Supplemental Materials for

Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events

Jinfeng Liu[1], William Lee[1], Zhaoshi Jiang[1], Zhongqiang Chen[1], Suchit Jhunjhunwala[1], Peter M. Haverty[1], Florian Gnad[1], Yinghui Guan[2], Houston Gilbert[3], Jeremy Stinson[2], Christiaan Klijn[1], Joseph Guillory[2], Deepali Bhatt[2], Steffan Vartanian[4], Kimberly Walter[5], Jocelyn Chan[6], Thomas Holcomb[5], Peter Dijkgraaf[2], Stephanie Johnson[7], Julie Koeman[8], John D. Minna[9], Adi F. Gazdar[9], Howard M. Stern[7], Klaus P. Hoeflich[6], Thomas D. Wu[1], Jeff Settleman[4], Frederic J. de Sauvage[2], Robert C. Gentleman[1], Richard M. Neve[4], David Stokoe[4], Zora Modrusan[2], Somasekar Seshagiri[2], David S. Shames[5], Zemin Zhang[1§]

Contents:

Supplemental Figure Legends

Supplemental Figure 1. Receiver operating characteristic curve for mutation validation.
Validation of somatic score for non-smoker patients (21214 and 34560). Somatic score was obtained using cgatools v1.3. 223 mutations from patient 21214, and 187 from 34560 were experimentally tested using the Sequenom technology. The patient specific ROC curves are shown.

Supplemental Figure 2. The fraction of C:G>A:T transversions is most negatively correlated with that of C:G>T:A transitions.
(A) A scatter plot showing the negative correlation between the fraction of C:G>A:T transversions and that of C:G>T:A transitions. (B) A heatmap showing pairwise Spearman's rank correlations between all mutation types. Although correlations are generally high between all mutation types due to their dependencies, the mutation type with the highest absolute correlation with C:G>A:T is C:G>T:A.

Supplemental Figure 3. Somatic missense mutations are more likely to be deleterious than germline variations.
Left panel, Polyphen predicted 59% damaging mutations in the somatic set compared to 15% in the germline set (p=5e-90, chi-squared test). Right panel, 52% of the somatic missense mutations are predicted to be deleterious according to SIFT in comparison to 18% for the germline variants from normal tissues (p=3e-57, chi-squared Test).

Supplemental Figure 4. GISTIC analysis of SNP Array-based DNA copy number showed recurrent gain and loss of multiple genes characteristic of NSCLC.
Negative $\log_{10}$ False Discovery Rates are plotted in red for Gain and in blue for Loss. Genes representative of the most significant peaks are labeled with dotted grey lines. *ADAR* is at the focus of the 1q21.3 peak.

Supplemental Figure 5. *ADAR* amplification and over-expression.
DNA copy number for *ADAR* was calculated from SNP Array data processed via PICNIC (see Methods). Segmented total copy number values for all SNPs bounding and within the boundaries of *ADAR* were averaged and plotted against RNA-Seq-based expression for *ADAR*.

Supplemental Figure 6. Fraction of putative chimera transcripts detected from transcriptome data supported by DNA-level evidence.

The percentage of fusion transcripts with supporting DNA reads for each sample is plotted. The dotted horizontal lines represent the mean percentages. Taken together, 44% of 1,097 putative chimera transcripts predicted by ChimeraScan from our transcriptome data have at least two genomic sequencing reads supporting the same gene pair (top panel). Among the 291 putative chimera transcripts with RNA-seq reads directly spanning the chimera junctions, 62% of them have at least two genomic sequencing reads supporting the same gene pair (bottom panel).

Supplemental Figure 7. Experimental validation of four predicted gene fusion events.

PCR of lung cancer cell lines containing genomic fusions. Primers pairs that anneal to sequence on either side of the fusion sites were used to amplify genomic DNA (Panel A) or cDNA (Panel B) from cell lines containing the fusions indicated. Lymphocytes represent normal genomic DNA. The negative control cell line H2009 does not contain the fusions. Amplification products were run on 2% agarose gels. (Lymph = lymphocytes, NTC = no template control)

Supplemental Figure 8. *CLTC-VMP1* fusion in H1299.

(A) Genome browser view showing the context of observed *CLTC-VMP1* fusion. The fusion is supported by 81 discordant RNA-seq reads (in red), and the supporting DNA reads from whole genome sequencing are shown in blue. We were able to obtain a *de novo* assembly (in green) of these reads. (B) Nucleotide sequence of the RNA fusion junction derived from *de novo* assembly of discordant RNA-seq reads (top sequence), and the alignment of a subset of RNA-seq reads directly mapped across the junction. (C) Putative amino acid sequence at the fusion junction, and impacts on protein domain organization.

Supplemental Figure 9. *HIF1A-SNAPC1* fusion and *MLLT3-TMIGD1* fusion.

(A) A cluster of discordant reads DNA-seq reads (in blue) suggests a 50 kb deletion in H1299, which juxtapose the *HIF1A* gene to its neighboring gene *SNAPC1*. Multiple discordant RNA-seq reads suggest that there are aberrant splicing events connecting the first few exons of HIF1A to the last few exons of *SNAPC1*. The predicted in-frame fusion protein product is 284 amino acids in length, missing the critical C-terminal transactivation domain of the wild-type HIF1A protein. (B) The *MLLT3-TMIGD1* fusion in H838 resulted from an inter-chromosomal translocation between chromosome 9 and chromosome 17 (discordant DNA reads in purple and RNA reads in red). The fusion was predicted to be out-of-frame, therefore, no productive fusion protein product was expected.

Supplemental Figure 10. Cytogenetic characterization of cell line samples.

For each of 19 cell lines, one representative image of spectral karyotyping (SKY) results was illustrated.

Supplemental Figure 11. Splice site mutation in *AKT3* and the associated aberrant splicing event

A splice site mutation is associated with an exon skipping event in *AKT3* in H1155. We identified a novel mutation in serine/threonine protein kinase *AKT3* that alters the GT essential splice donor sequence after exon 6. From the RNA-seq data, we observed four reads spanning a novel exon-exon junction between exons 5 and 8, skipping the three exons in between. The resulting protein product has a 45 amino acid in-frame deletion in the essential protein kinase domain.

Supplemental Figure 12. Examples of differential isoform expression in tumors.

RNA-seq depth coverage for four genes from the tissue samples was plotted using Integrated Genome Browser. Tracks for tumors (in red) and their respective adjacent normal tissues (in blue) were arranged next to each other. Shaded rectangles represented the exons that have consistent differential expression between normal and tumor samples in all three patients.

Supplemental Figure 13. Number of variants in lung cancer samples after removing the known germline variants.

All single nucleotide variations (SNVs) called by the Complete Genomics pipeline were filtered against known germline variants, including those in the dbSNP database, the 1000 genome project, the 69 fully sequenced genomes by Complete Genomics, and the NHLBI GO Exome Sequencing Project (ESP). Shown in the plots are the numbers of remaining protein-altering variants only (A) and all remaining variants (B) in each sample. The variants in the normal genomes represent 'private' SNPs that are not in current database of known germline variants. We observed an average of 315 such private protein-altering SNPs per normal genome. By contrast, most of our cell line samples harbor a much higher number of filtered variants.

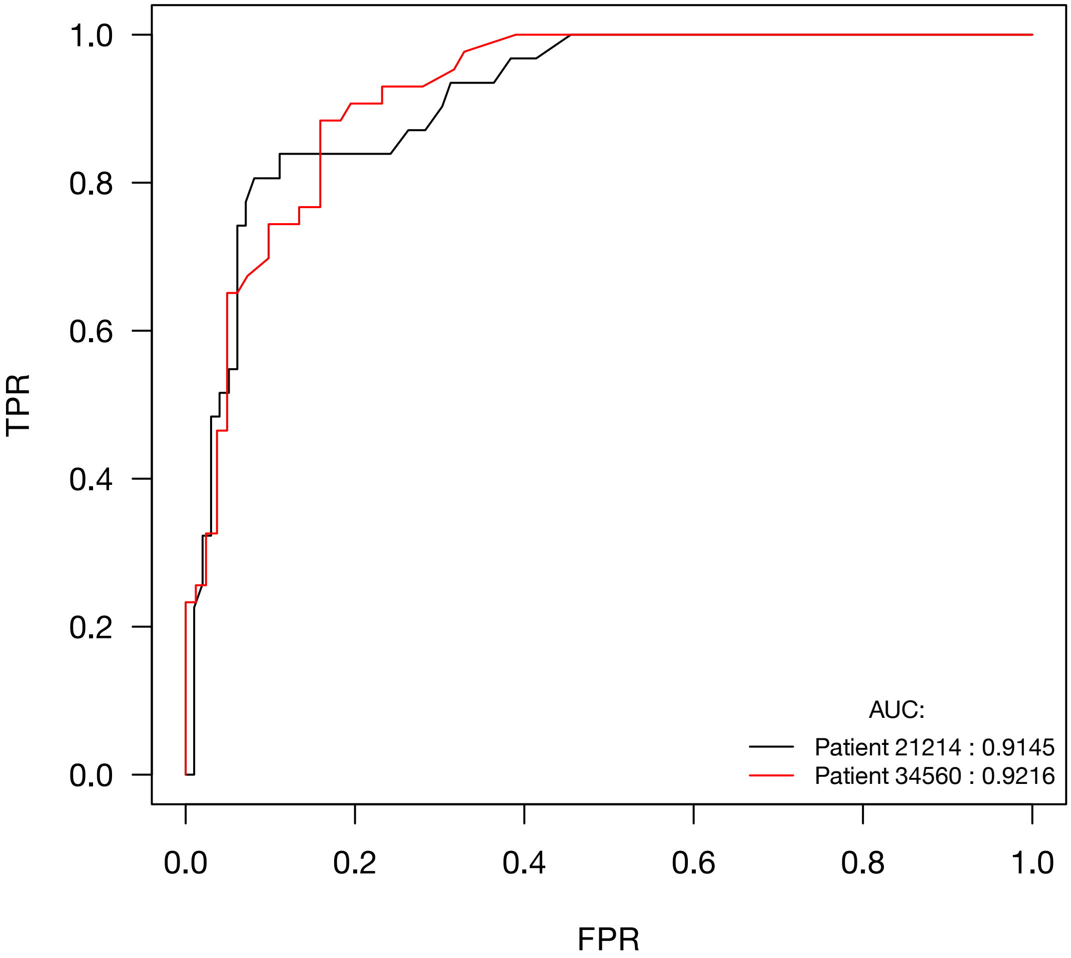
Supplemental Figure 14. Summary of genomic alterations by Circos plots in lung tumor and cell line genomes.

Various types of genomic alterations in the lung tumor and cell line genomes using Circos plots (Krzywinski et al. 2009) (A–D). High-confidence candidate structural variations (SVs) are shown as lines, with red lines representing inter-chromosomal SVs and blue lines indicating intra-chromosomal SVs. Regions of loss of heterozygosity (LOH) and allelic imbalance (AI) are shown in green. Copy number alterations are shown as bar plots with copy number gain shown in red and copy number loss in blue (the scale ranges from −2 to 4). The copy number alteration and LOH/AI data of tissue samples were derived from whole genome sequencing data while that of cell lines were derived from Illumina Human Omni2.5-4 SNP array results. For cell lines, copy number and LOH/AI results of chromosome Y were not processed and were not shown. Each surrounding red dot represents the number of high-confidence filtered SNVs within a 1 million base pair window. Cell line names or tumor genome ids are shown at the center of each circular view. Cell lines tend to exhibit higher number of SVs compared to tumor tissue samples, for the lack of matched normal samples in our germline filter. The pattern of structural variations is very diverse among these samples, similar to what was reported previously for breast cancer cell lines. Of note, H2073 contains a large number of SVs, particularly for small inversions, which is consistent with that fact that this cell line was isolated from the primary tumor after chemotherapy.

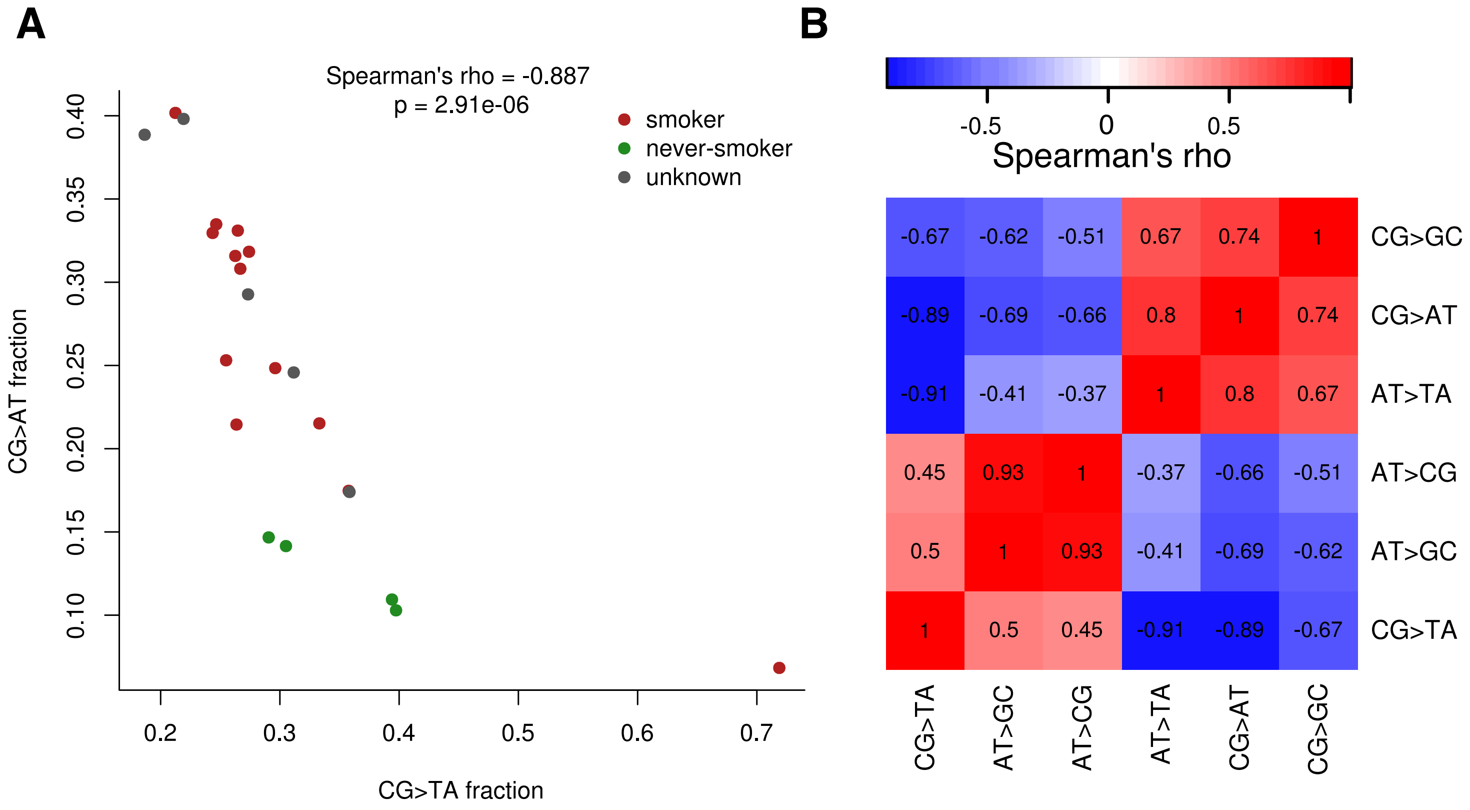
Supplemental Figure 15. Validation of called genomic mutations by transcriptome sequencing data.
The top panel shows the numbers of SNVs that are covered by 10 or more RNA-seq reads, and they display significant variability. The lower panel shows the percentage of candidate SNVs that are supported by RNA Seq data. On average, 75% of the SNV (the dotted line) that are covered by10 or more RNA-seq reads are supported by the presence of variant reads.

Supplemental Figure 16. Comparison of variant calls between Complete Genomics whole-genome sequencing and CCLE targeted sequencing.
14 cell lines in our collection were also analyzed in the recently published targeted-sequencing data from the Cancer Cell Line Encyclopedia (CCLE). On average, 94% (the dotted line) of all the mutations published by CCLE were also called by our Complete Genomics data, representing a possible 6% false negative rate in our data.

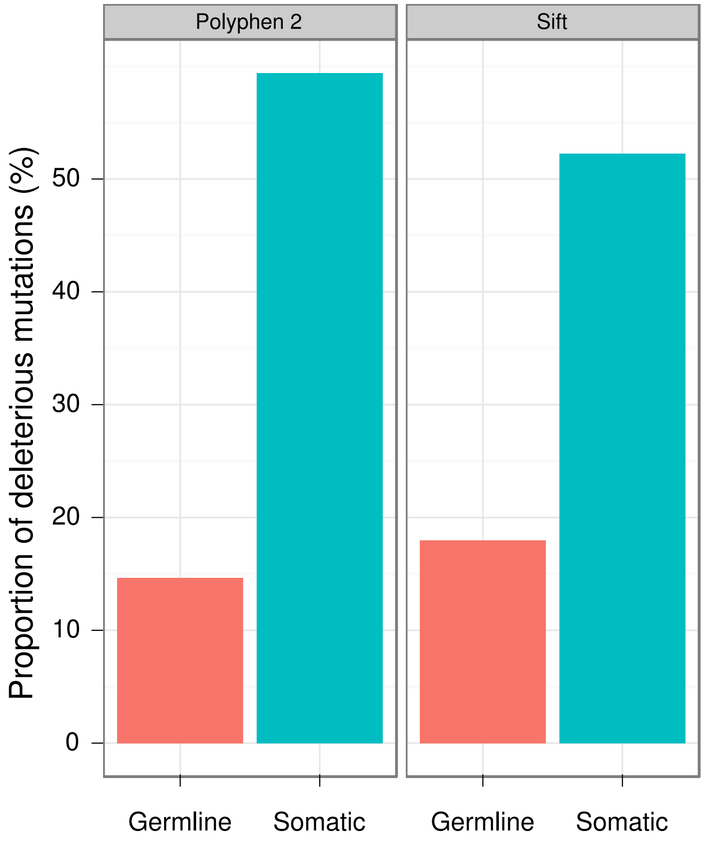
Supplemental Figure 1



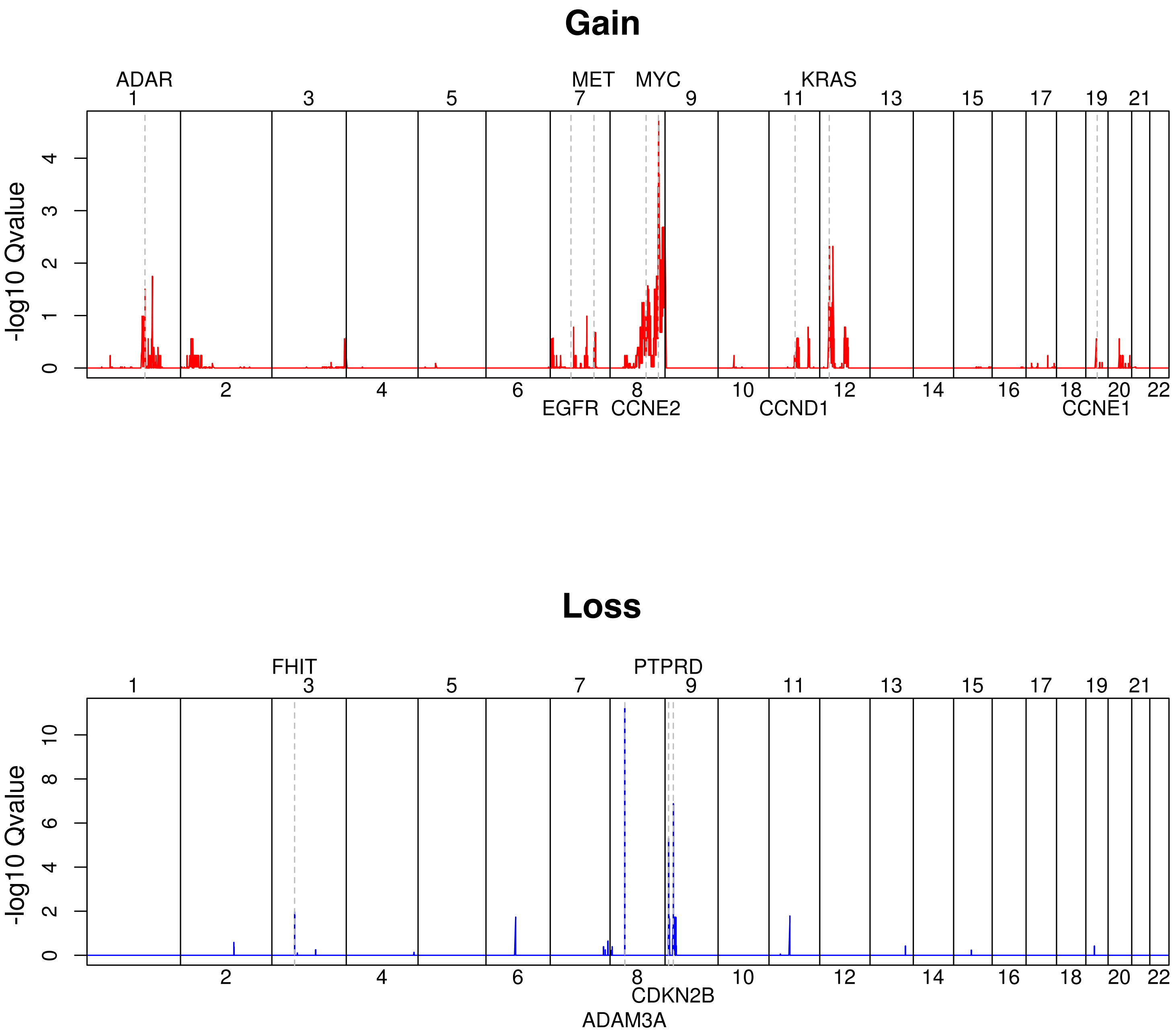
Supplemental Figure 2



Supplemental Figure 3

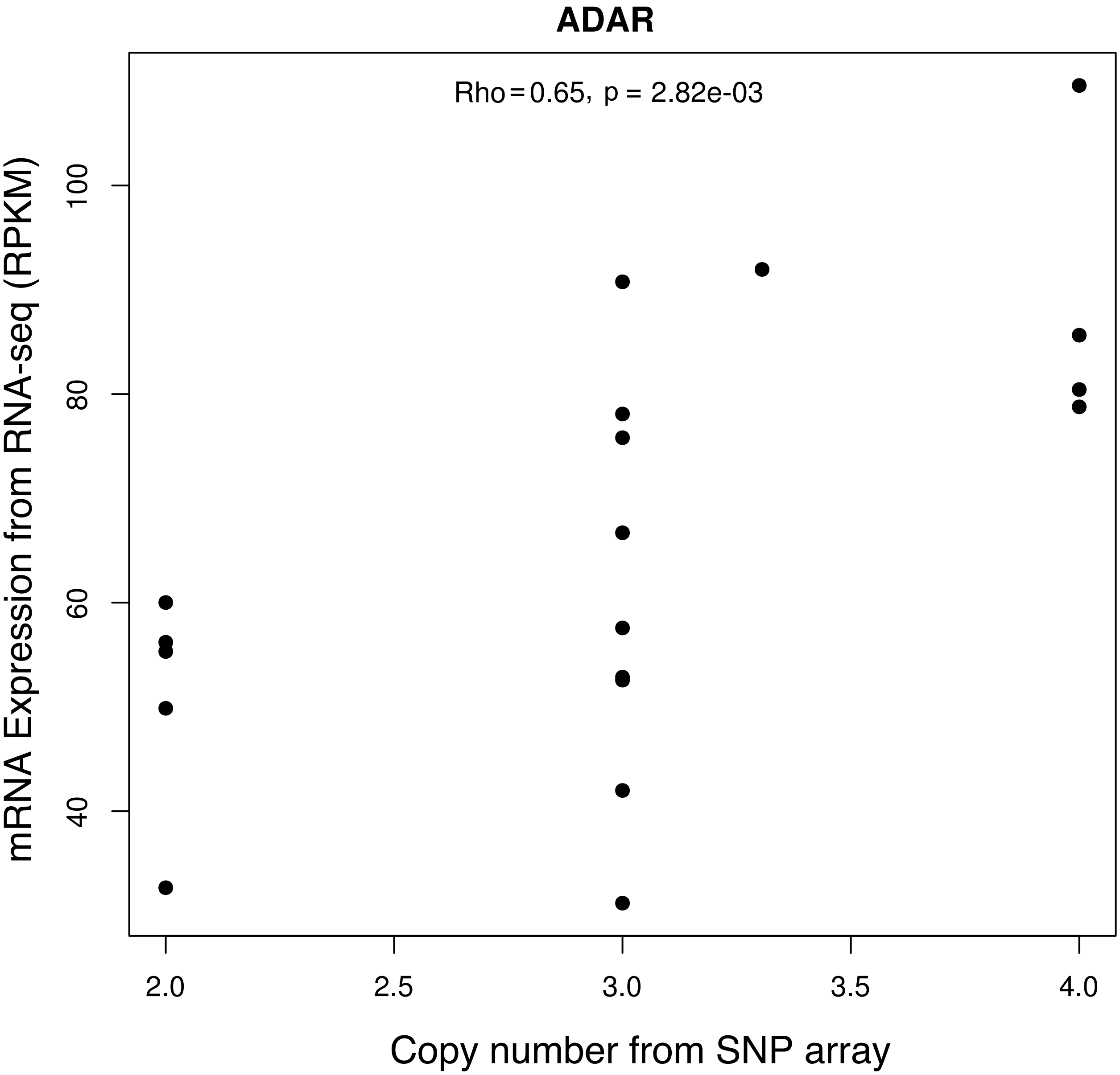


Supplemental Figure 4

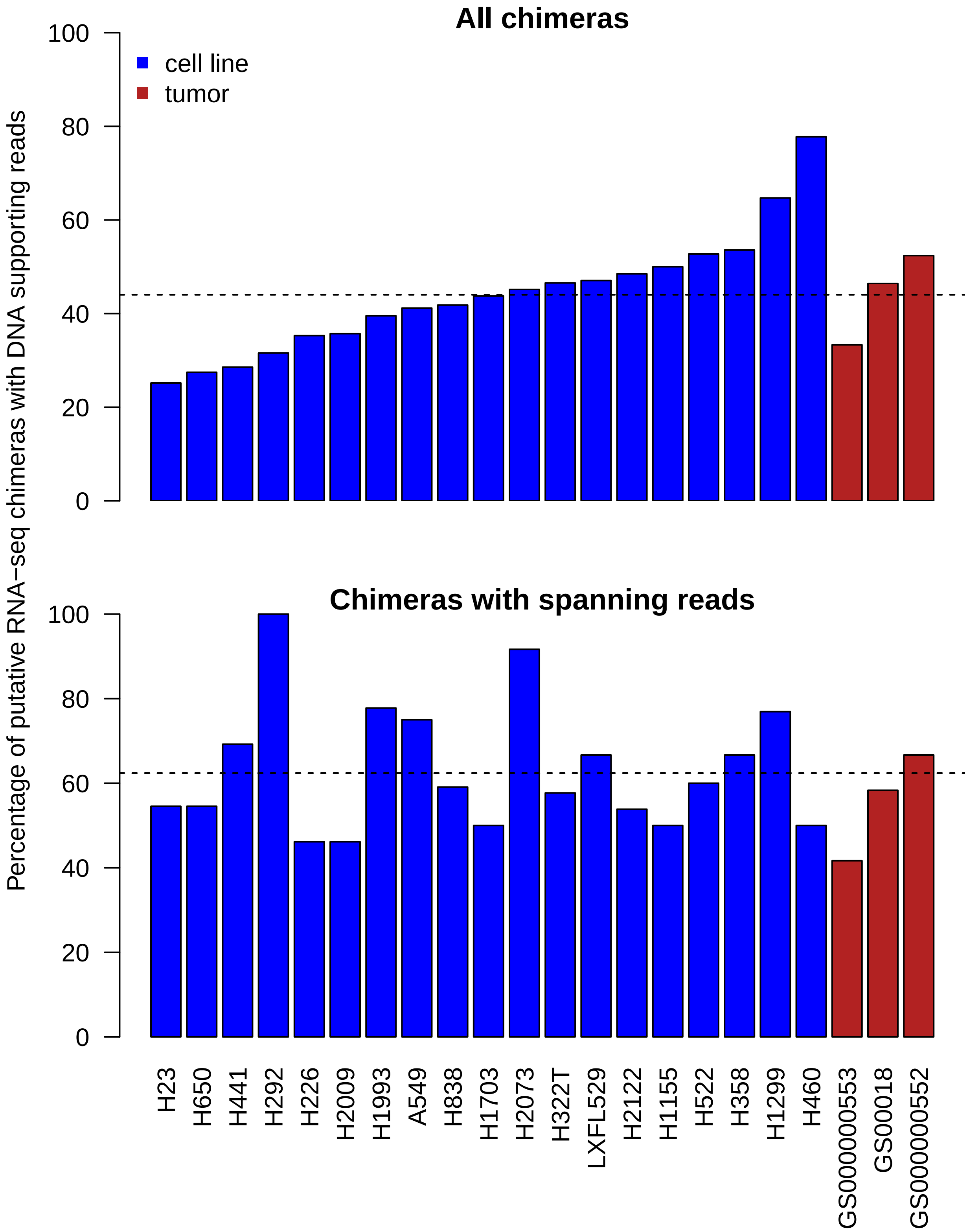

Gain

Loss

Supplemental Figure 5



Supplemental Figure 6

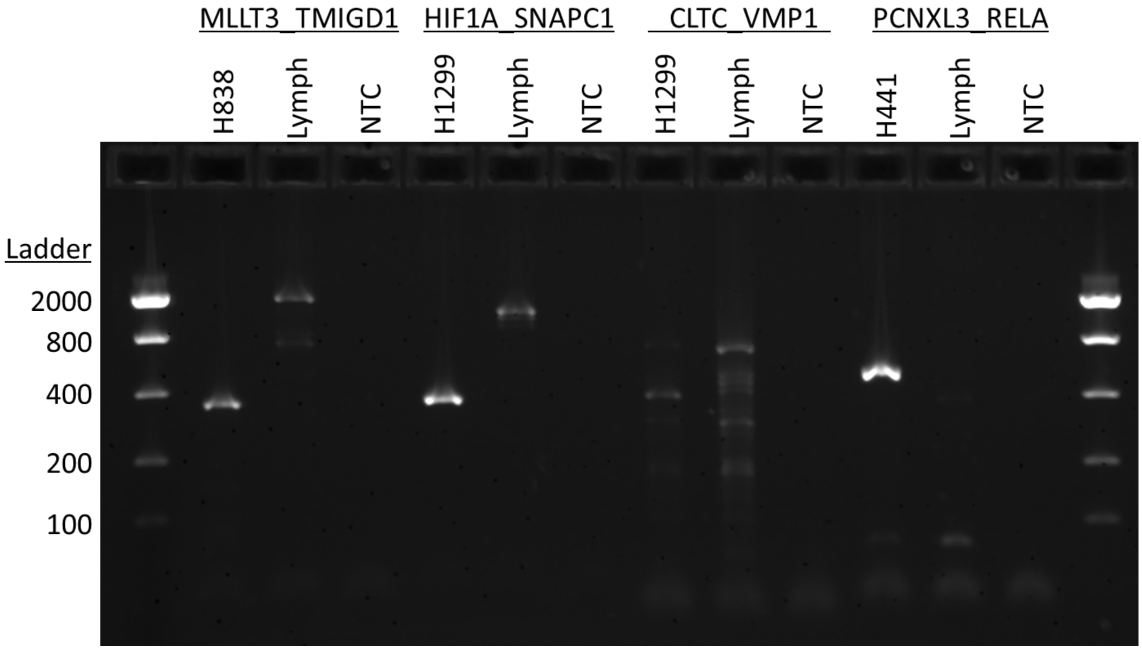


All chimeras

Chimeras with spanning reads

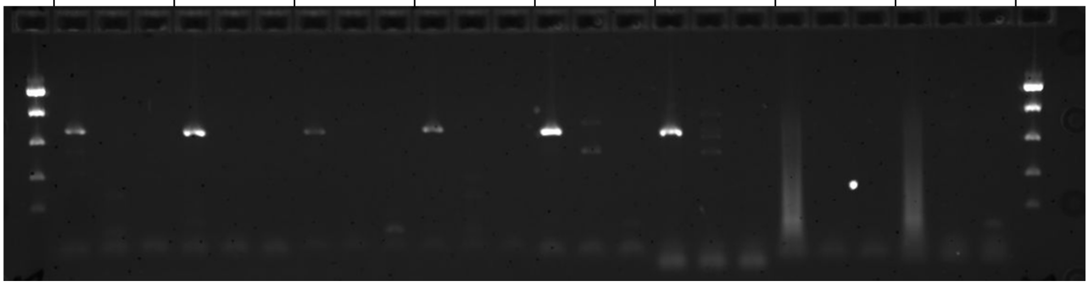
Supplemental Figure 7

A

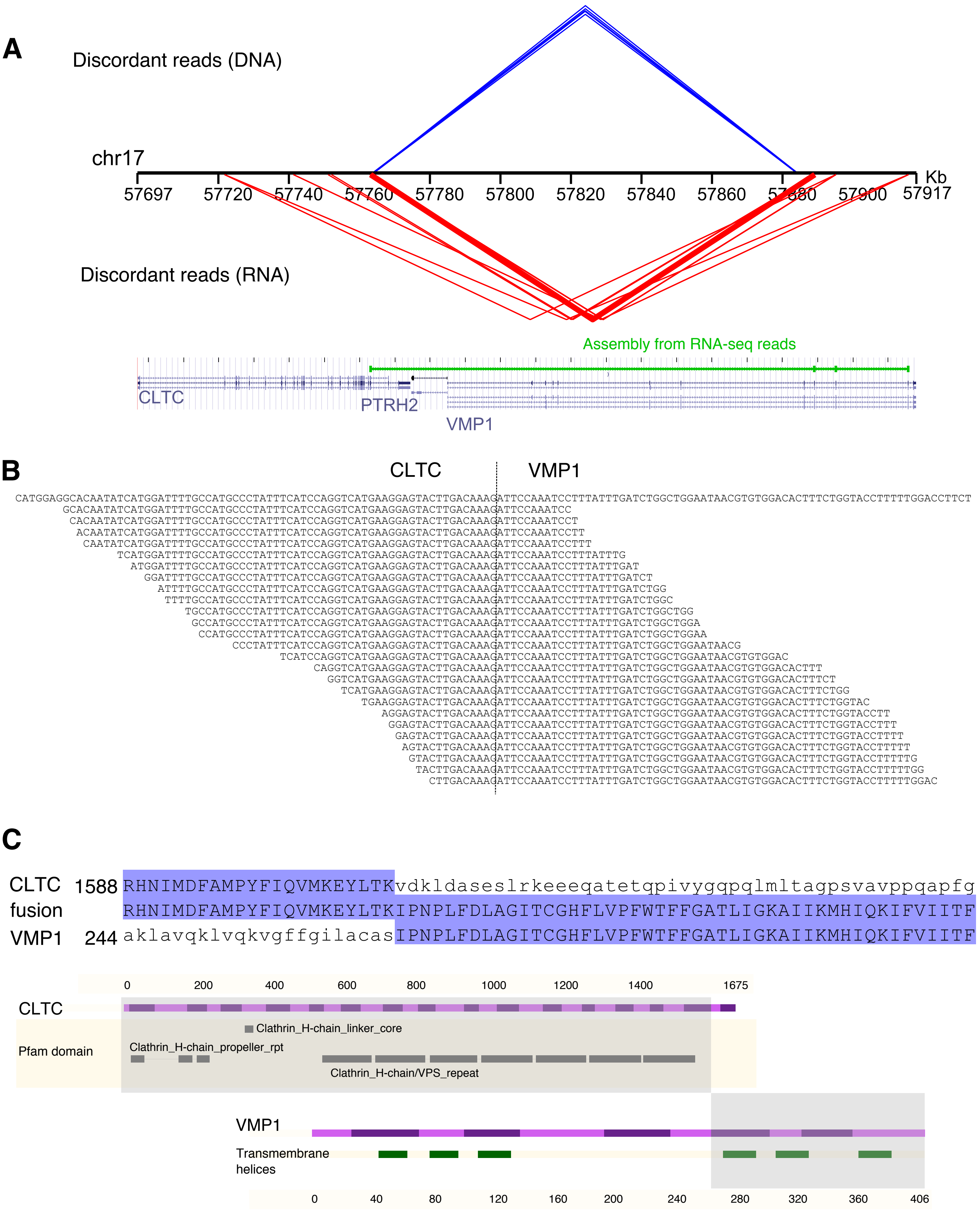


B

	MLLT3_TMIGD1		HIF1A_SNAPC1		CLTC_VMP1		PCNXL3_RELA	
	RT1	RT2	RT1	RT2	RT1	RT2	RT1	RT2
H838	+	+						
H1299			+	+	+	+		
H441							+	+
H2009	+	+	+	+	+	+	+	+
NTC	+	+	+	+	+	+	+	+

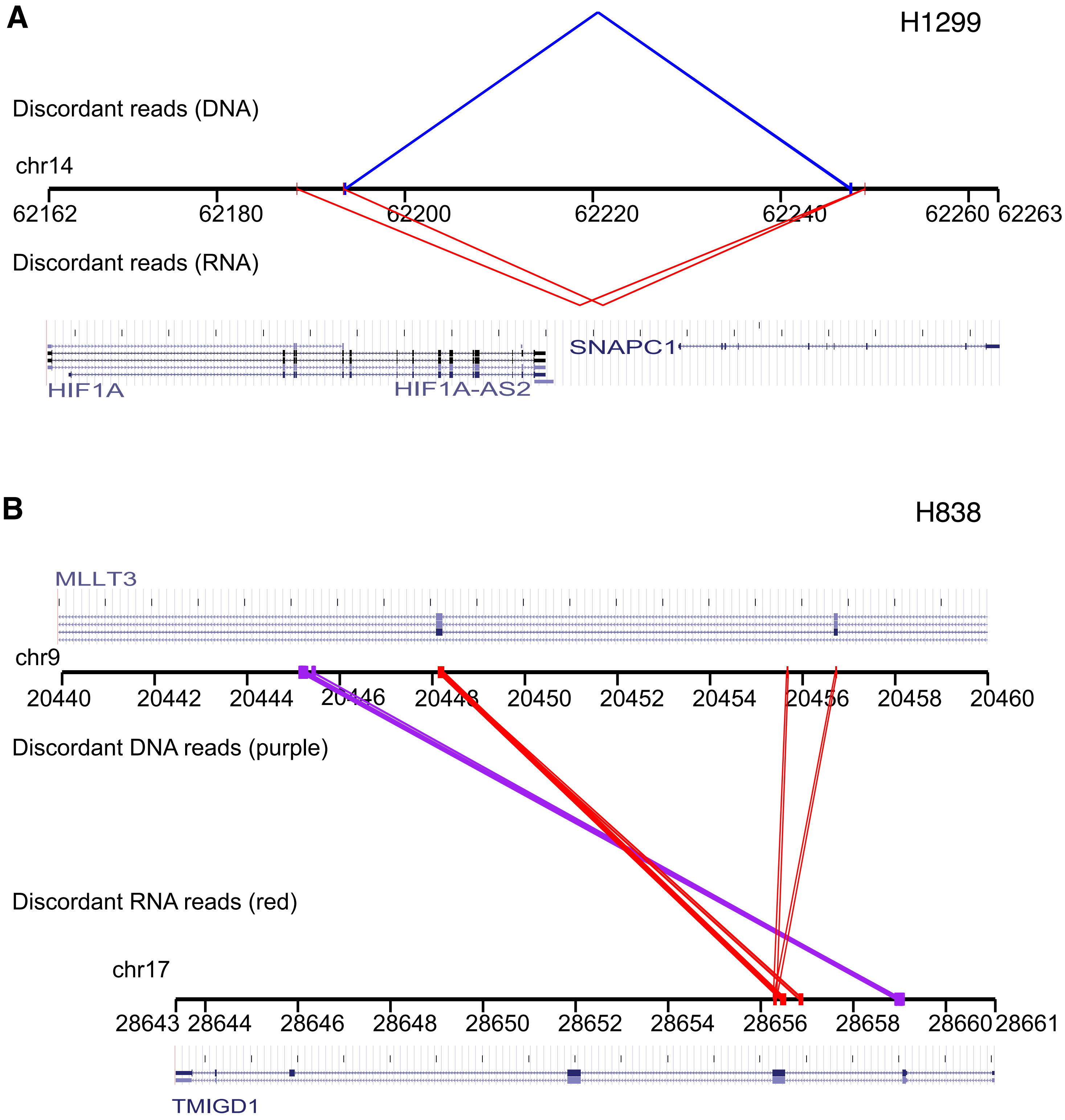


Supplemental Figure 8

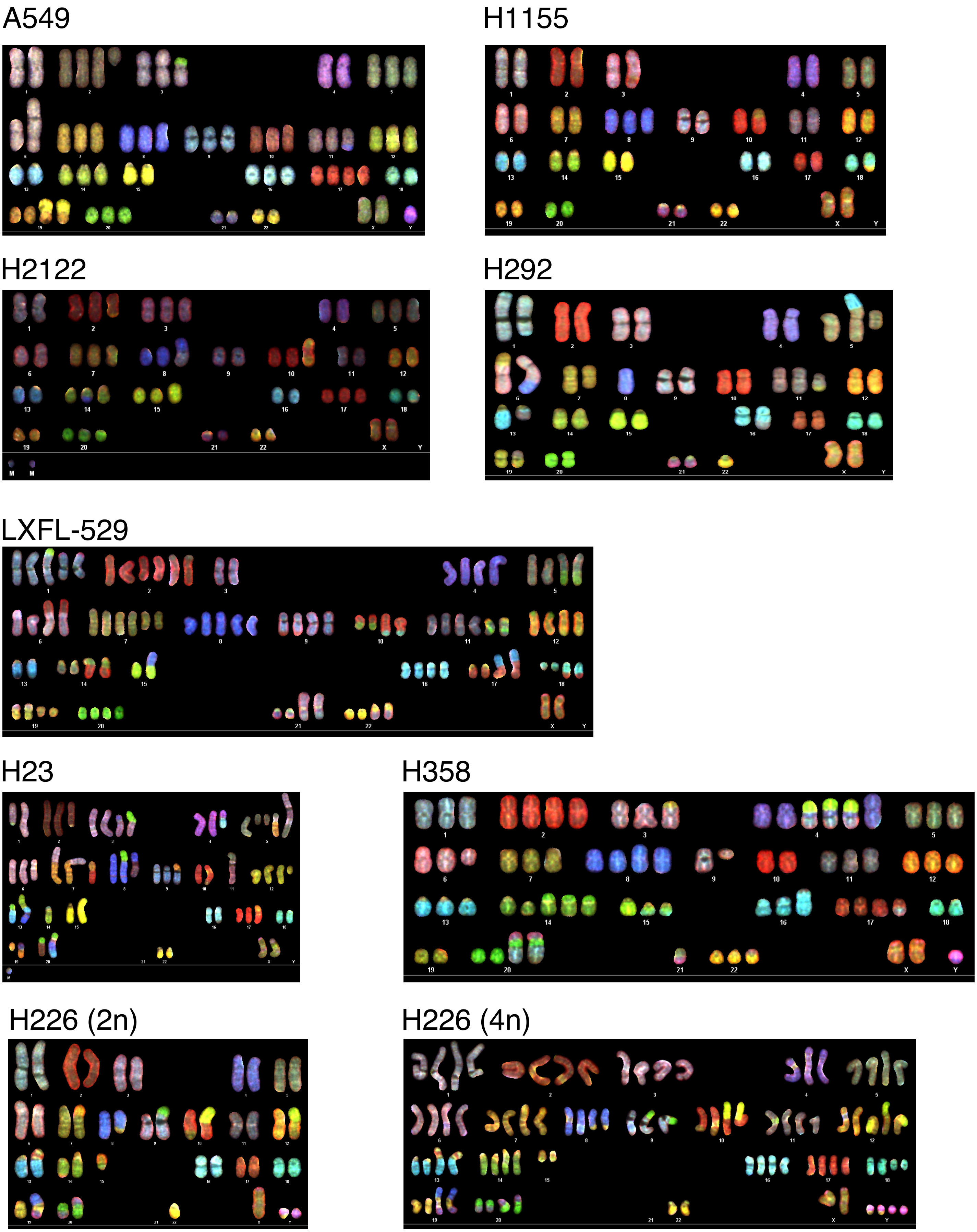

A
Discordant reads (DNA)
chr17
57697
57720
57740
57760
57780
57800
57820
57840
57860
57880
57900
57917
Kb
Discordant reads (RNA)
Assembly from RNA-seq reads
CLTC
PTRH2
VMP1
B
CLTC
VMP1
C
CLTC
fusion
VMP1
1588
244
RHNIMDFAMPYFIQVMKEYLTKvdkldaseslrkeeeqatetqpivygqpqlmltagpsvavppqapfg
RHNIMDFAMPYFIQVMKEYLTKIPNPLFDLAGITCGHFLVPFWTFFGATLIGKAIIKMHIQKIFVIITF
aklavqklvqkvgffgilacasIPNPLFDLAGITCGHFLVPFWTFFGATLIGKAIIKMHIQKIFVIITF
0
200
400
600
800
1000
1200
1400
1675
CLTC
Pfam domain
Clathrin_H-chain_linker_core
Clathrin_H-chain_propeller_rpt
Clathrin_H-chain/VPS_repeat
VMP1
Transmembrane
helices
0
40
80
120
160
200
240
280
320
360
406

```
CATGGAGGCACAATATCATGGATTTTGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCTGGTACCTTTTTGGACCTTCT
       GCACAATATCATGGATTTTGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCC
        CACAATATCATGGATTTTGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCT
         ACAATATCATGGATTTTGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTT
          CAATATCATGGATTTTGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTT
               TCATGGATTTTGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTG
                 ATGGATTTTGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGAT
                   GGATTTTGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCT
                     ATTTTGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGG
                      TTTTGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGC
                         TGCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGG
                          GCCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGA
                           CCATGCCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAA
                                CCCTATTTCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACG
                                       TCATCCAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGAC
                                            CAGGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTT
                                              GGTCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCT
                                                TCATGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCTGG
                                                   TGAAGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCTGGTAC
                                                      AGGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCTGGTACCTT
                                                       GGAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCTGGTACCTTT
                                                        GAGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCTGGTACCTTTT
                                                         AGTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCTGGTACCTTTTT
                                                          GTACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCTGGTACCTTTTTG
                                                           TACTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCTGGTACCTTTTTGG
                                                             CTTGACAAAGATTCCAAATCCTTTATTTGATCTGGCTGGAATAACGTGTGGACACTTTCTGGTACCTTTTTGGAC
```

Supplemental Figure 9



Supplemental Figure 10 A



A549

H1155

H2122

H292

LXFL-529

H23

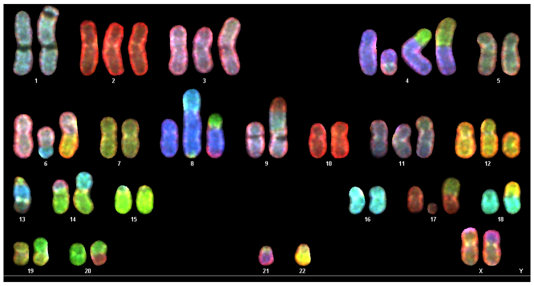
H358

H226 (2n)

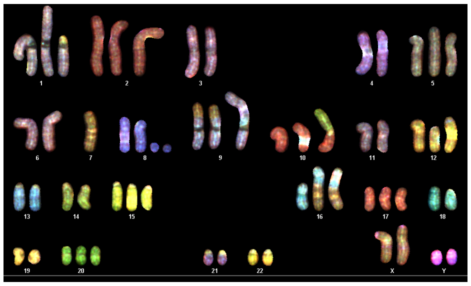
H226 (4n)

Supplemental Figure 10 B

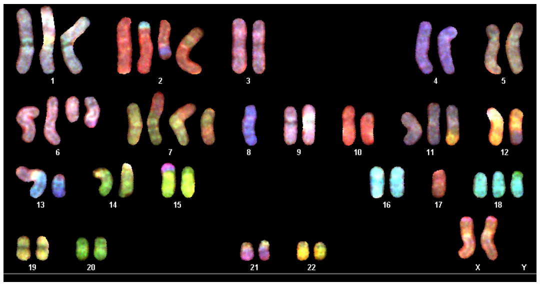
H441



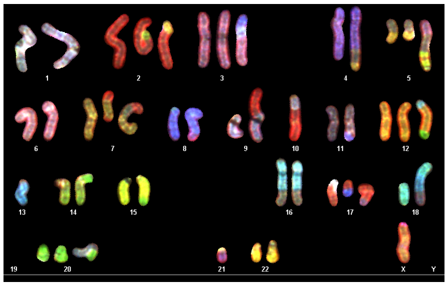
H460



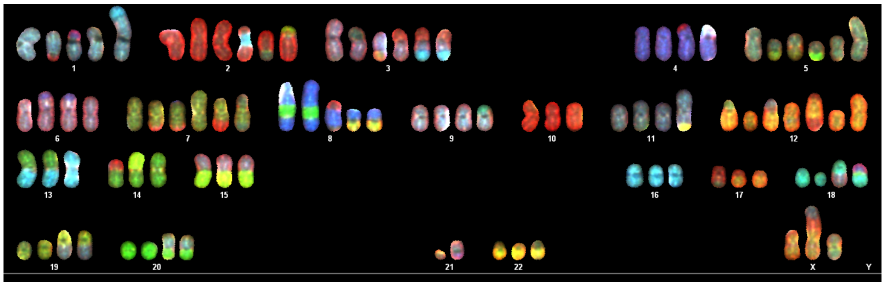
H522



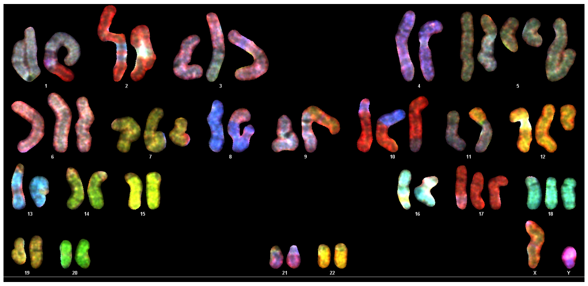
H838



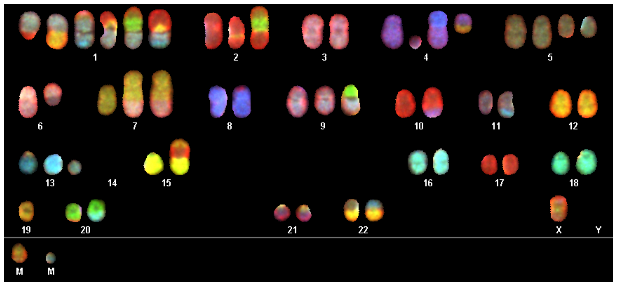
H650



H1703

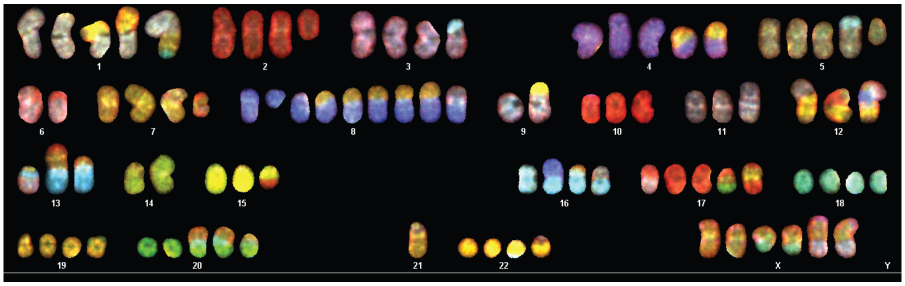


H2009

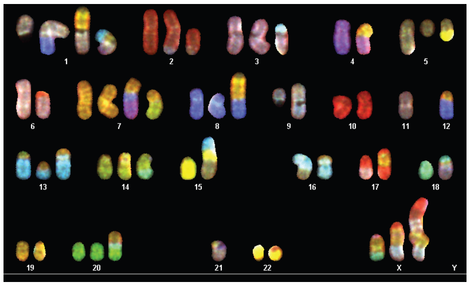


Supplemental Figure 10 C

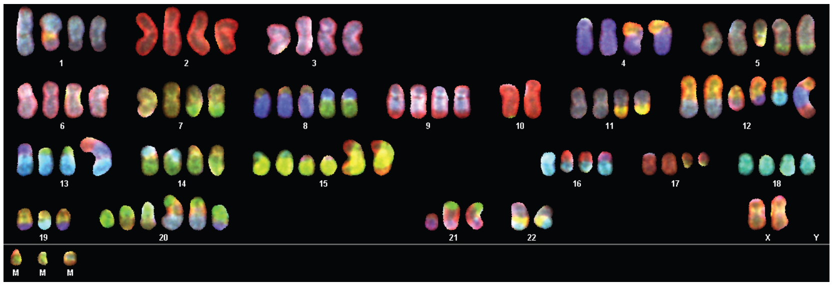
H1993



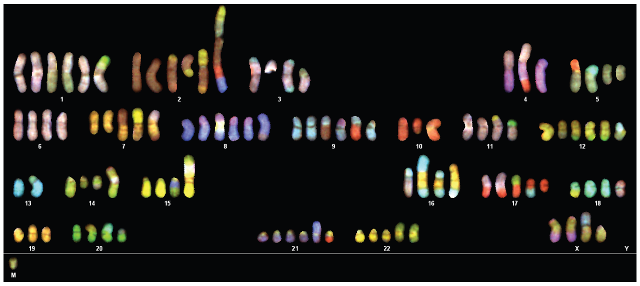
H2073



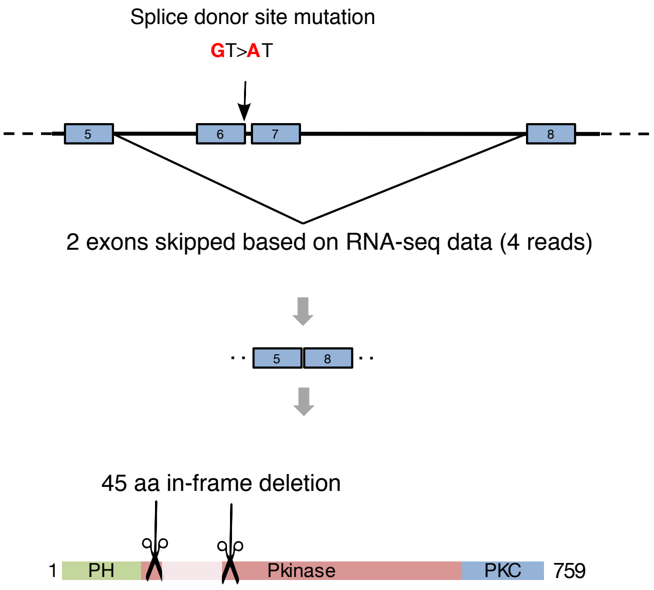
H322T



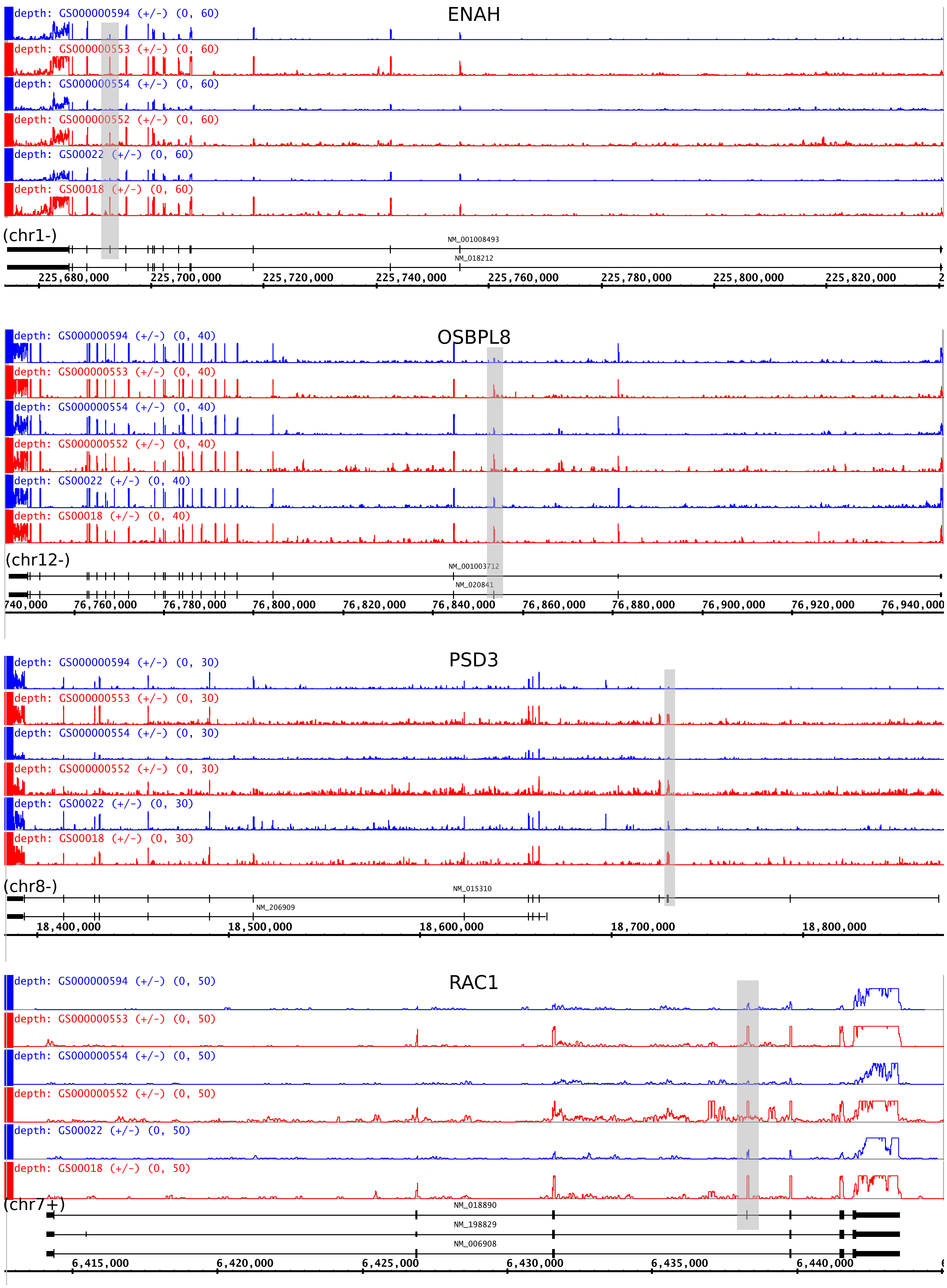
H1299



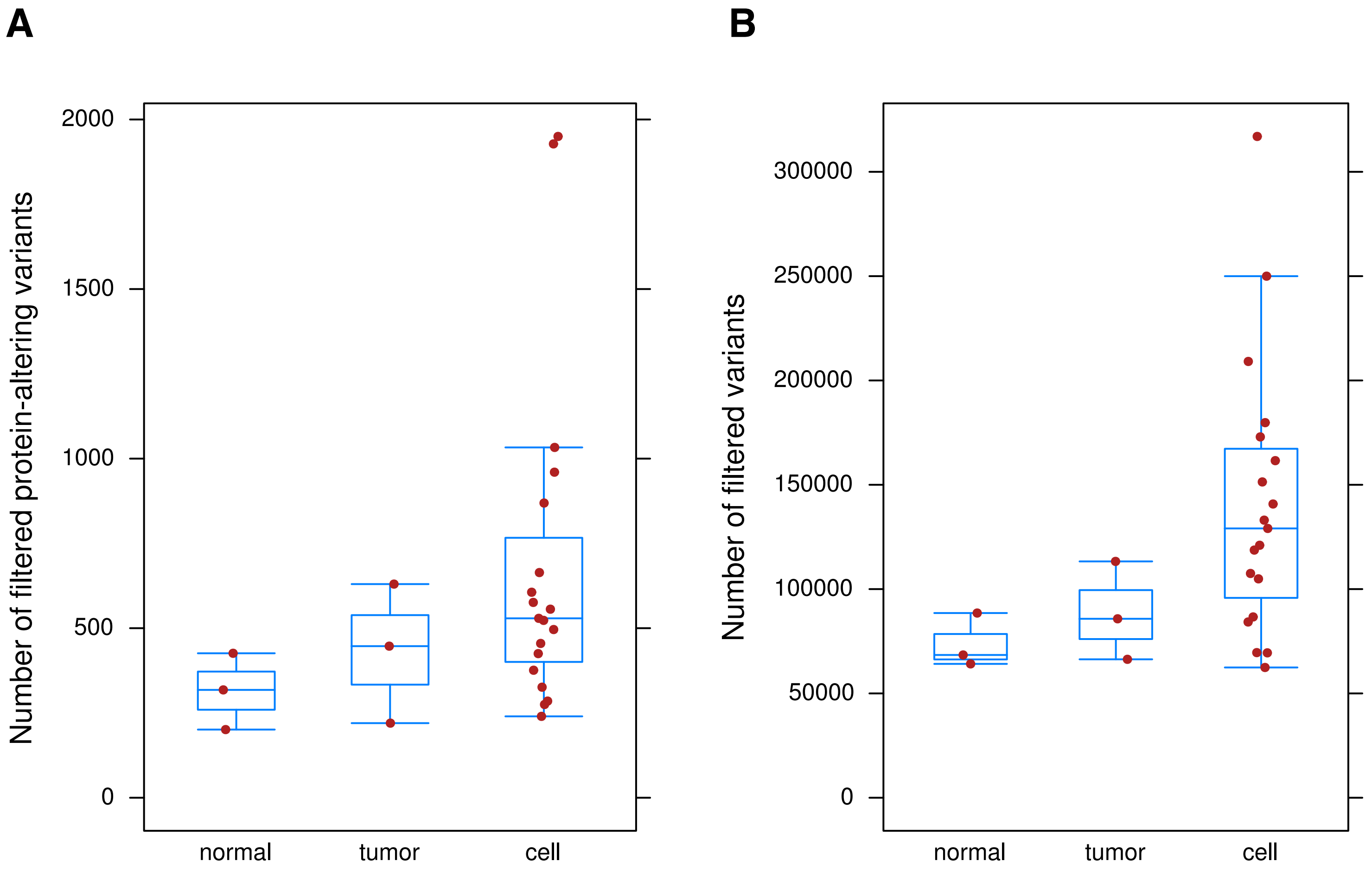
Supplemental Figure 11


Splice donor site mutation
GT>AT
5
6
7
8
2 exons skipped based on RNA-seq data (4 reads)
5
8
45 aa in-frame deletion
1
PH
Pkinase
PKC
759

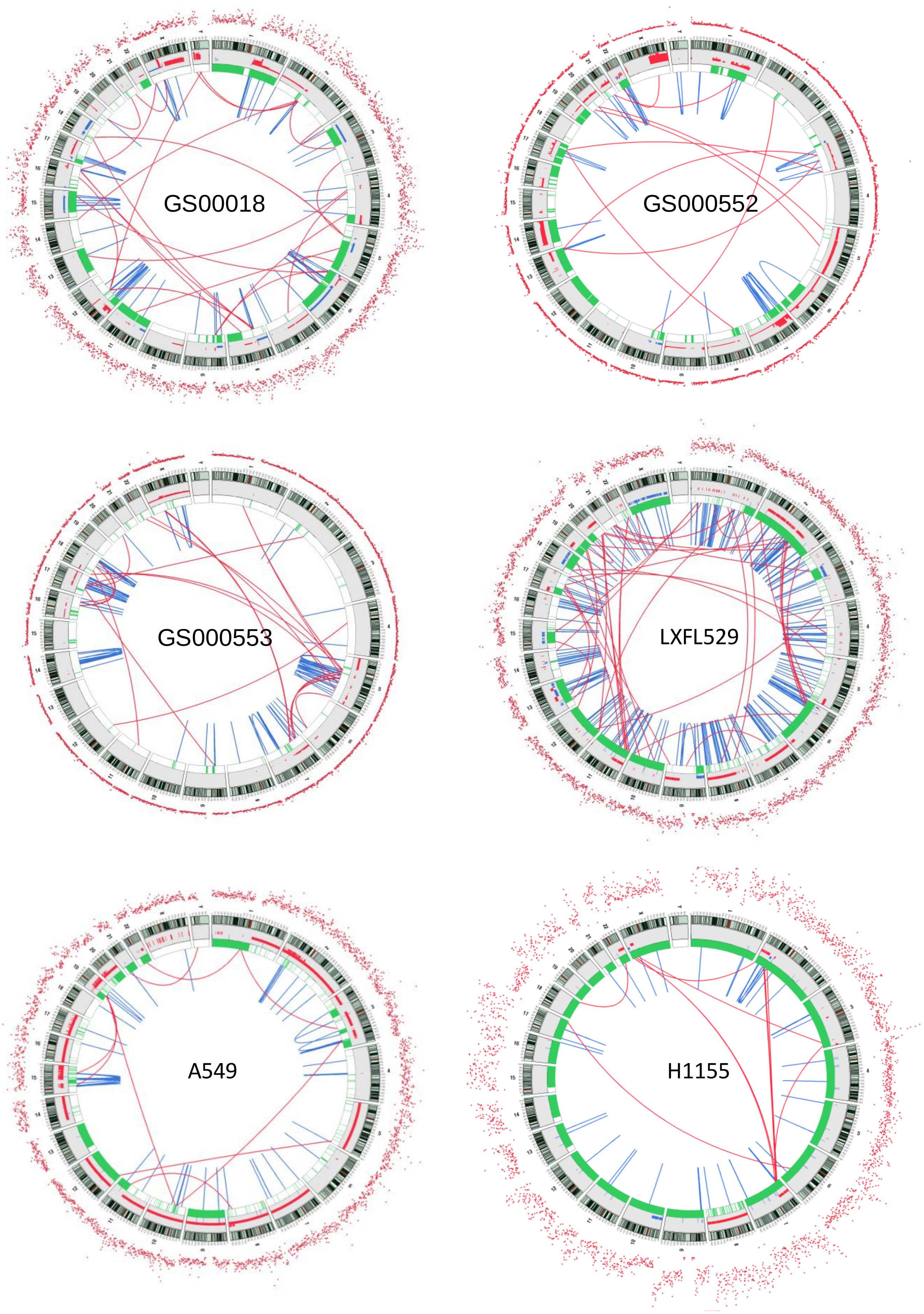
Supplemental Figure 12



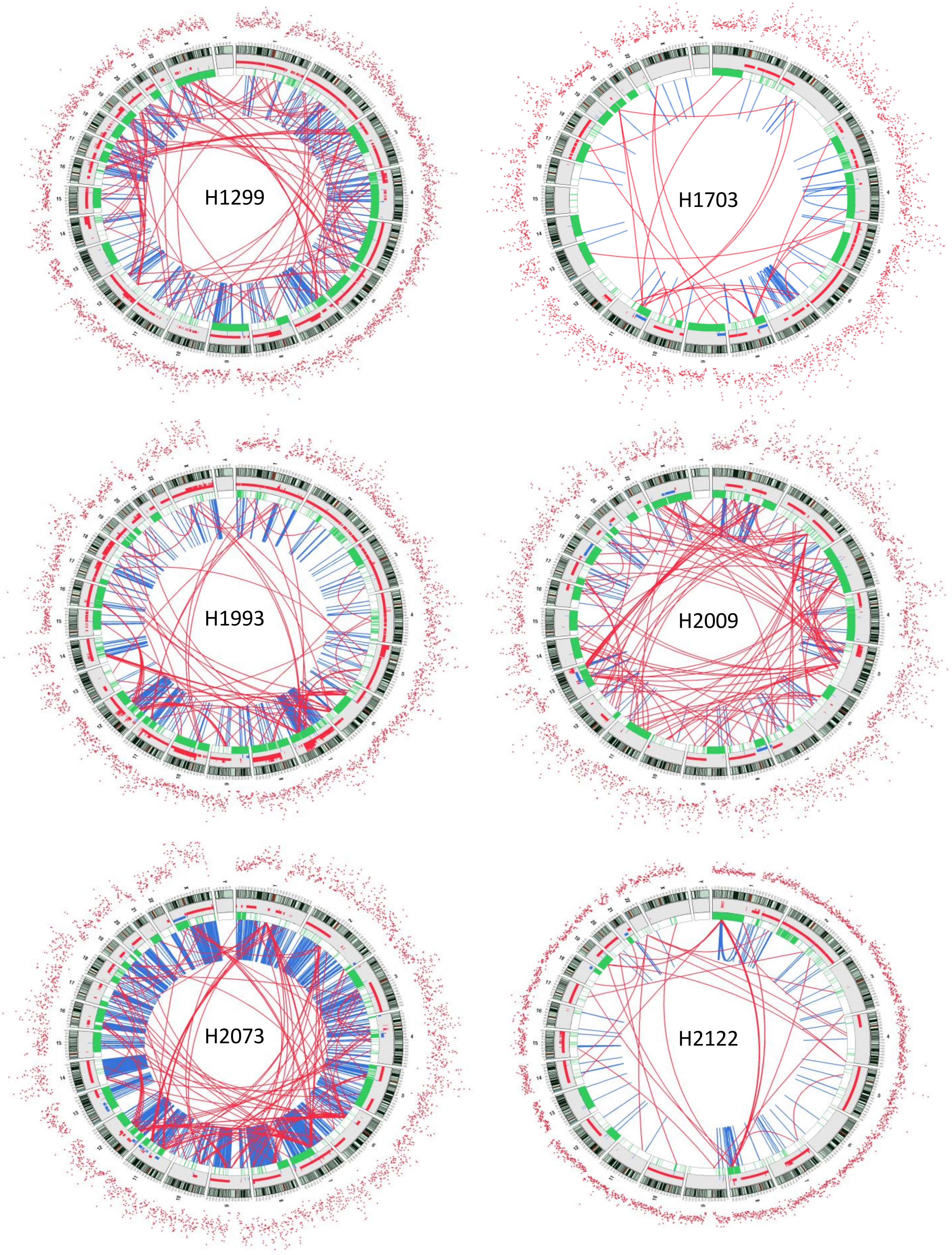
Supplemental Figure 13



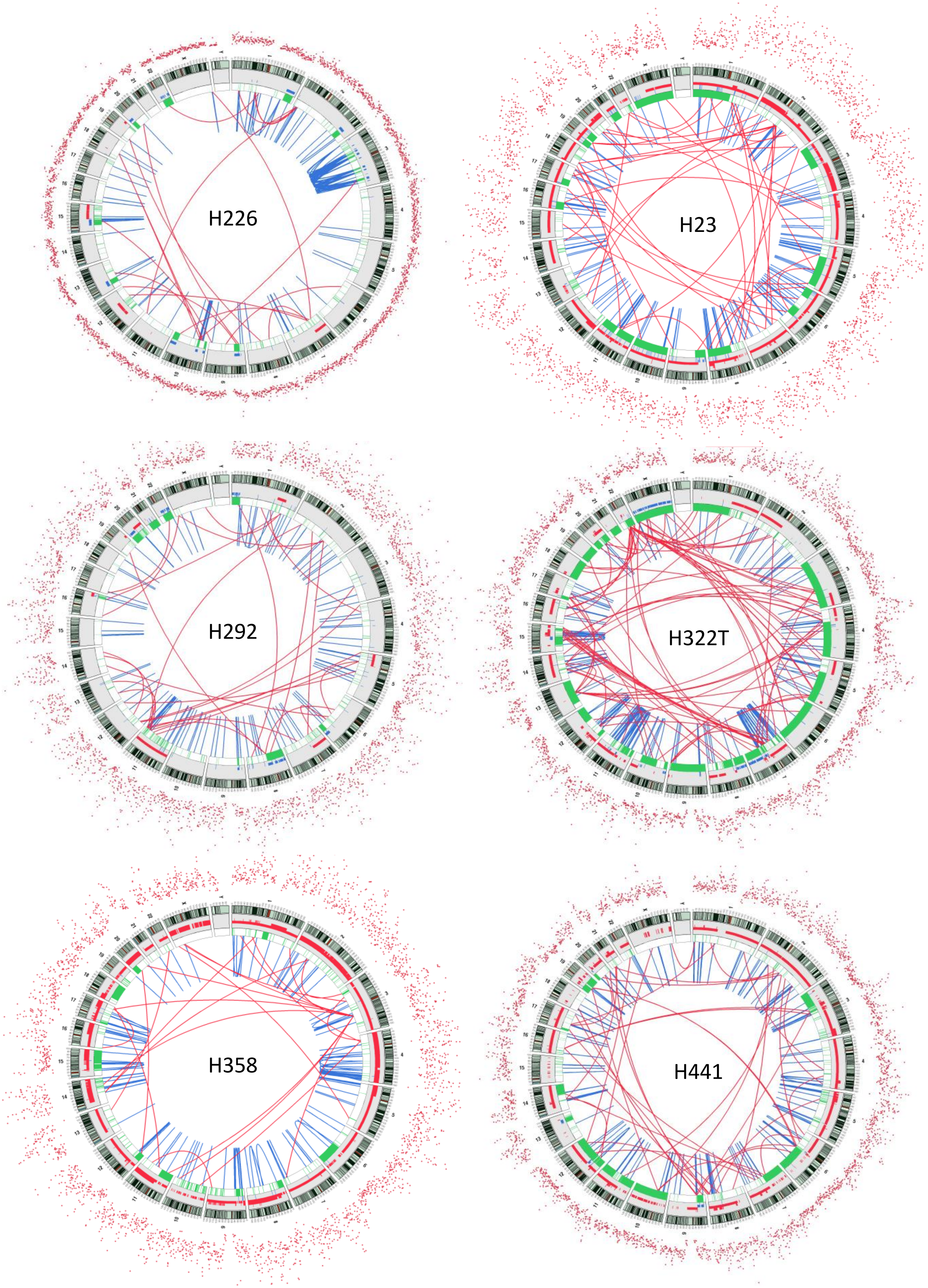
Supplemental Figure 14 A



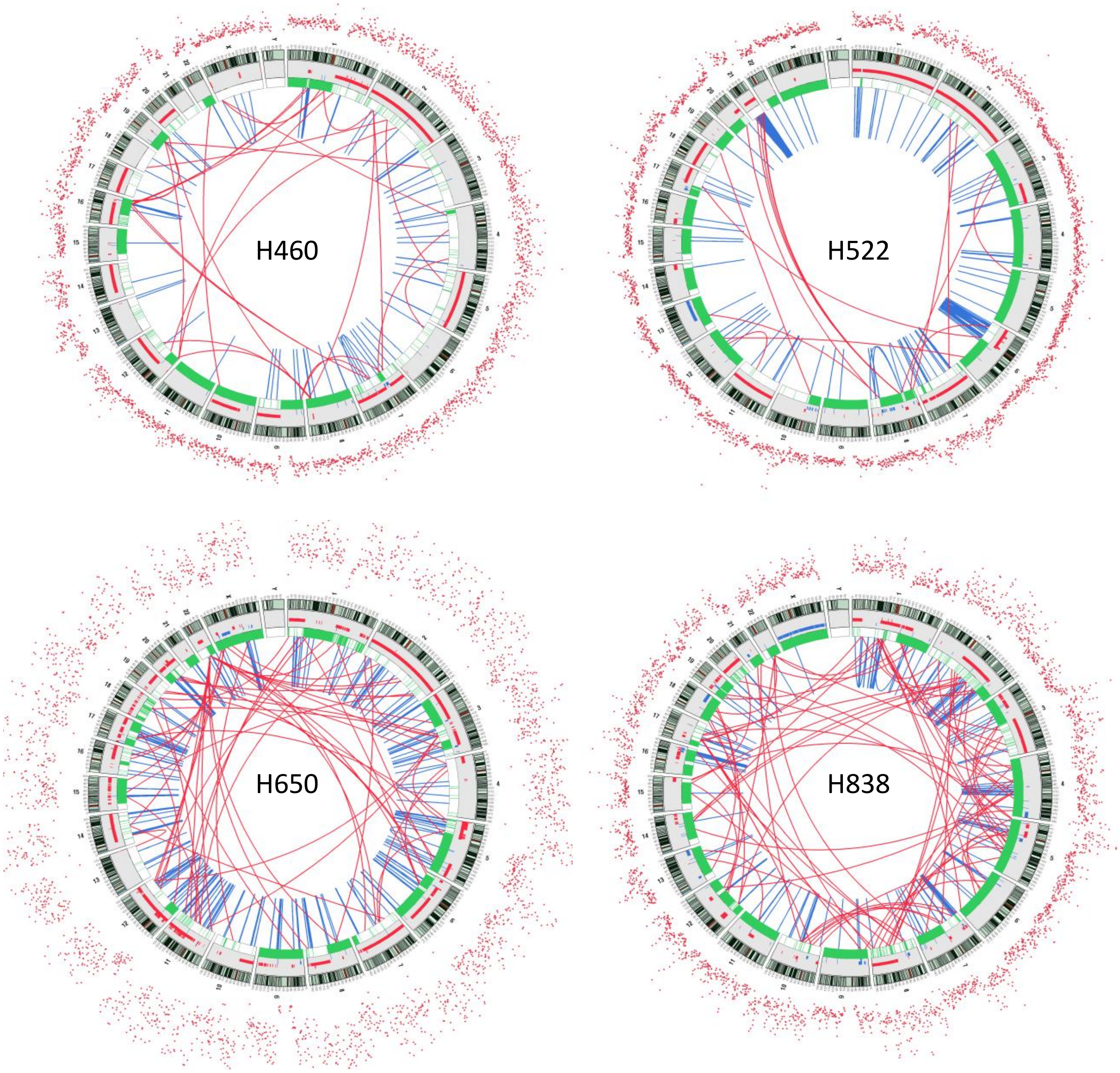
Supplemental Figure 14 B



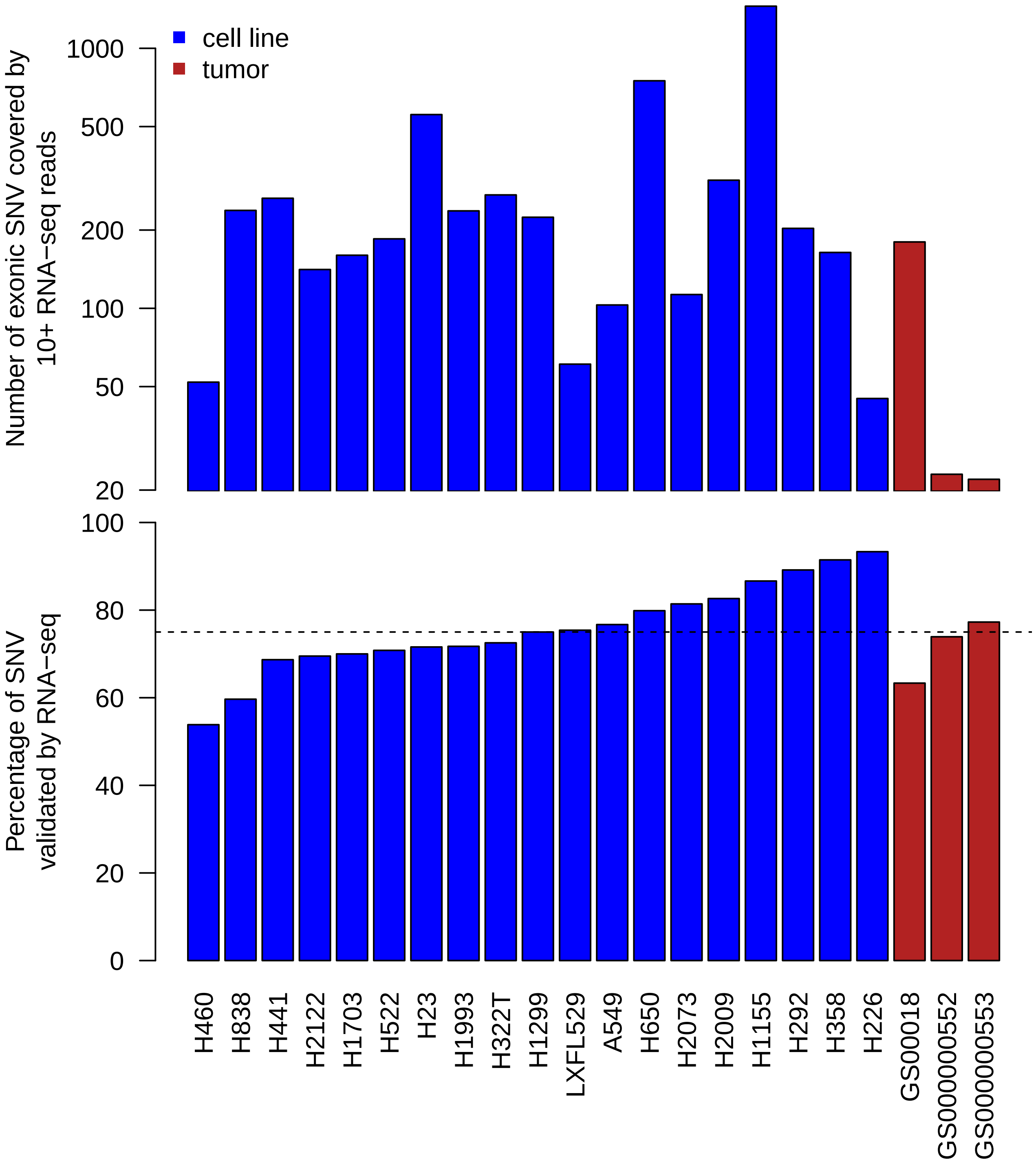
Supplemental Figure 14 C



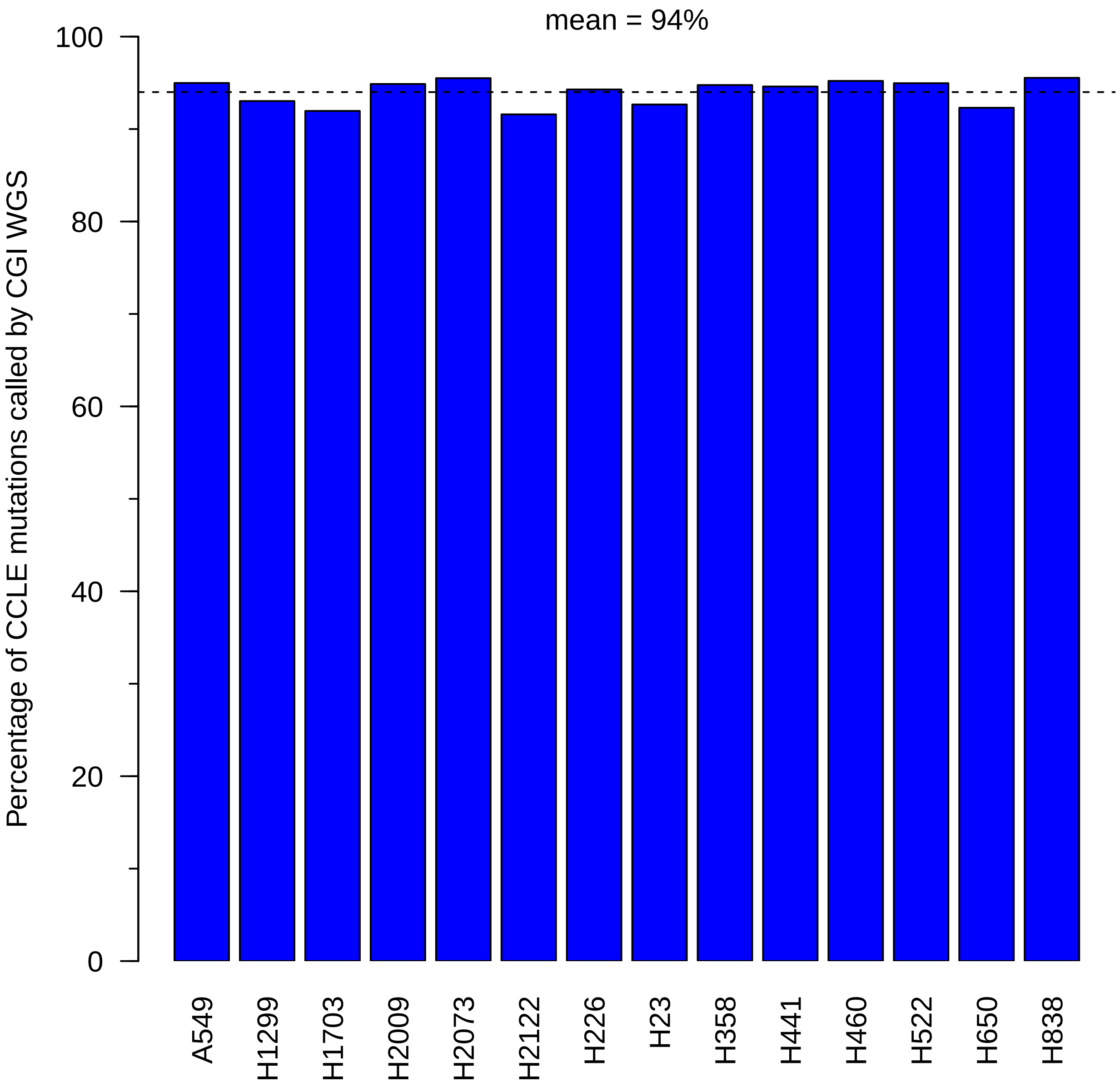
Supplemental Figure 14 D



Supplemental Figure 15



Supplemental Figure 16


mean = 94%
Percentage of CCLE mutations called by CGI WGS
A549
H1299
H1703
H2009
H2073
H2122
H226
H23
H358
H441
H460
H522
H650
H838