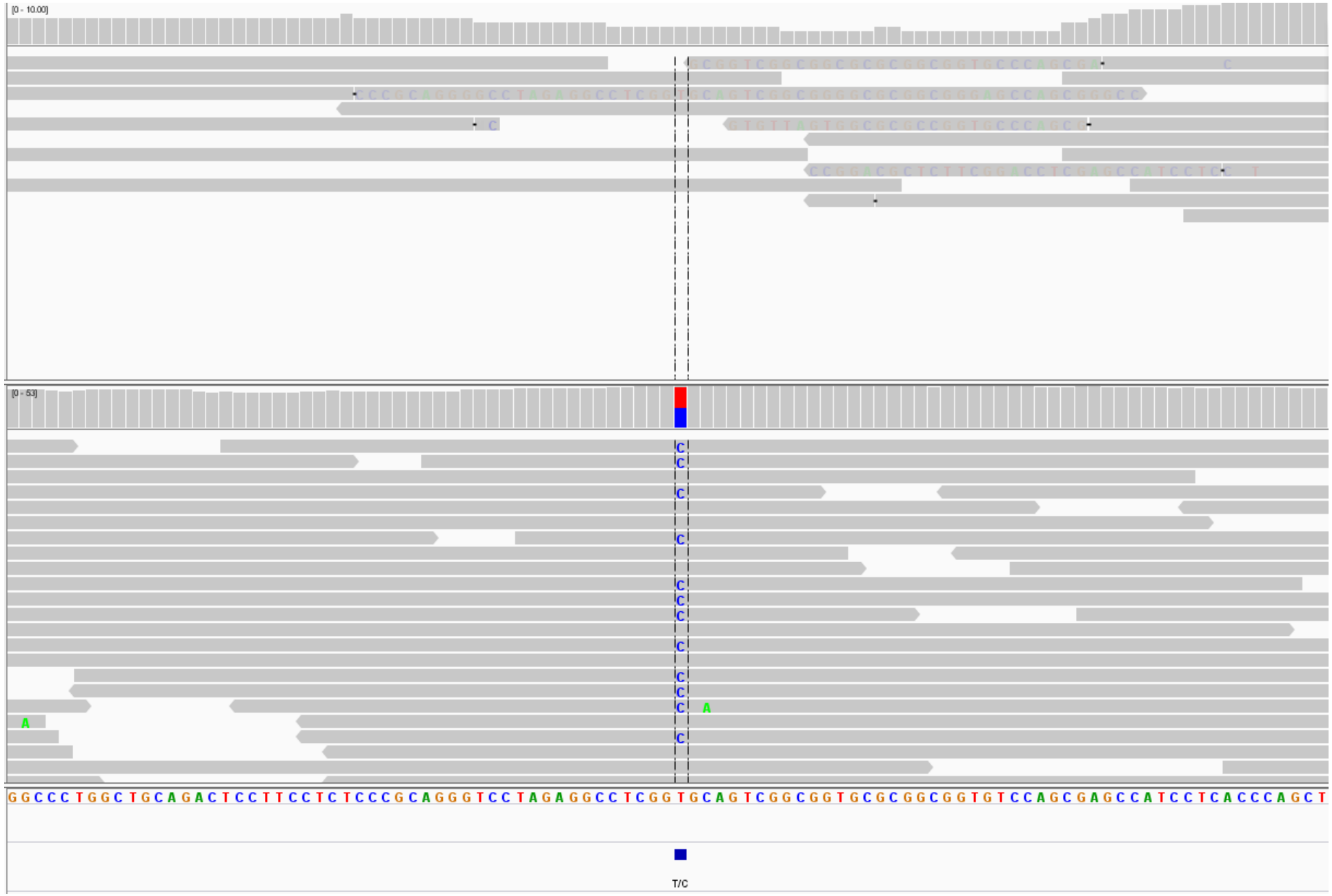

[0 - 47]
[0 - 169]
AGAGTGTTGCCAGGACCCAGGCACAGGCATTAGTGCCCGTTGGAGAAAACAGGGGAATCCCGAAGAAATGGTGGGTCCTGGCCATCCGTGAGATCTT
DDX11L1

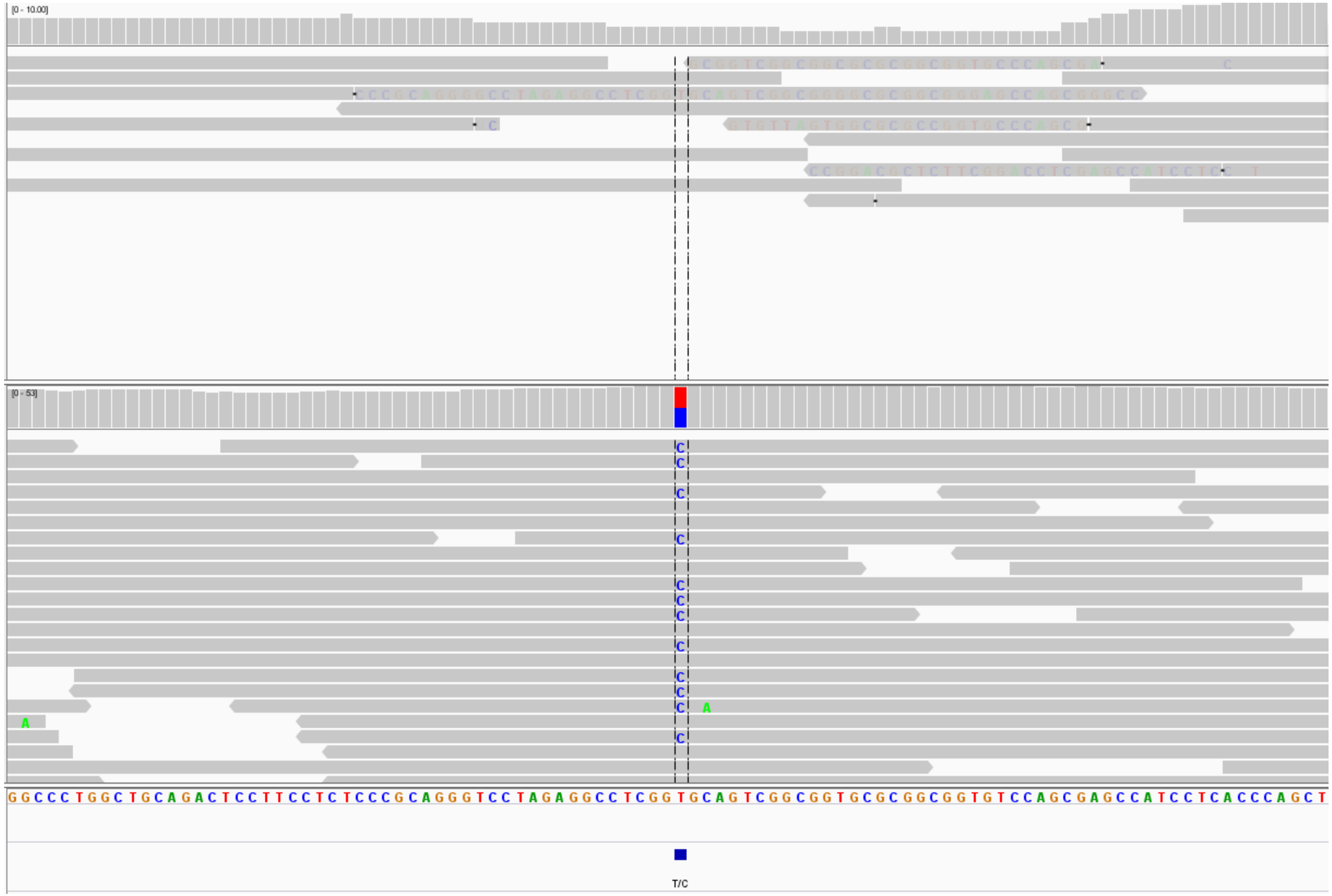

[0 - 10.00]
[0 - 53]
GGCCCTGGCTGCAGACTCCTTCCTCTCCCGCAGGGTCCTAGAGGCCTCGGTGCAGTCGGCGGTGCGCGGCGGTGTCCAGCGAGCCATCCTCACCCAGCT
T/C

- Normal – sequenced in Aug 2009
  - Mix of 75bp and 100bp PE reads

- Tumor – sequenced in Nov 2015
  - 125bp PE reads

dx: poorly matched controls


[0 - 10.00]
[0 - 53]
GGCCCTGGCTGCAGACTCCTTCCTCTCCCGCAGGGTCCTAGAGGCCTCGGTGCAGTCGGCGGTGCGCGGCGGTGTCCAGCGAGCCATCCTCACCCAGCT
T/C


[0 - 10.00]
[0 - 53]
GGCCCTGGCTGCAGACTCCTTCCTCTCCCGCAGGGTCCTAGAGGCCTCGGTGCAGTCGGCGGTGCGCGGCGGTGTCCAGCGAGCCATCCTCACCCAGCT
T/C

Damage Control

- Resequence the normal with matching read lengths

- Match your data as closely as possible!
  - Read lengths
  - Capture kits
  - Sample preparation

General Tips

- Visualize your data
  - A picture is worth a thousand p-values
- Hone, and then trust your instincts
  - If something seems unusual, it's often either a big problem or a big finding
- Be relentless
  - don't stop digging until you convince yourself that nothing is wrong

Expertise

An expert is a man who has made all the mistakes which can be made, in a narrow field.

--Niels Bohr

Spring 2025

Spring presenters and topics are still tentative

Date	Topic	Presenter
01/13/25	10:00am	Intro/prereqs, Genomic Intervals and Bedtools
01/20/25		NO SEMINAR - MLK DAY
01/27/25	10:00am	Epigenomics, ChIP/ATAC/WGBS
02/03/25	10:00am	Genome Assembly, Pangenome
02/10/25	10:00am	Data visualization with R and ggplot2 - part 1
02/17/25	10:00am	Data visualization with R and ggplot2 - part 2
02/24/25	10:00am	Single-cell RNAseq part 1
03/03/25	10:00am	Single-cell RNAseq part 2
03/10/25		NO SEMINAR - SPRING BREAK
03/17/25	10:00am	Genomic Workflows/Cloud Computing
03/24/25	10:00am	Microbial Genomics
03/31/25	10:00am	Machine Learning/AI in Genomics
04/07/25		NO SEMINAR
04/14/25	10:00am	Long Read Sequencing
04/21/25	10:00am	Genomic Medicine, course wrap-up