SUPPLEMENTARY INFORMATION

FOR

Systematic Discovery of Complex Indels in Human Cancers

Kai Ye[1,2], Jiayin Wang[1], Reyka Jayasinghe[1,3], Eric-Wubbo Lameijer[4], Joshua F. McMichael[1], Jie Ning[1], Michael D. McLellan[1], Mingchao Xie[1,3], Song Cao[1], Venkata Yellapantula[1,3], Kuan-lin Huang[1,3], Adam Scott[1,3], Steven Foltz[1,3], Beifang Niu[1], Kimberly J. Johnson[5], Matthijs Moed[4], P. Eline Slagboom[4], Feng Chen[3,6], Michael C. Wendl[1,2,7], Li Ding[1,2,3,6#]

[1]McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO, USA

[2]Department of Genetics, Washington University in St. Louis, St. Louis, MO, USA

[3]Department of Medicine, Washington University in St. Louis, St. Louis, MO, USA

[4]Leiden University Medical Center, Leiden, the Netherlands

[5]Brown School Master of Public Health Program, Washington University in St. Louis, St. Louis, MO, USA

[6]Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO, USA

[7]Department of Mathematics, Washington University in St. Louis, St. Louis, MO, USA

Corresponding Author:

Li Ding, Ph.D

Email: lding@genome.wustl.edu

Complex somatic indels supported by alternative DNA/RNA-seq data

Our analysis showed that some complex indels in cancer genes appear to be in the founding clones, while others are in subclones. We searched alternative sequence data generated within TCGA and examined whether complex indels we reported are also supported there. Independently generated WGS data from different sequencing libraries have been recruited to validate complex indel events detected using exome sequencing data in a bias-free way. Out of the 285 somatic coding complex indels identified in 624 cancer genes using TCGA exome sequencing data, 15 events were from samples having both matching exome and WGS data and coverage higher than 10x. We examined these 15 sites using IGV and identified complex indel supporting evidence in all 15, suggesting they are bona-fide complex indel events (**Supplementary Table 8** and **Supplementary Documents**). We further investigated expression of complex indels using RNA-seq data, which were available for 248 of the 285 sites discussed above. Of these sites, 45 had no coverage, but 138, 112, and 72 sites had coverage more than 10x, 20x and 50x, respectively (**Supplementary Table 4**). Despite the low transcription level of those sites, we observed RNA-seq read support for 57 sites. In fact, for 51 sites with $\geq$ 100x coverage, 18 of them (35.3%) had supporting evidence for the complex indels in RNA-seq data. We next examined the gene and tumor type pair *EGFR* in lung adenocarcinoma, described in the previous VAF analysis section as present in the founding clone. We found that among six sites with RNA-seq data, all of them had total coverage below 80x and had various numbers of RNA-seq reads supporting the mutant allele using cDNA wide-type and mutant competitive mapping (Methods). On the other hand, for tumor suppressors we only observed 4 out of 18 for *PIK3R1* and 1 out of 8 for *PTEN* with RNA-seq read support. We speculate that the relatively low

percentage of complex variants with RNA-seq read support is caused by the dominating number of tumor suppressor genes with mostly frameshift variants.

SUPPLEMENTARY DATA

Supplementary Figures

Supplementary Figure 1: The impacts of insert size, read length and mapping condition on complex indel detection. a) Insert size 300bp, read length 100bp and BWA aln as the alignment; b) Insert size 600bp, read length 250bp and BWA aln as the alignment; c) Insert size 300bp, read length 100bp, and BWA mem as the alignment; d) Insert size 600bp, read length 250bp and BWA mem as the alignment.

Supplementary Figure 2: The deletion size vs insertion size for predicted (green) and missed (black) complex indels in Venter's genome.

Supplementary Figure 3. Histogram of percentage of reads carrying an indel according to CIGAR string. The reads per sample were screened one by one and count the total number as well as the number of reads with either "I" or "D" in the CIGAR string. At the end, each sample has one value of percentage of reads carrying an indel.

Supplementary Documents:

Sanger sequencing of validated complex indels from COLO 829: IGV, PCR and sanger trace of somatic and germline variants in supplementary table 3.

Screenshots of WGS validation of complex indels with good coverage: IGV screenshots of variants with sufficient coverage (>10x) in supplementary table 8.

Supplementary Tables

Supplementary Table 1. In total 1,128 complex indels on Venter chr1 were spiked in.

Supplementary Table 2. The complex indels reported by Pindel-C taking the bam file with spiked in Venter complex indels as input.

Supplementary Table 3. Somatic and germline complex indel validation result in COLO829 cell lines.

Supplementary Table 4. Exome-wide complex indels.

Supplementary Table 5. MUSIC correlation analysis.

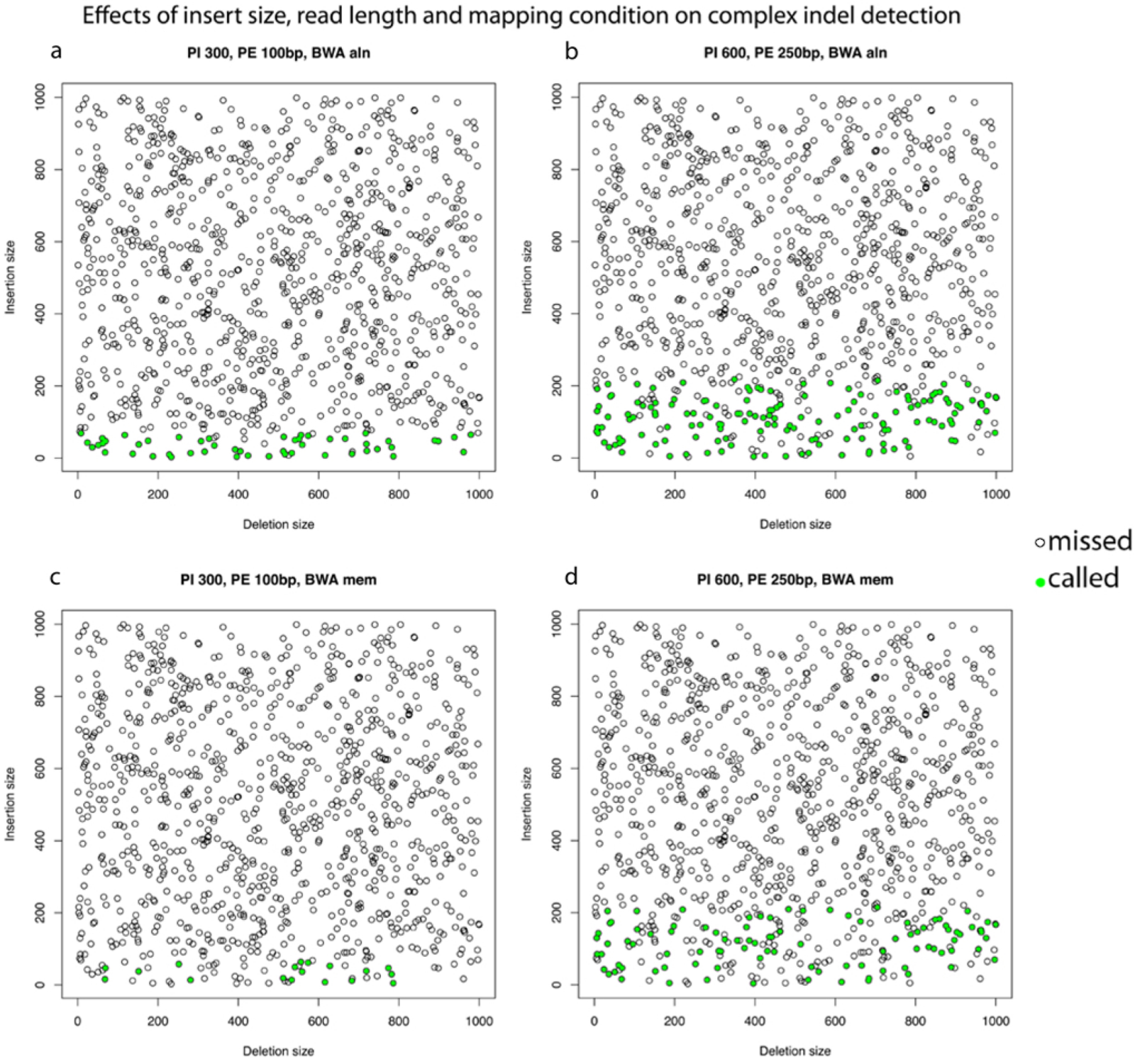
Supplementary Table 6. A list of 624 cancer-associated genes compiled from literature.

Supplementary Table 7. A list of complex indels in cancer-associated genes.

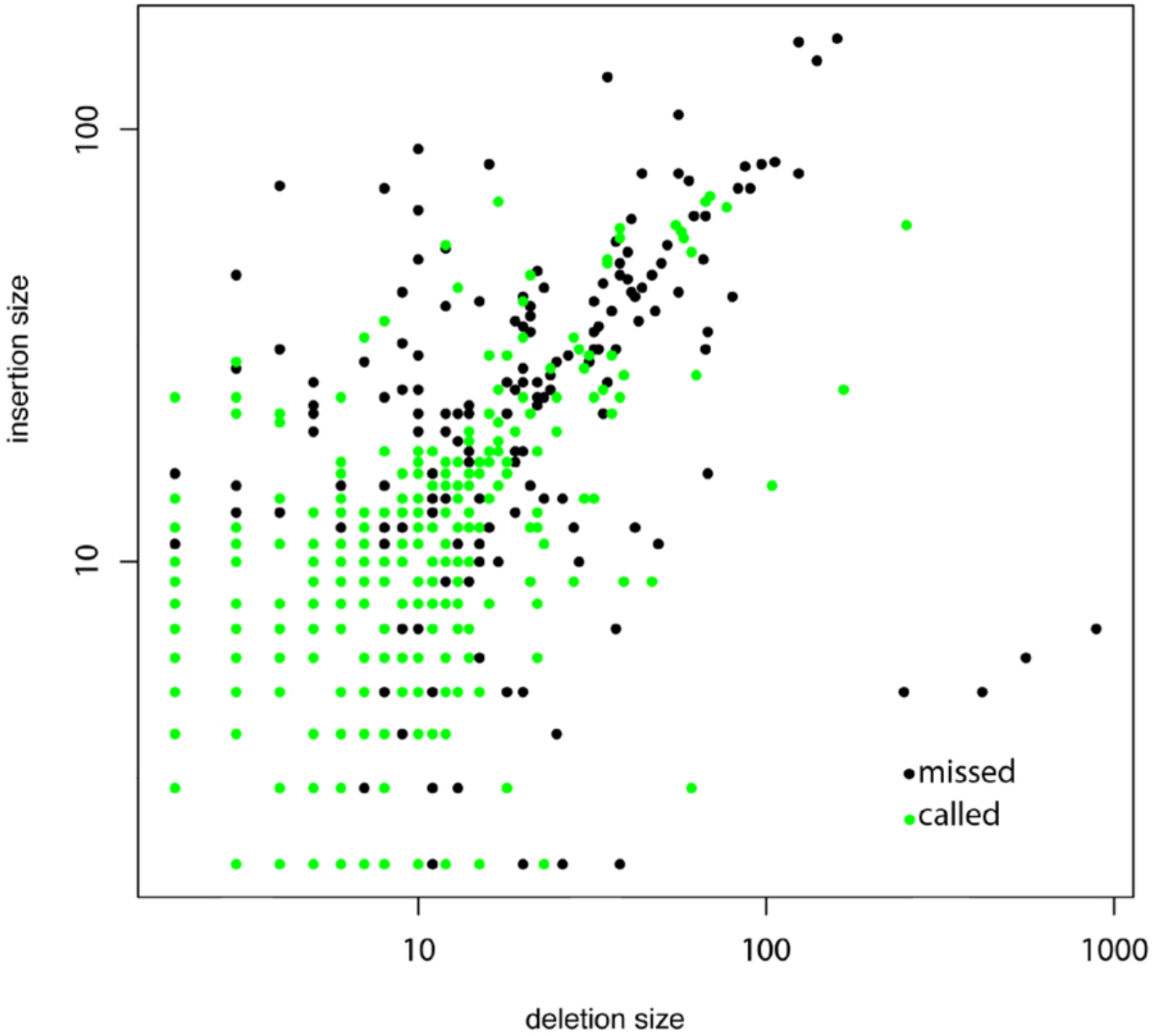
Supplementary Table 8. Validation of somatic complex indels discovered in exome data using whole genome sequence data.

Supplementary Table 9. Complex indel variant allele fraction and VAF of simple variants.

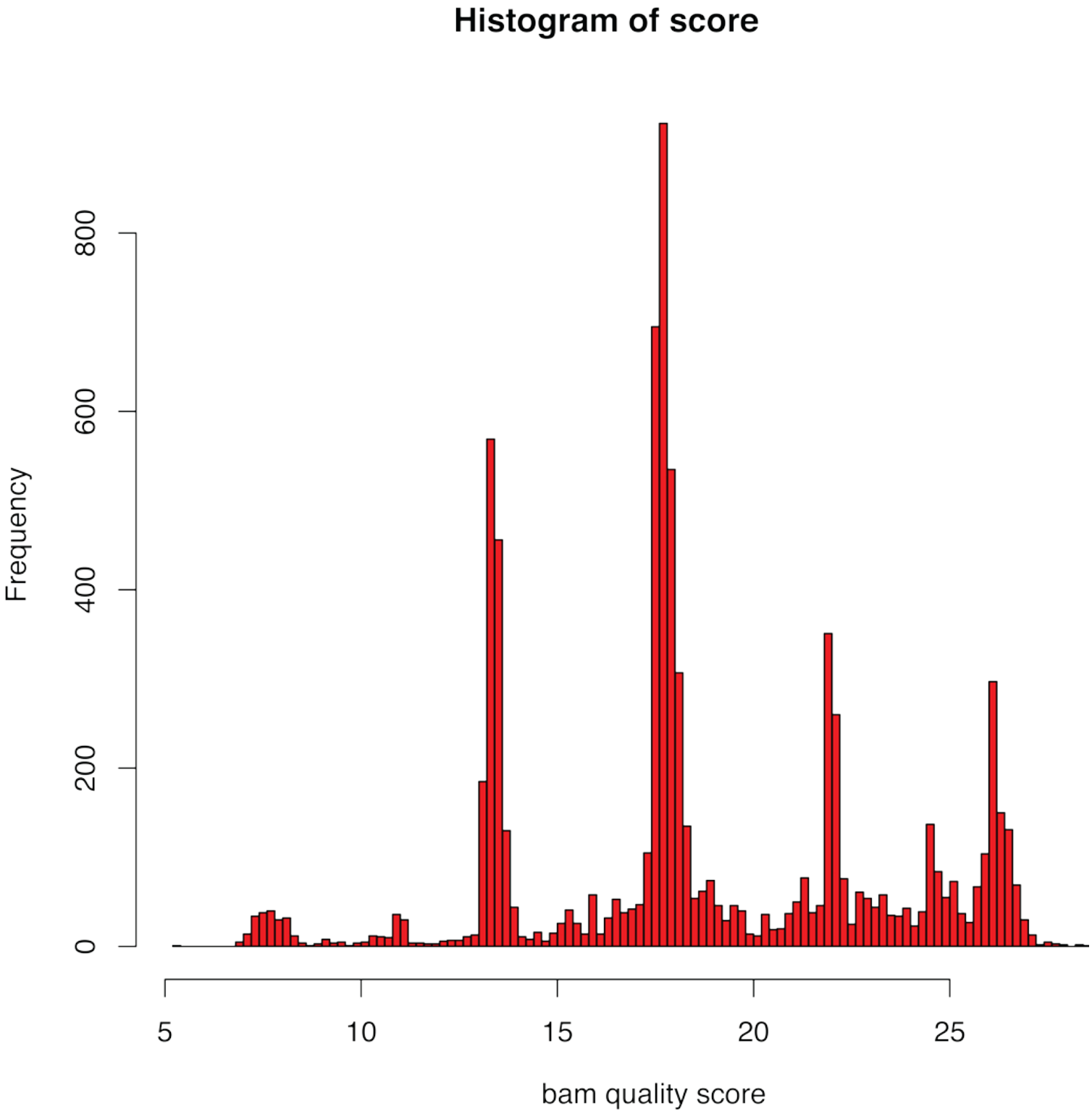
Supplemental Figure 1



Supplemental Figure 2



Supplemental Figure 3



Supplemental Materials: Sanger Sequencing of Validated Complex Indels from COLO829

General Layout of Supplementary Material

Germline/Somatic # (Name of Event in Bold)
Information about complex event:
D Deletion Size NT Insertion Size Inserted Sequence Chromosome Start Position End Position
Picture: Screenshot from Integrative Genomics Viewer, tumor bam on top and normal bam on bottom.
Picture: Screenshot of sanger trace

Sanger Sequencing Identifier Legend:
JN = Identifier
COLO829=Cell Line Identifier
BL=Normal Cell Line (when lacking the BL identifier = Tumor Cell Line)
G#F=Germline # Forward Primer
G#R= Germline # Reverse Primer
S#F=Somatic # Forward Primer
S#R= Somatic # Reverse Primer
_[A-Z]##.seq: Sanger sequencing identifiers

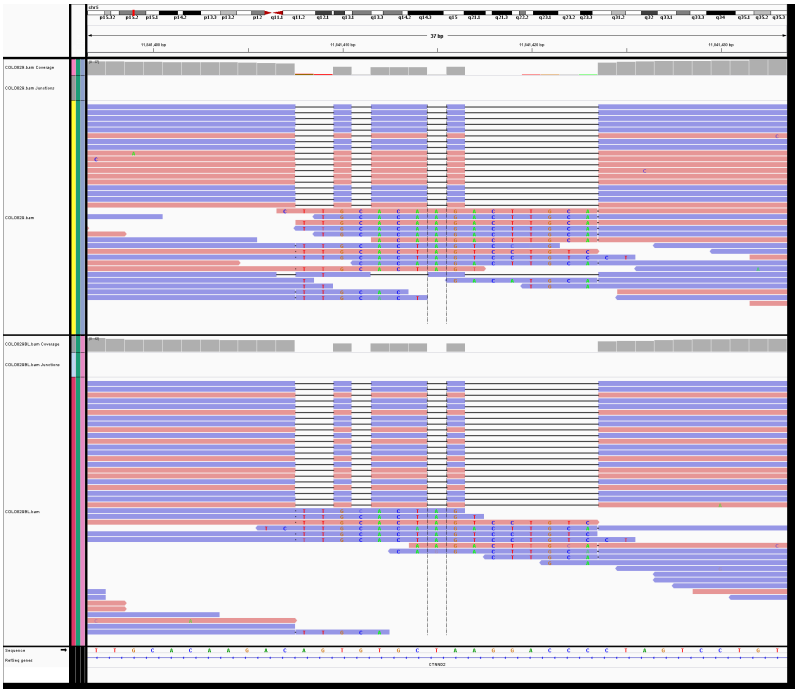
Example:
JNCOLO829BLG1F_A01.seq: Normal Cell Line, Germline 1, Forward Primer

Summary

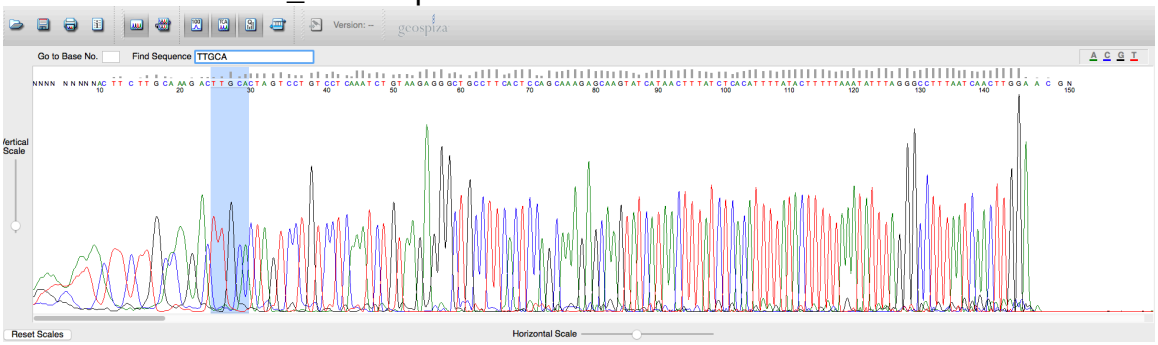
	Somatic	**Germline**
Starting # events to be sequenced	17	58
Transposable Elements (not to be sequenced)	*4*	*5*
Multiple Bands PCR Result (Not to be sequenced)	*0*	*11*
Sequencing Result Un-interpretable	1	3
Total Sites Sequenced (w/o un-interpretable seq results)	12	39
Validated	9	39
No Support	3	0
Validation Rate	**75%**	**100%**

Germline 1

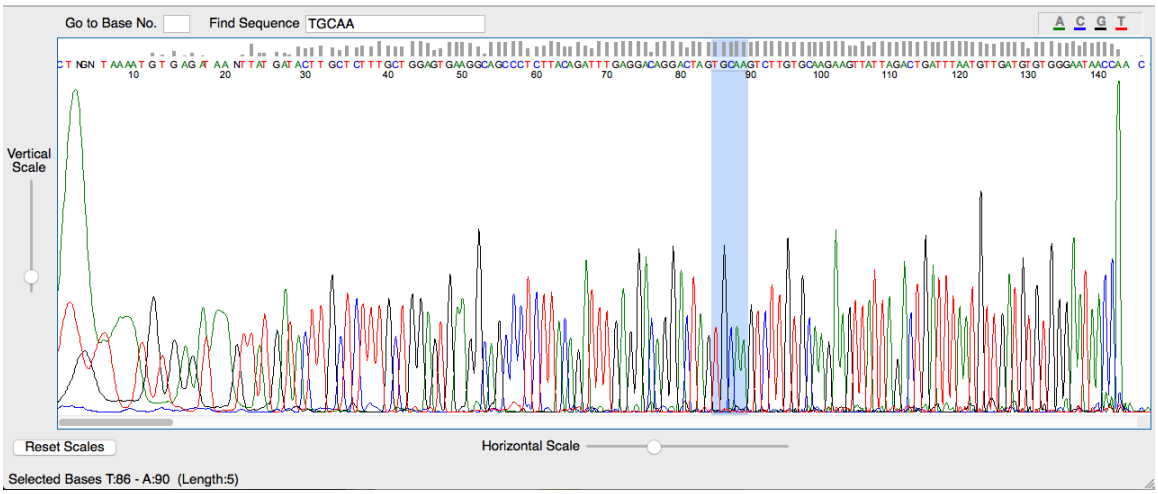
D 16 NT 5 TTGCA 5 11841407 11841424



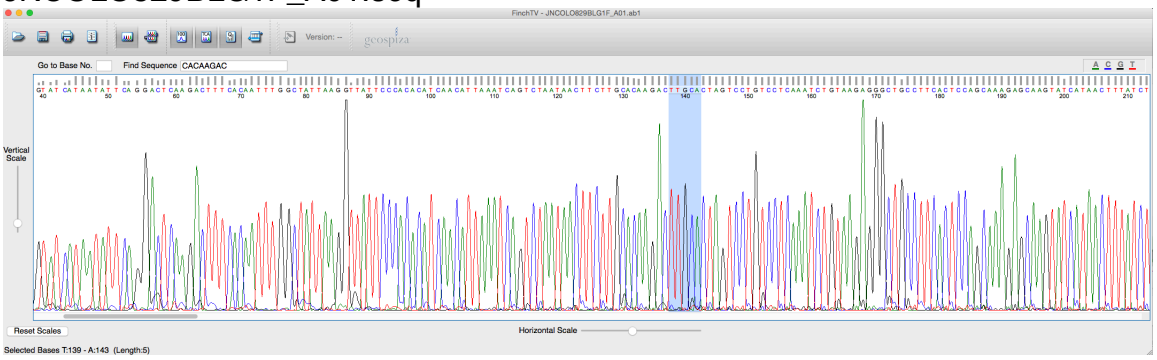
JNCOLO829BLG1F_A01.seq



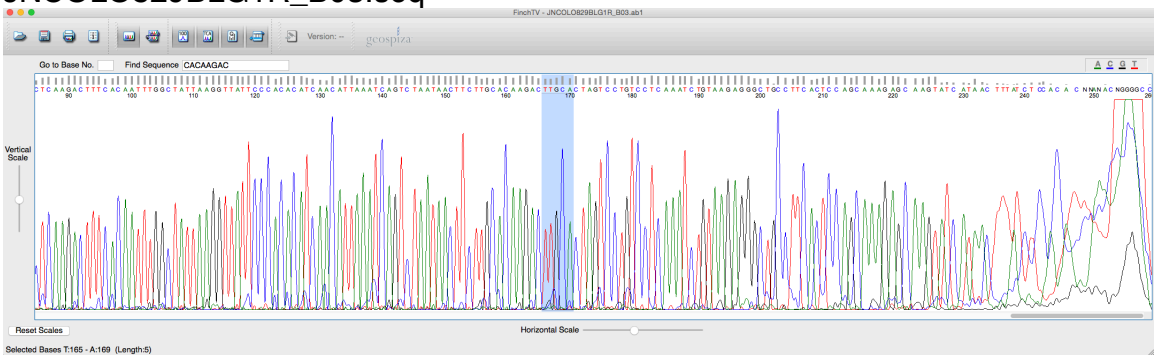
JNCOLO829BLG1R_C02.seq



JNCOLO829BLG1F_A01.seq

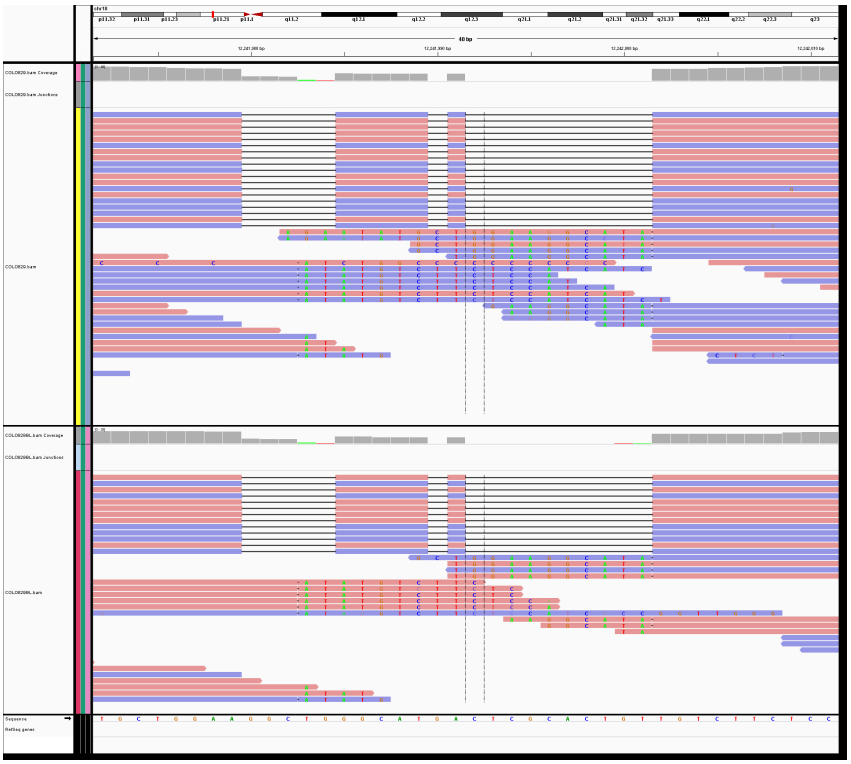


JNCOLO829BLG1R_B03.seq

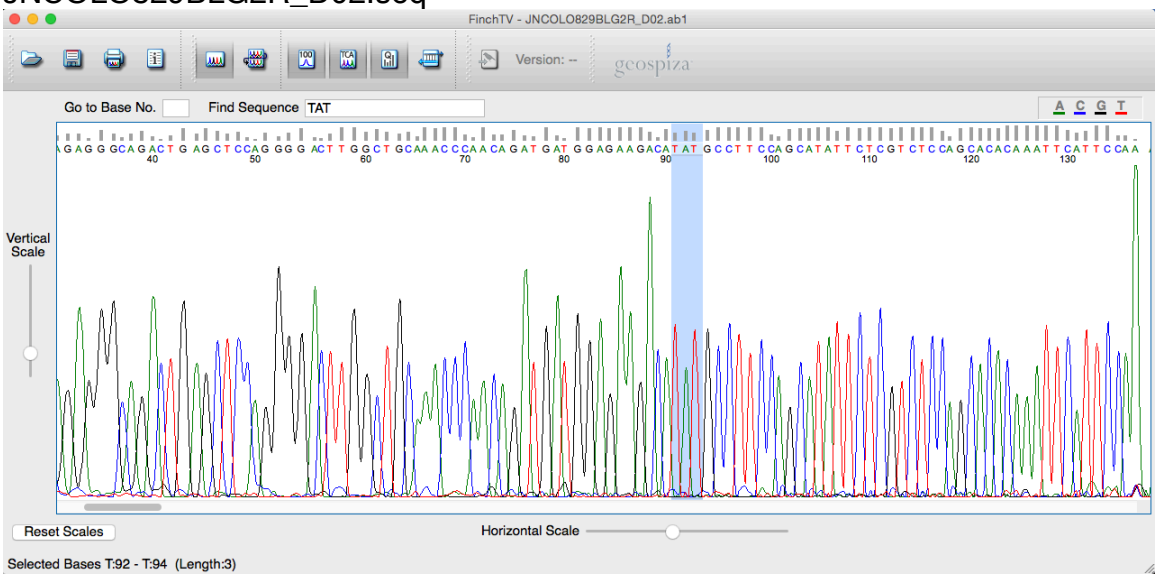


Germline 2

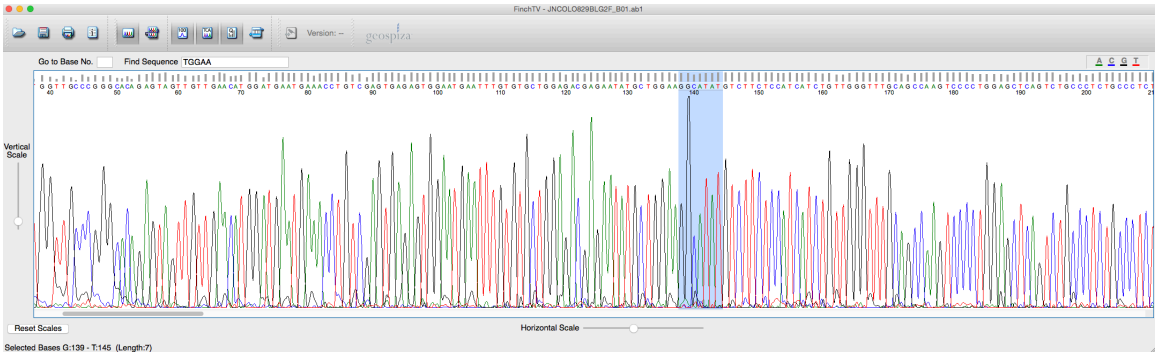
D 19 NT 3 ATA 18 12241982 12242002



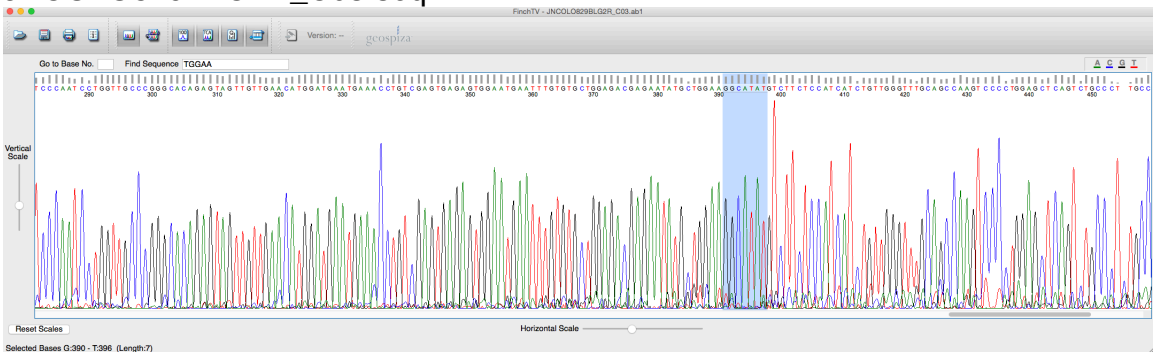
JNCOLO829BLG2R_D02.seq



JNCOLO829BLG2F_B01.seq



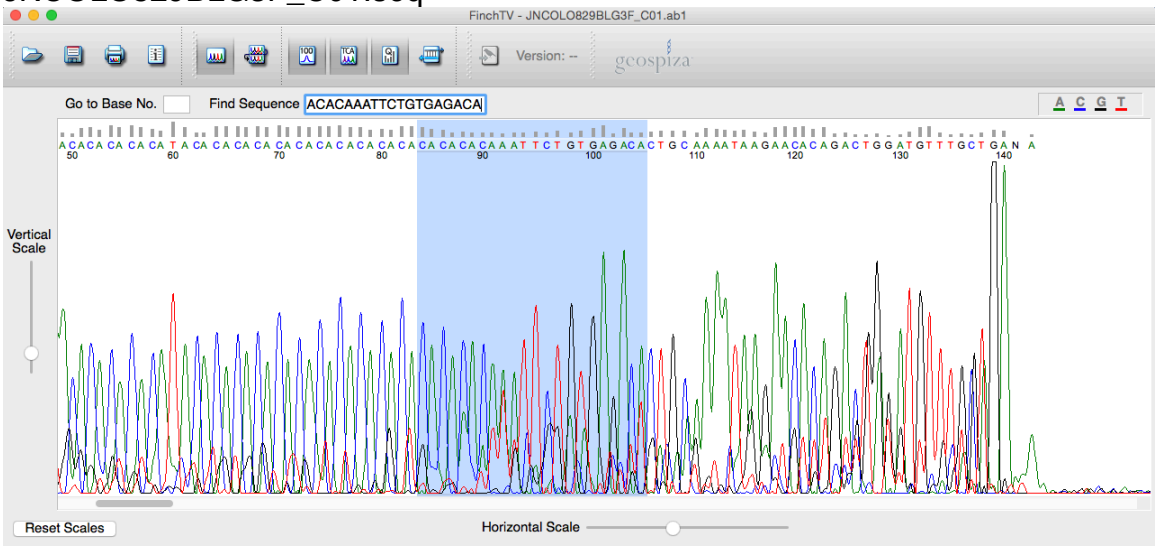
JNCOLO829BLG2R_C03.seq



Germline 3

D 16 NT 22 CACACACAAATTCTGTGAGACA 10 18551365 18551382

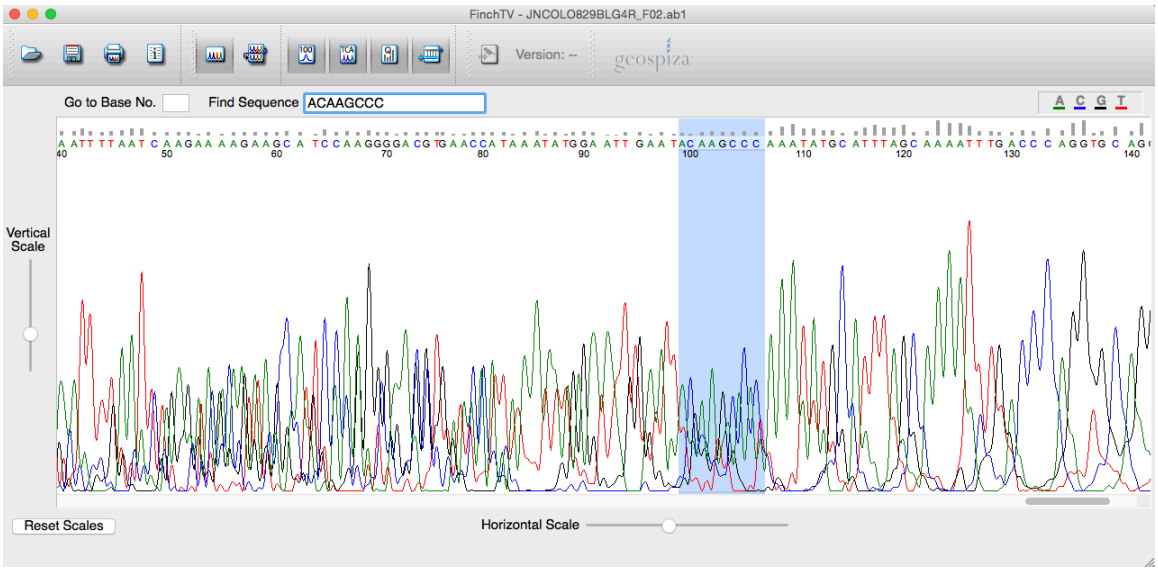
JNCOLO829BLG3F_C01.seq



Germline 4

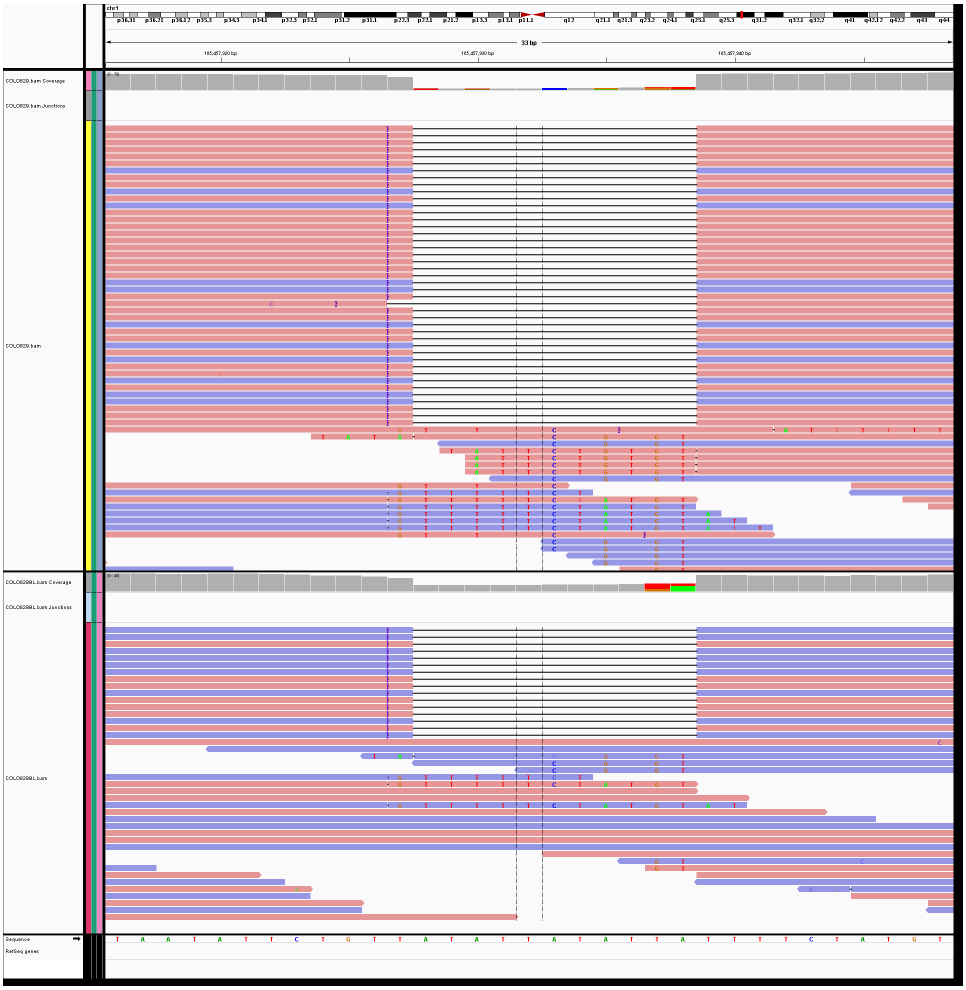
D	14	NT	8	ACAAGCCC	2	171897783	171897798

JNCOLO829BLG4R_F02.seq

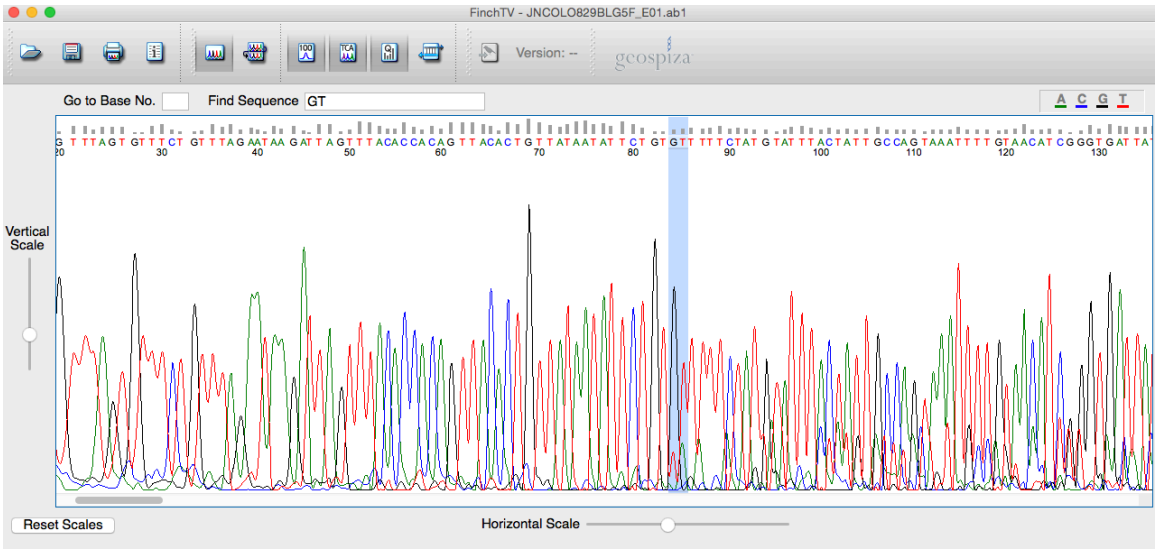


Germline 5

D 12 NT 2 GT 1 185457926 185457939



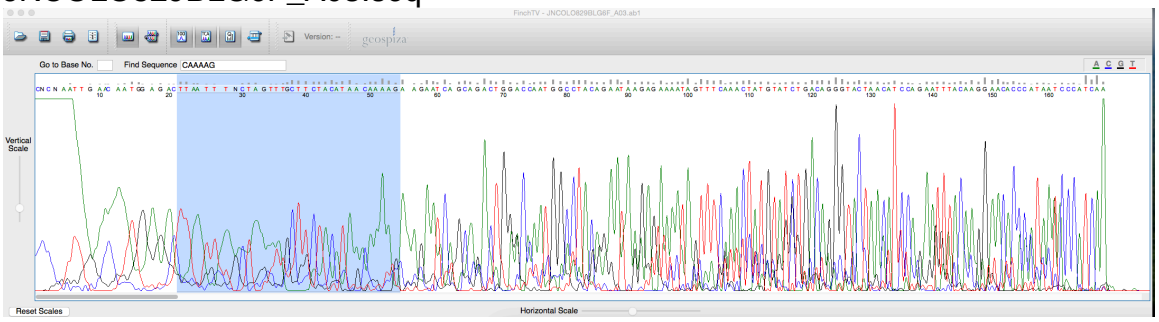
JNCOLO829BLG5F_E01.seq



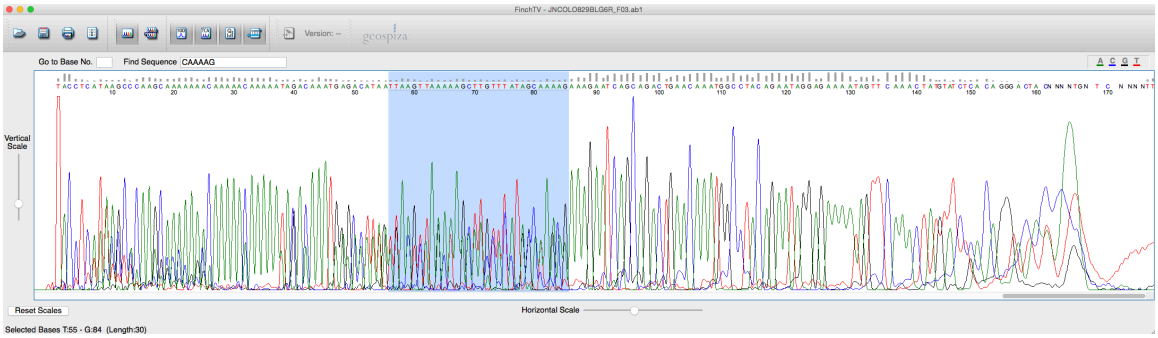
Germline 6

D 22 NT 30 TTAAGCTAAAAAGCTTCTACATAGCAAAAG 7 151730626 151730649

JNCOLO829BLG6F_A03.seq



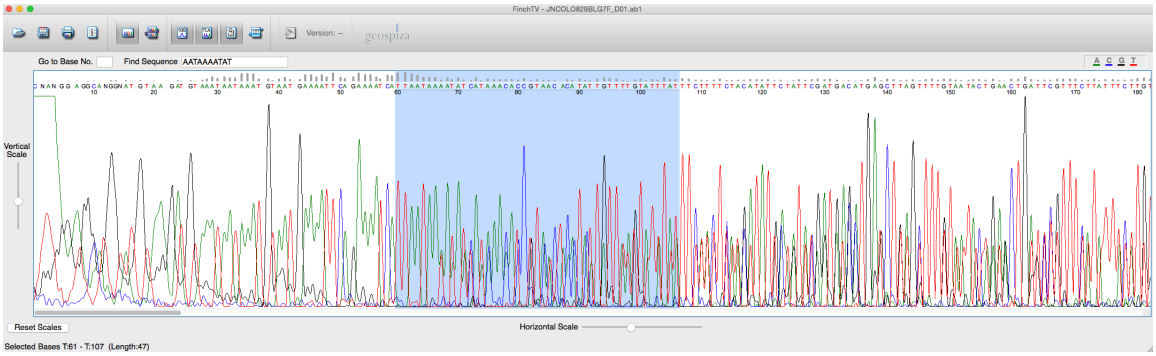
JNCOLO829BLG6R_F03.seq



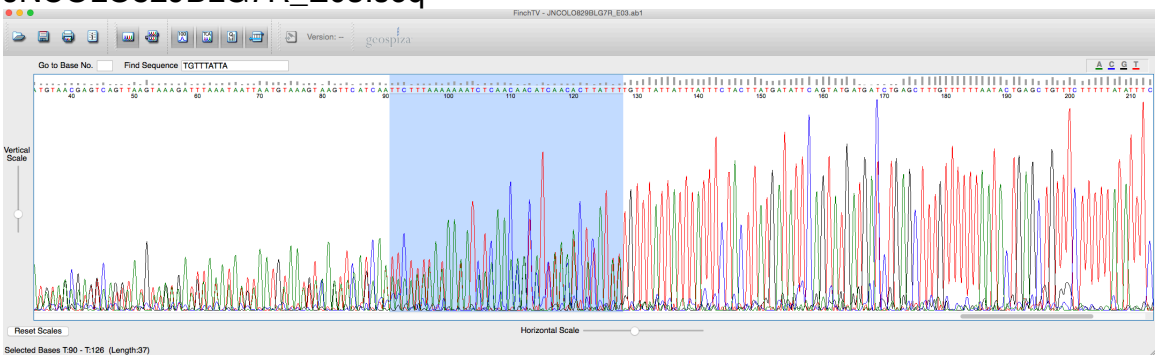
Germline 7

D 31 NT 26 AATATCAAAAACATCTTAACTCATAT 10 92055665 92055697

JNCOLO829BLG7F_D01.seq

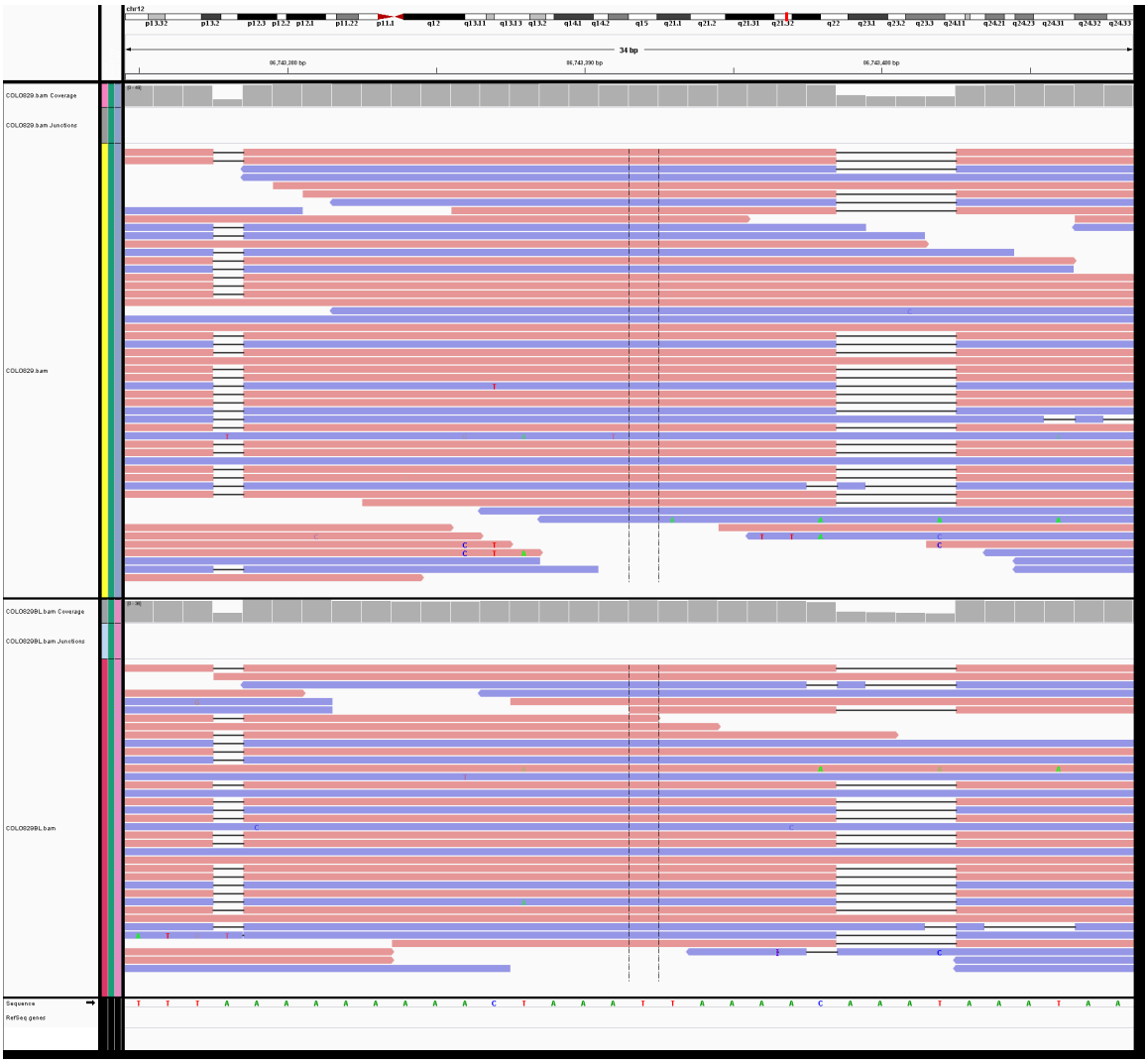


JNCOLO829BLG7R_E03.seq

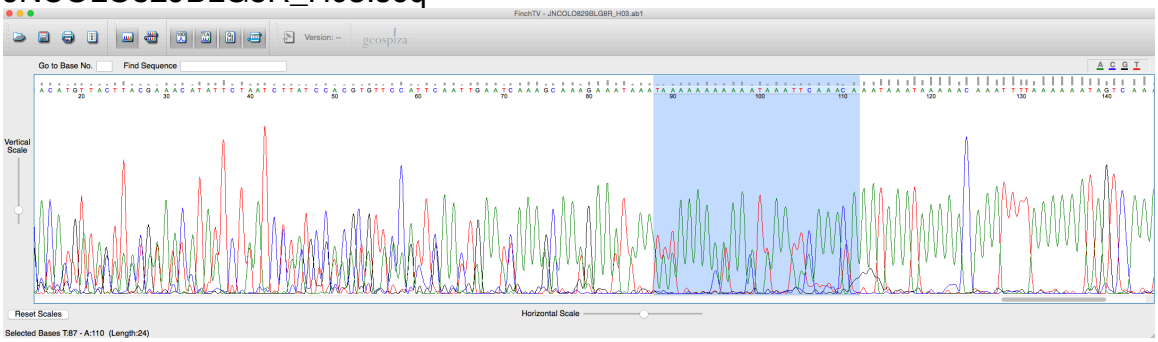


Germline 8

D	13	NT	8	CTAAATTA	12	86743385	86743399



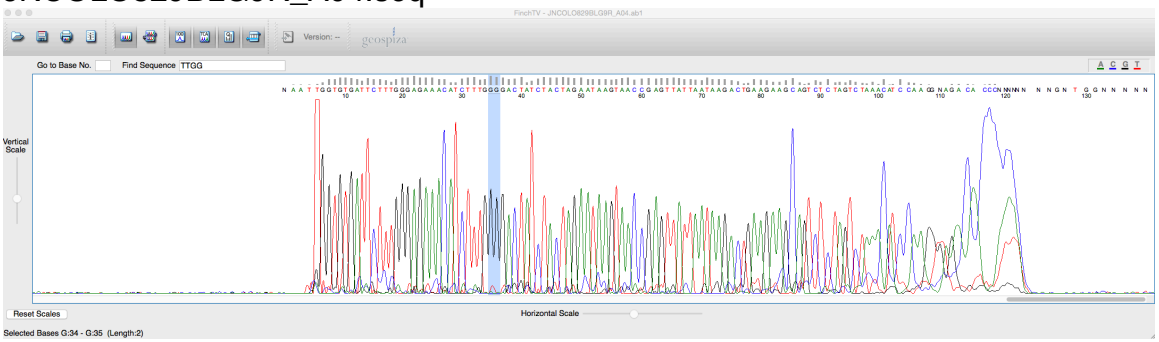
JNCOLO829BLG8R_H03.seq



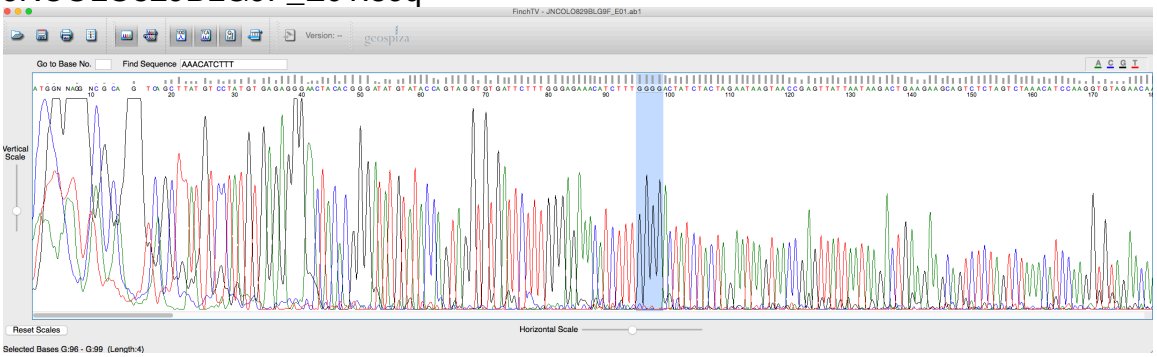
Germline 9

D 17 NT 2 GG 12 83961554 83961572

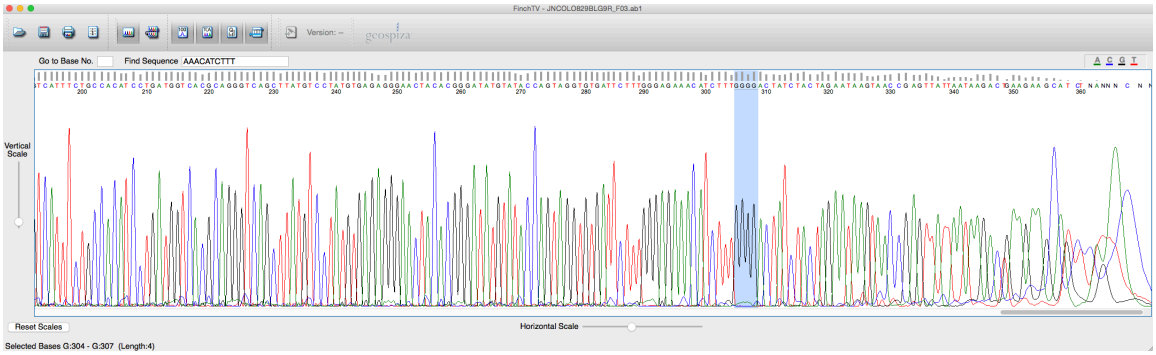
JNCOLO829BLG9R_A04.seq



JNCOLO829BLG9F_E01.seq

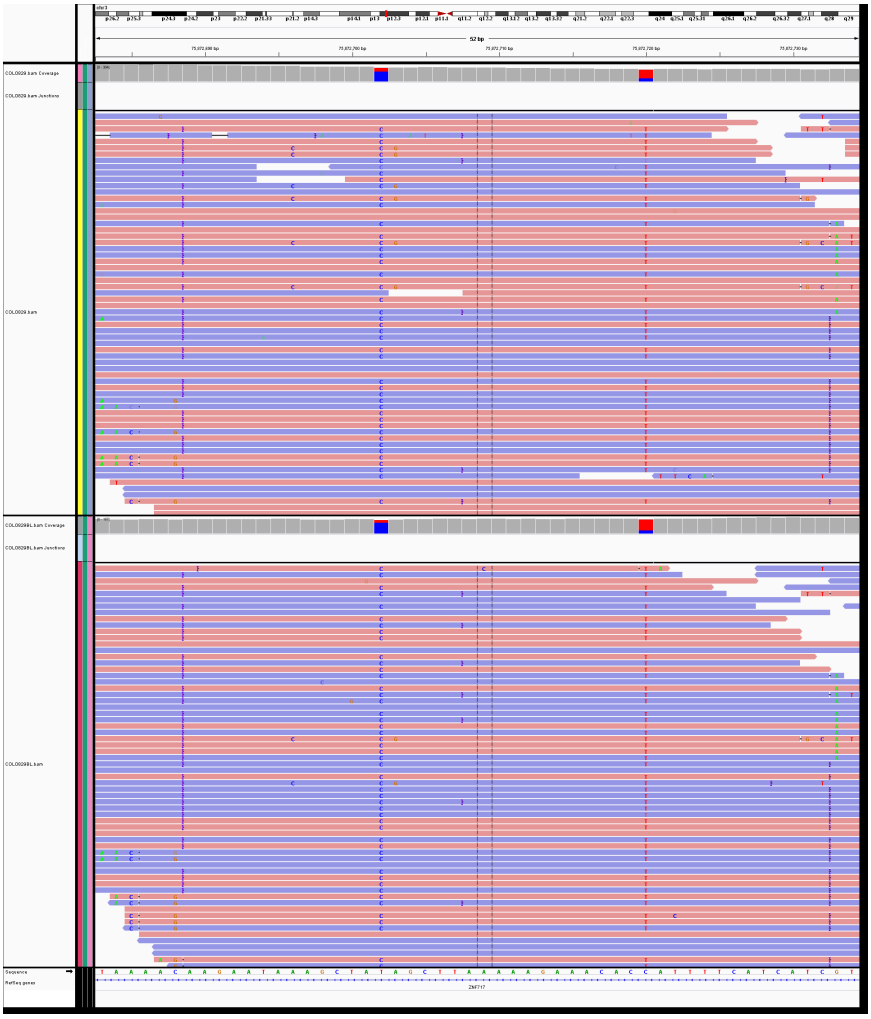


JNCOLO829BLG9R_F03.seq

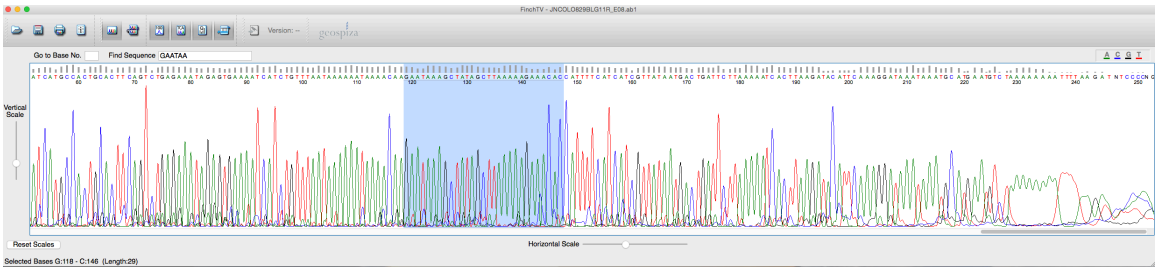


Germline 11

D	31	NT	37	GAATAAAGCTACAGCTTAAAAAGAAACACTATTTTCA
	3	75872693	75872725	



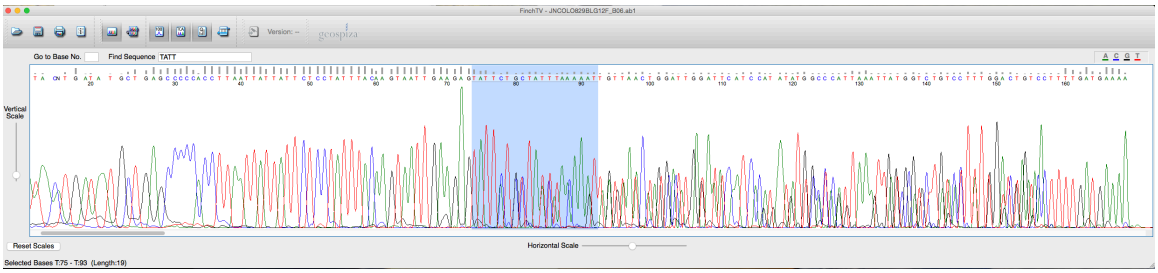
JNCOLO829BLG11R_E08.seq



Germline 12

D 25 NT 13 ATGTTAACAAAAA 4 103723977 103724003

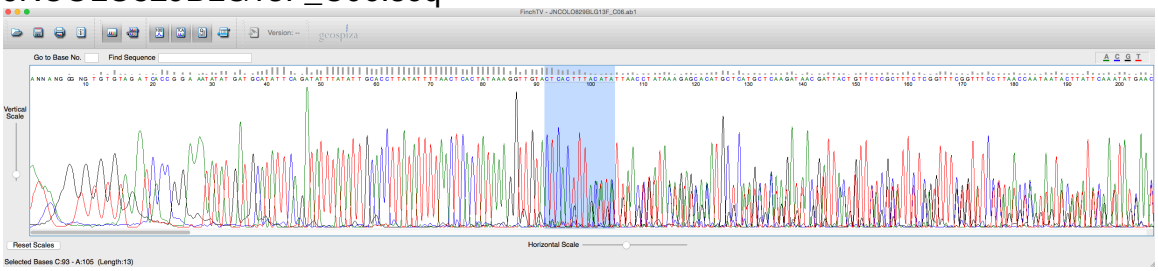
JNCOLO829BLG12F_B06.seq



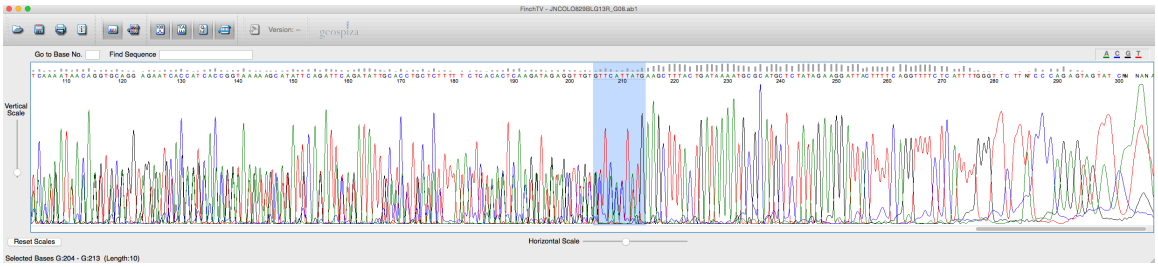
Germline 13

D 2 NT 8 GTTACTTA 14 57844474 57844477

JNCOLO829BLG13F_C06.seq



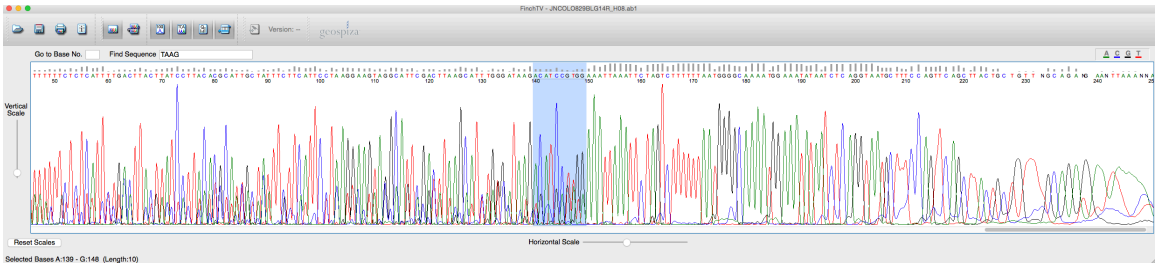
JNCOLO829BLG13R_G08.seq



Germline 14

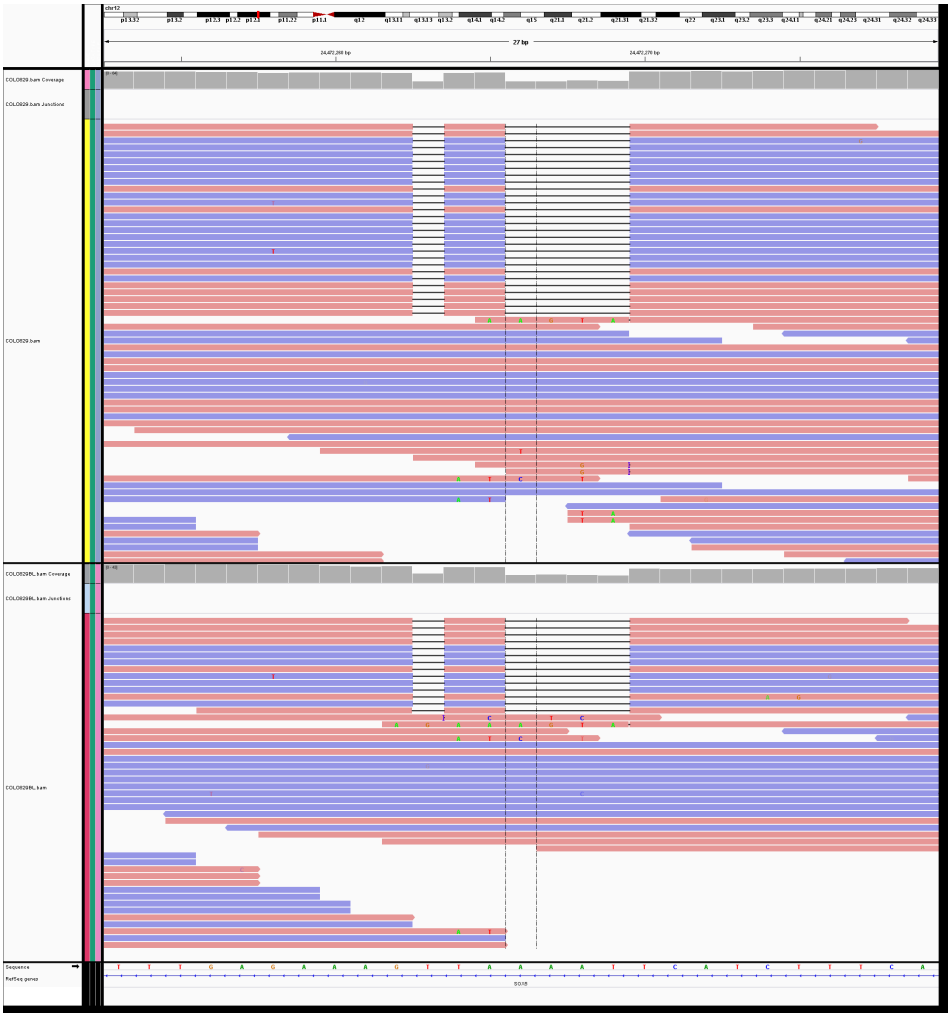
D 2 NT 9 CATCCTTGG 4 190833430 190833433

JNCOLO829BLG14R_H08.seq

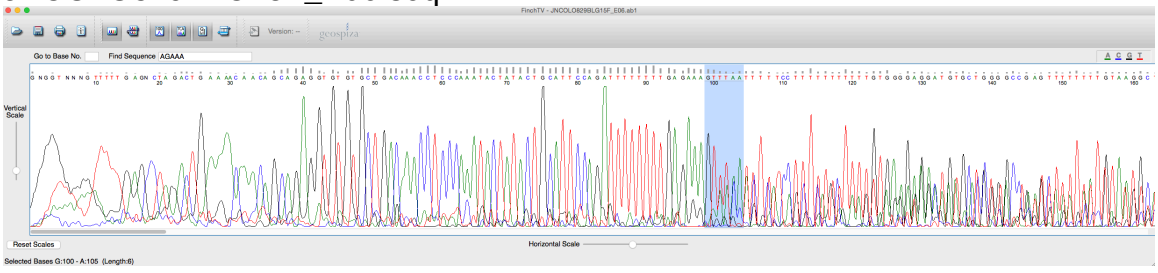


Germline 15

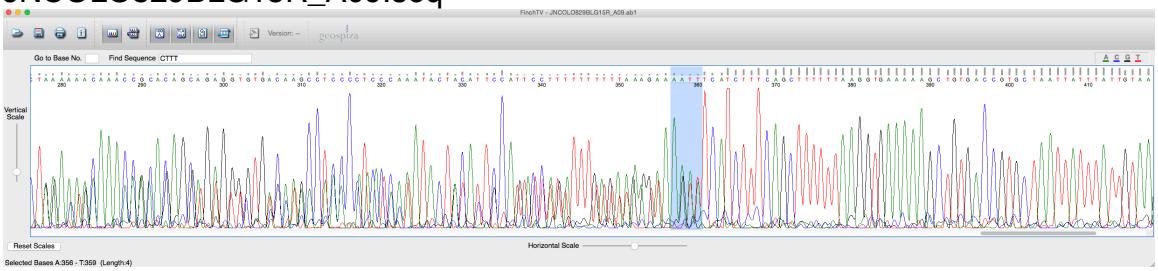
D 6 NT 1 A 12 24472263 24472270



JNCOLO829BLG15F_E06.seq



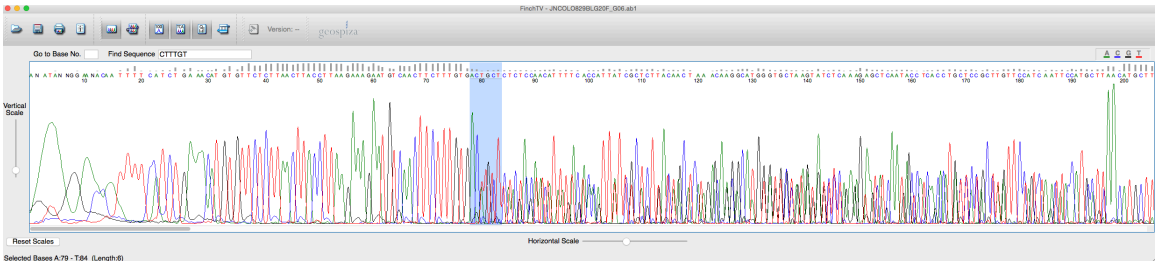
JNCOLO829BLG15R_A09.seq



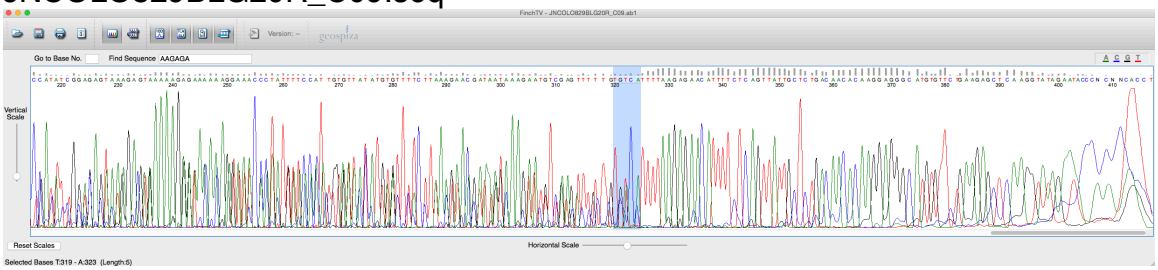
Germline 20

D 10 NT 1 A 2 129914554 129914565

JNCOLO829BLG20F_G06.seq



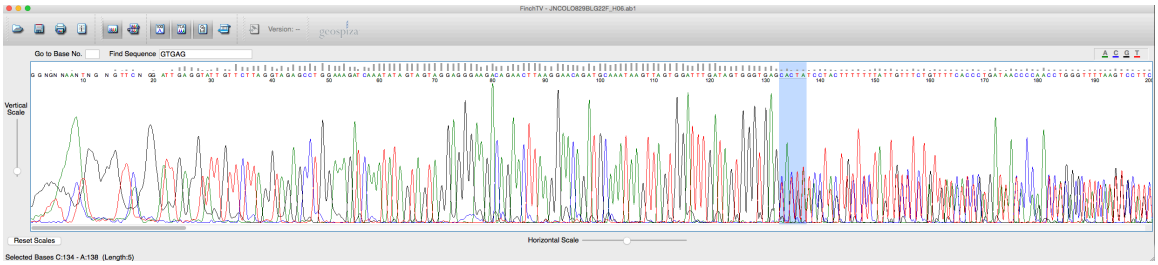
JNCOLO829BLG20R_C09.seq



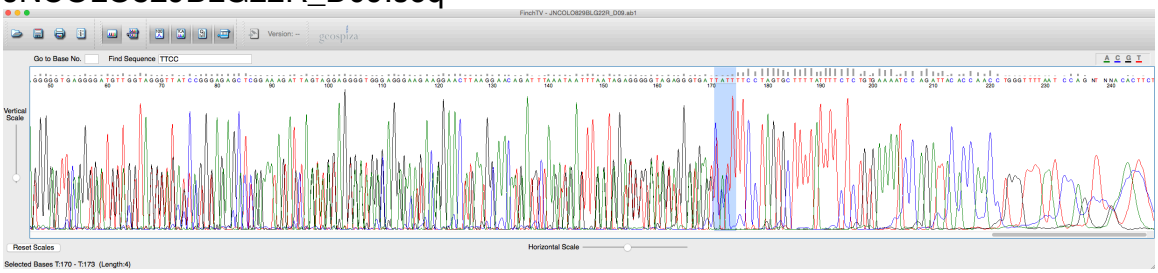
Germline 22

D 12 NT 3 CAC 12 92874068 92874081

JNCOLO829BLG22F_H06.seq



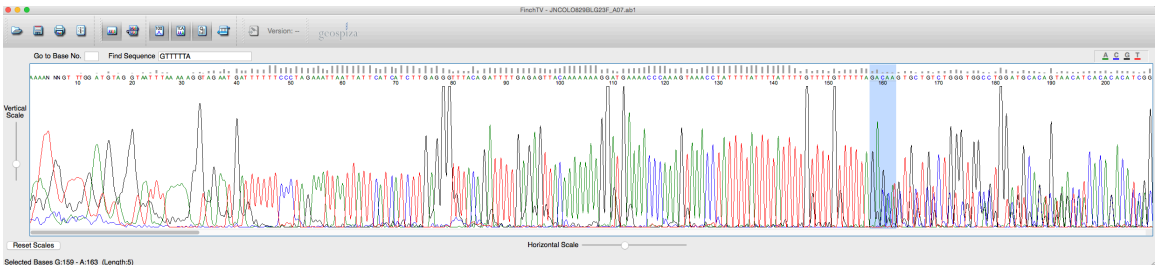
JNCOLO829BLG22R_D09.seq



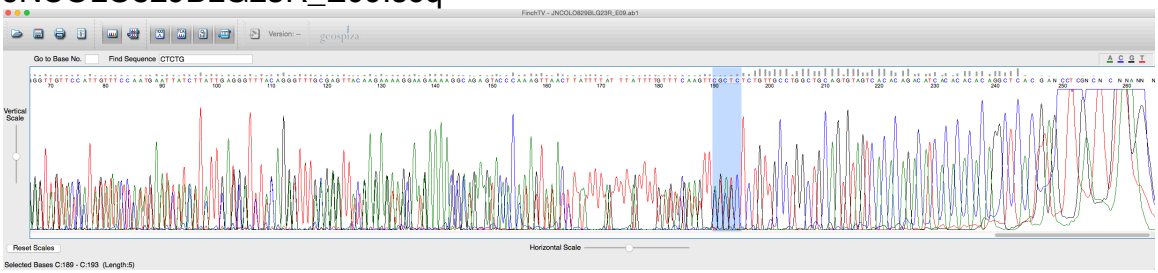
Germline 23

D 10 NT 1 G 22 24421642 24421653

JNCOLO829BLG23F_A07.seq

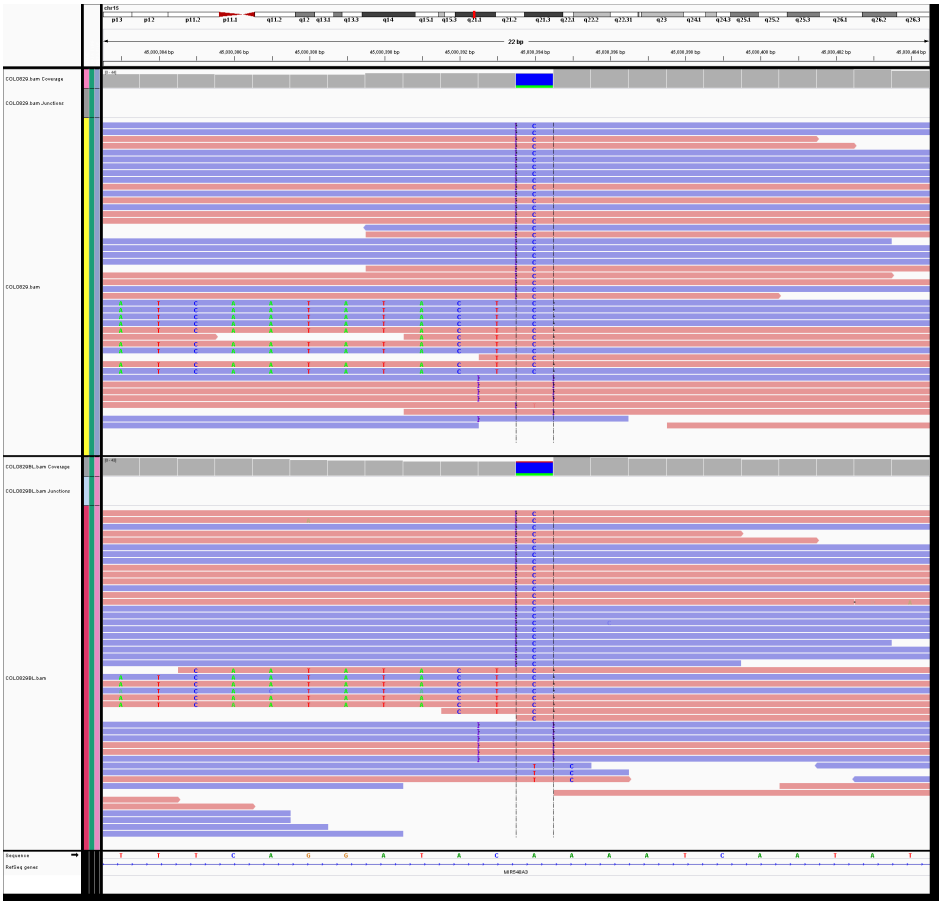


JNCOLO829BLG23R_E09.seq

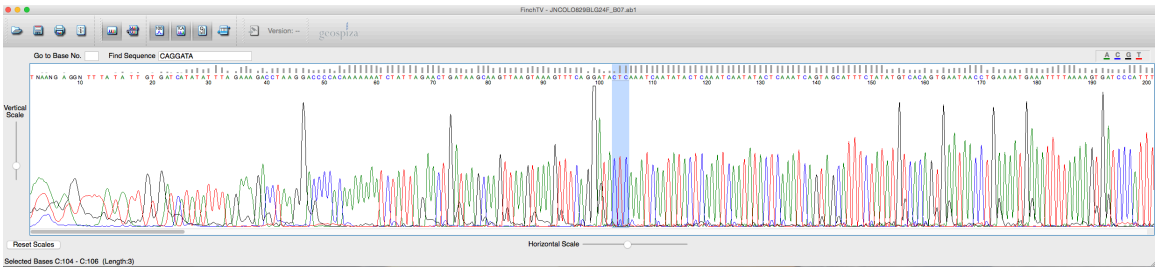


Germline 24

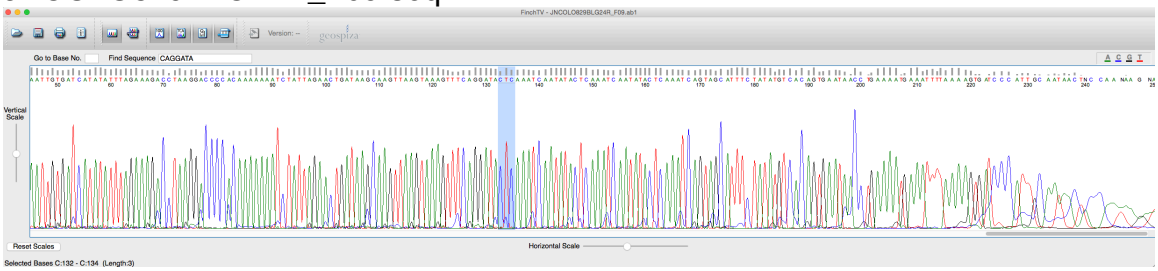
D 1 NT 16 TCAAATCAATATACTC 15 45030393 45030395



JNCOLO829BLG24F_B07.seq



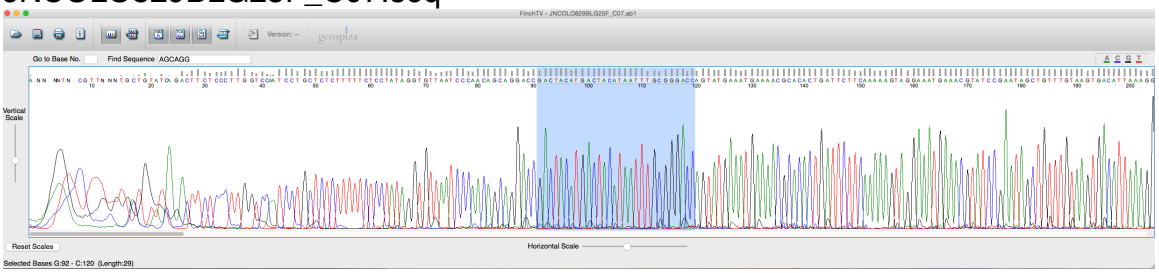
JNCOLO829BLG24R_F09.seq



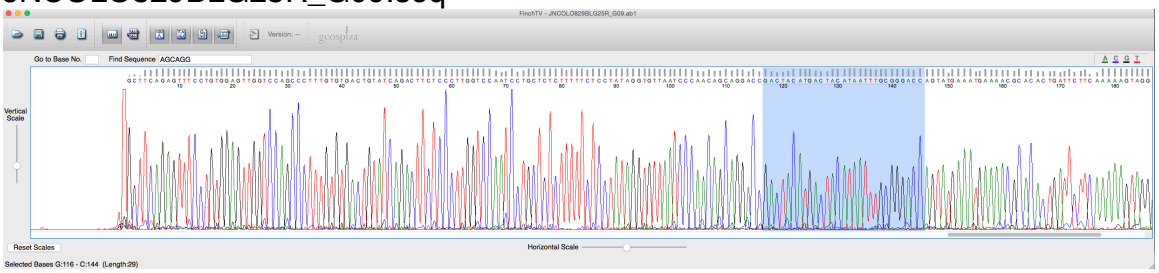
Germline 25

D 13 NT 20 GACTACATAATTTGCGGGAC 2 192873104 192873118

JNCOLO829BLG25F_C07.seq



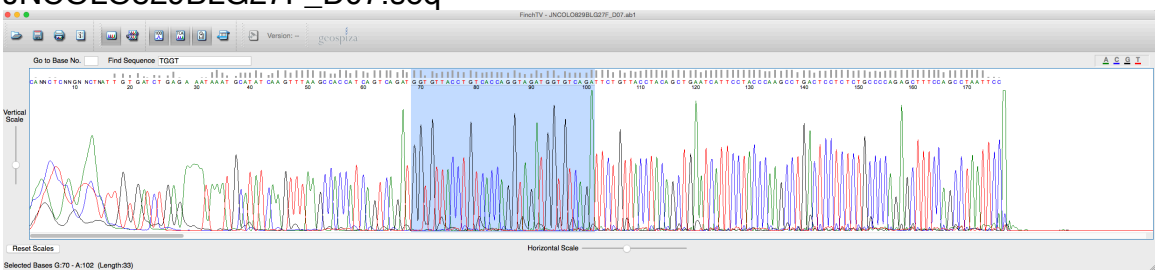
JNCOLO829BLG25R_G09.seq



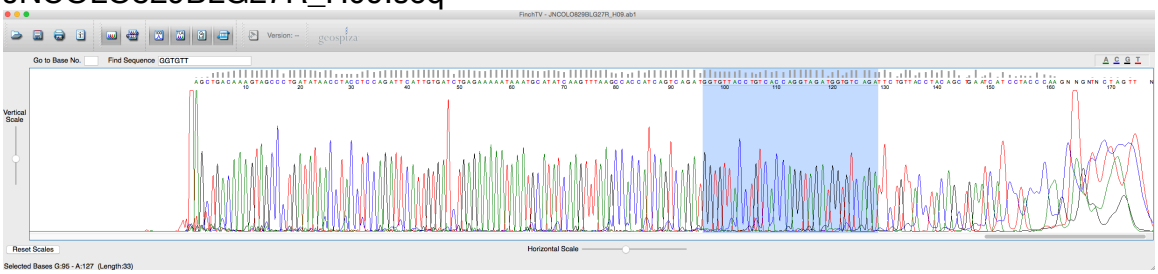
Germline 27

D 1 NT 33 GGTGTTACCTGTCACCAGGTAGATGGTGTCAGA 6 36908476 36908478

JNCOLO829BLG27F_D07.seq



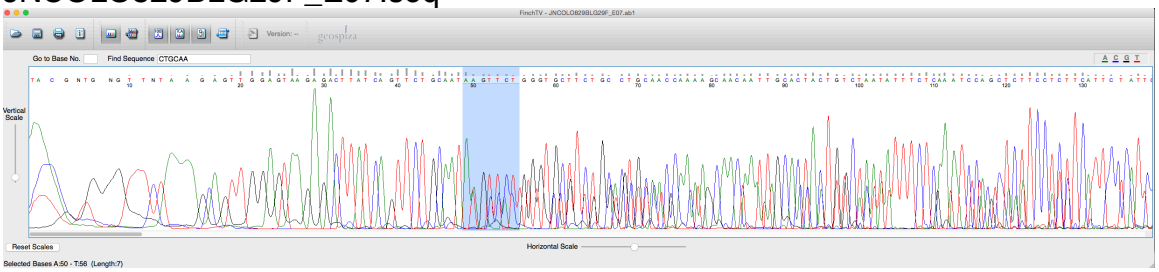
JNCOLO829BLG27R_H09.seq



Germline 29

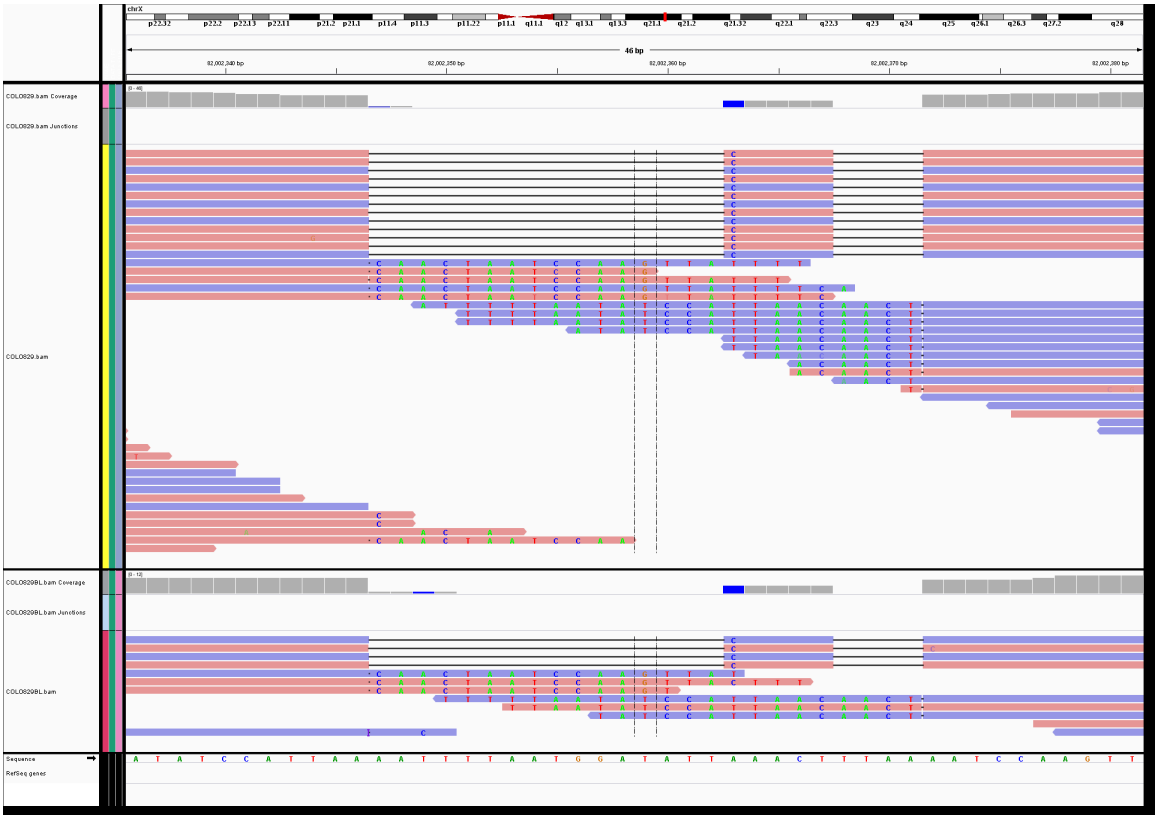
D 13 NT 4 AGTT 10 37392070 37392084

JNCOLO829BLG29F_E07.seq

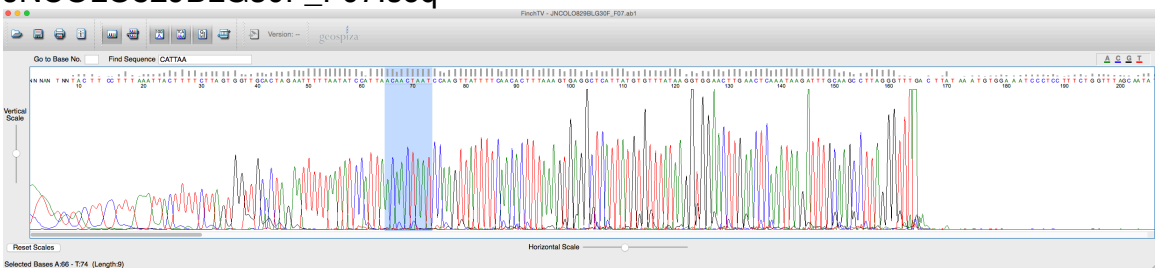


Germline 30

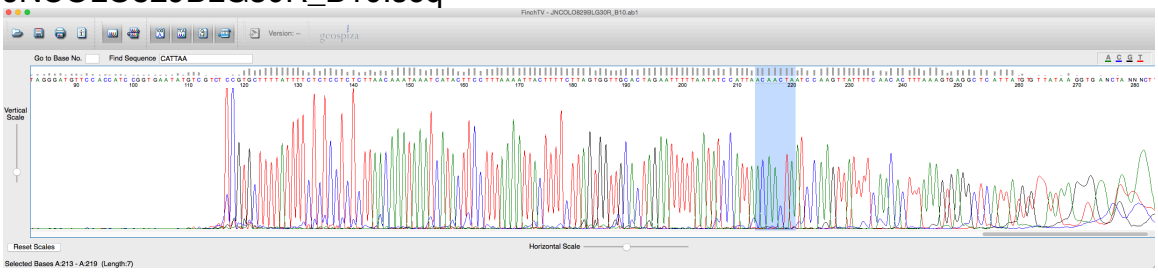
D 25 NT 5 CAACT X 82002346 82002372



JNCOLO829BLG30F_F07.seq



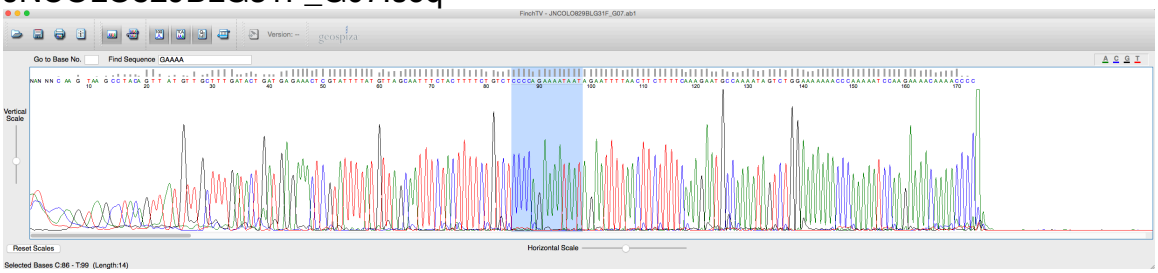
JNCOLO829BLG30R_B10.seq



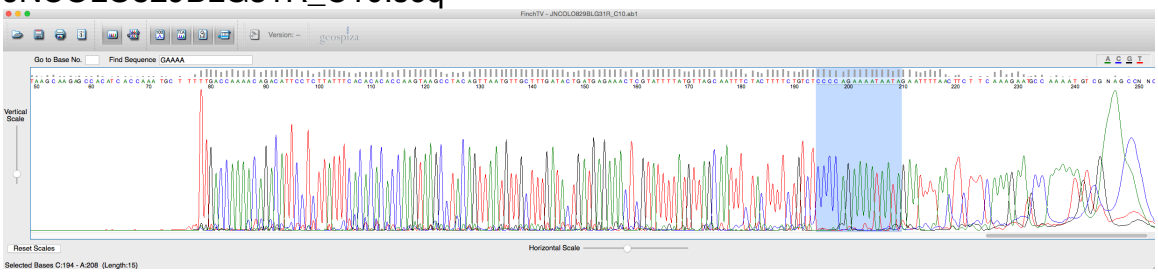
Germline 31

D	20	NT	13	CCCAGAAAATAAT	4	148852782	148852803

JNCOLO829BLG31F_G07.seq

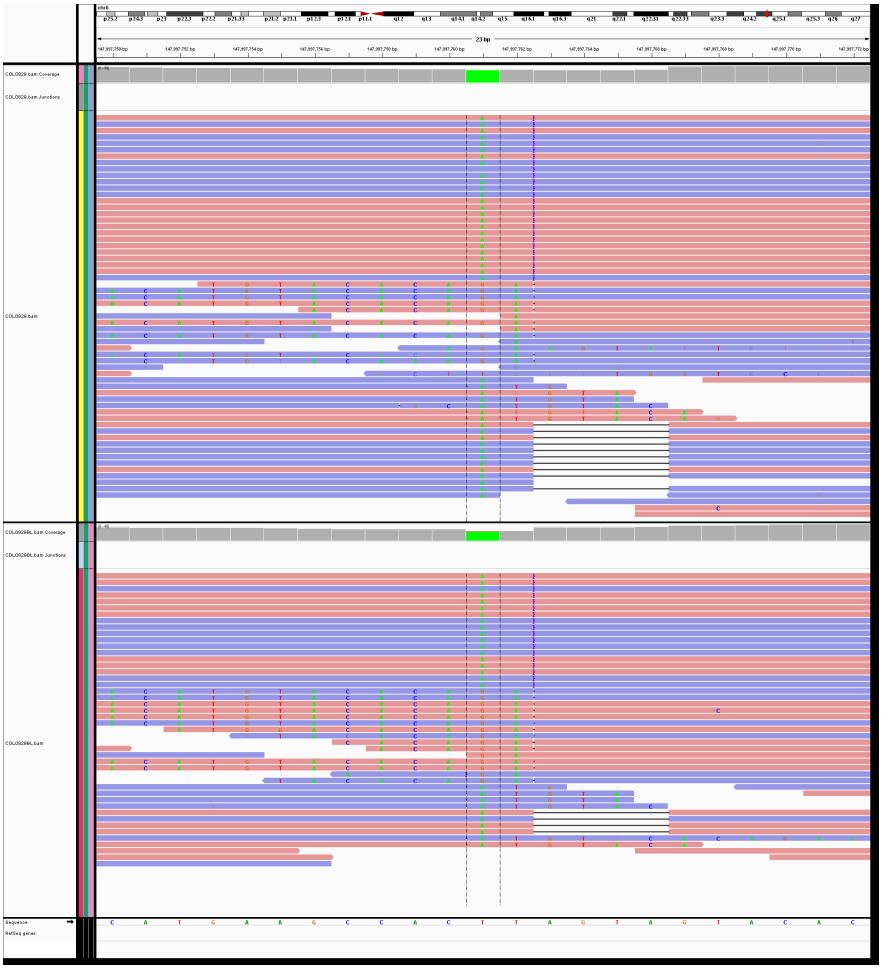


JNCOLO829BLG31R_C10.seq

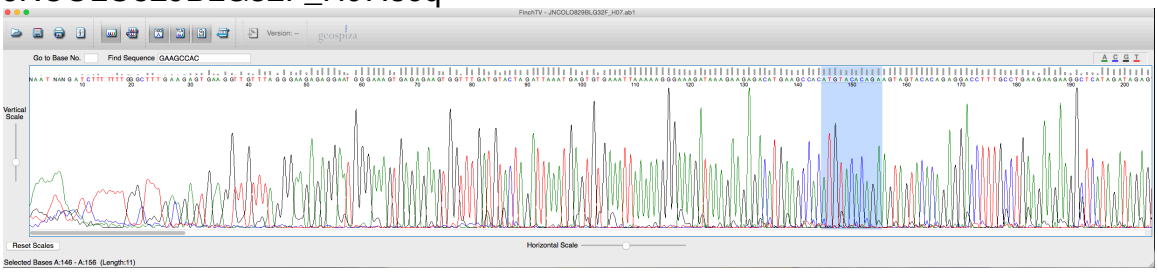


Germline 32

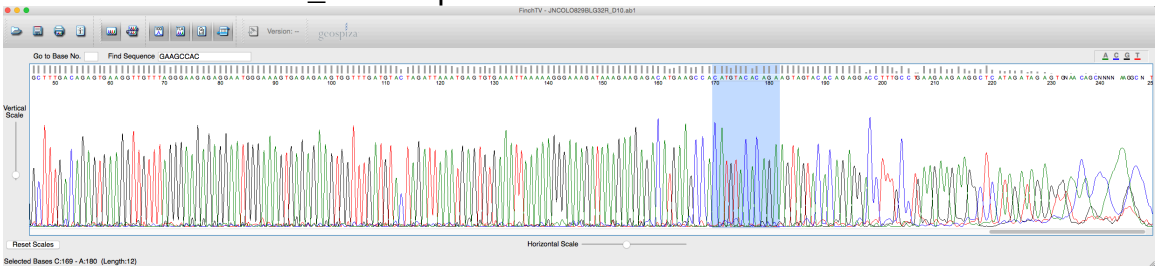
D	2	NT	11	ATGTACACAGA	6	147997760	147997763



JNCOLO829BLG32F_H07.seq



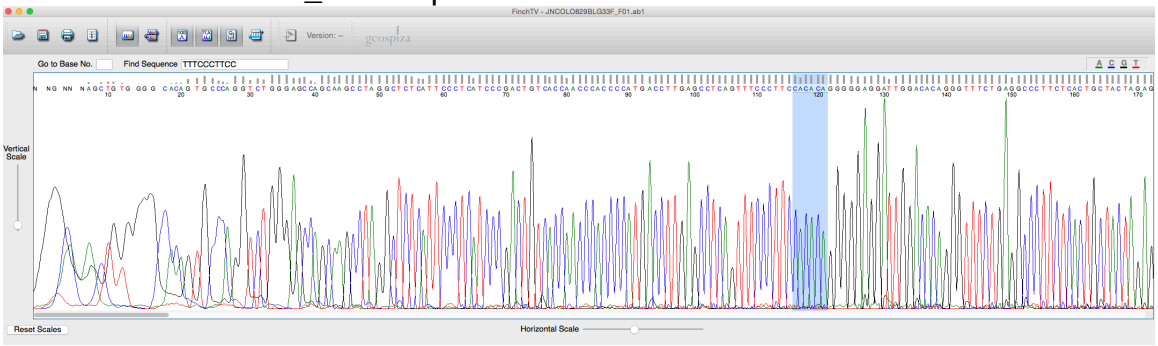
JNCOLO829BLG32R_D10.seq



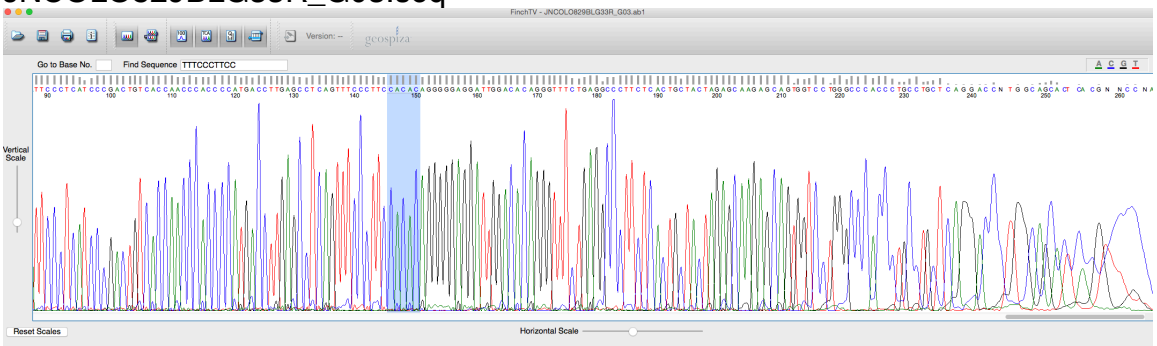
Germline 33

D 9 NT 3 ACA 1 47425457 47425467

JNCOLO829BLG33F_F01.seq

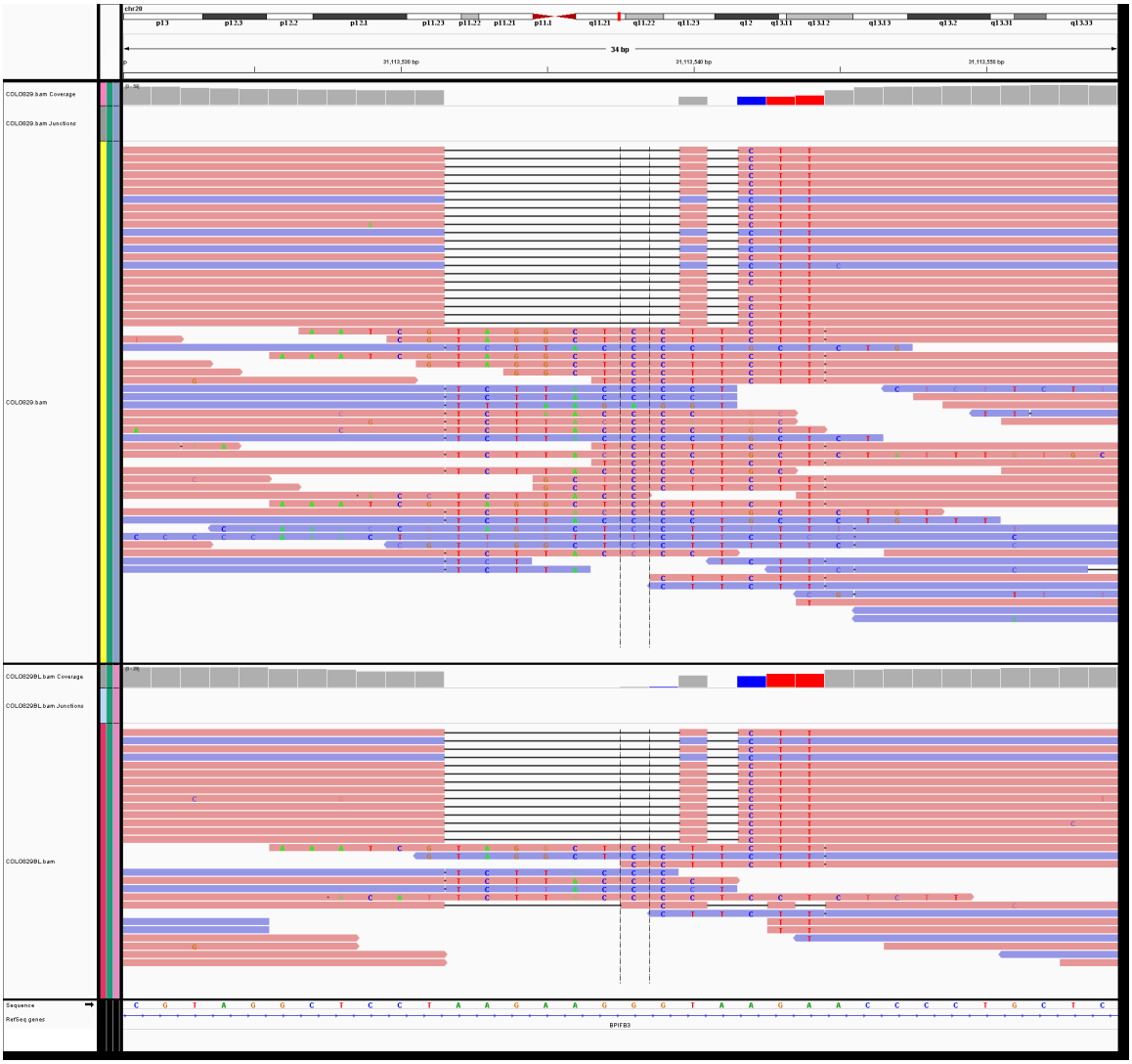


JNCOLO829BLG33R_G03.seq

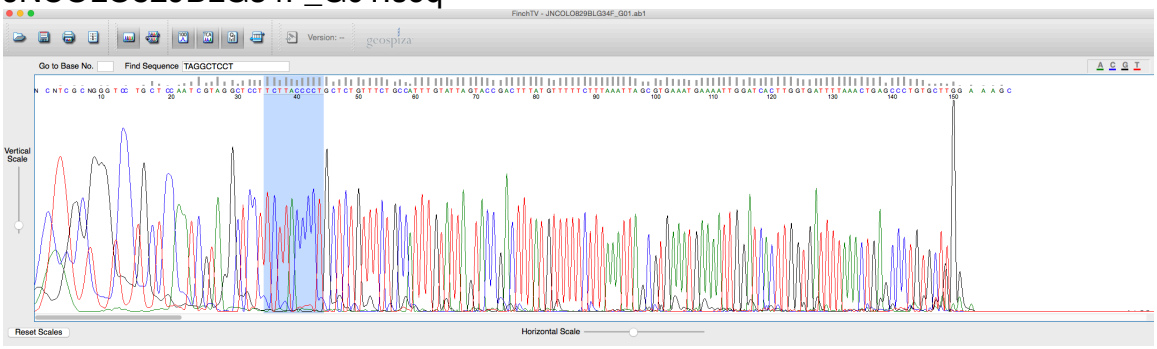


Germline 34

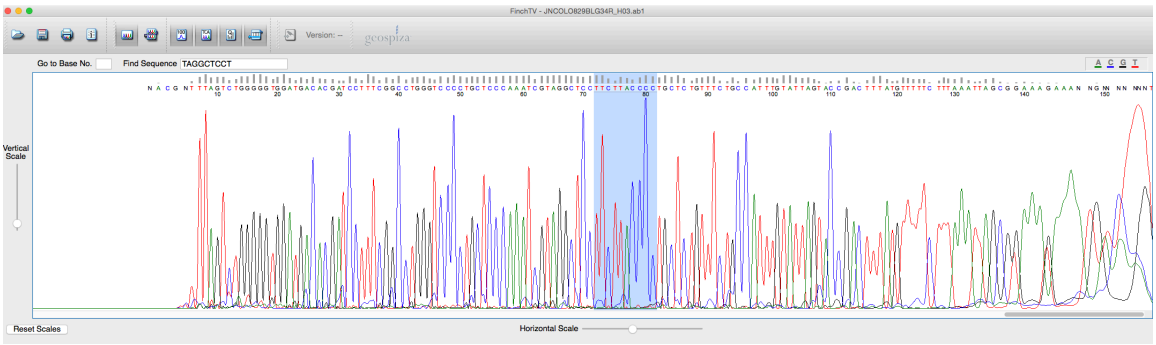
D 13 NT 4 TCTT 20 31113531 31113545



JNCOLO829BLG34F_G01.seq

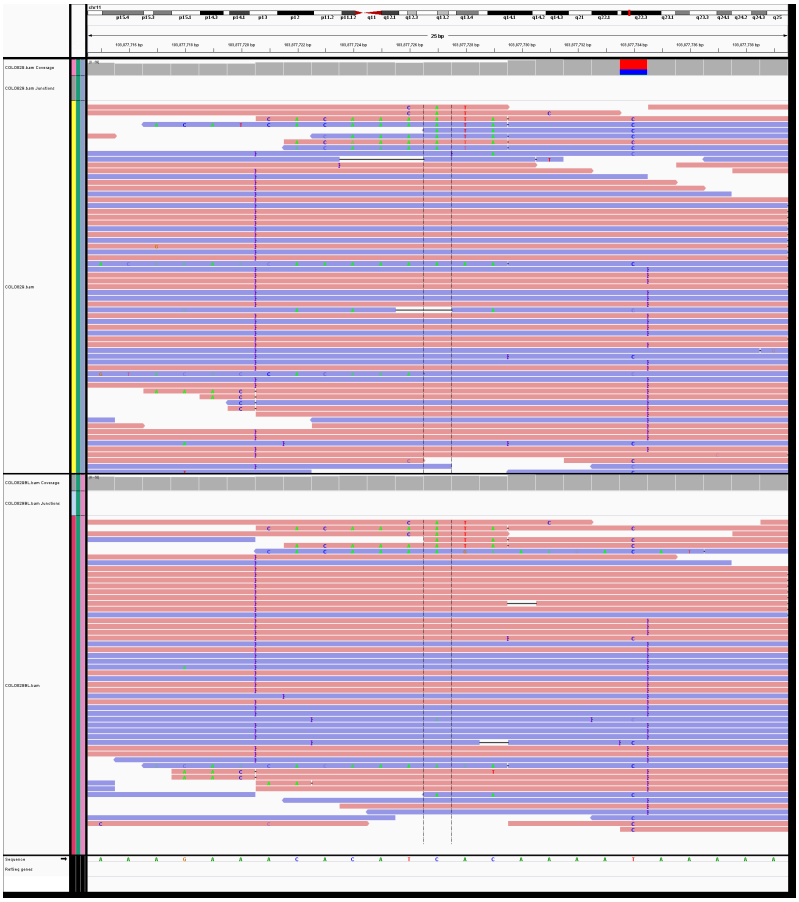


JNCOLO829BLG34R_H03.seq

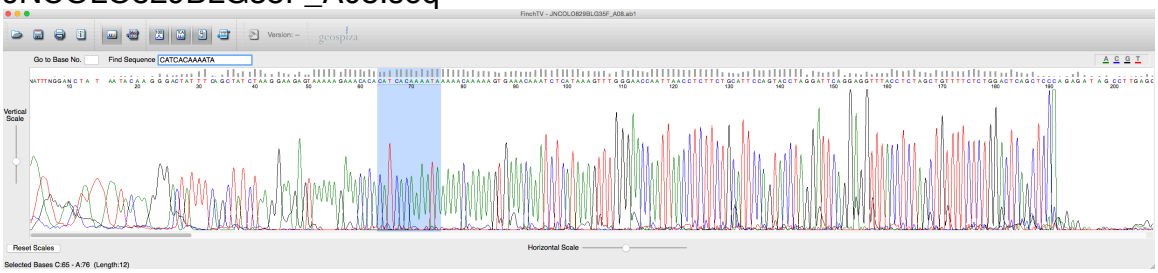


Germline 35

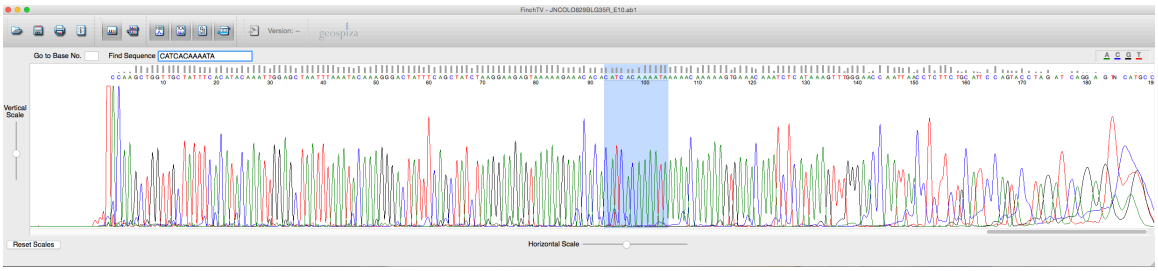
D 4 NT 12 CATCACAAAATA 11 103877725 103877730



JNCOLO829BLG35F_A08.seq

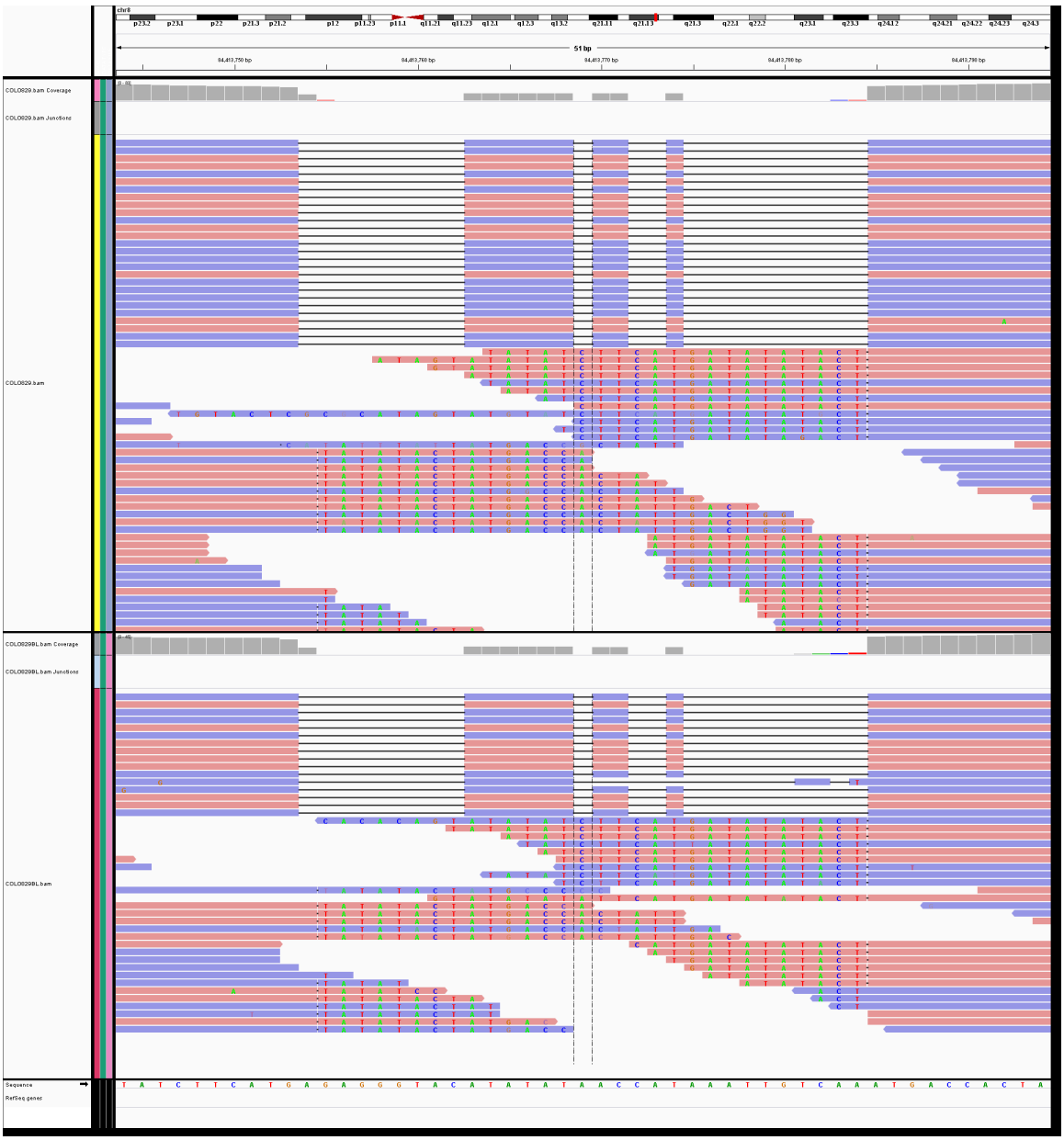


JNCOLO829BLG35R_E10.seq

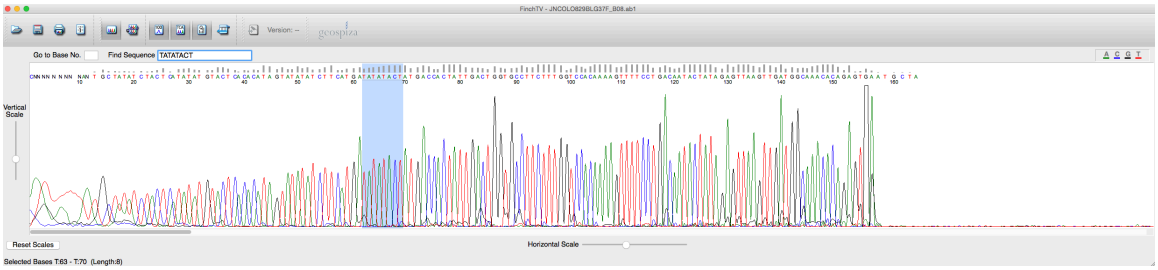


Germline 37

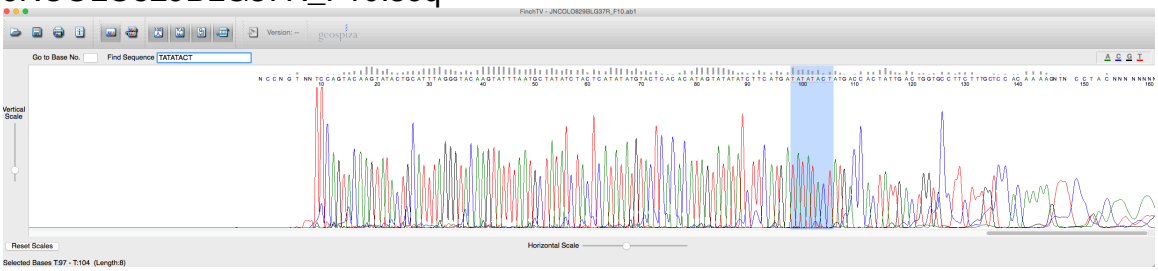
D 30 NT 8 TATATACT 8 84413754 84413785



JNCOLO829BLG37F_B08.seq

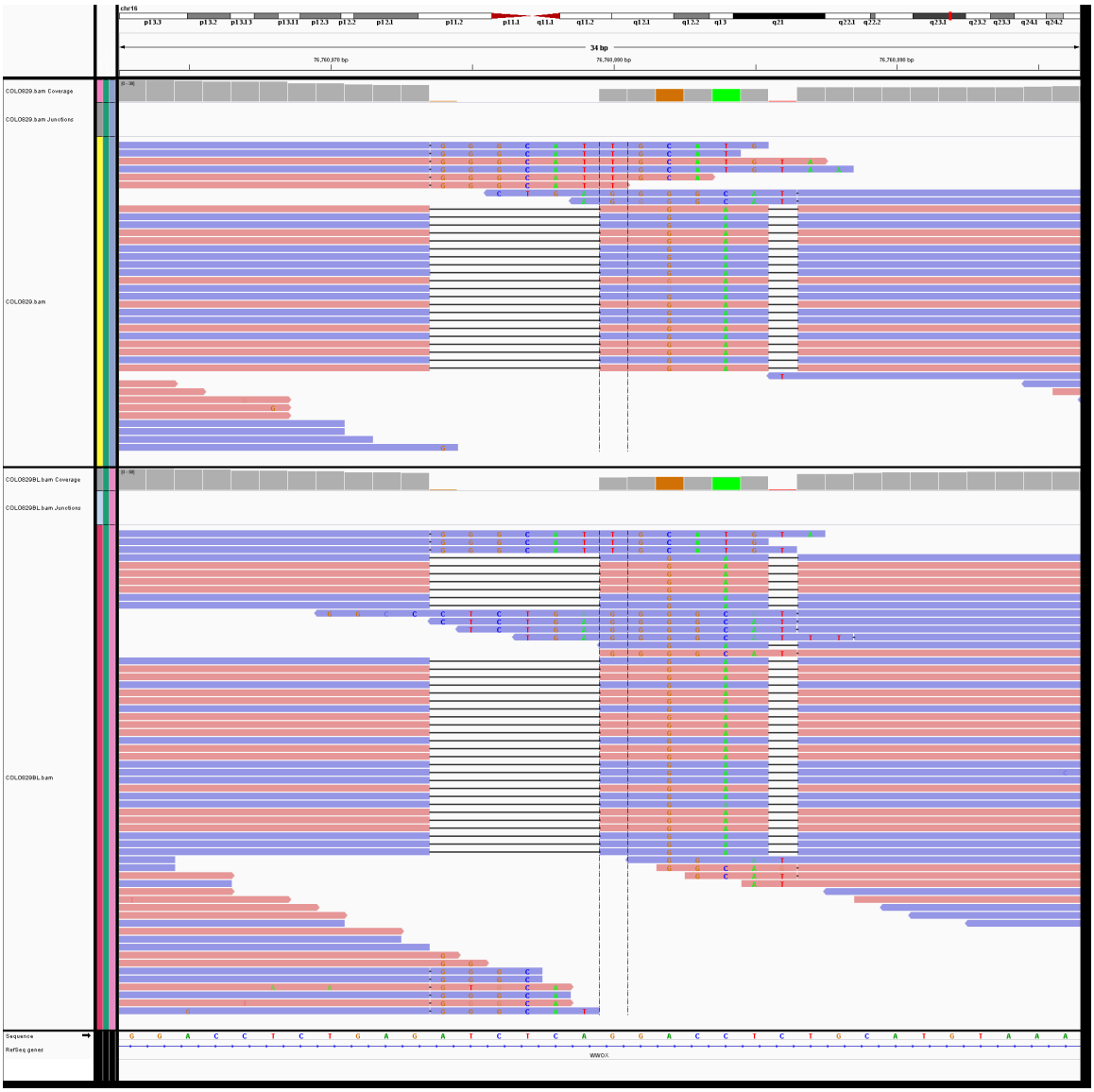


JNCOLO829BLG37R_F10.seq

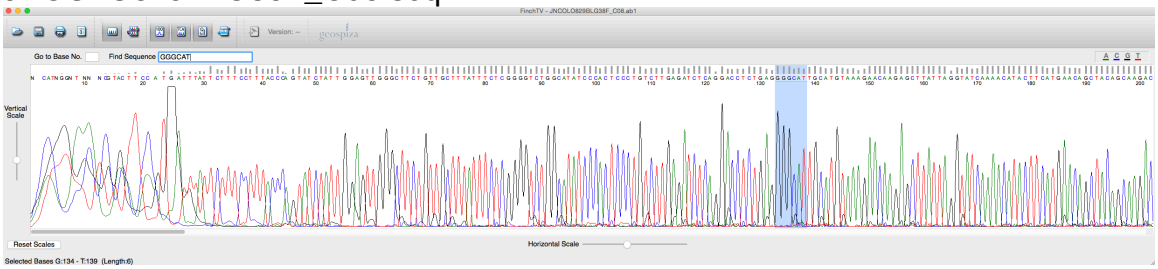


Germline 38

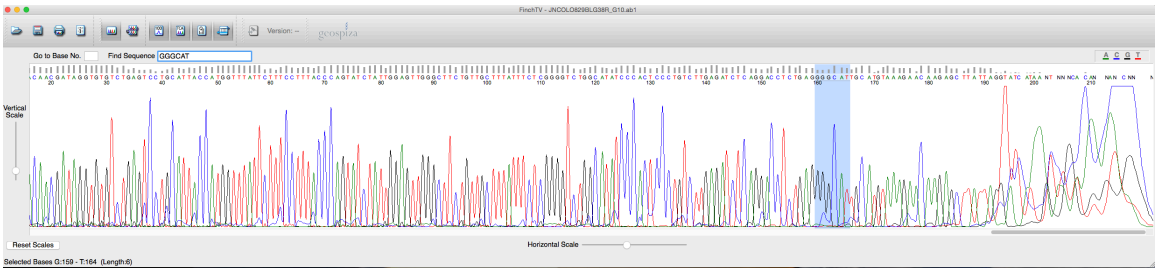
D 13 NT 6 GGGCAT 16 76760873 76760887



JNCOLO829BLG38F_C08.seq

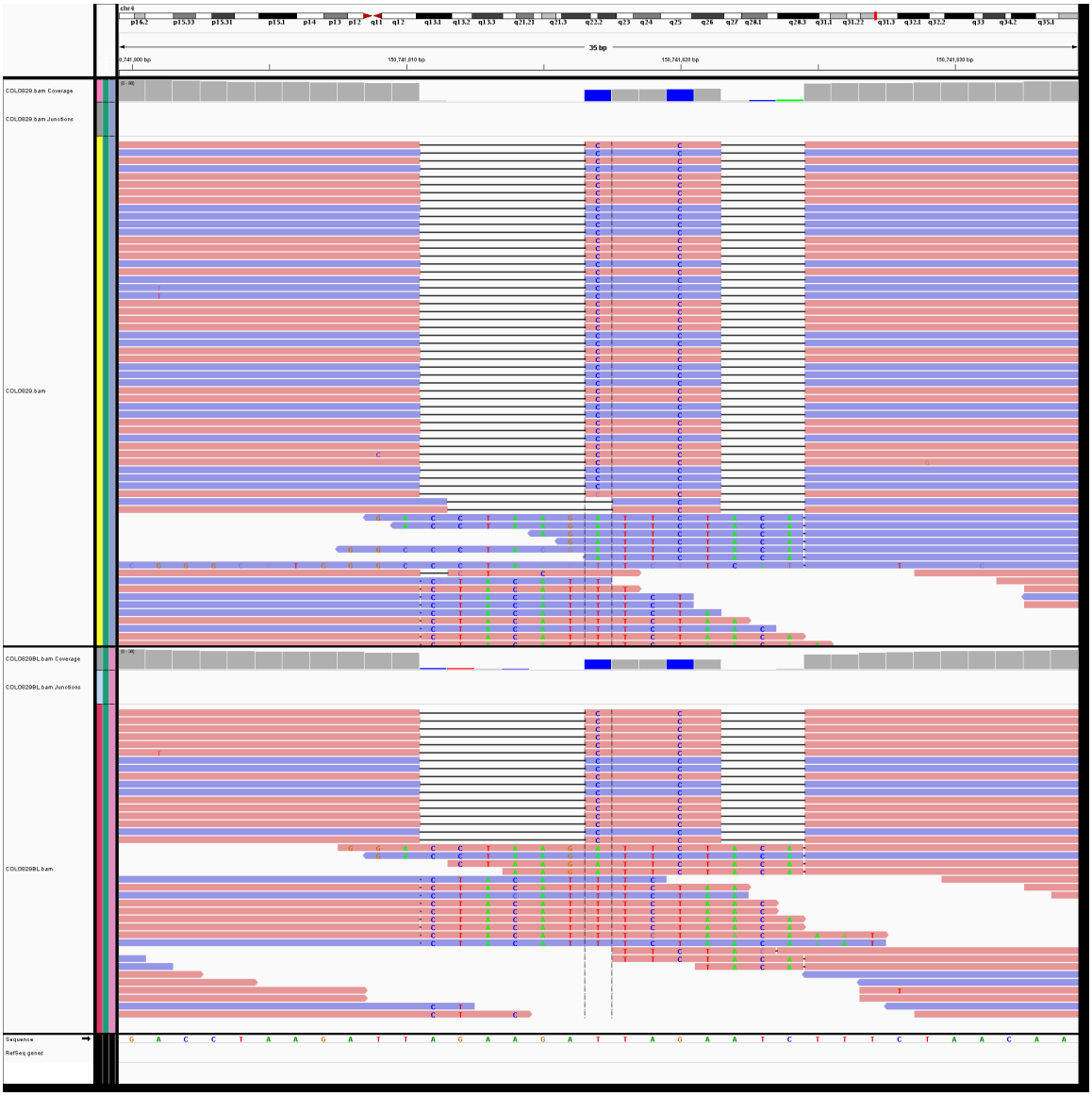


JNCOLO829BLG38R_G10.seq

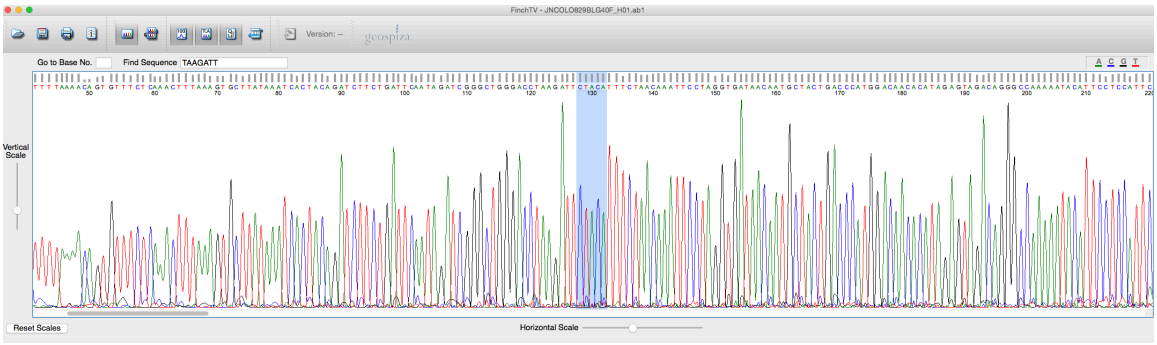


Germline 40

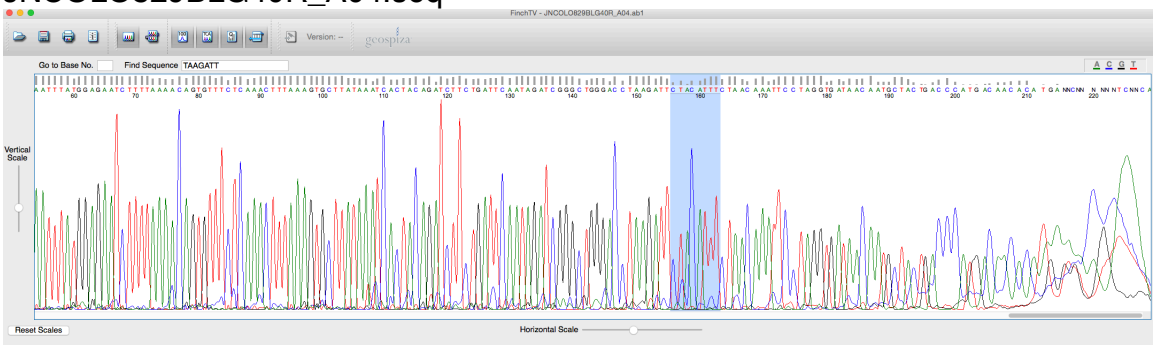
D 14 NT 5 CTACA 4 150741810 150741825



JNCOLO829BLG40F_H01.seq

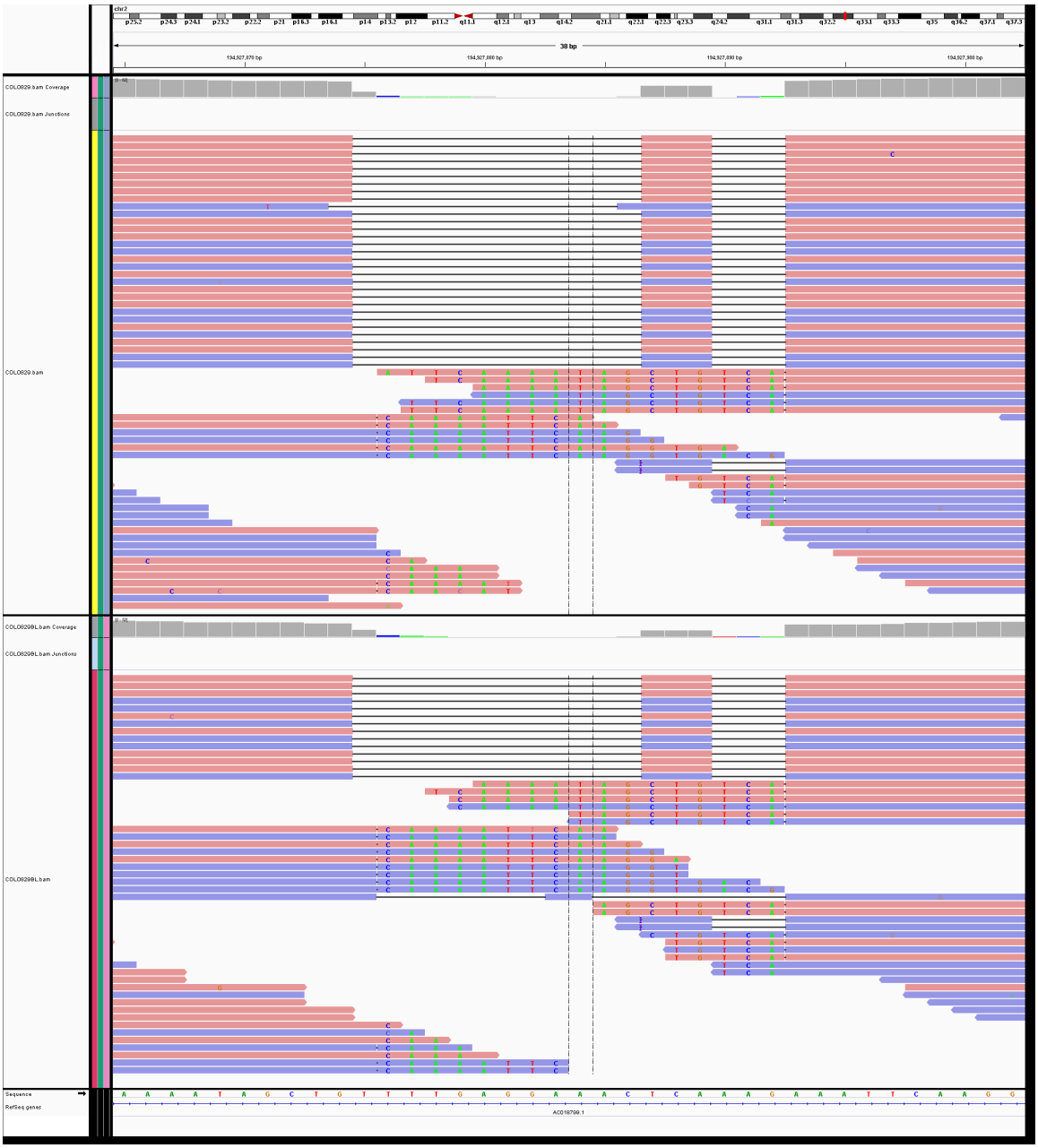


JNCOLO829BLG40R_A04.seq

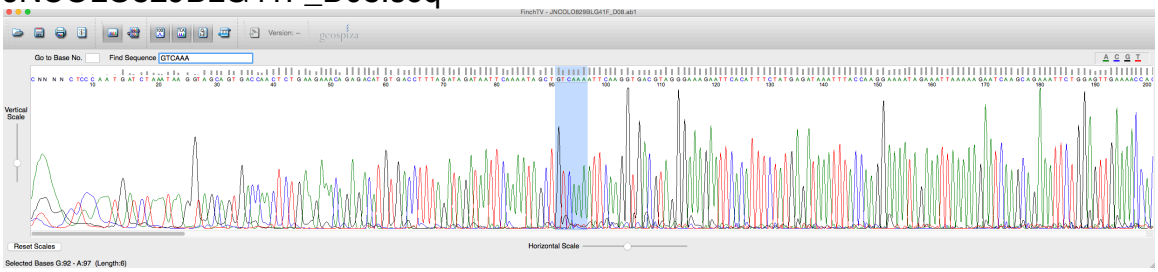


Germline 41

D	17	NT	2	CA	2	194927875	194927893



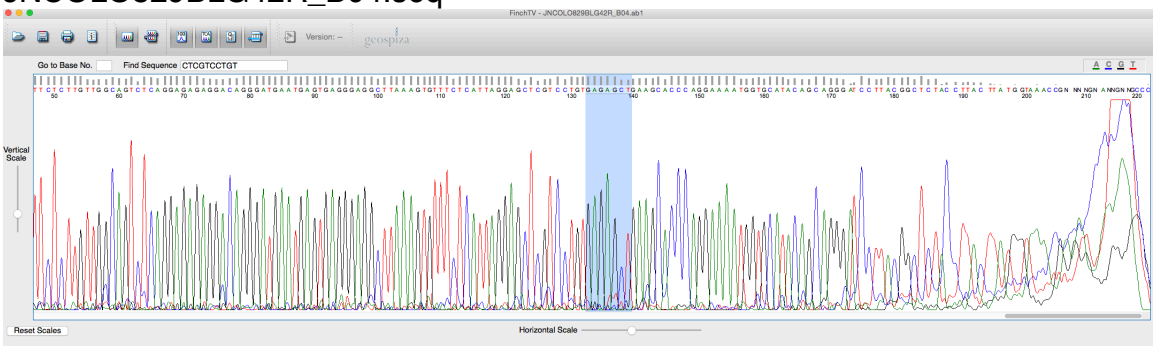
JNCOLO829BLG41F_D08.seq



Germline 42

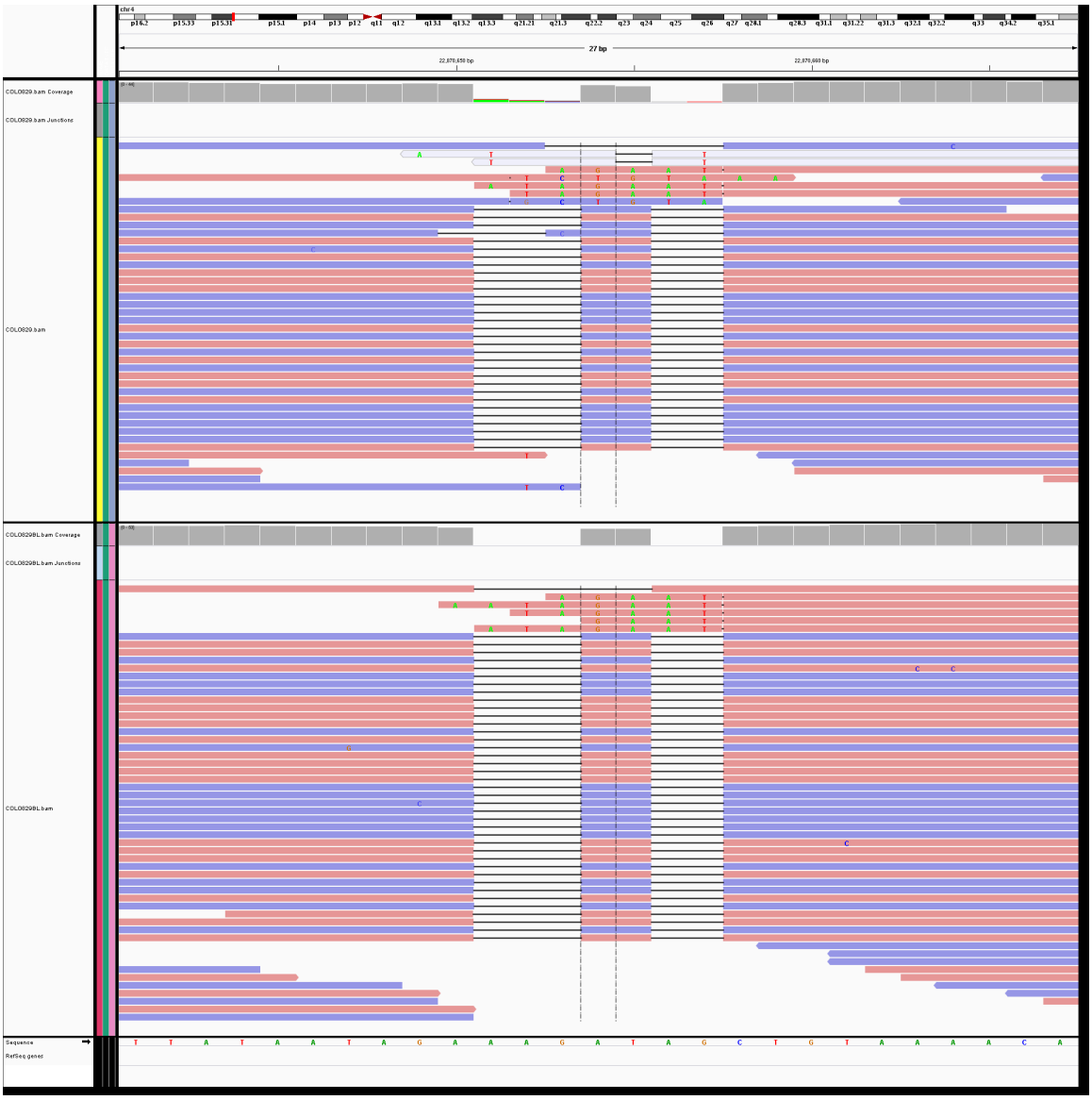
D	9	NT	3	AGA	3	118791327	118791337

JNCOLO829BLG42R_B04.seq

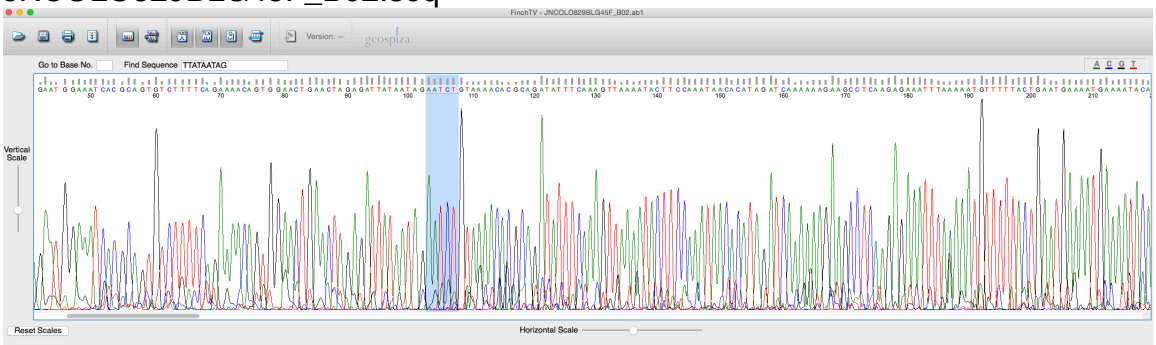


Germline 45

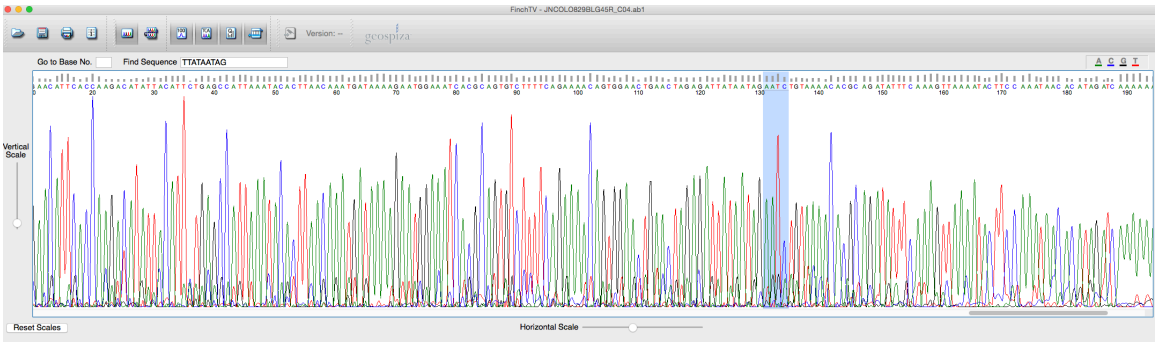
D 6 NT 1 T 4 22870651 22870658



JNCOLO829BLG45F_B02.seq

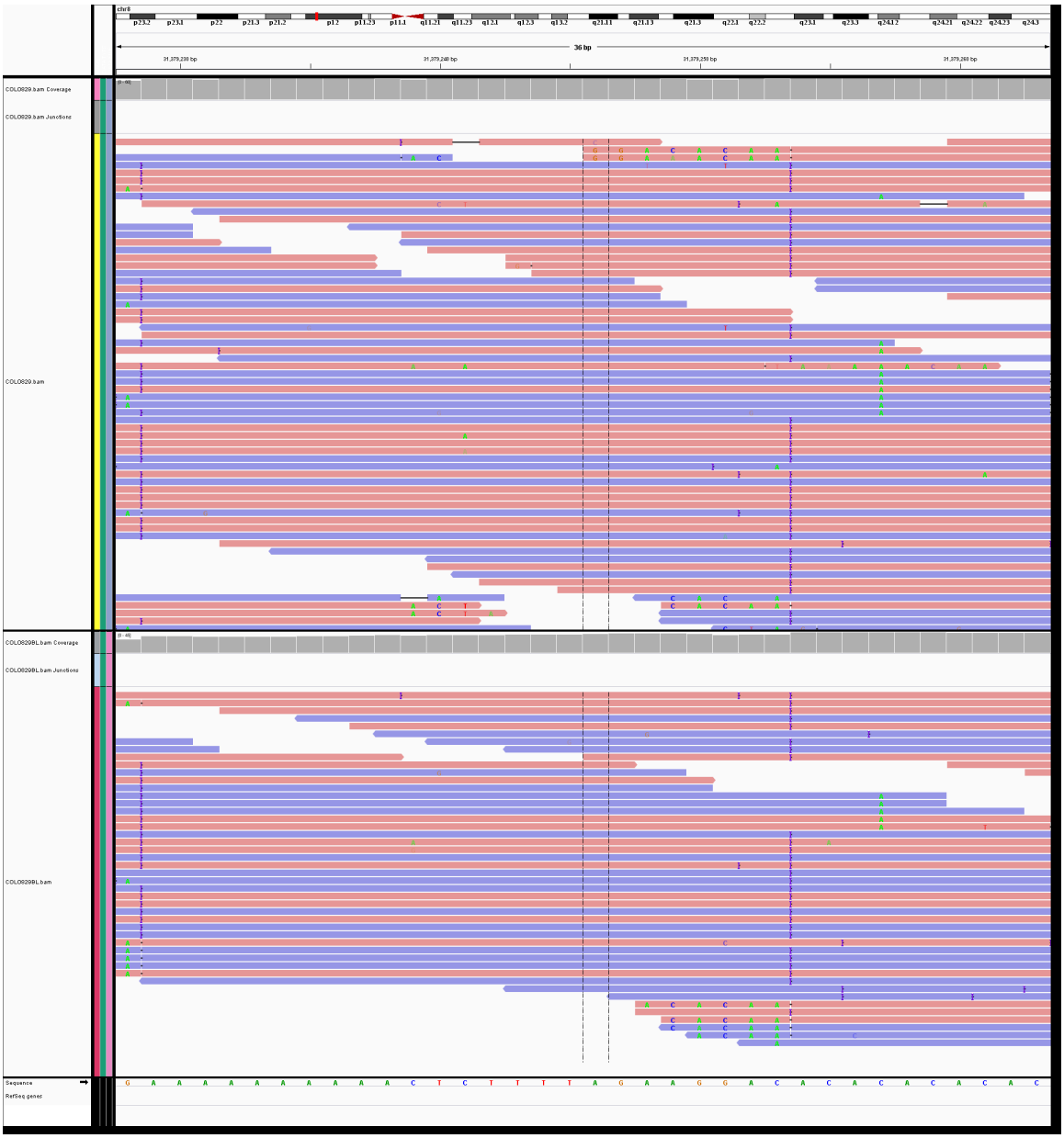


JNCOLO829BLG45R_C04.seq

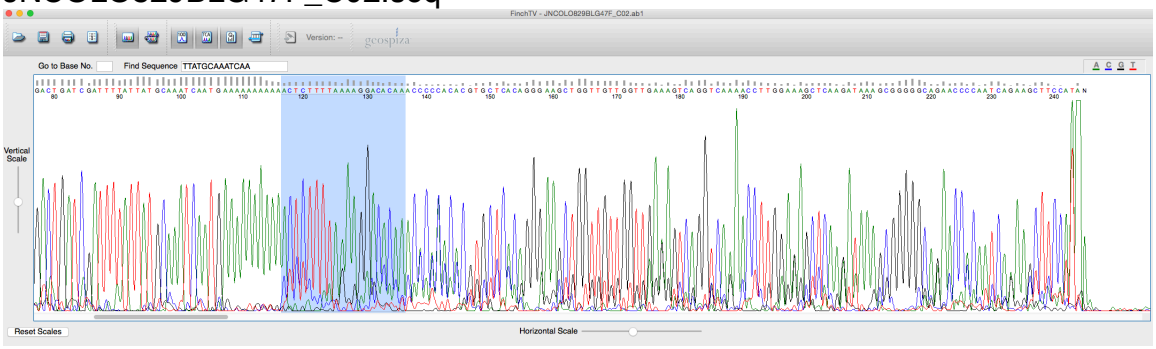


Germline 47

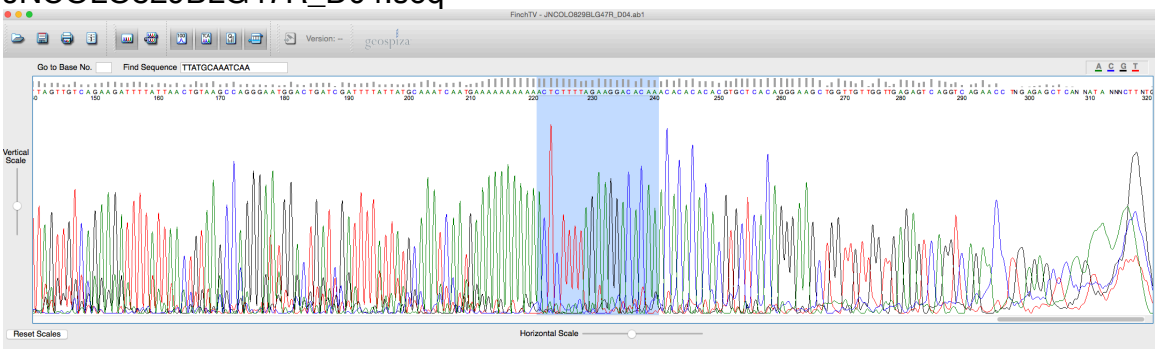
D 15 NT 20 ACTCTTTTAGAAGGACACAA 8 31379238 31379254



JNCOLO829BLG47F_C02.seq

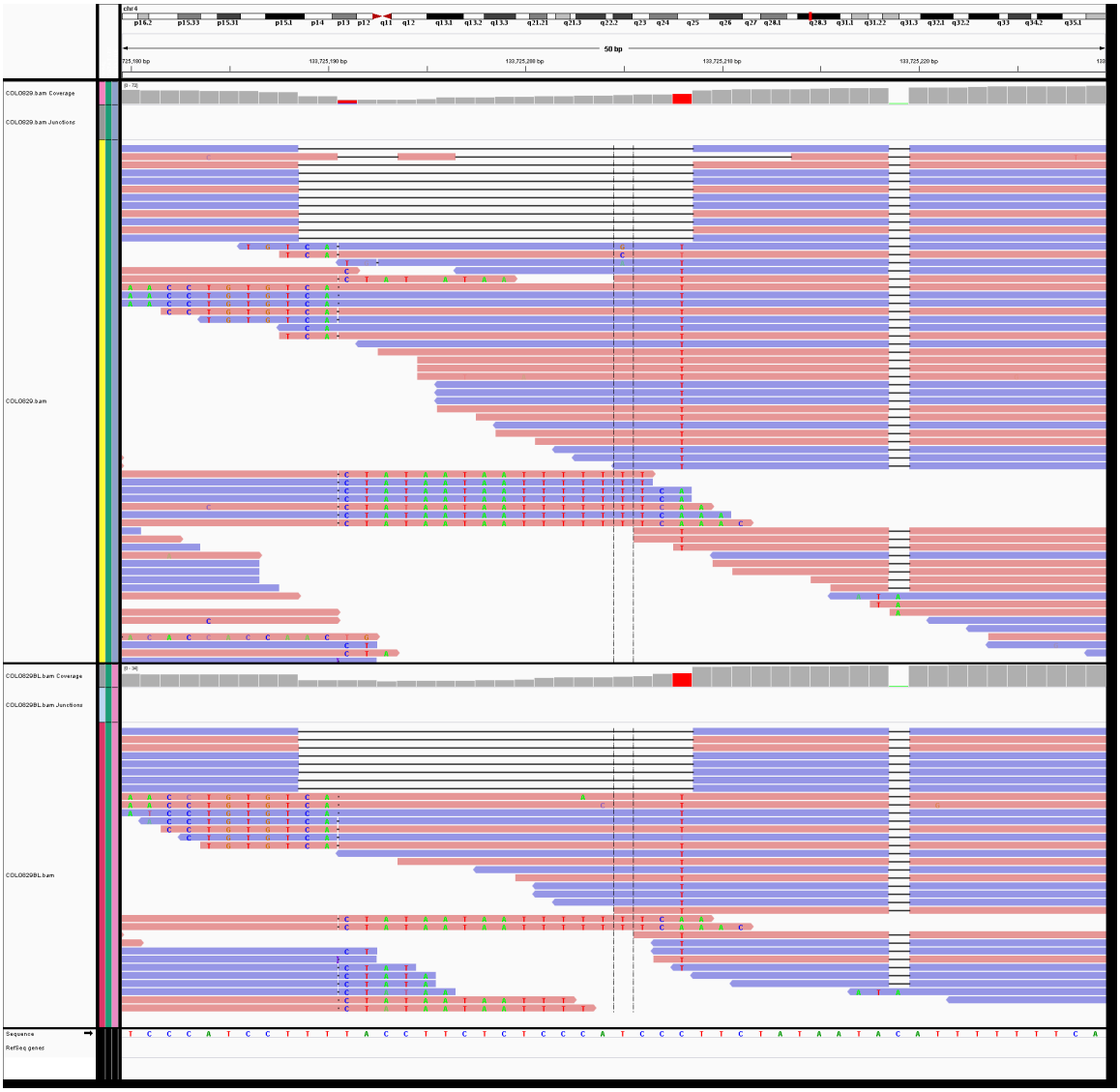


JNCOLO829BLG47R_D04.seq

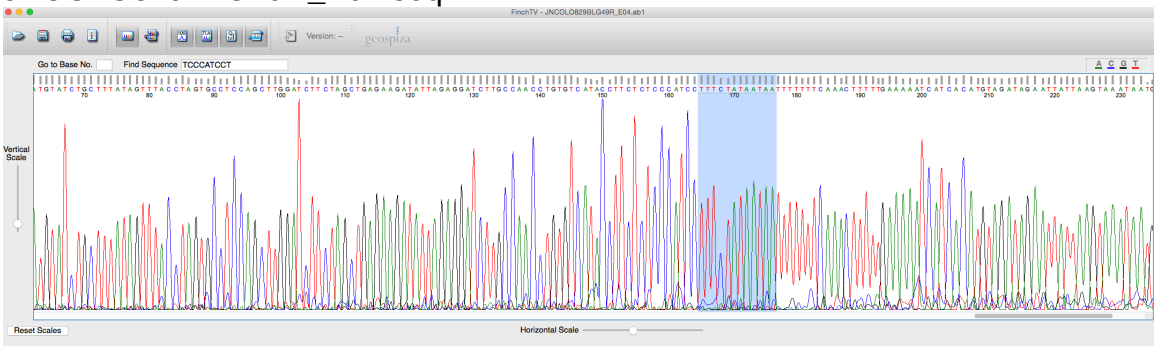


Germline 49

D	29	NT	8	CTATAATA	4	133725190	133725220

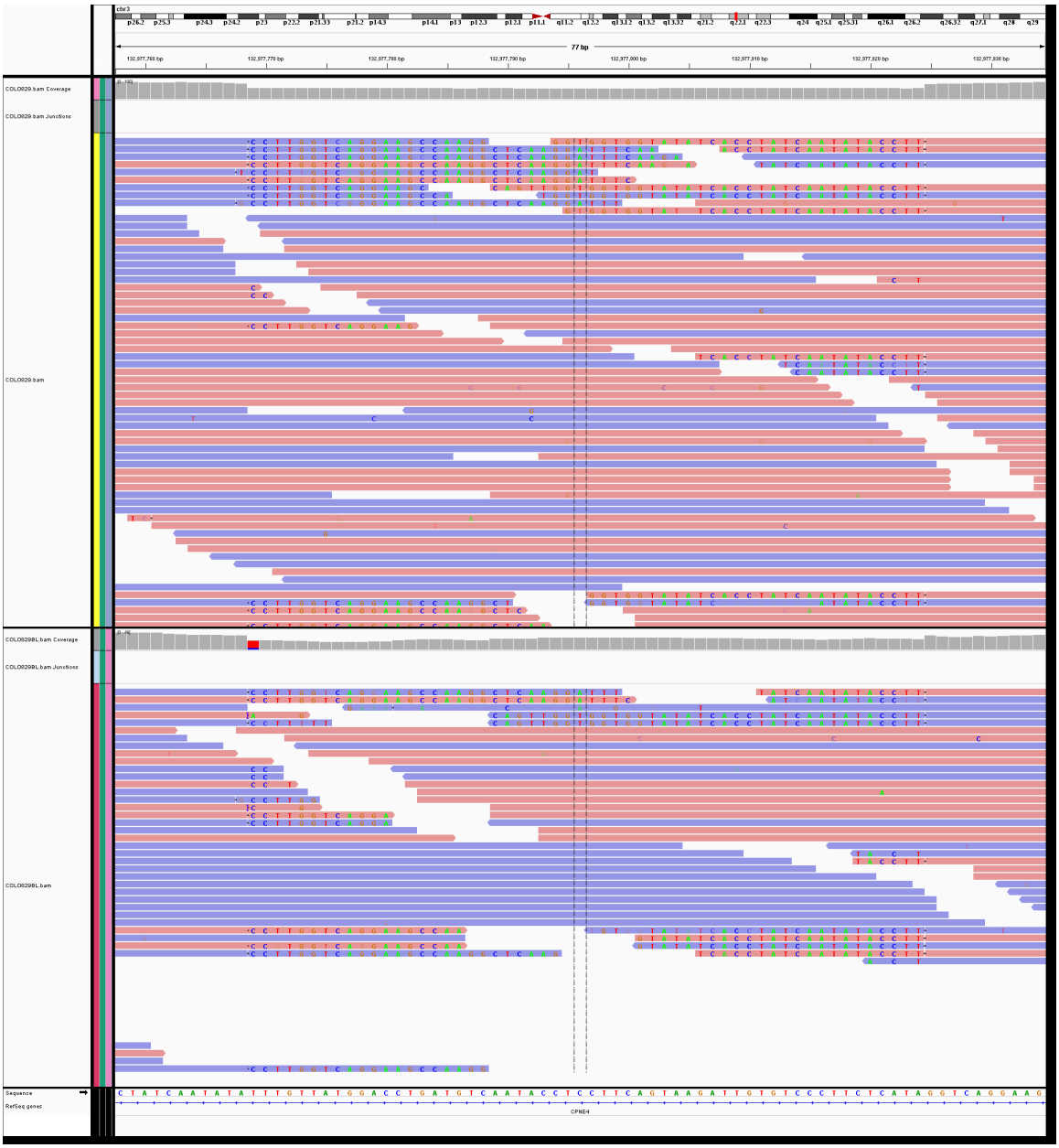


JNCOLO829BLG49R_E04.seq

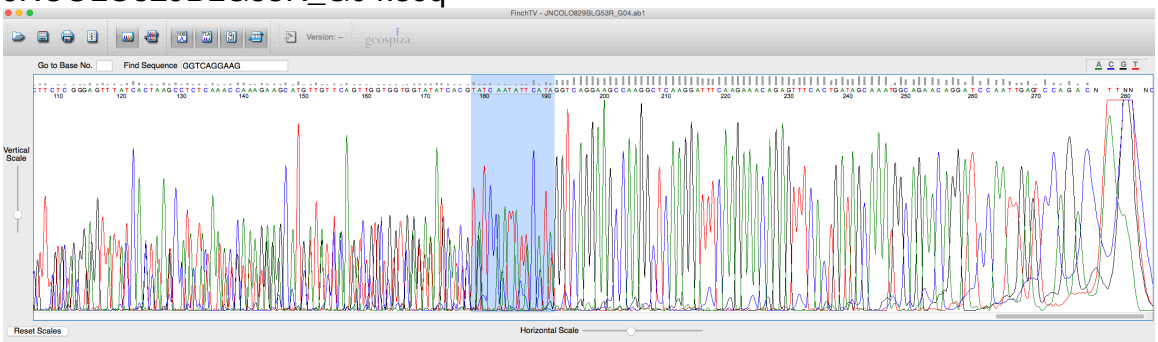


Germline 53

D 56 NT 4 CCTT 3 132977768 132977825



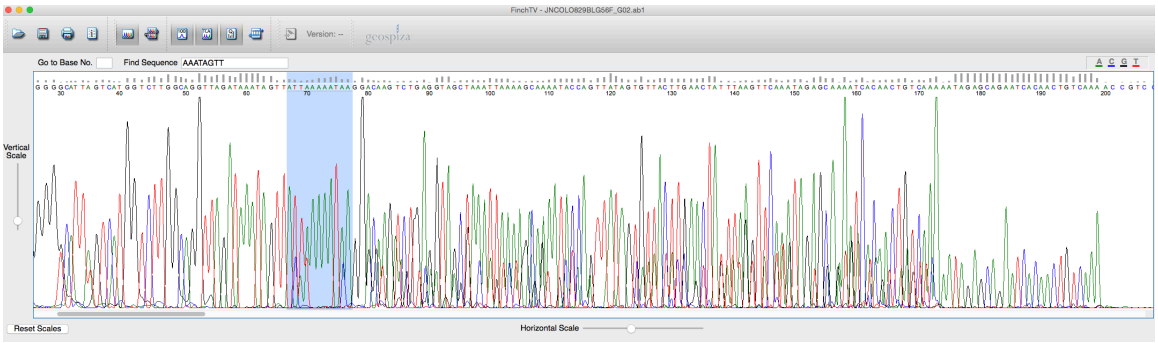
JNCOLO829BLG53R_G04.seq



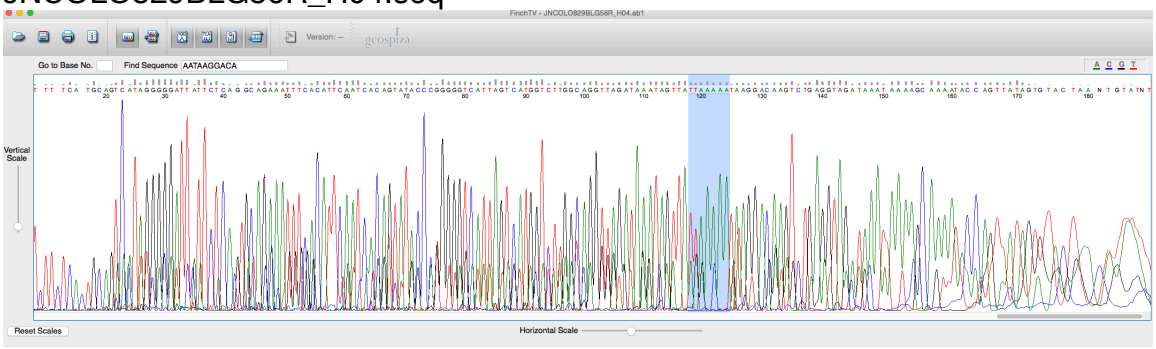
Germline 56

D 22 NT 5 TTAAA 13 62523004 62523027

JNCOLO829BLG56F_G02.seq

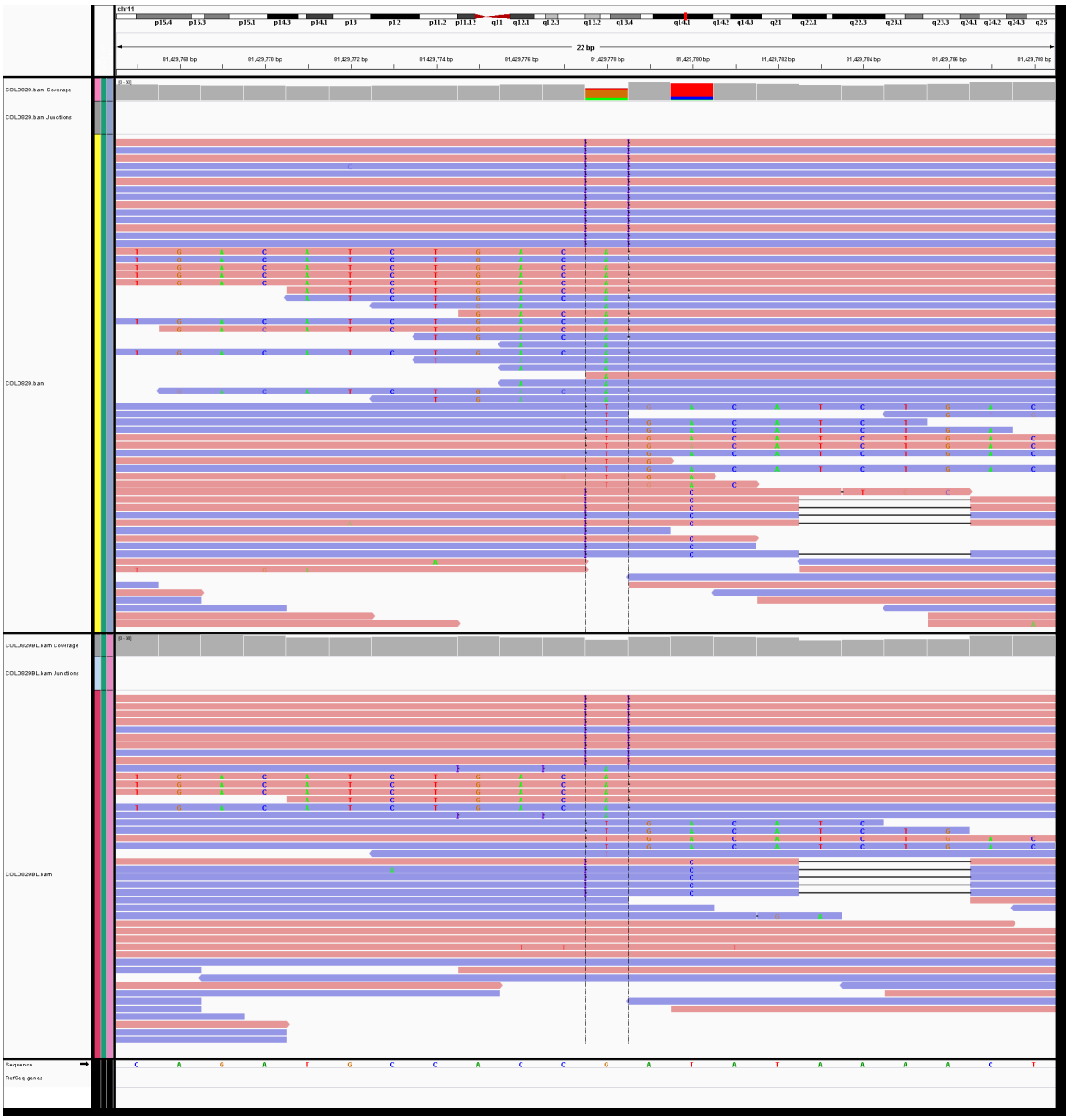


JNCOLO829BLG56R_H04.seq

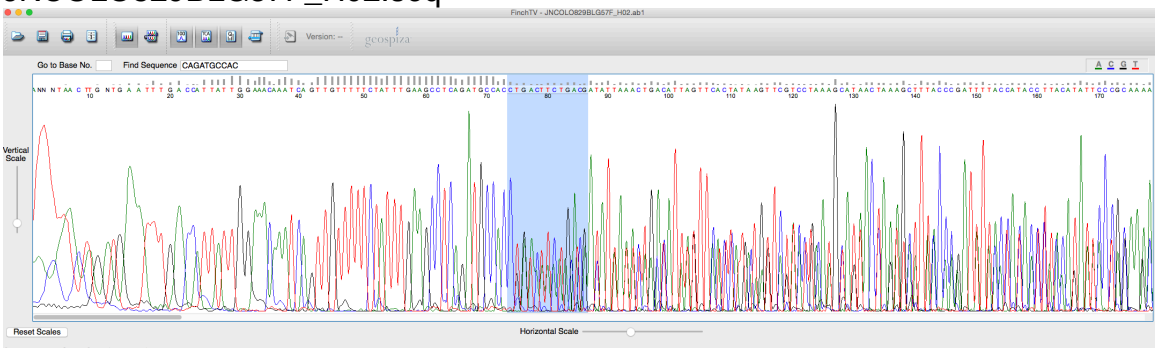


Germline 57

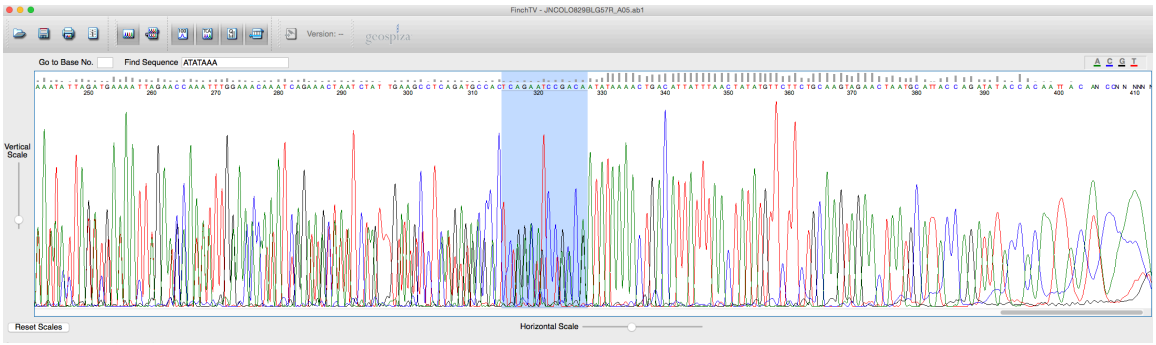
D	1	NT	12	TGACATCTGACA	11	81429777	81429779



JNCOLO829BLG57F_H02.seq

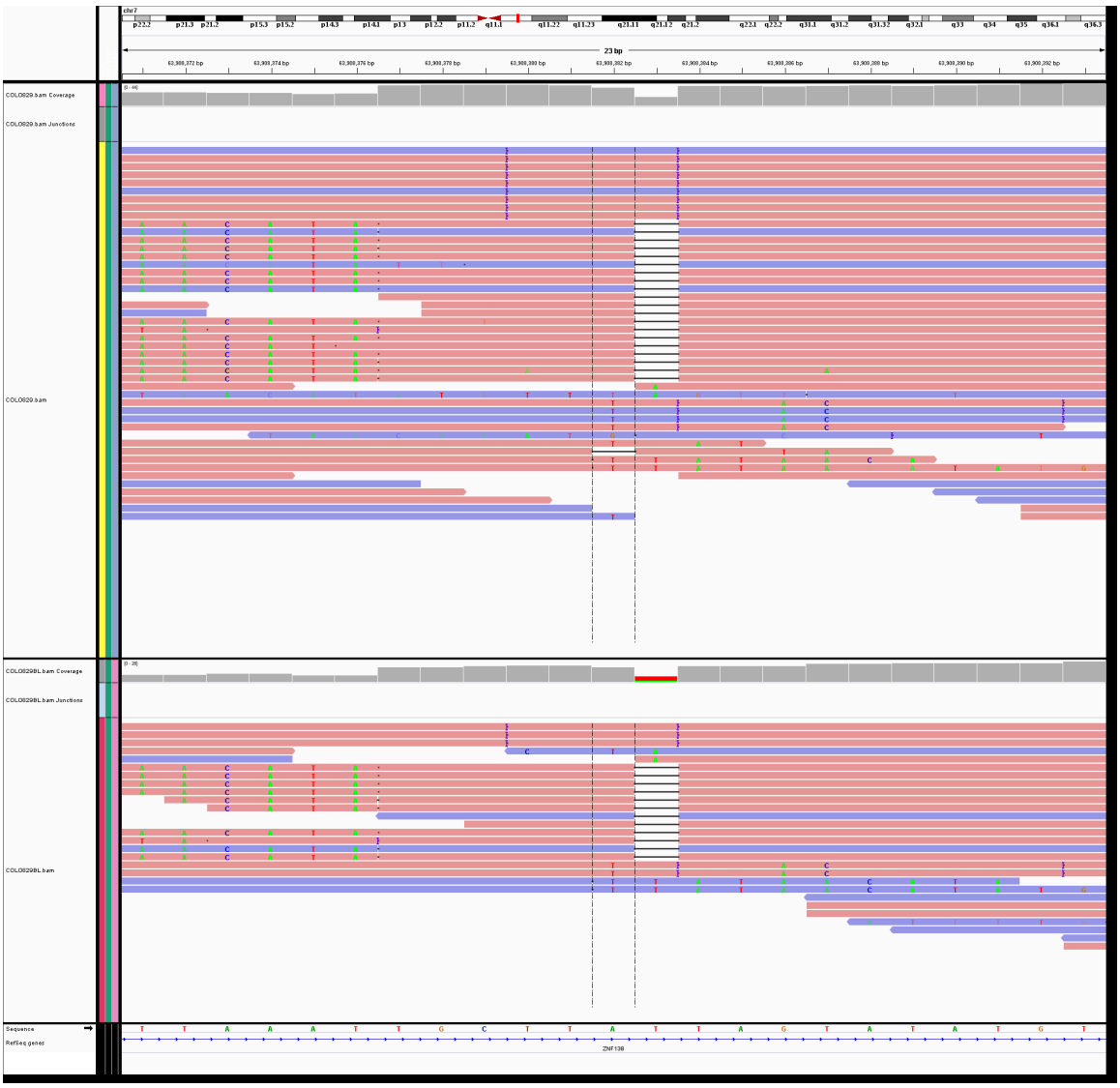


JNCOLO829BLG57R_A05.seq

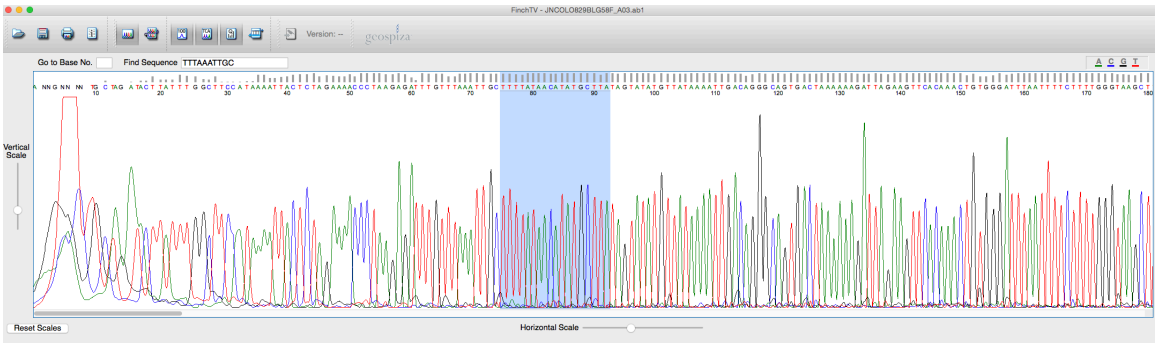


Germline 58

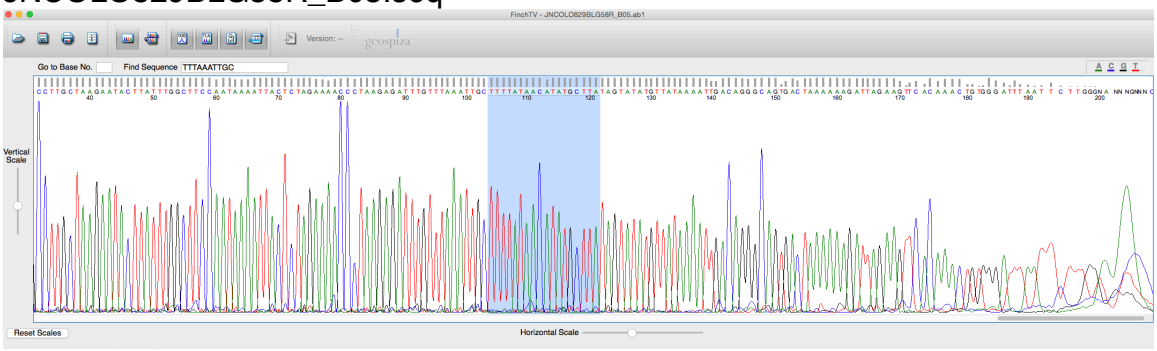
D 2 NT 16 TTATAACATATGCTTA 7 63908381 63908384



JNCOLO829BLG58F_A03.seq

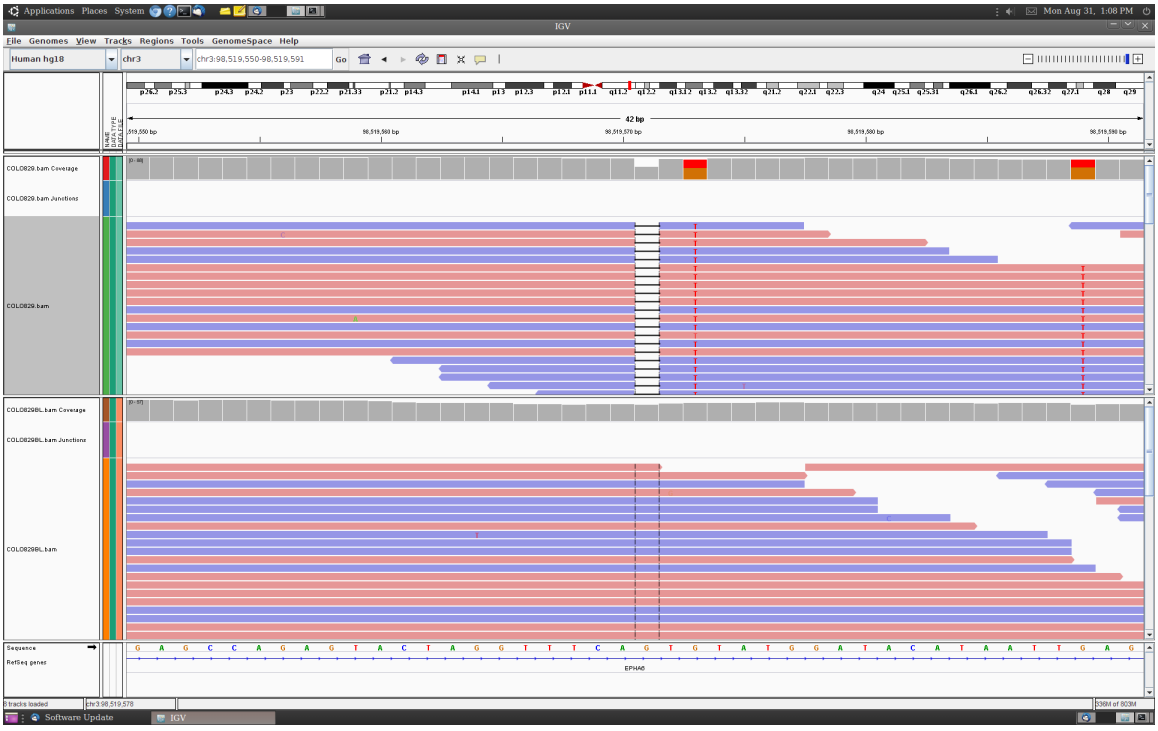


JNCOLO829BLG58R_B05.seq

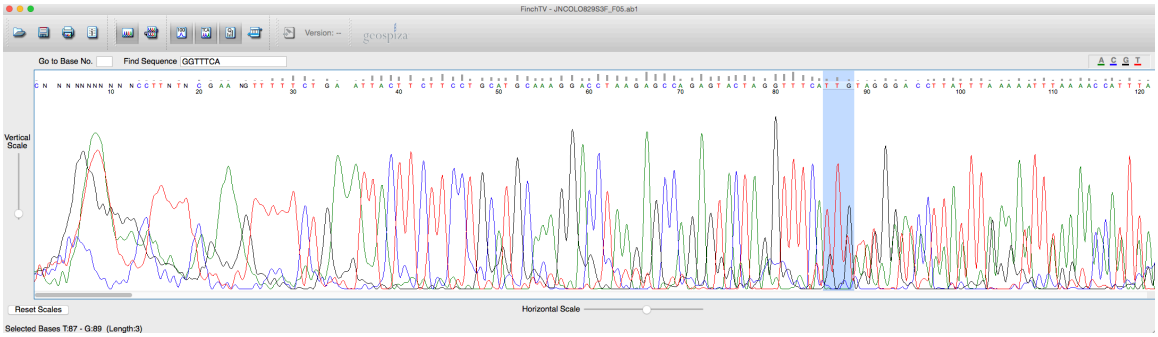


Somatic 3

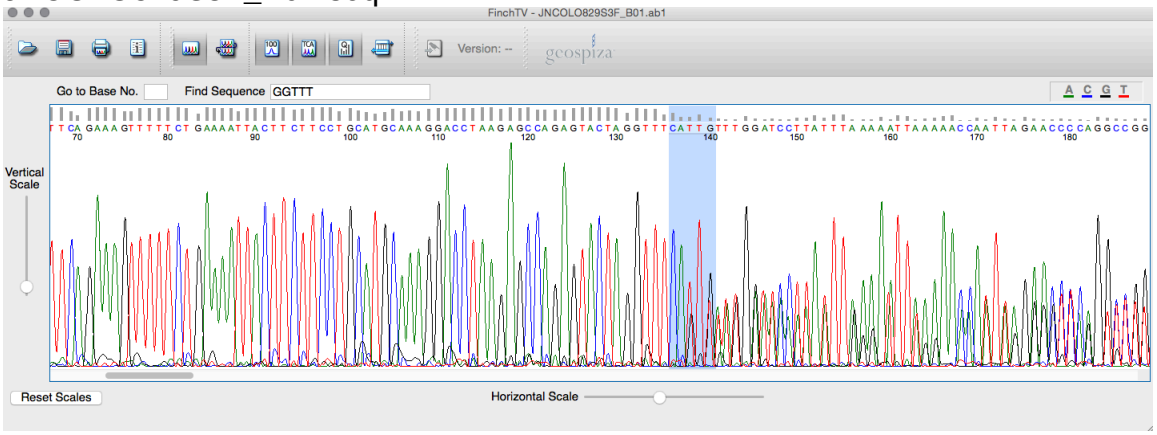
D 3 NT 2 TT 3 98519570 98519573



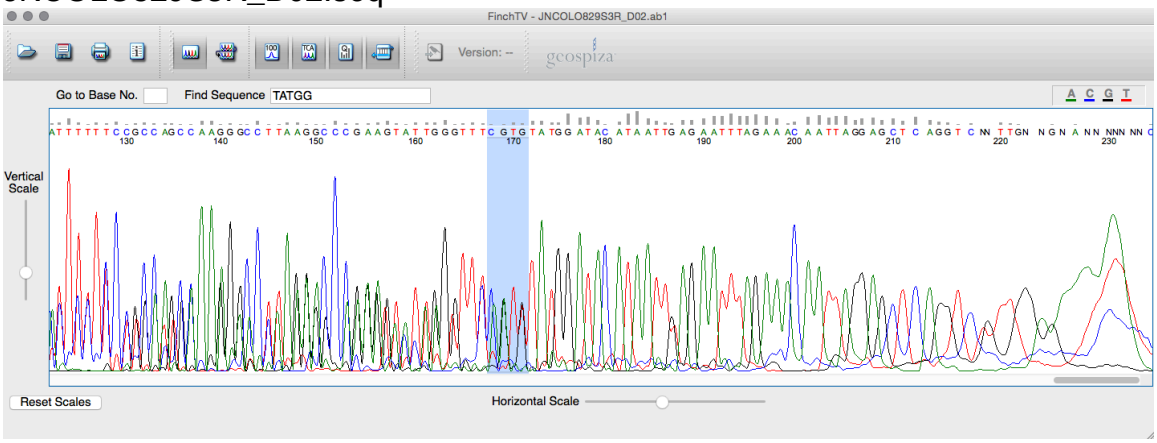
JNCOLO829S3F_F05.seq



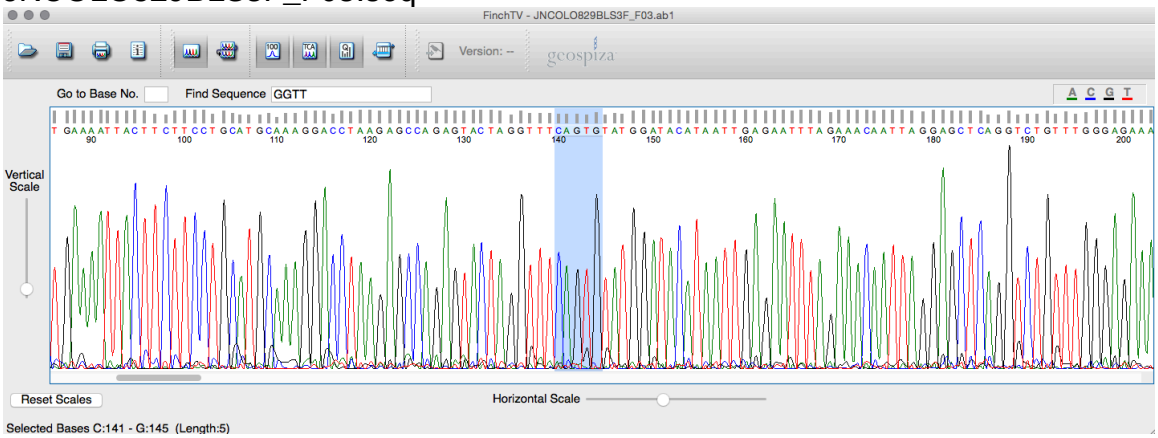
JNCOLO829S3F_B01.seq



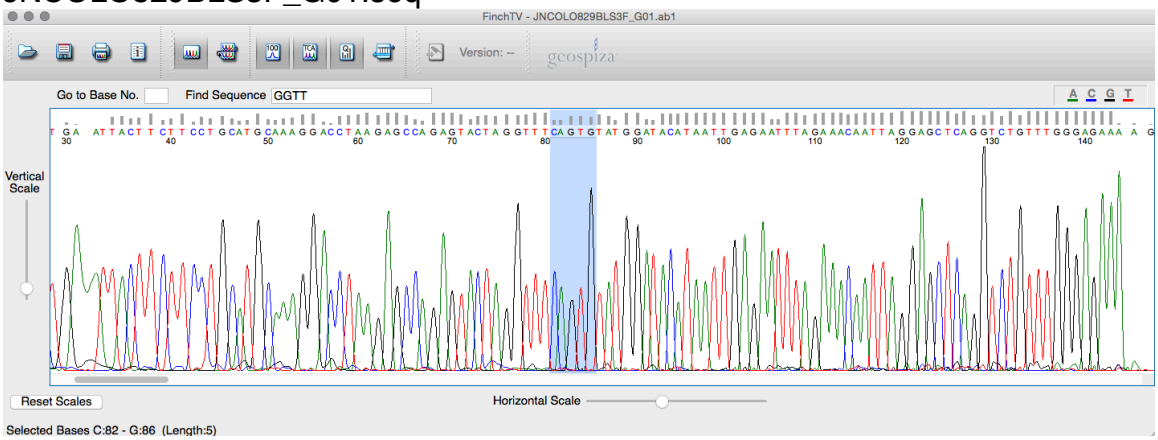
JNCOLO829S3R_D02.seq



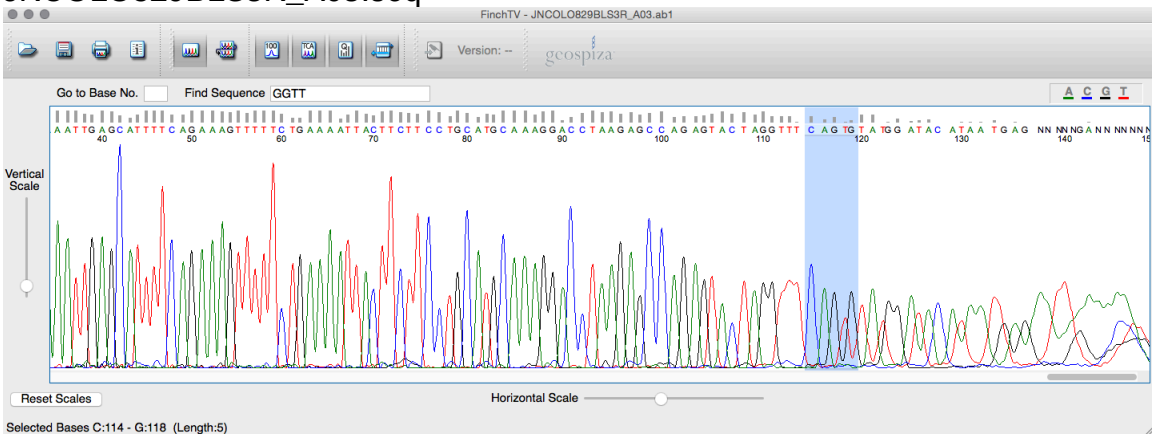
JNCOLO829BLS3F_F03.seq



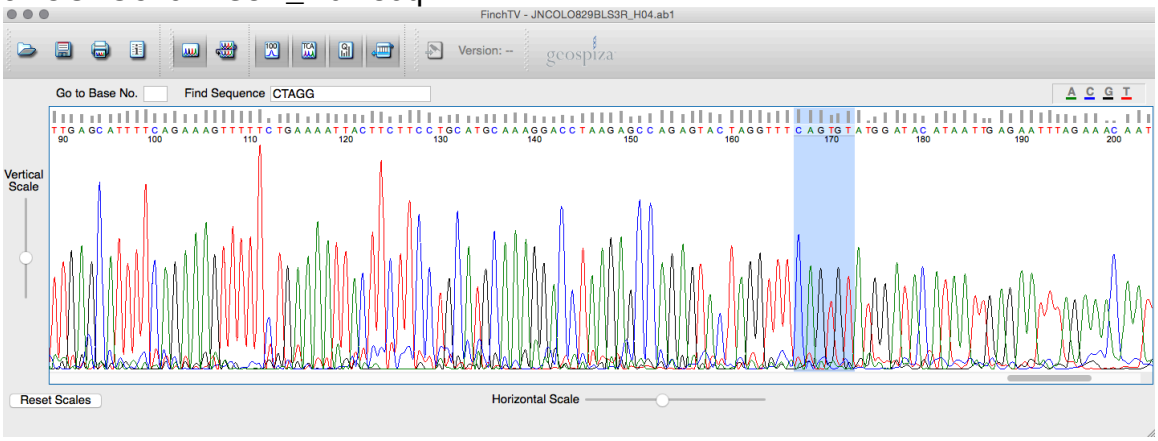
JNCOLO829BLS3F_G01.seq



JNCOLO829BLS3R_A03.seq

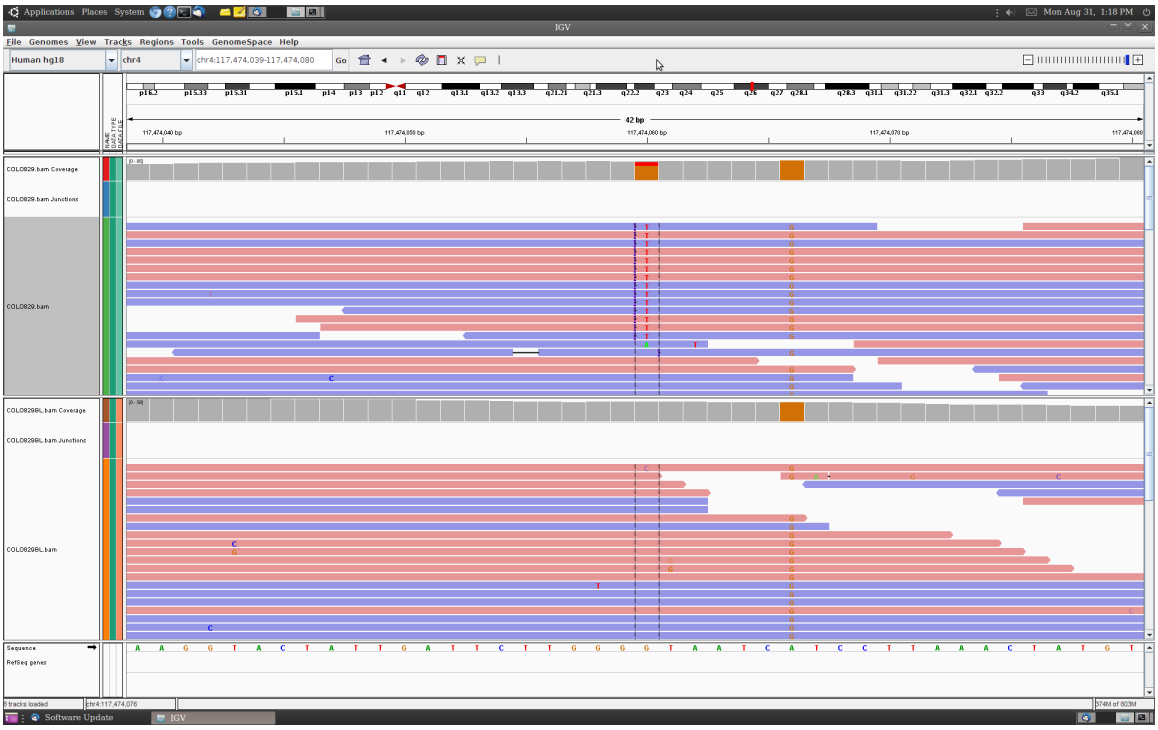


JNCOLO829BLS3R_H04.seq

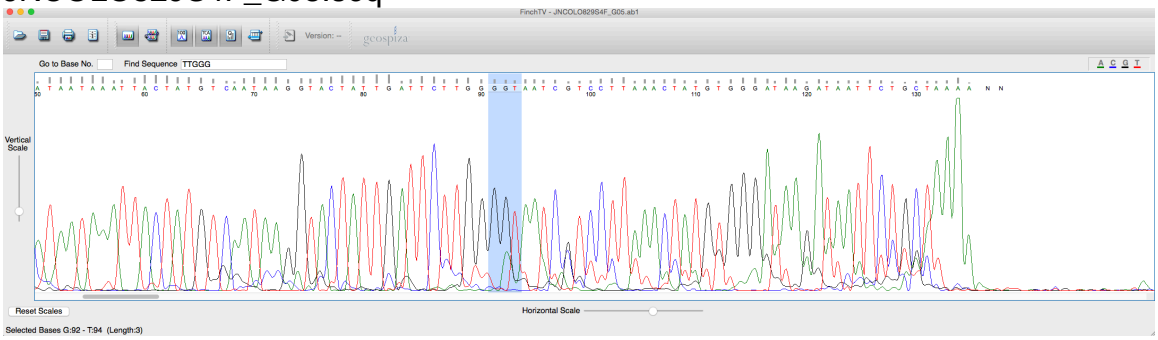


Somatic 4

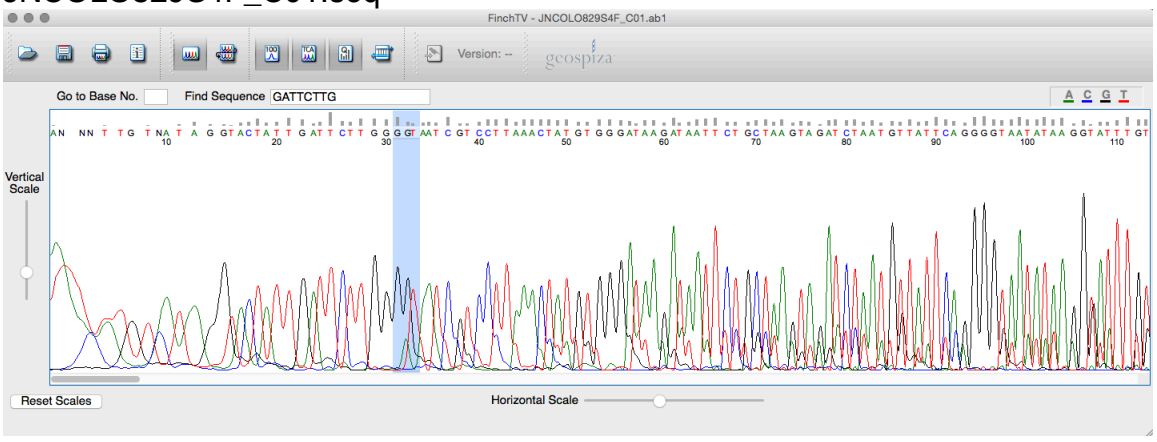
D 1 NT 2 AT 4 117474059 117474060



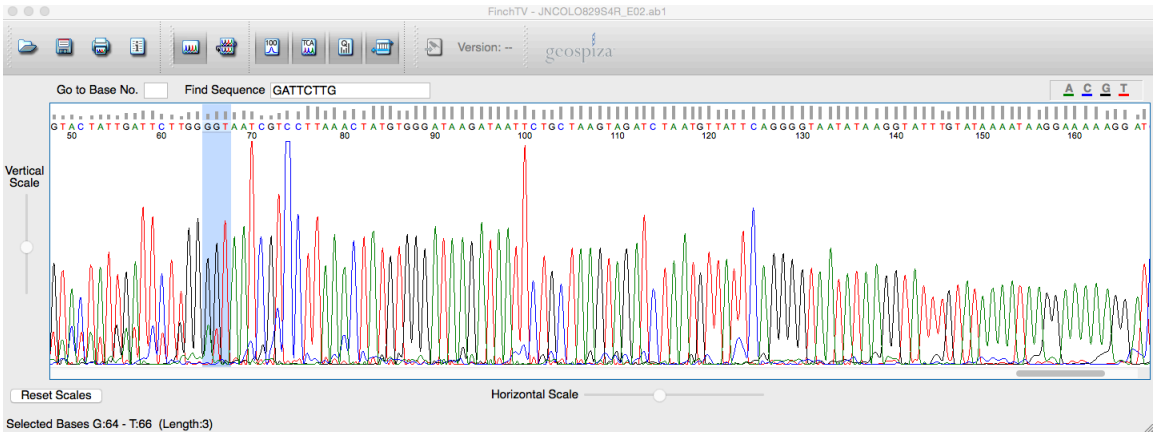
JNCOLO829S4F_G05.seq



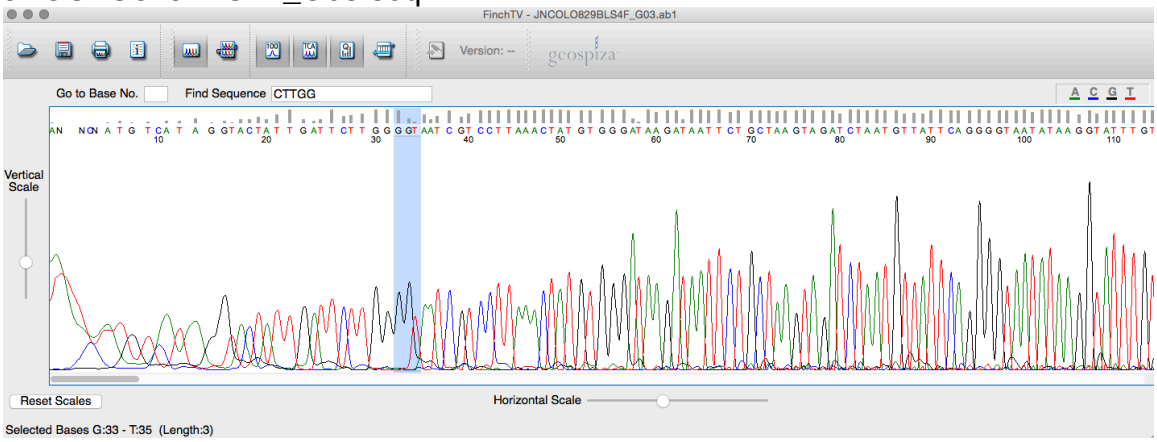
JNCOLO829S4F_C01.seq



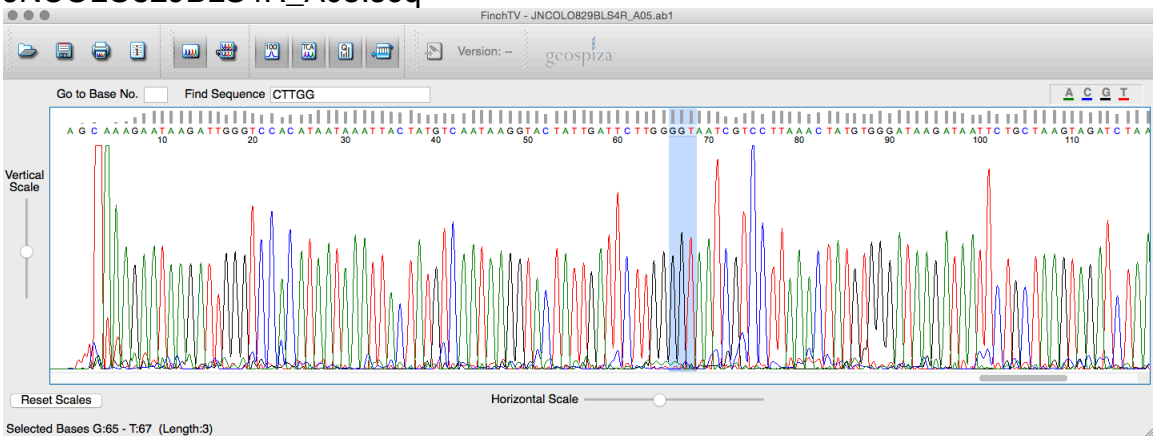
JNCOLO829S4R_E02.seq



JNCOLO829BLS4F_G03.seq

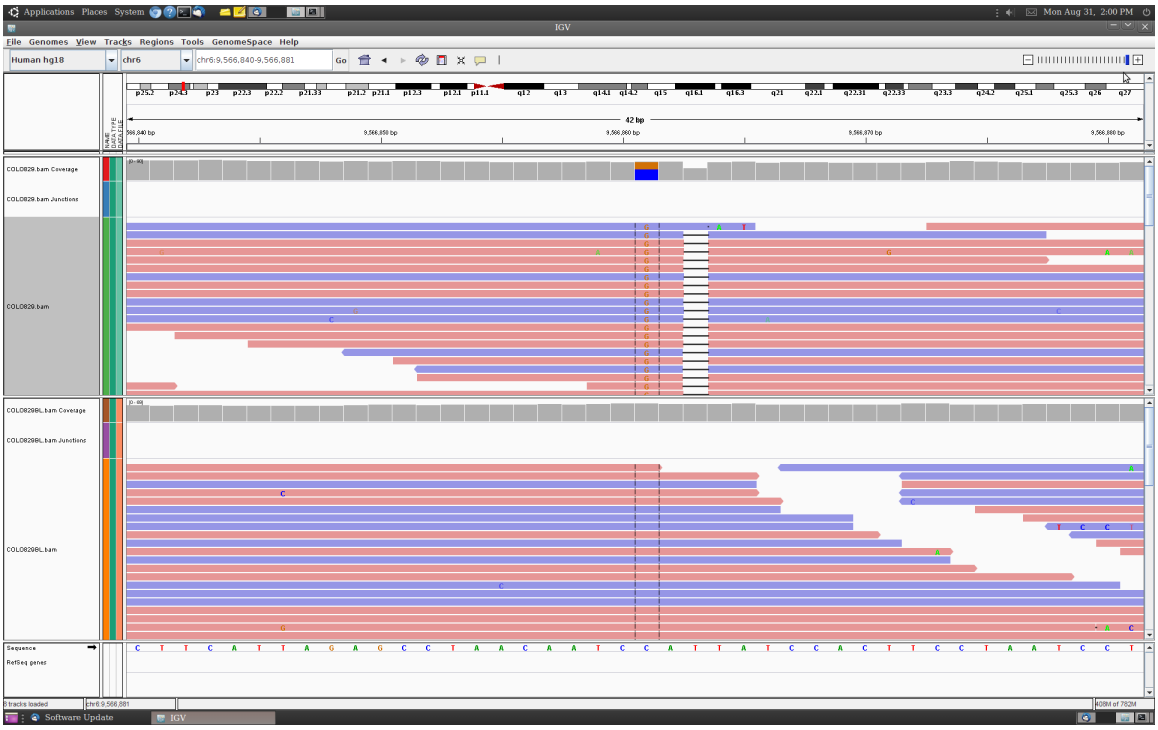


JNCOLO829BLS4R_A05.seq

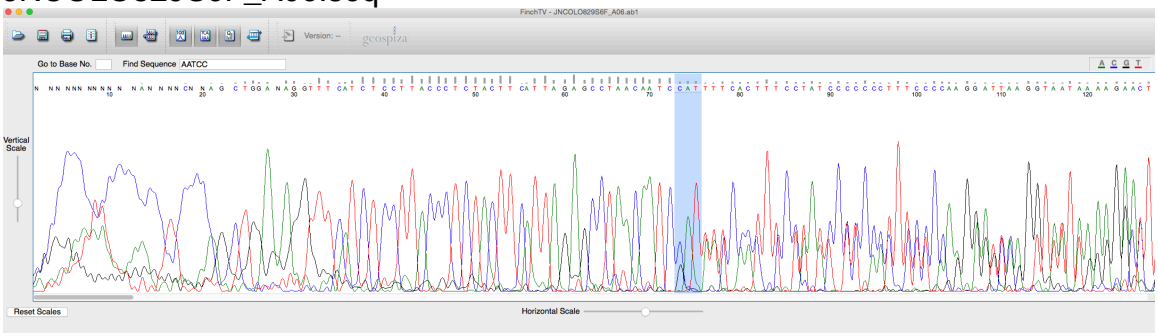


Somatic 6

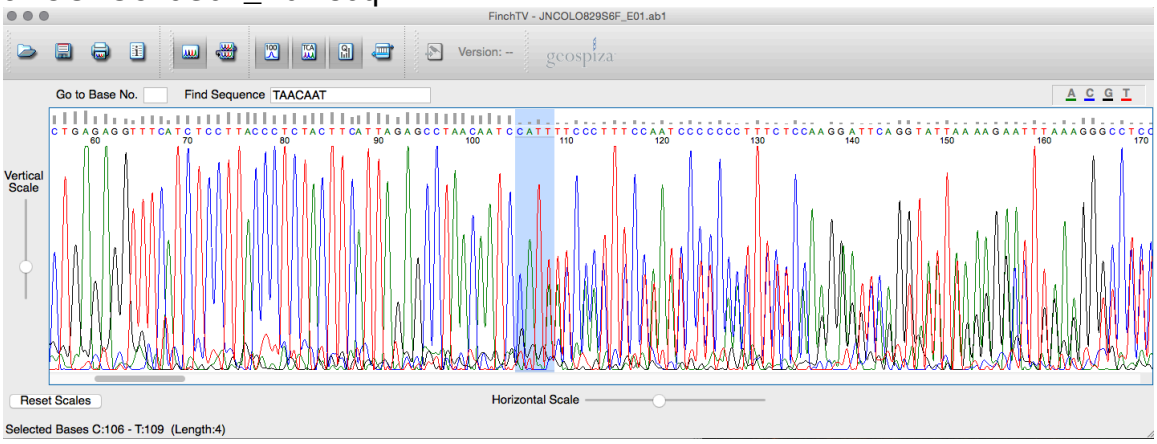
D 3 NT 2 GA 6 9566860 9566863



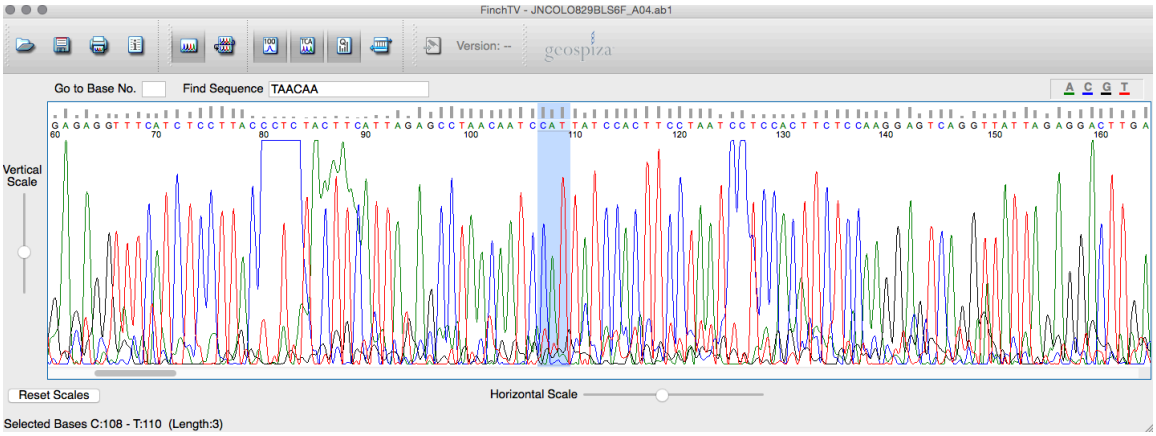
JNCOLO829S6F_A06.seq



JNCOLO829S6F_E01.seq

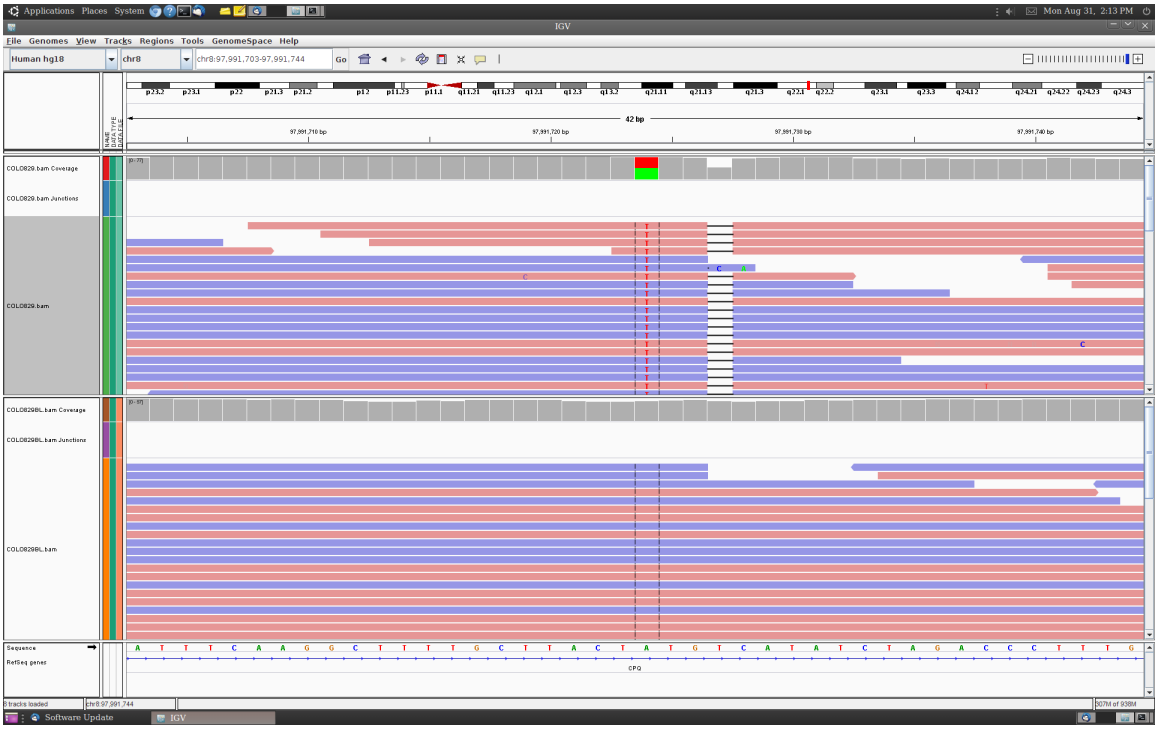


JNCOLO829BLS6F_A04.seq

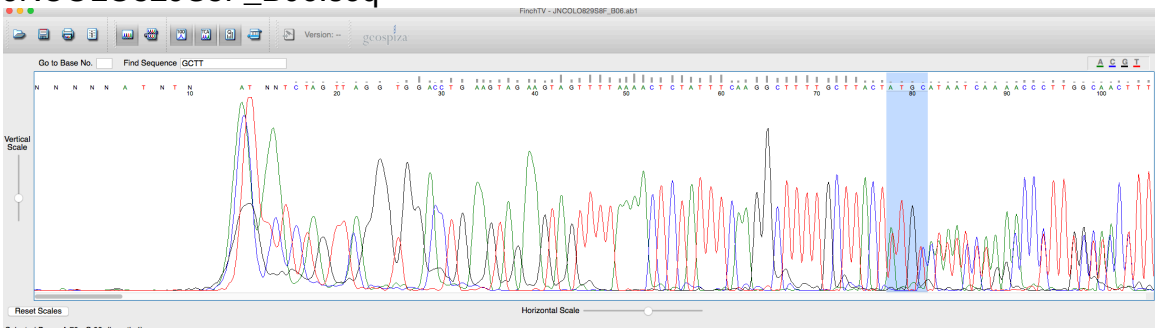


Somatic 8

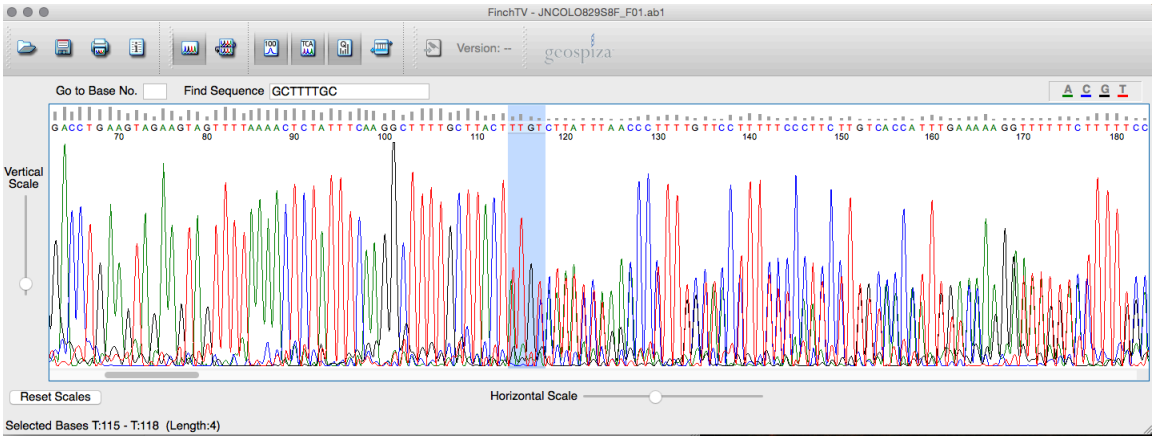
D 4 NT 3 TTG 8 97991723 97991727



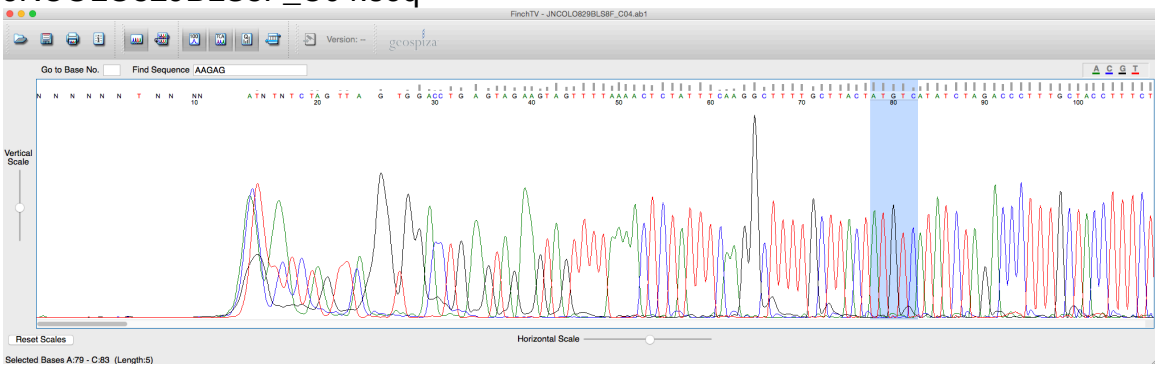
JNCOLO829S8F_B06.seq



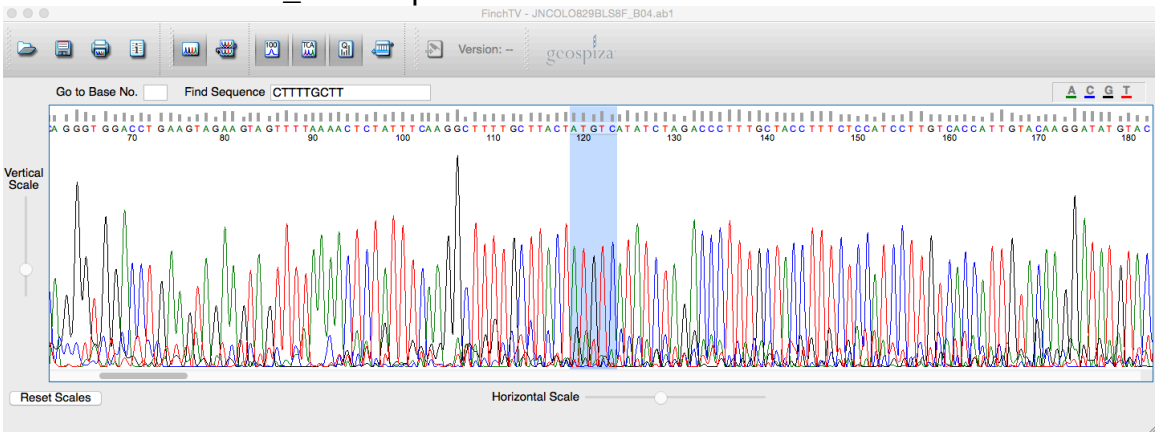
JNCOLO829S8F_F01.seq



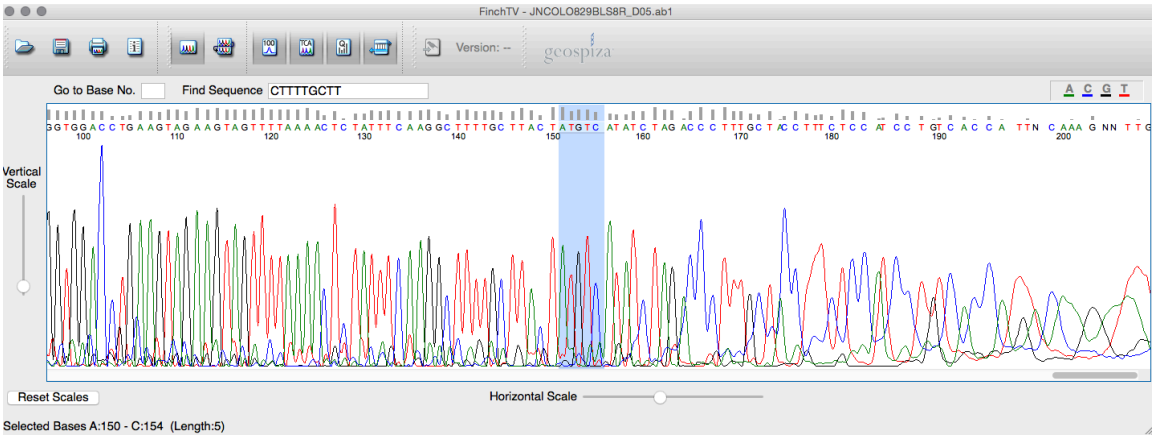
JNCOLO829BLS8F_C04.seq



JNCOLO829BLS8F_B04.seq

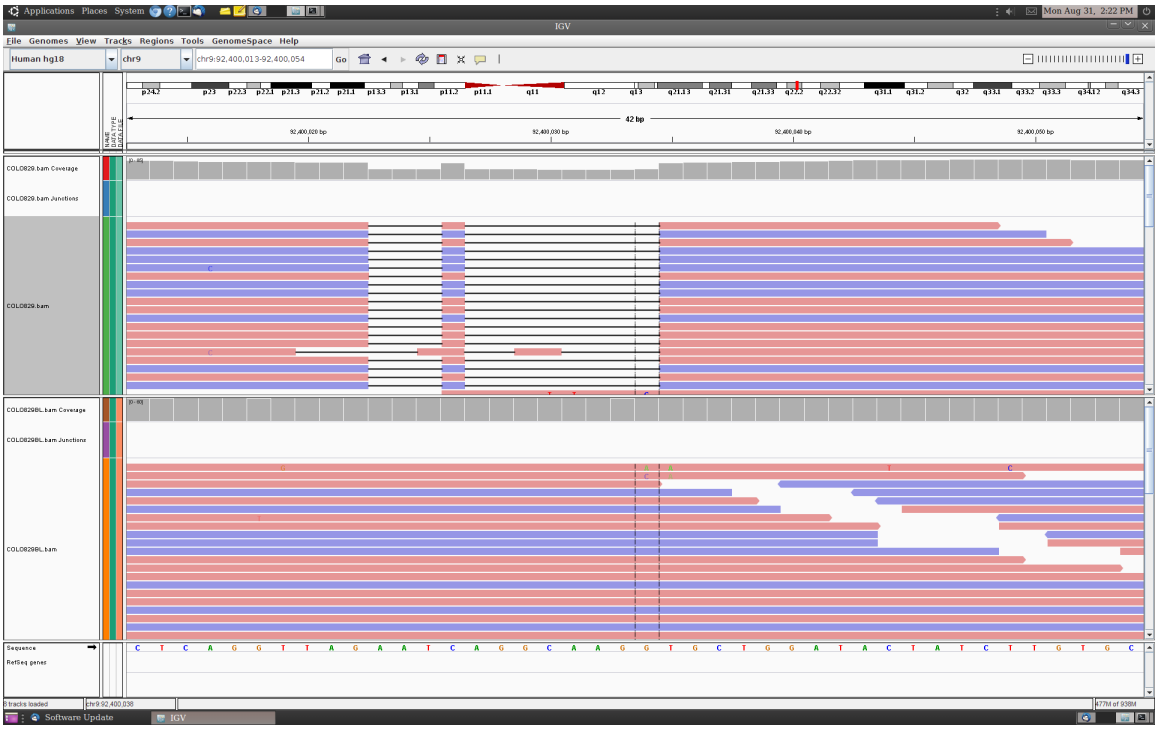


JNCOLO829BLS8R_D05.seq

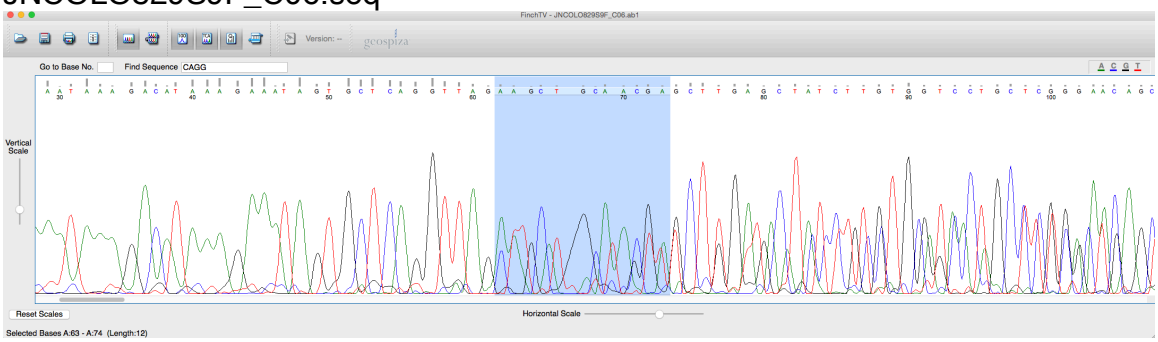


Somatic 9

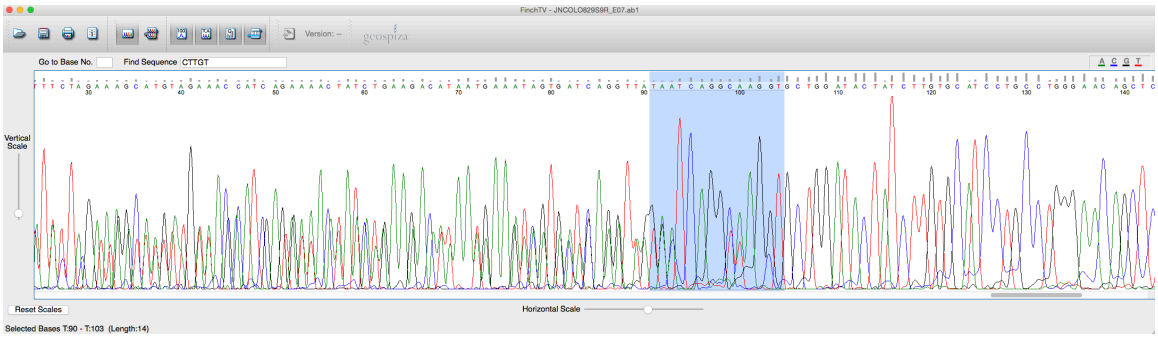
D 12 NT 1 C 9 92400022 92400034



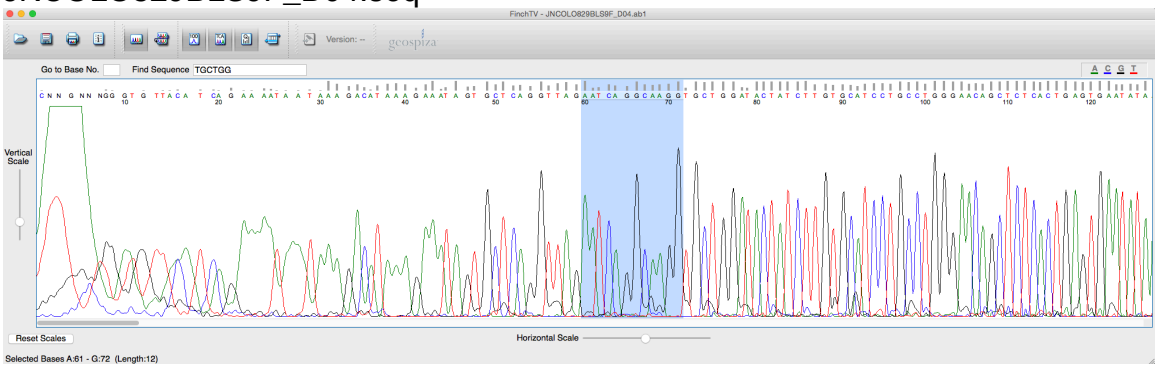
JNCOLO829S9F_C06.seq



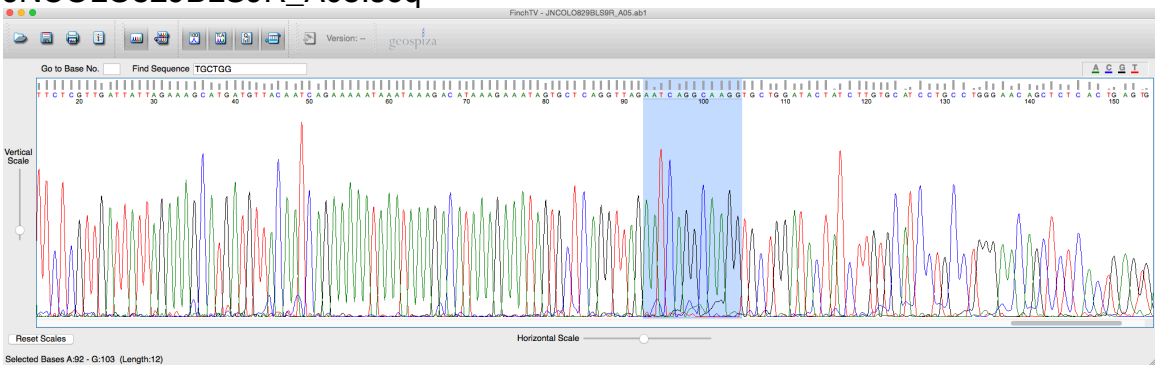
JNCOLO829S9R_E07.seq



JNCOLO829BLS9F_D04.seq

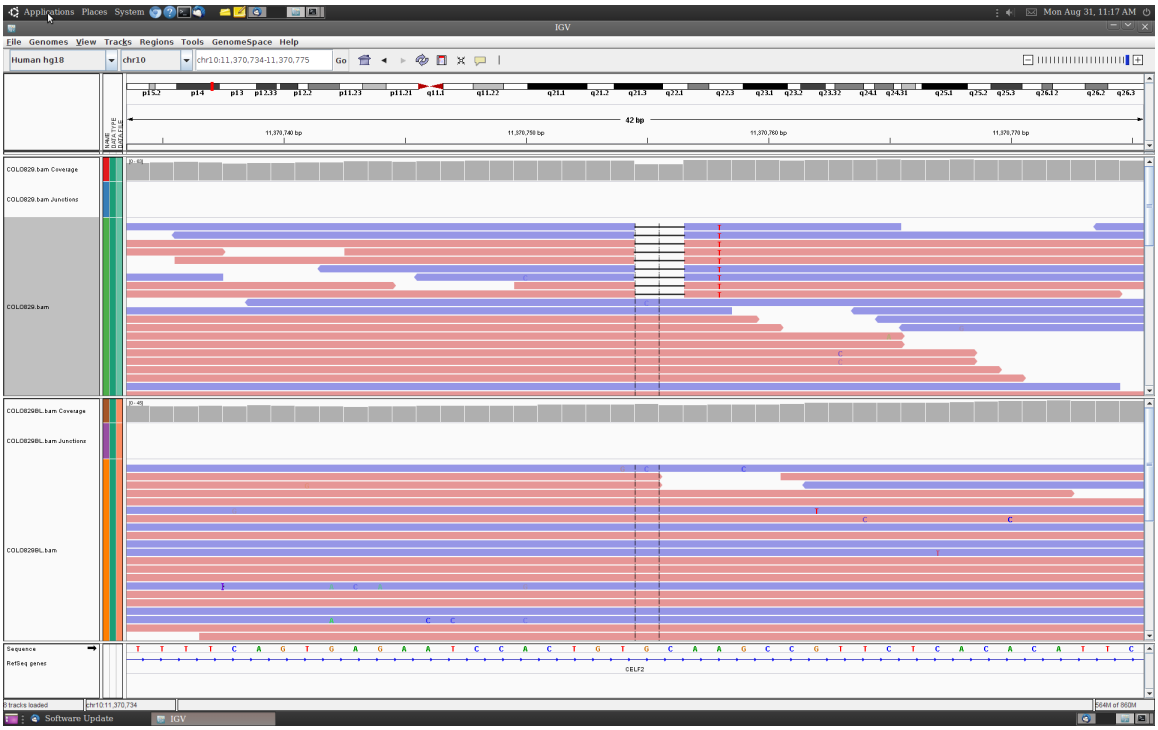


JNCOLO829BLS9R_A05.seq

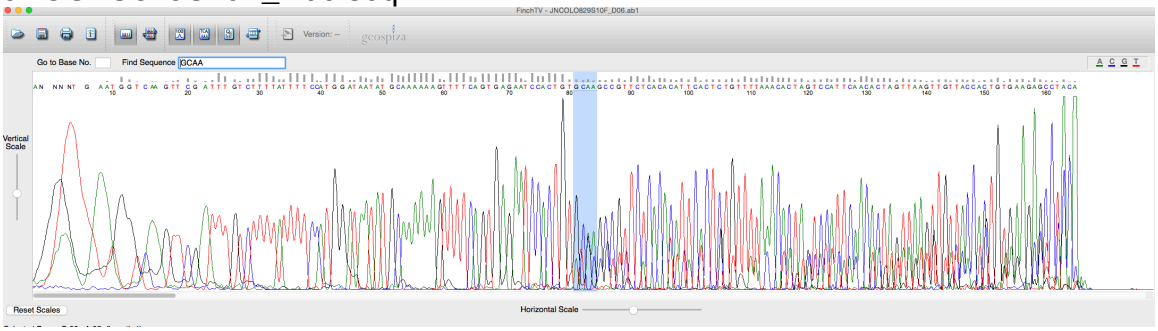


Somatic 10

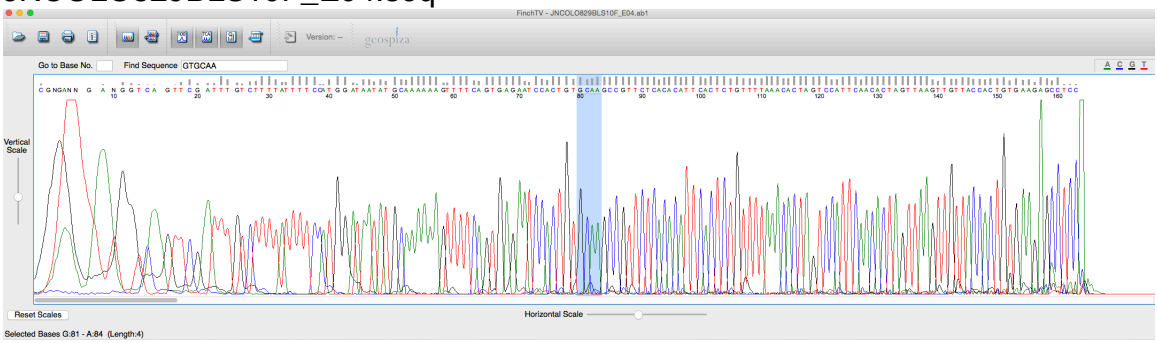
D	4	NT	2	AT	10	11370754	11370758



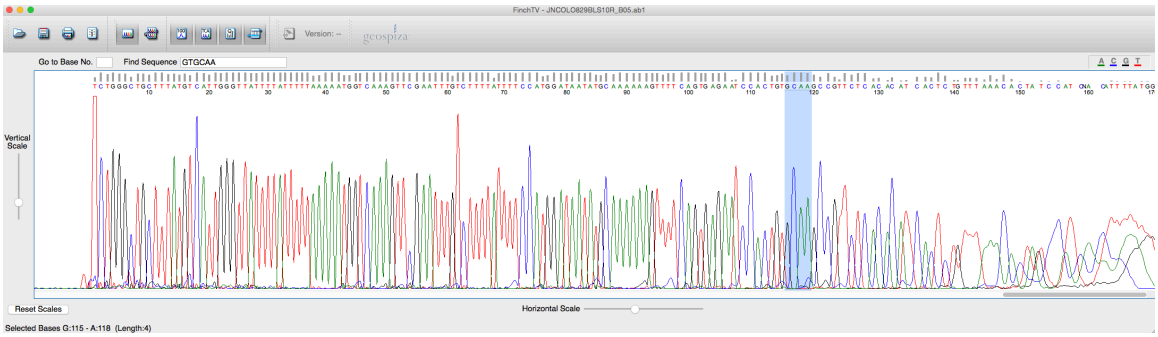
JNCOLO829S10F_D06.seq



JNCOLO829BLS10F_E04.seq

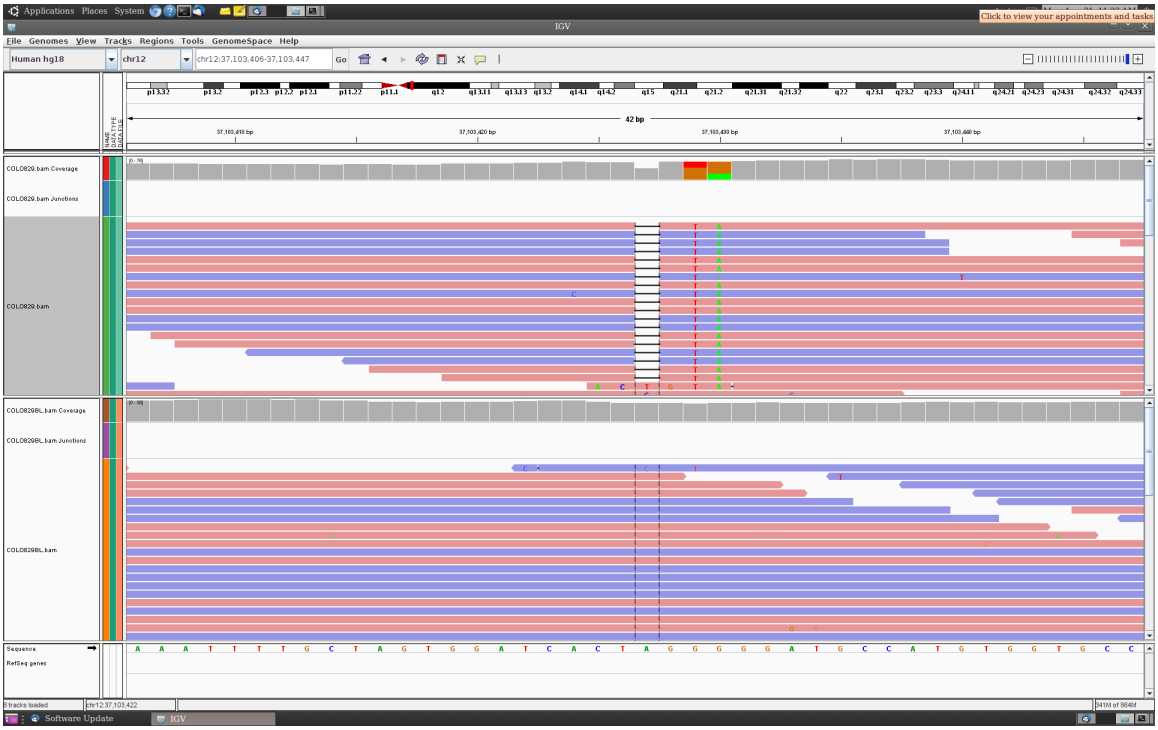


JNCOLO829BLS10R_B05.seq

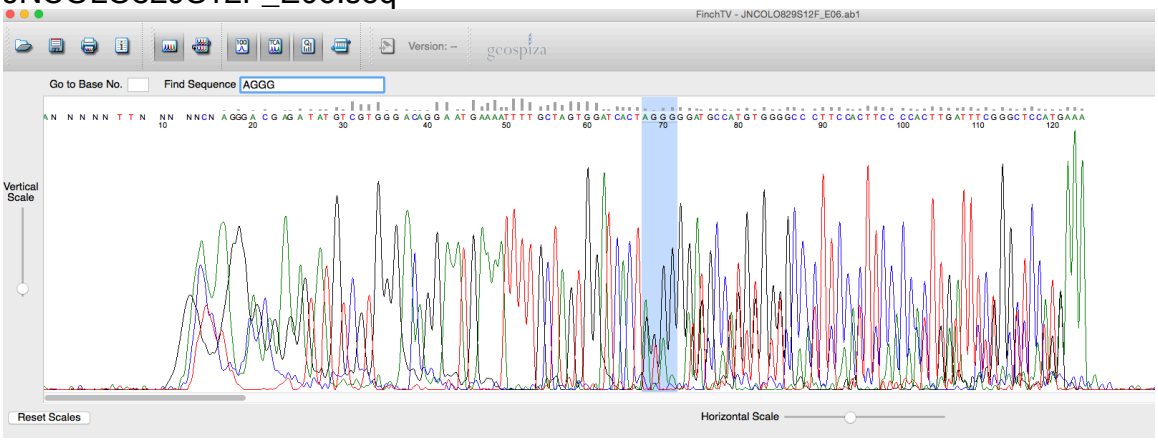


Somatic 12

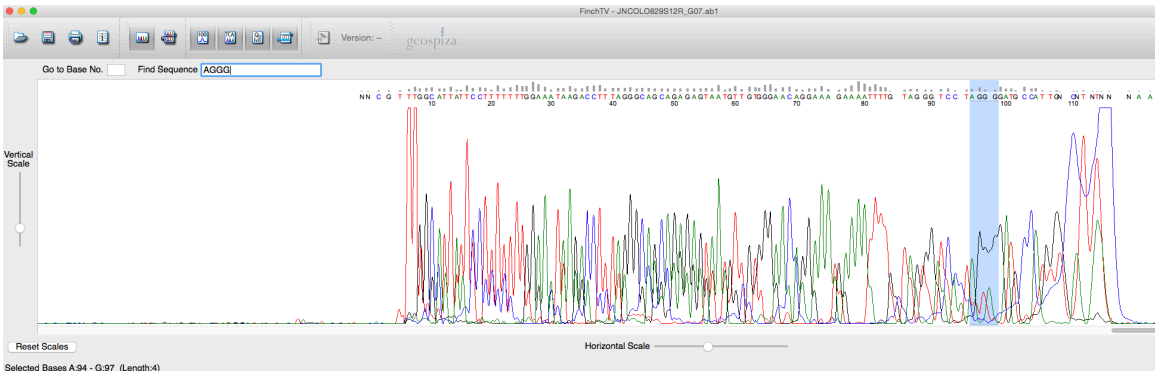
D 4 NT 3 GTA 12 37103426 37103430



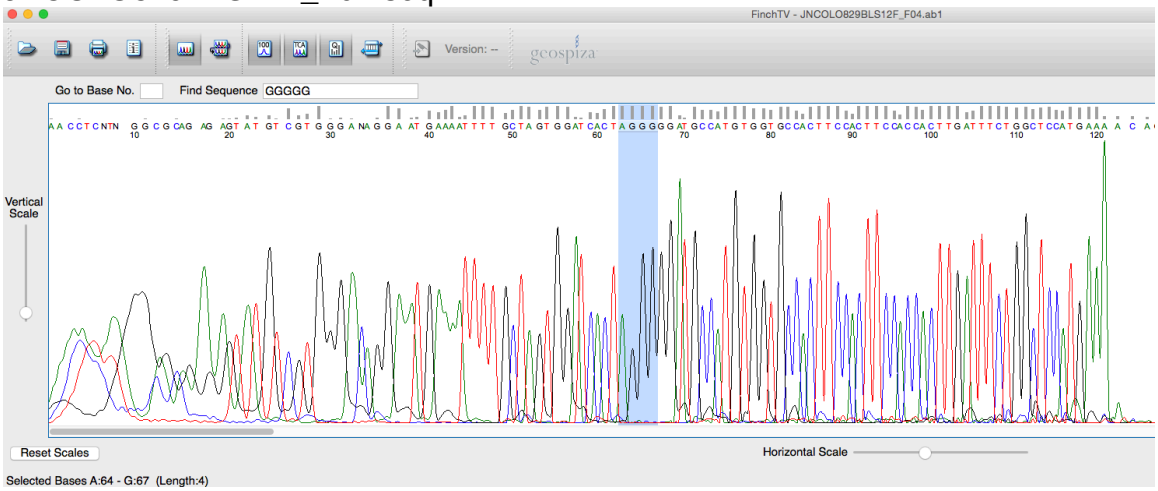
JNCOLO829S12F_E06.seq



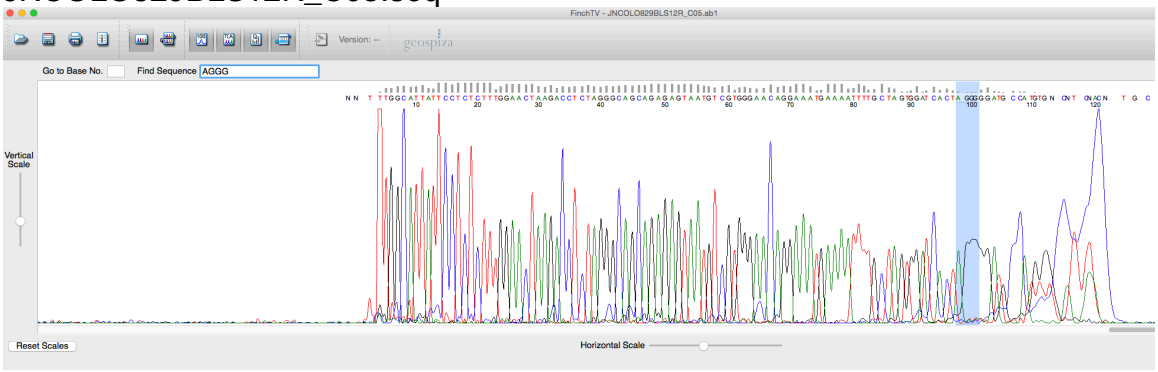
JNCOLO829S12R_G07.seq



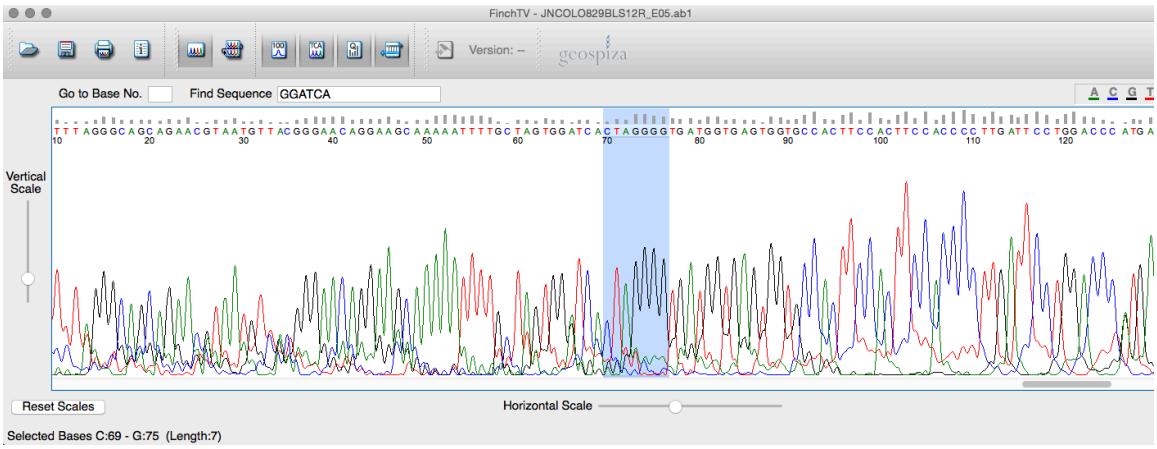
JNCOLO829BLS12F_F04.seq



JNCOLO829BLS12R_C05.seq

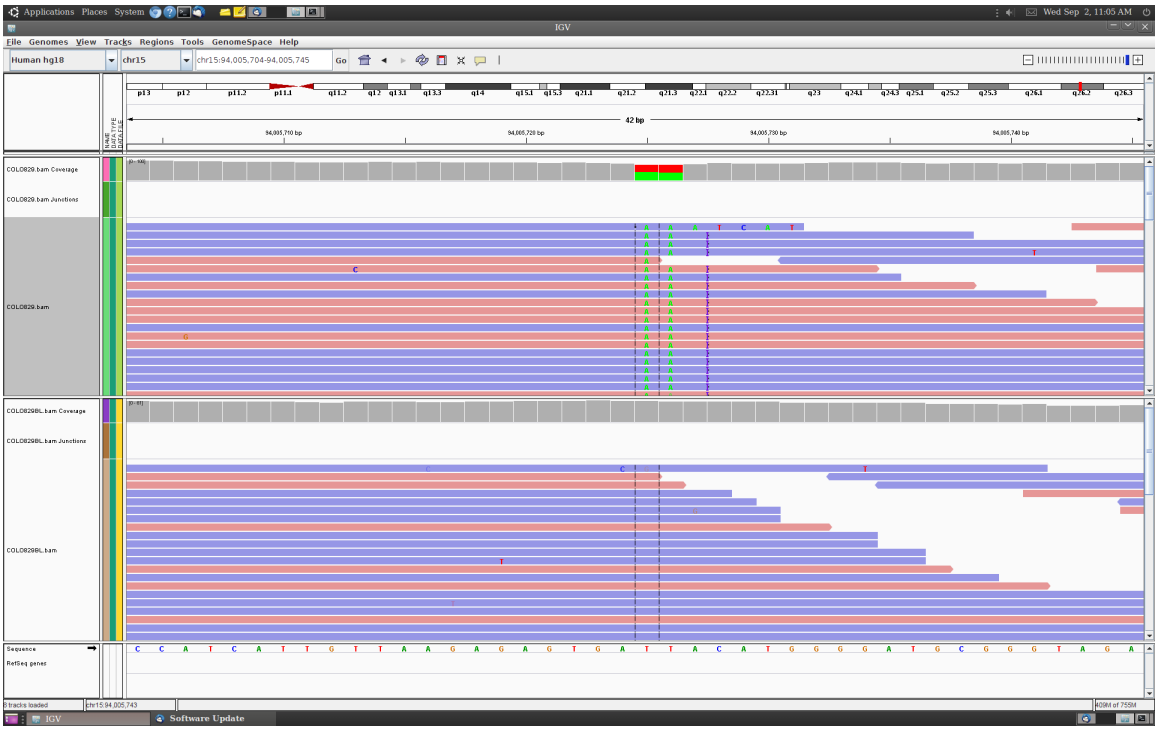


JNCOLO829BLS12R_E05.seq

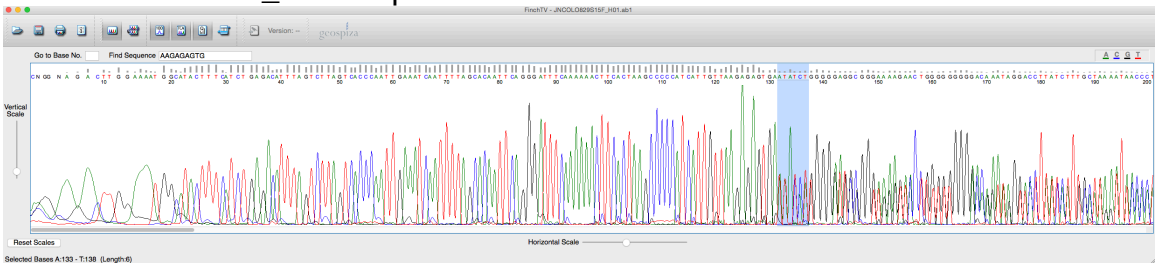


Somatic 15

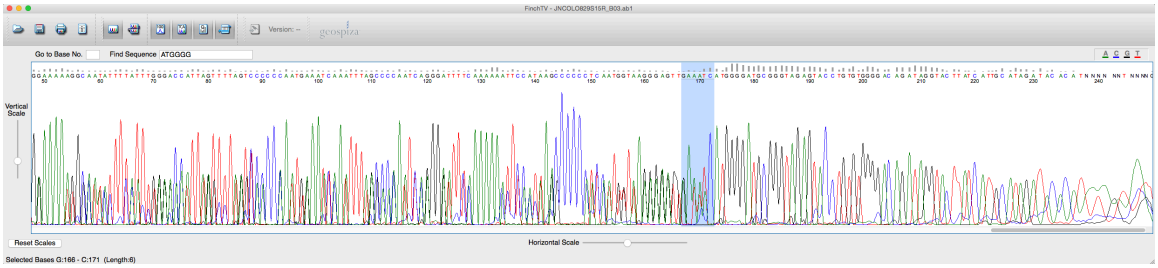
D 3 NT 4 AAAT 15 94005724 94005727



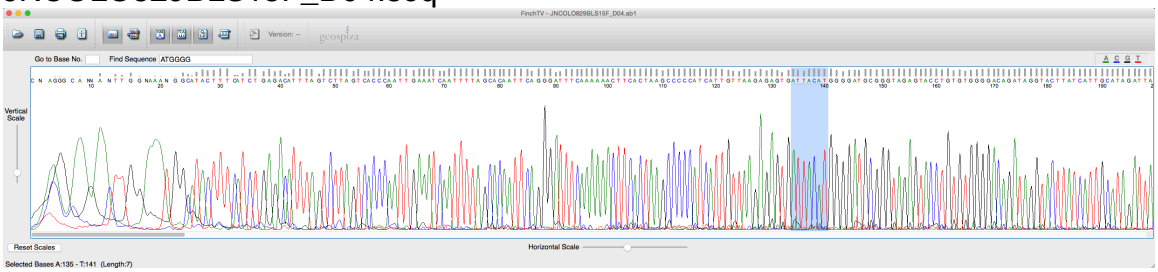
JNCOLO829S15F_H01.seq



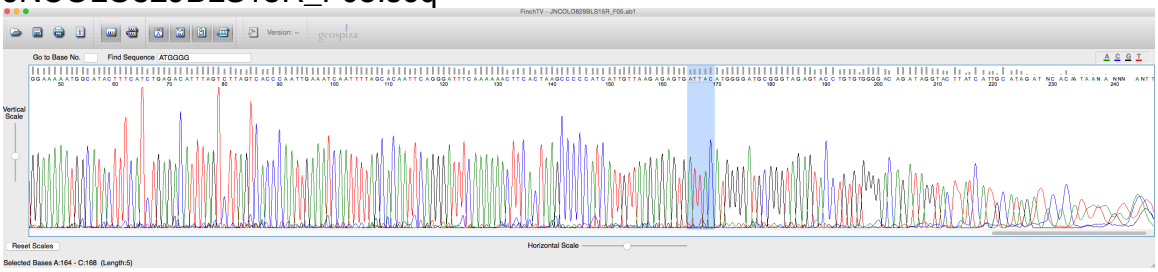
JNCOLO829S15R_B03.seq



JNCOLO829BLS15F_D04.seq

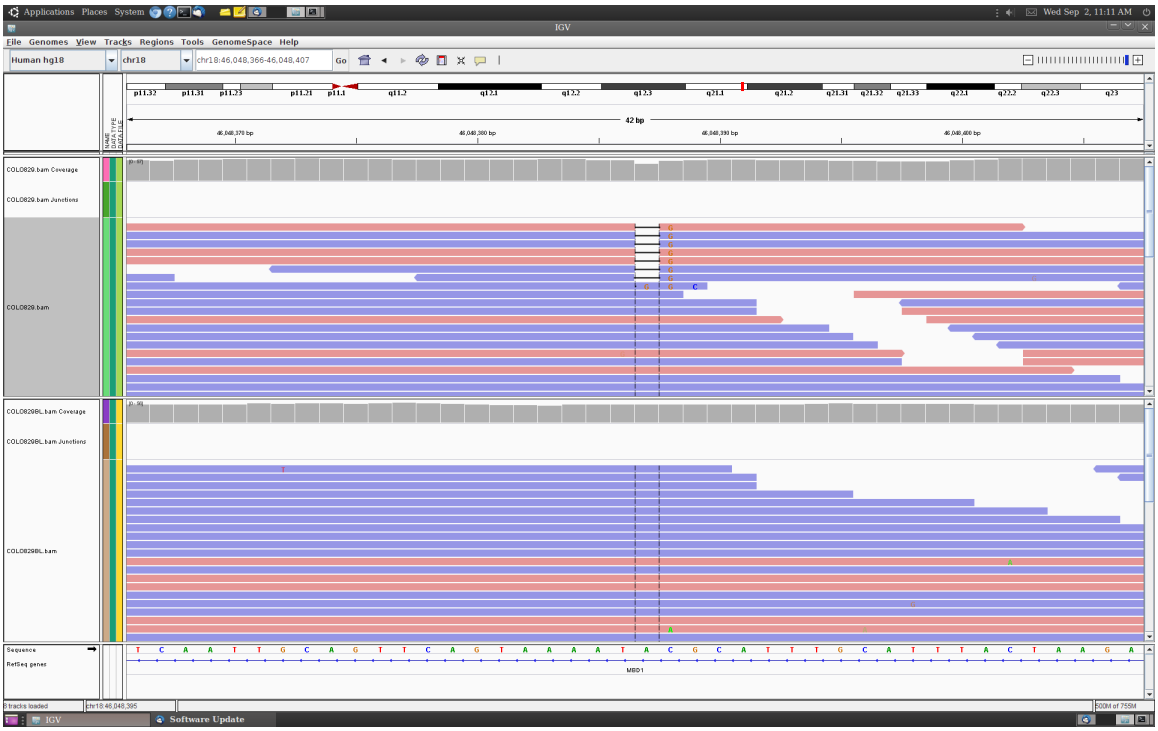


JNCOLO829BLS15R_F05.seq

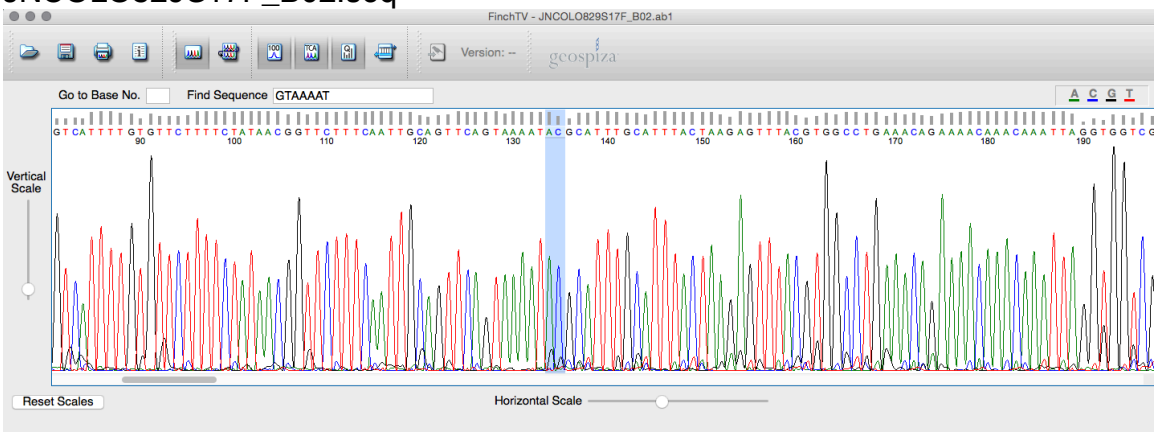


Somatic 17

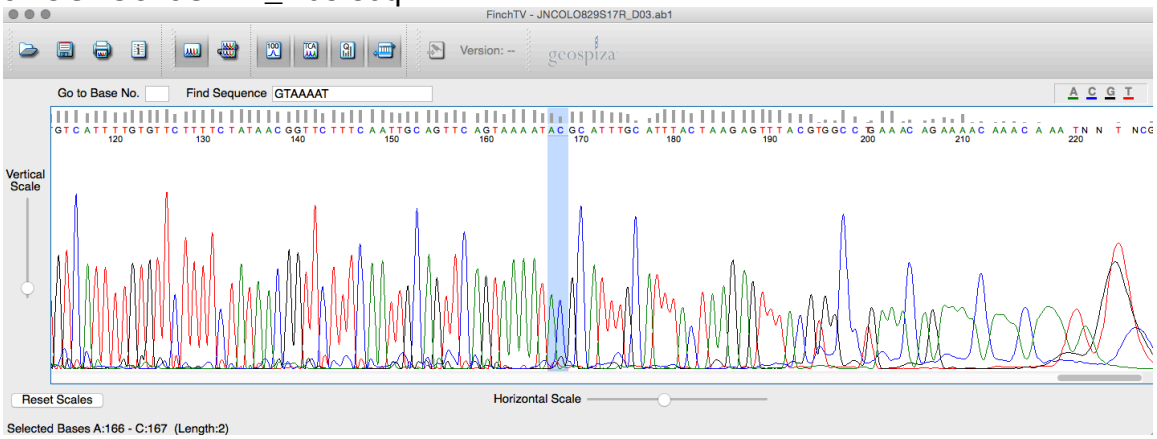
D	2	NT	1	G	18	46048386	46048388



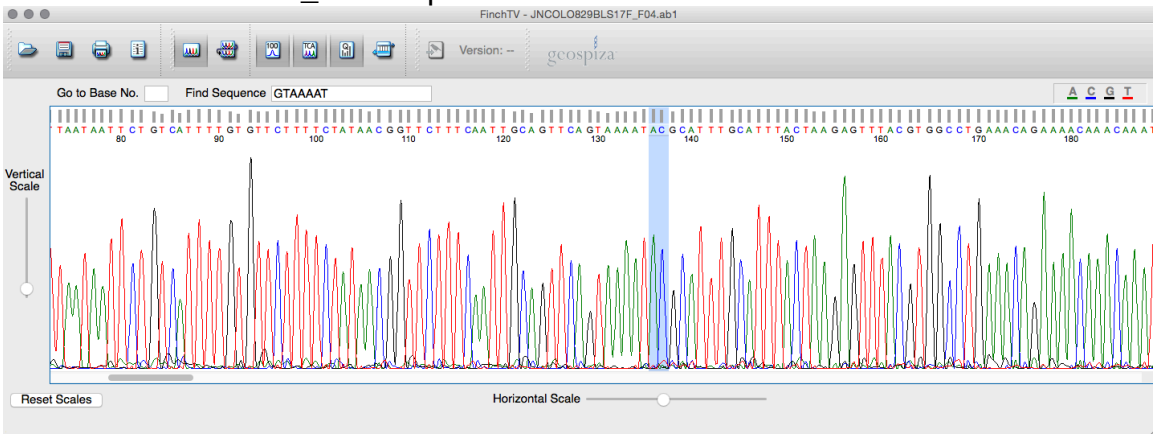
JNCOLO829S17F_B02.seq



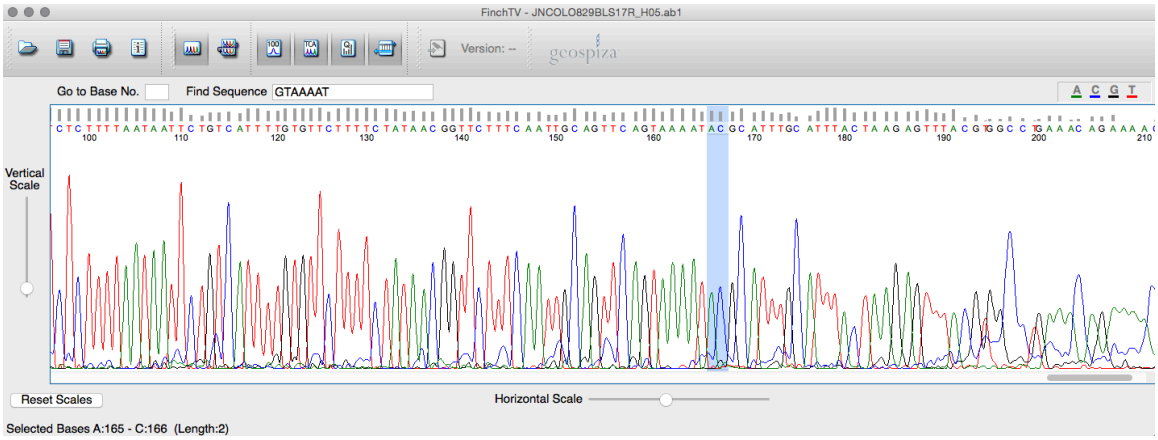
JNCOLO829S17R_D03.seq



JNCOLO829BLS17F_F04.seq



JNCOLO829BLS17R_H05.seq


FinchTV - JNCOLO829BLS17R_H05.ab1
Version: --
geospiza
Go to Base No.
Find Sequence
GTAAAAT
A C G T
Vertical Scale
Reset Scales
Horizontal Scale
Selected Bases A:165 - C:166 (Length:2)

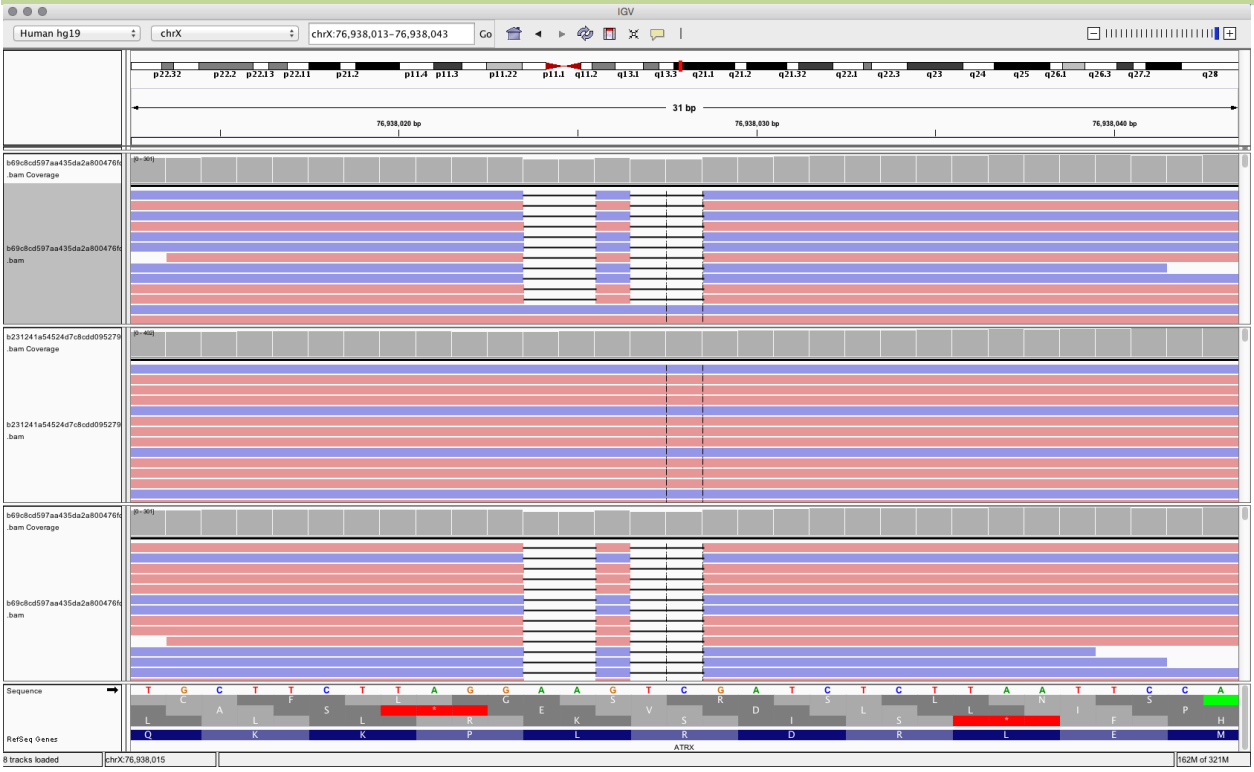
Supplemental Document: Screenshots of WGS validation of complex indels with good coverage

Following are screenshots of complex insertions and deletions identified in TCGA samples that had associated whole genome screen data with coverage > 10 in the tumor tissue.

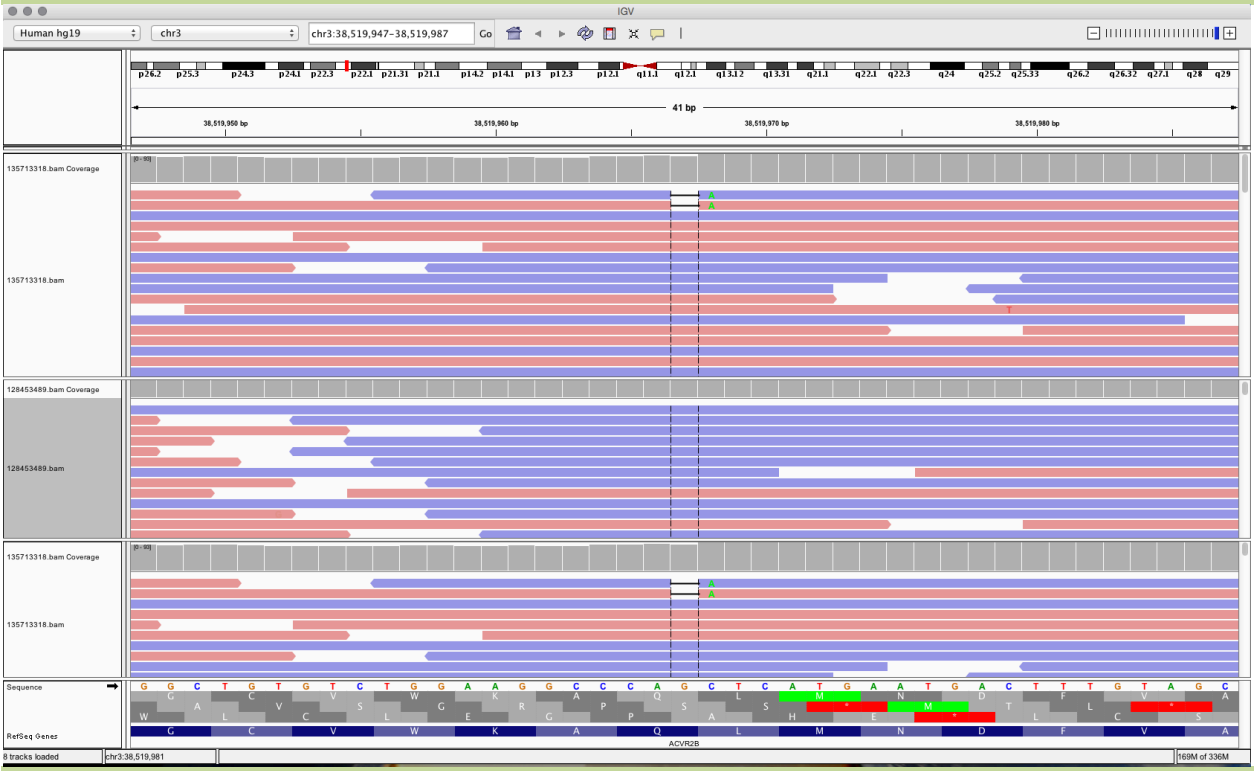
Table Information:

cancer_type	sample	gene	chromosome	position	deleted	inserted
GBM	TCGA-02-2483	ATRX	X	76938023	5	1

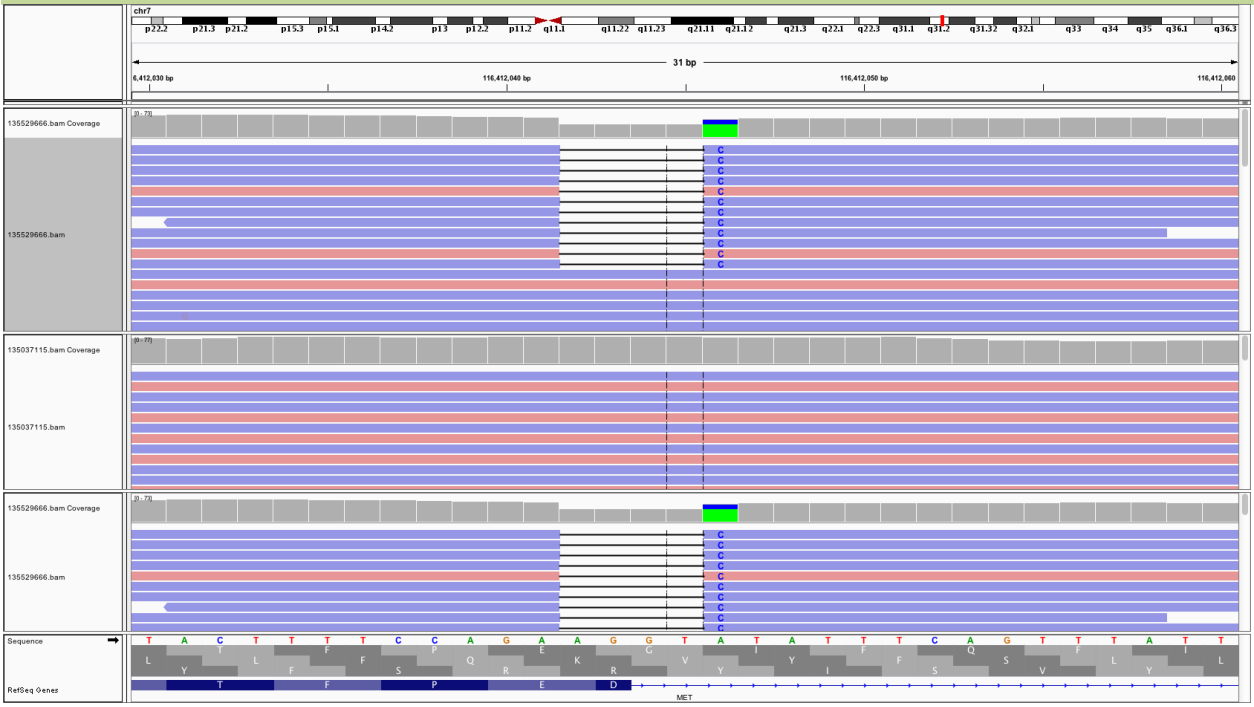
Screenshot from Integrative Genomics Viewer:
In each screenshot, the three bams are loaded in the following order: WGS tumor bam (top), WGS normal bam (middle), WXS tumor bam (bottom - where available).



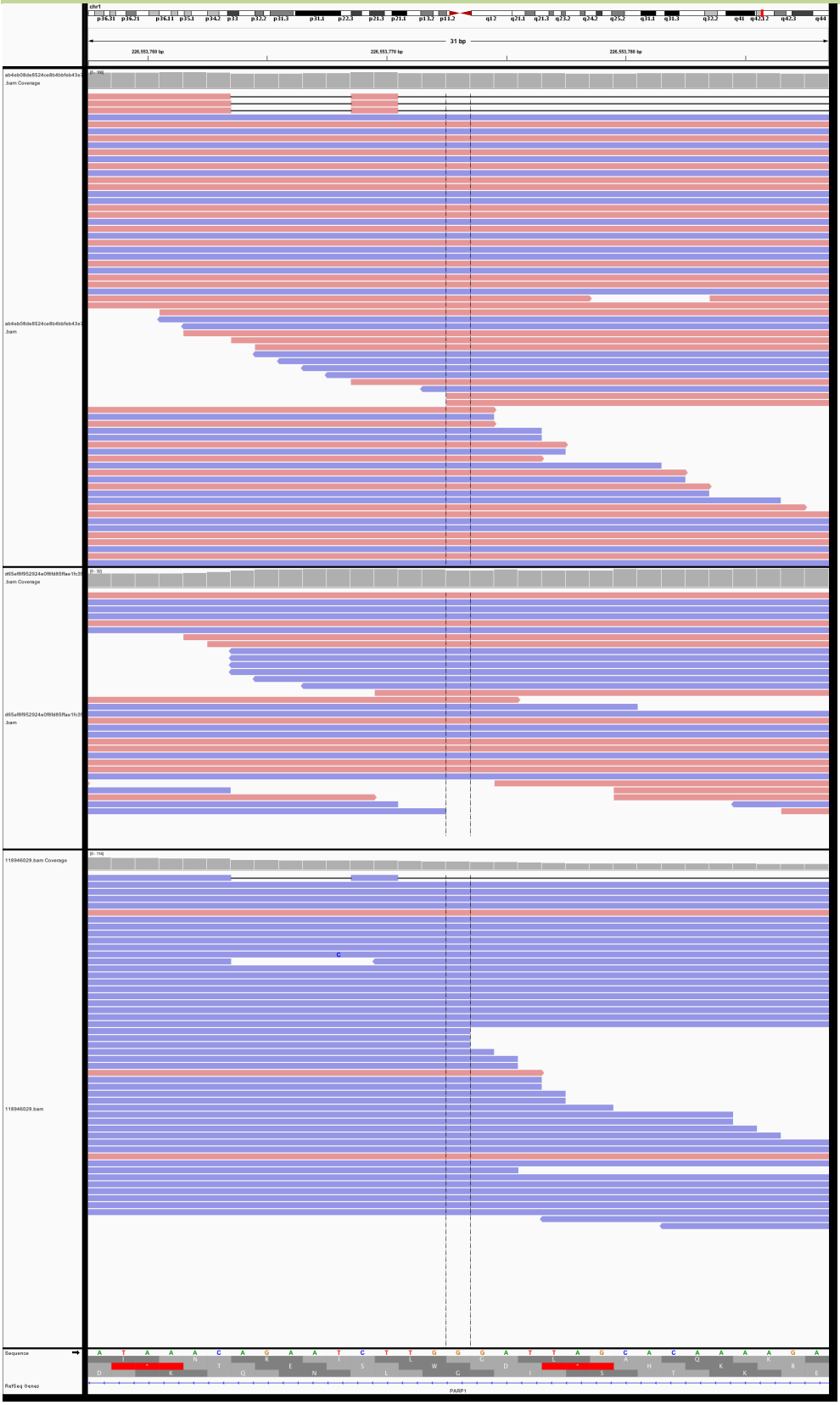
LUSC	TCGA-21-1083	ACVR2B	3	38519966	2	1



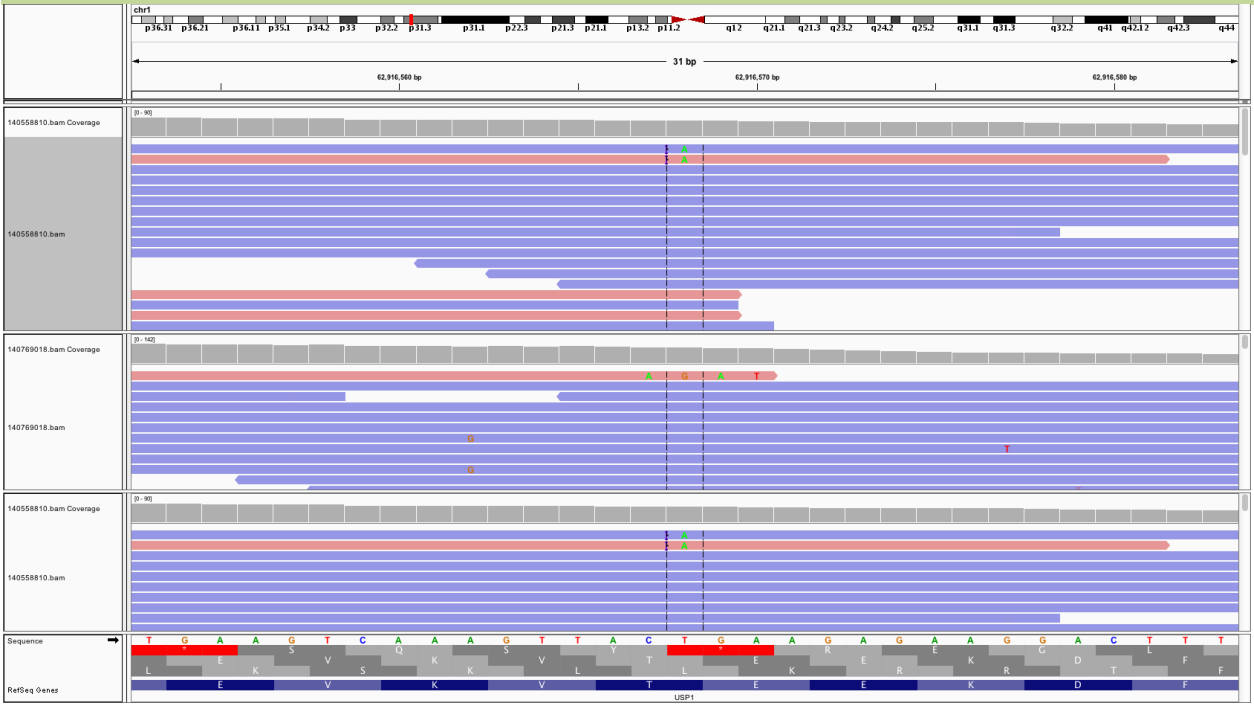
LUAD	TCGA-50-6597	MET	7	116412041	5	1



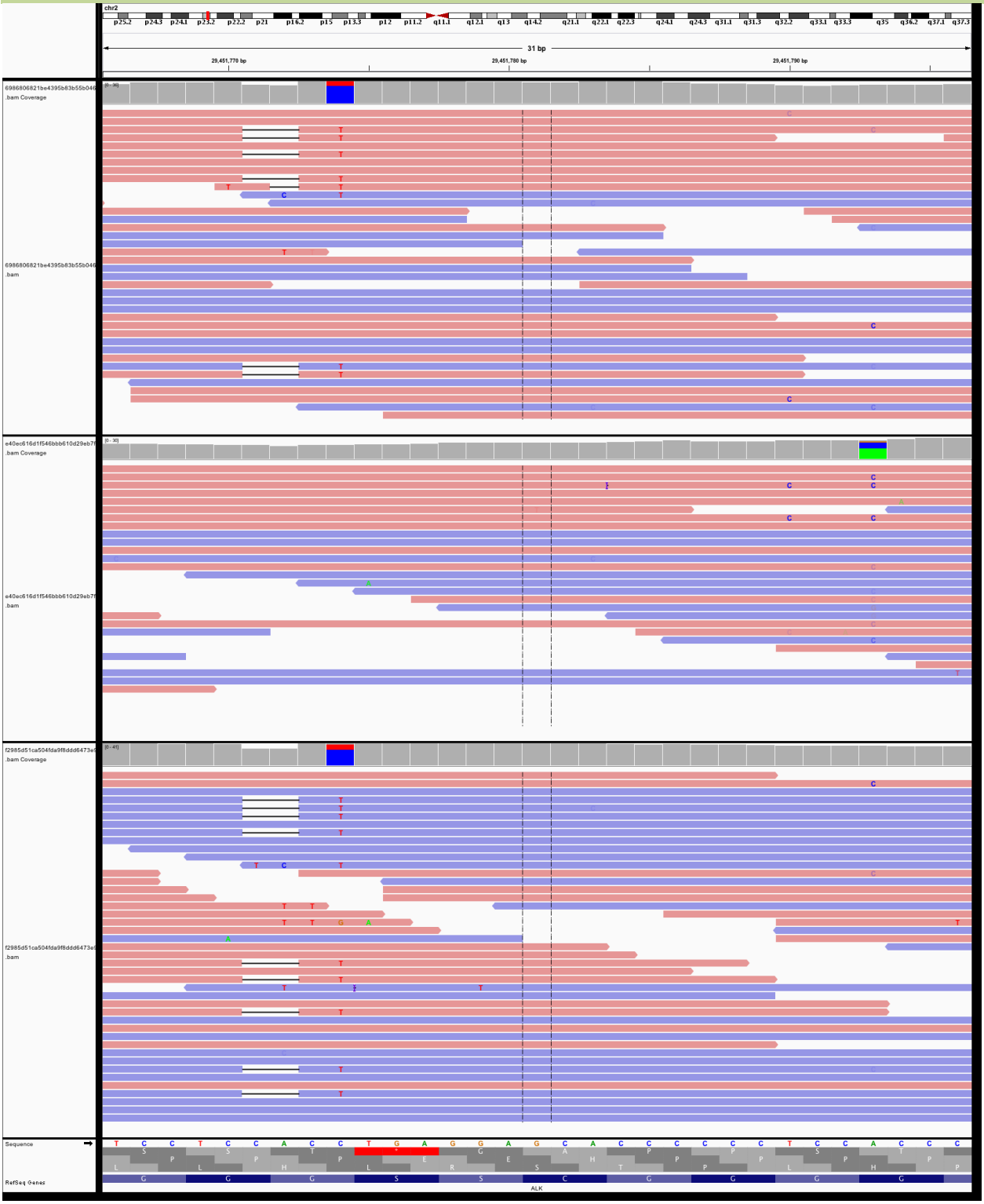
BRCA	TCGA-AR-A256	PARP1	1	226553763	42	2



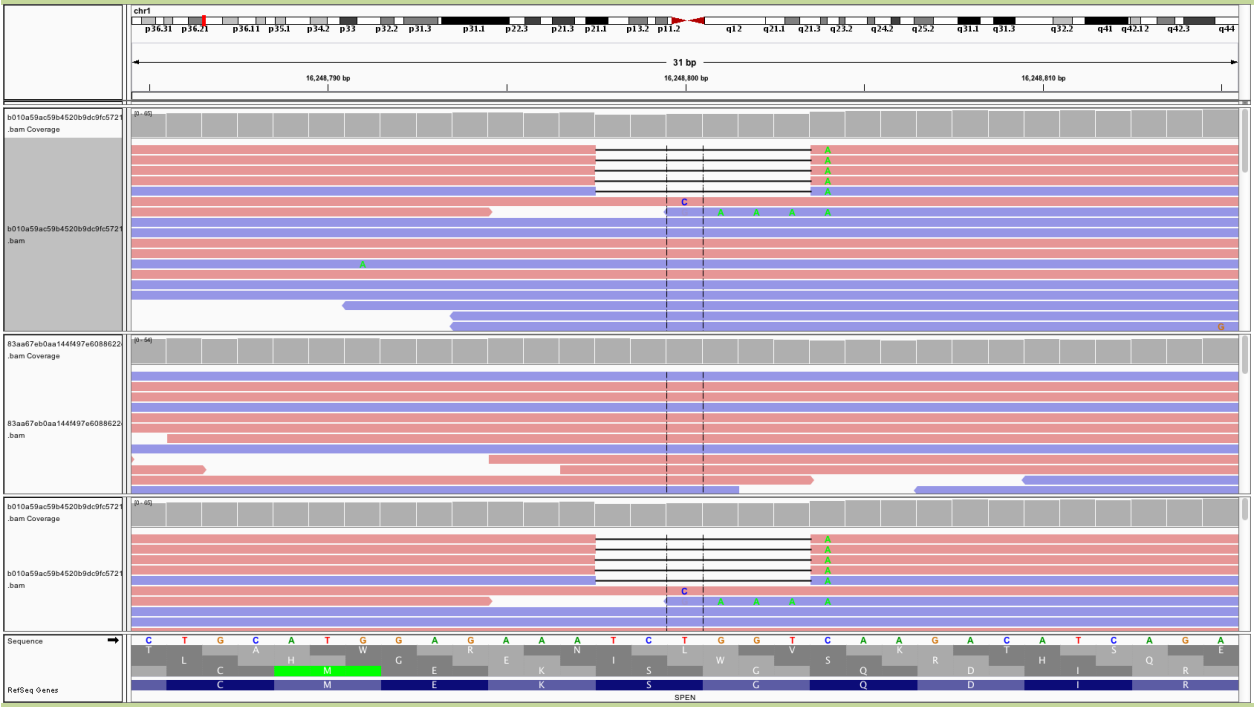
COAD	TCGA-AZ-6601	USP1	1	62916567	1	2



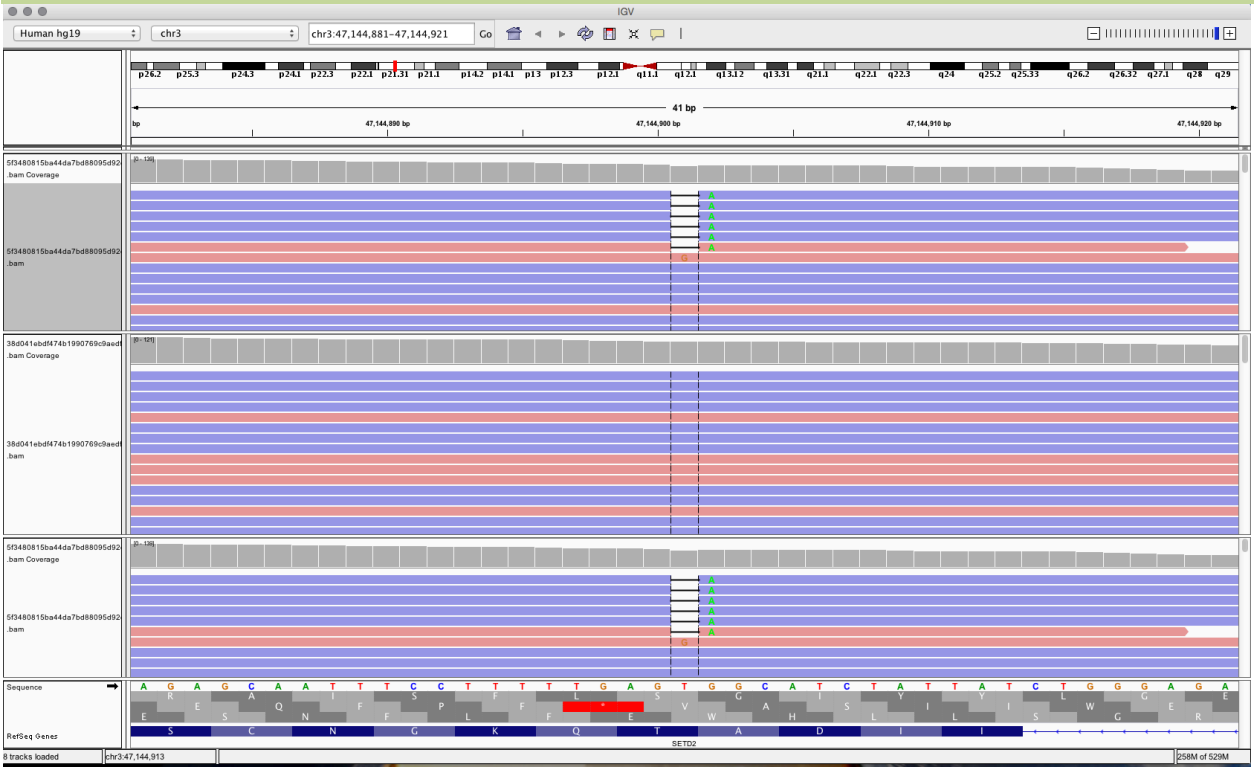
HNSC	TCGA-CV-7100	ALK	2	29451771	3	1



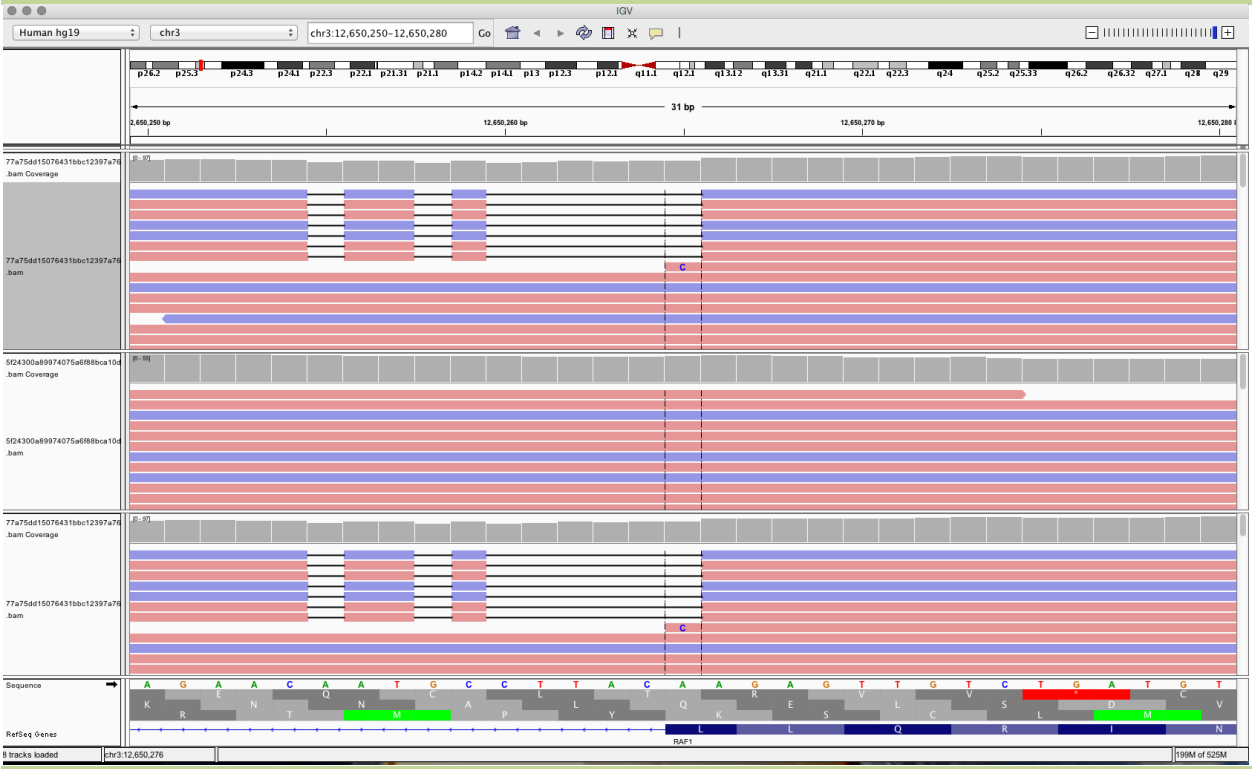
LIHC	TCGA-DD-A3A8	SPEN	1	16248797	7	1



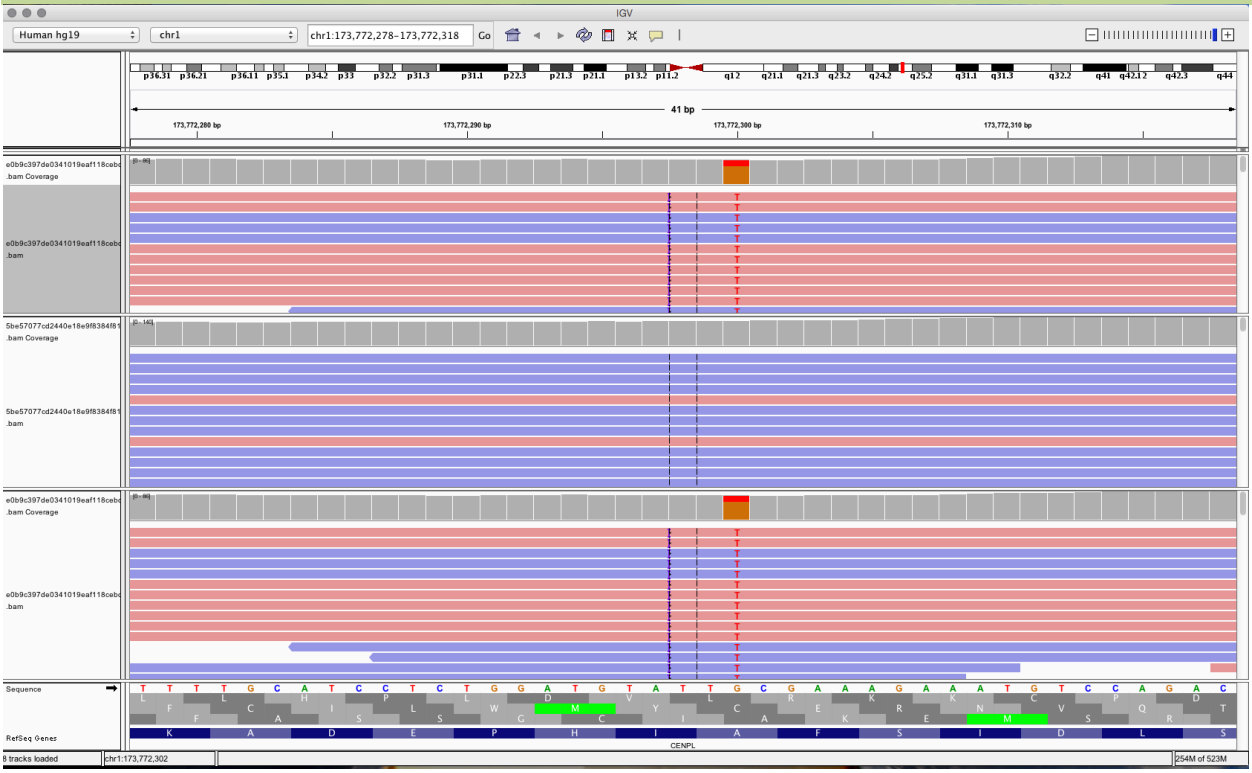
LIHC	TCGA-ED-A459	SETD2	3	47144900	2	1



STAD	TCGA-FP-7998	RAF1	3	12650255	10	2

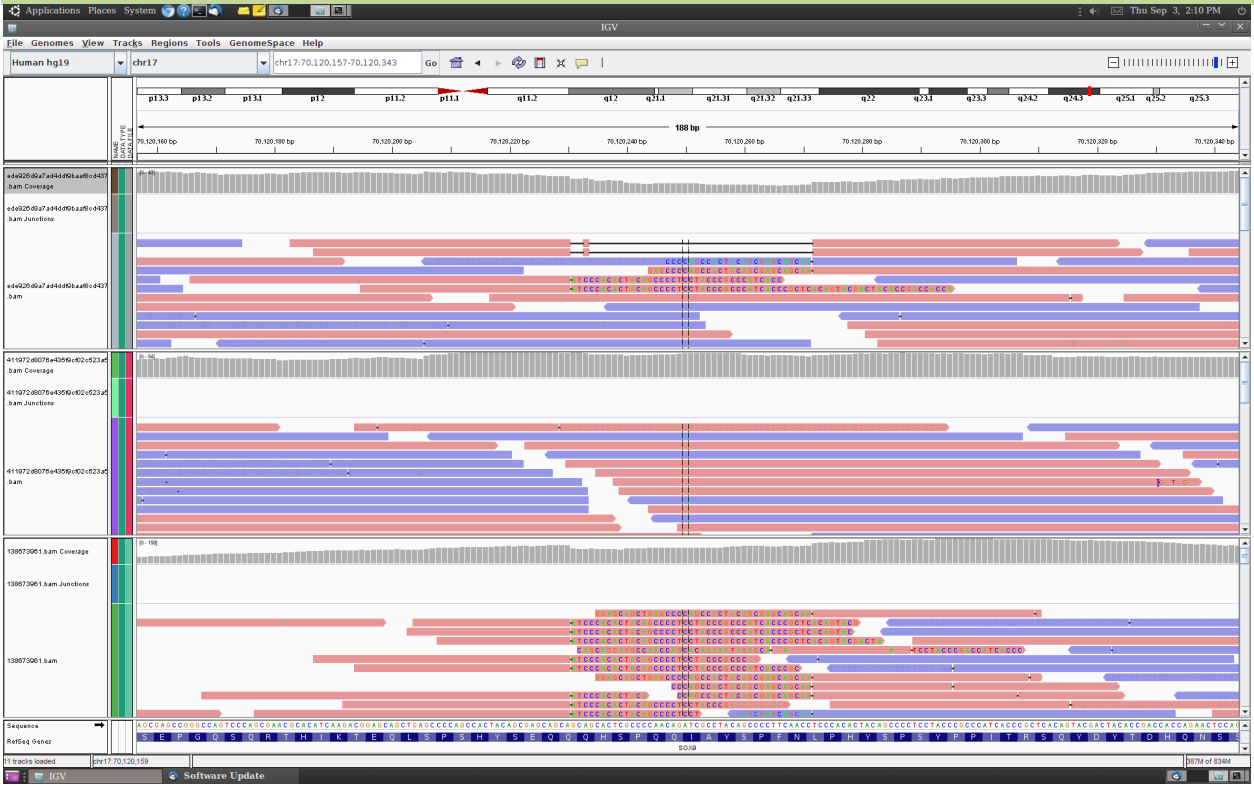


KIRP	TCGA-MH-A561	CENPL	1	173772299	1	2

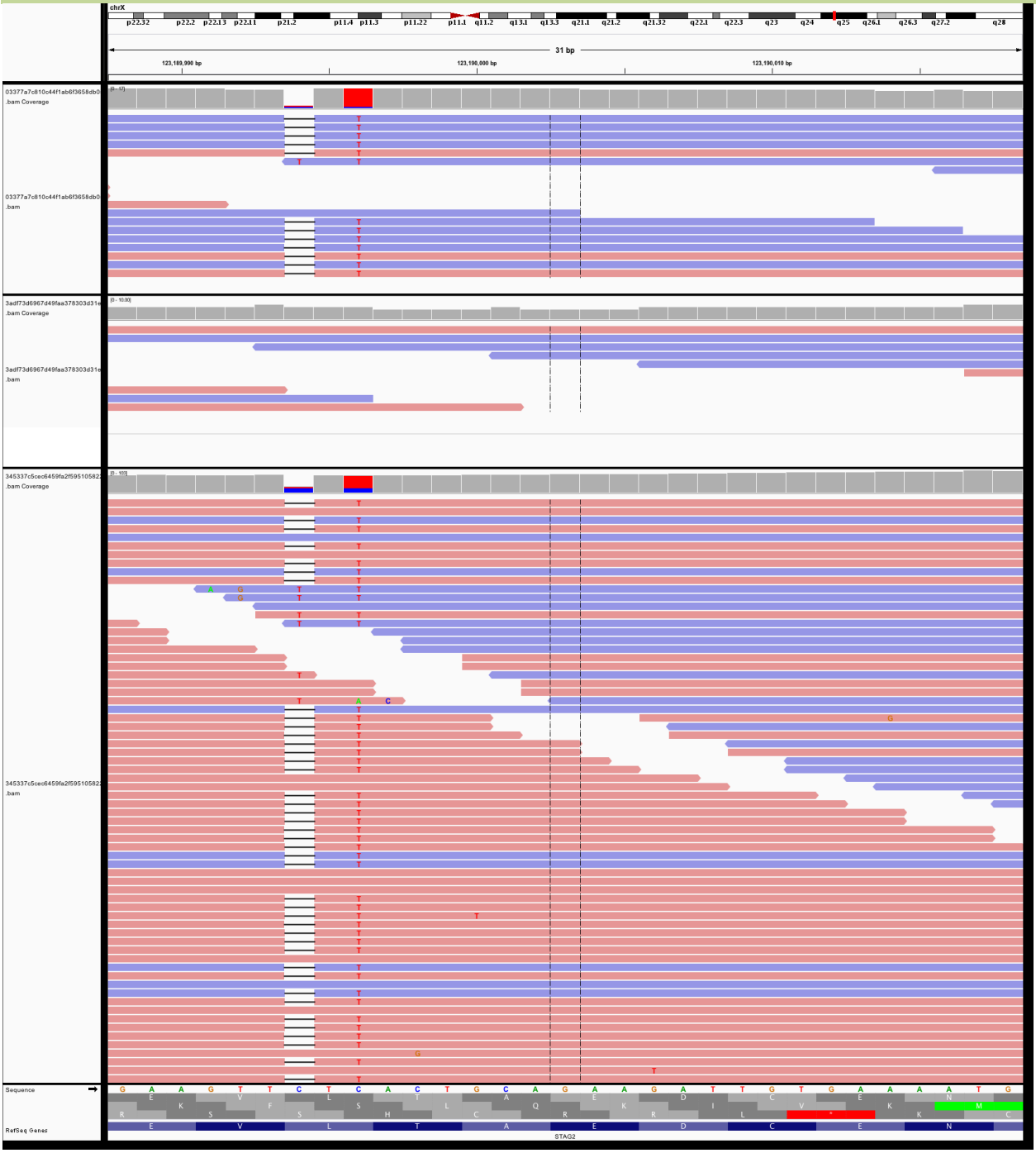


SARC	TCGA-MO-A47R	KIT	4	55593600	16	1

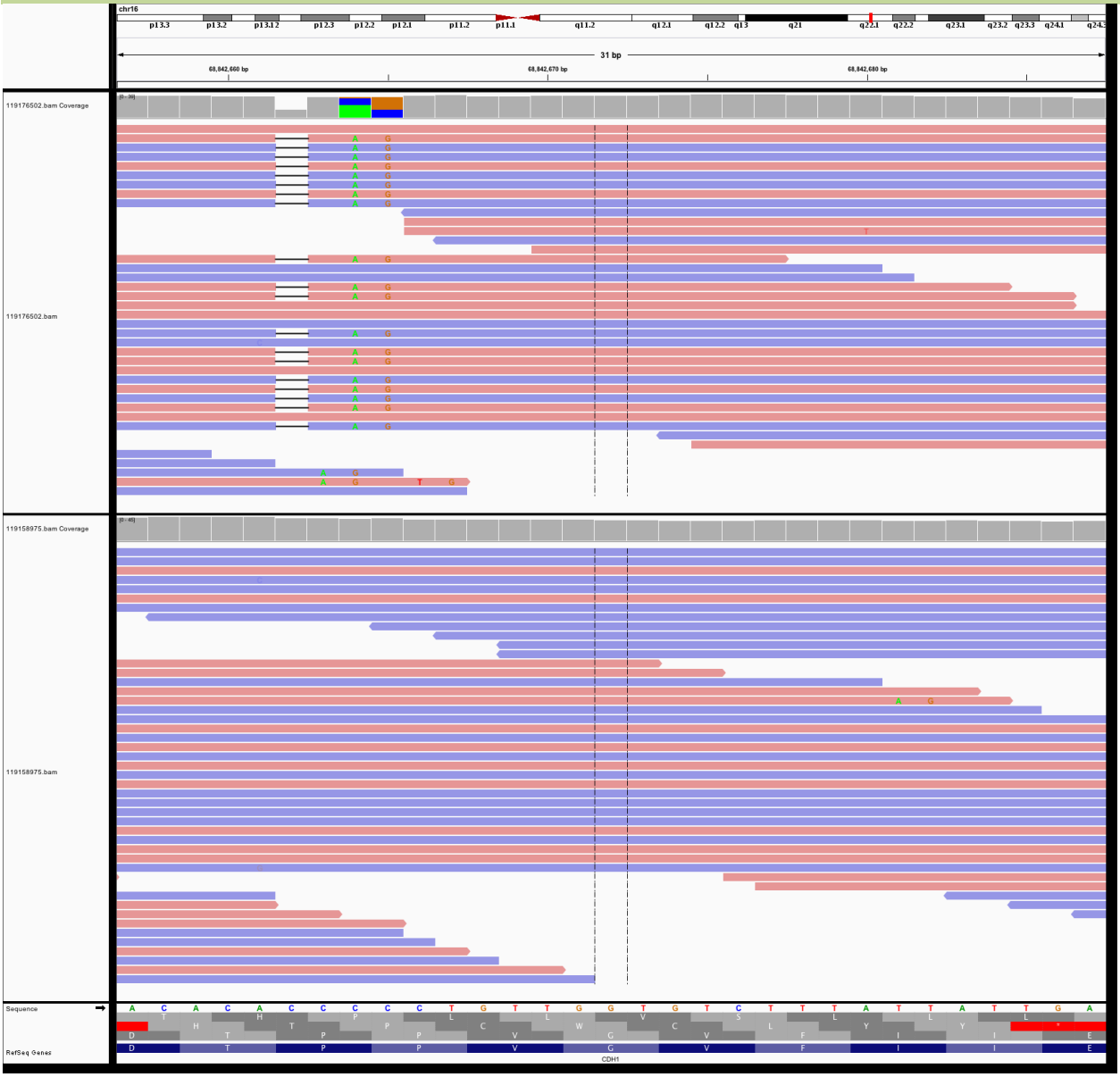
STAD	TCGA-BR-4280	SOX9		17	70120230	41	1



BLCA	TCGA-H4-A2HQ	STAG2	X	123189993	3	2



BRCA	TCGA-A8-A07B	CDH1	16	68842662	3	2



UCEC	TCGA-BS-A0U9	PIK3R1	5	67591142	12	2