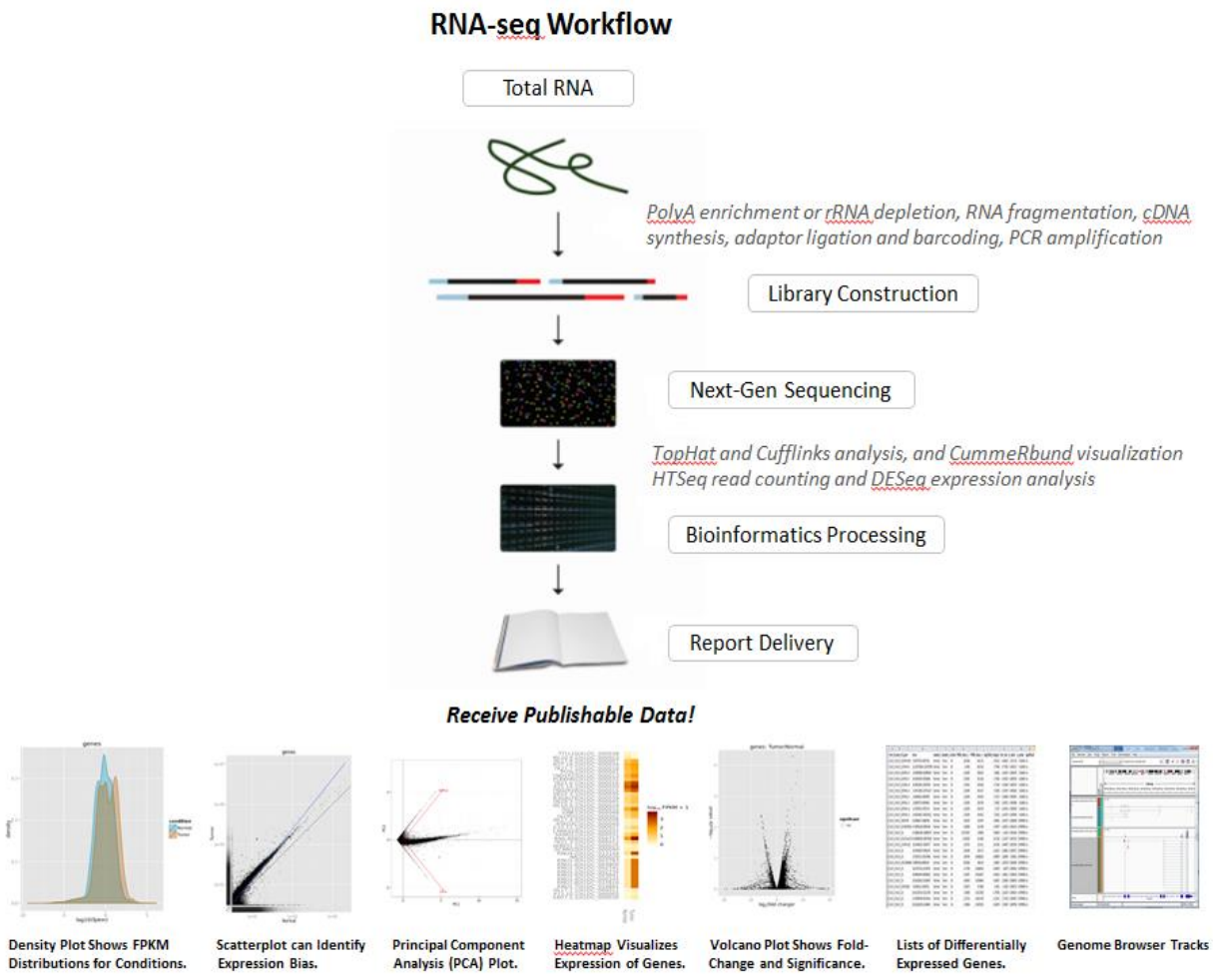
Services Performed

The following checklist confirms the steps of the RNA-Seq Service that were performed on your samples.

SERVICE	
Sample Received	✓
Sample Quality Evaluated	✓
Sample Prepared for Sequencing	✓
Next-Gen Sequencing	✓
Sequence Quality Check	✓
Bioinformatics Processing	✓
Data/Results	✓



Data Report

Sample IDs

1. ZR xxx-1 = Control
2. ZR xxx-2 = Sample

Method

HiSeq 2 x 50bp paired-end reads from RNA-Seq of a control-sample pair samples were analyzed using the TopHat and Cufflinks software. TopHat (v2.0.9) was utilized for alignment of short reads to human genome hg19, Cufflinks (v2.1.1) for isoform assembly and quantification, and commeRbund (v2.0.0) for visualization of differential analysis. Default parameters were used.

Result

The statistics of the sequencing run was shown here:

Index	Description	Control	Project	Yield (Mbases)	% PF	# Reads	% of raw clusters per lane	% Perfect Index Reads	% One Mismatch Reads (Index)	% of >= Q30 Bases (PF)	Mean Quality Score (PF)
GTGAAGC	2068-XS-0001	N	PS9434	2,119	95.21	44,517,622	10.47	100.00	0.00	95.71	37.61
CAGTAGG	2068-XS-0002	N	PS9434	2,548	95.39	53,410,732	12.57	100.00	0.00	95.63	37.57

The statistics of the reference alignment was shown here:

	Read1_Total	Read1_PassQC	Read2_Total	Read2_PassQC	Total_AlignInput	Unmapped	Mapped	%Mapped
Normal	22258811	22245262	22258811	22197926	44443188	3103771	41339417	92.90%
Tumor	26705366	26690162	26705366	26634086	53324248	3466008	49858240	93.30%

Note: The values of Read_Total and Read_PassQC were obtained from the prep_reads.info file outputed by TopHat.
Total_AlignInput = Read1_PassQC + Read2_PassQC
The value of Unmapped was obtained based on the unmapped.bam file outputed by TopHat.
Mapped = Total_AlignInput - Unmapped
%Mapped = Mapped / (Read1_Total + Read2_Total)

1. The TopHat and Cufflinks analysis detected 182 differentially expressed loci with p-value < 0.05 (see DiffExprResults.xlsx).

2. The following are the delivered result files with descriptions at the end of this section.

Gene Expression Files:

DiffExprResults.xlsx (derived from the file Gene_exp.diff)

The heading titles in this Excel file are described below:

Column number	Column name	Example	Description
1	test id	`XLOC_000001`	A unique identifier describing the transcipt, gene, primary transcript, or CDS being tested
2	gene	`Lypla1`	The `gene_name`(s) or `gene_id`(s) being tested
3	locus	`1:4797771-4835363`	Genomic coordinates for easy browsing to the genes or transcripts being tested.
4	sample 1	`Liver`	Label (or number if no labels provided) of the first sample being tested
5	sample 2	`Brain`	Label (or number if no labels provided) of the second sample being tested
6	status	`NOTEST`	Can be one of OK (test successful), NOTEST (not enough alignments for testing), LOWDATA (too complex or shallowly sequenced), HIDATA (too many fragments in locus), or FAIL, when an ill-conditioned covariance matrix or other numerical exception prevents testing.
7	value 1 (FPKM)	`8.01089`	FPKM of the gene in sample 1
8	value 2 (FPKM)	`8.551545`	FPKM of the gene in sample 2
9	log2(foldchange)	`0.06531`	The (base2) log of the fold change of samples 2/1
10	test stat	`0.860902`	The value of the test statistic used to compute significance of the observed change in FPKM
11	p value	`0.389292`	The **uncorrected** *p*-value of the test statistic
12	q value	`0.985216`	The **FDR-adjusted** *p*-value of the test statistic
13	significant	`no`	Can be either "yes" or "no", depending on whether *p* is greater than the FDR **after** Benjamini-Hochberg correction for multiple-testing

Isoform_exp.diff
Gene_exp.diff
Tss_group_exp.diff
Cds_exp.diff
Splicing.diff
Cds.diff
Promoters.diff
Cds.count_tracking
Genes.count_tracking
Isoforms.count_tracking
Tss_groups.count_tracking
Cds.fpkm_tracking
Genes.fpkm_tracking
Isoform.fpkm_tracking
Tss_groups.fpkm_tracking
Cds.read_group_tracking
Genes.read_group_tracking
Isoforms. read_group_tracking
Tss_groups. read_group_tracking

Plot Files:

DensityPlot.png
Boxplot.png
ScatterPlot.png
Dendrogram.png
Dispersion.png
MAplot.png
Top40Genes_Heatmap.png
VolcanoPlot.png
PCAplot.png

Visualization Files:

Accepted_hits_Normal.bam
Accepted_hits_Normal.bam.bai
Accepted_hits_Tumor.bam
Accepted_hits_Tumor.bam.bai
Junctions_Normal.bed
Junctions_Tumor.bed
Transcripts_Normal.gtf

Transcripts_Tumor.gtf

Description of the Result Files

isoform_exp.diff: Transcript differential FPKM.

gene_exp.diff: Gene differential FPKM. Tests differences in the summed FPKM of transcripts sharing each gene_id

tss_group_exp.diff: Primary transcript differential FPKM. Tests differences in the summed FPKM of transcripts sharing each tss_id

cds_exp.diff: Coding sequence differential FPKM. Tests differences in the summed FPKM of transcripts sharing each p_id independent of tss_id

splicing.diff: This tab delimited file lists, for each primary transcript, the amount of overloading detected among its isoforms, i.e. how much differential splicing exists between isoforms processed from a single primary transcript. Only primary transcripts from which two or more isoforms are spliced are listed in this file.
cds.diff: This tab delimited file lists, for each gene, the amount of overloading detected among its coding sequences, i.e. how much differential CDS output exists between samples. Only genes producing two or more distinct CDS (i.e. multi-protein genes) are listed here.

promoters.diff: This tab delimited file lists, for each gene, the amount of overloading detected among its primary transcripts, i.e. how much differential promoter use exists between samples. Only genes producing two or more distinct primary transcripts (i.e. multi-promoter genes) are listed here.

FPKM Tracking Files: record estimated expression values for an object with a unique tracking_id, including genes.fpkm_tracking, isoforms.fpkm_tracking, cds.fpkm_tracking, and tss_groups.fpkm_tracking.

Count Tracking Files: record estimated fragment count values for an object with a unique tracking_id, including genes.count_tracking, isoforms.count_tracking, cds.count_tracking, and tss_groups.count_tracking.

Read Group Tracking Files: record per-replicate expression and count data for an object with a unique tracking_id, including genes.read_group_tracking,

isoforms.read_group_tracking, cds.read_group_tracking, and tss_groups.read_group_tracking.

DensityPlot.png: shows density of FPKM distributions for individual conditions.

Boxplot.png: exposes FPKM distributions for individual conditions.

ScatterPlot.png: Pairwise scatterplots can identify biases in gene expression between two particular conditions.

Dendrogram.png: Dendrograms with replicates=TRUE can identify outlier replicates.

Dispersion.png: plots count vs dispersion by condition for all genes.

MAplot.png: MA plots can identify systematic biases across ranges of FPKM intensity and fold-change.

Top40Genes_Heatmap.png: Heatmaps provide a convenient way to visualize the expression of entire gene sets at once. Shown are the top 40 differentially expressed genes.

VolcanoPlot.png: Volcano plots explore the relationship between fold-change and significance.

PCAplot.png: PCA plot for gene-level features. Dimensionality reduction such as principal component analysis (PCA) is an informative approach for clustering and exploring the relationships between conditions. It can be useful for feature selection as well as identifying the sources of variability within your data.

Accepted_hits.bam: A list of read alignments in SAM format. SAM is a compact short read alignment format that is increasingly being adopted. BAM format is the binary version of SAM.

Accepted_hits.bam.bai: An index file of accepted_hits.bam, which is also required when viewing alignment in a genomic viewer such as IGV.

Junctions.bed: A UCSC BED track of junctions reported by TopHat. Each junction consists of two connected BED blocks, where each block is as long as the maximal overhang of any read spanning the junction. The score is the number of alignments spanning the junction.

Insertions.bed and deletions.bed: UCSC BED tracks of insertions and deletions reported by TopHat.

Insertions.bed - chromLeft refers to the last genomic base before the insertion.
Deletions.bed - chromLeft refers to the first genomic base of the deletion.

Transcripts.gtf: This GTF file contains Cufflinks' assembled isoforms. The first 7 columns are standard GTF, and the last column contains attributes, some of which are also standardized ("gene_id", and "transcript_id"). There one GTF record per row, and each record represents either a transcript or an exon within a transcript. The columns are defined as follows: (from http://cufflinks.cbcb.umd.edu/manual.html)

Column number	Column name	Example	Description
1	seqname	chrX	Chromosome or contig name
2	source	Cufflinks	The name of the program that generated this file (always 'Cufflinks')
3	feature	exon	The type of record (always either "transcript" or "exon".
4	start	77696957	The leftmost coordinate of this record (where 1 is the leftmost possible coordinate)
5	end	77712009	The rightmost coordinate of this record, inclusive.
6	score	77712009	The most abundant isoform for each gene is assigned a score of 1000. Minor isoforms are scored by the ratio (minor FPKM/major FPKM)
7	strand	+	Cufflinks' guess for which strand the isoform came from. Always one of "+", "-", "."
7	frame	.	Cufflinks does not predict where the start and stop codons (if any) are located within each transcript, so this field is not used.
8	attributes	...	See below.

Each GTF record is decorated with the following attributes:

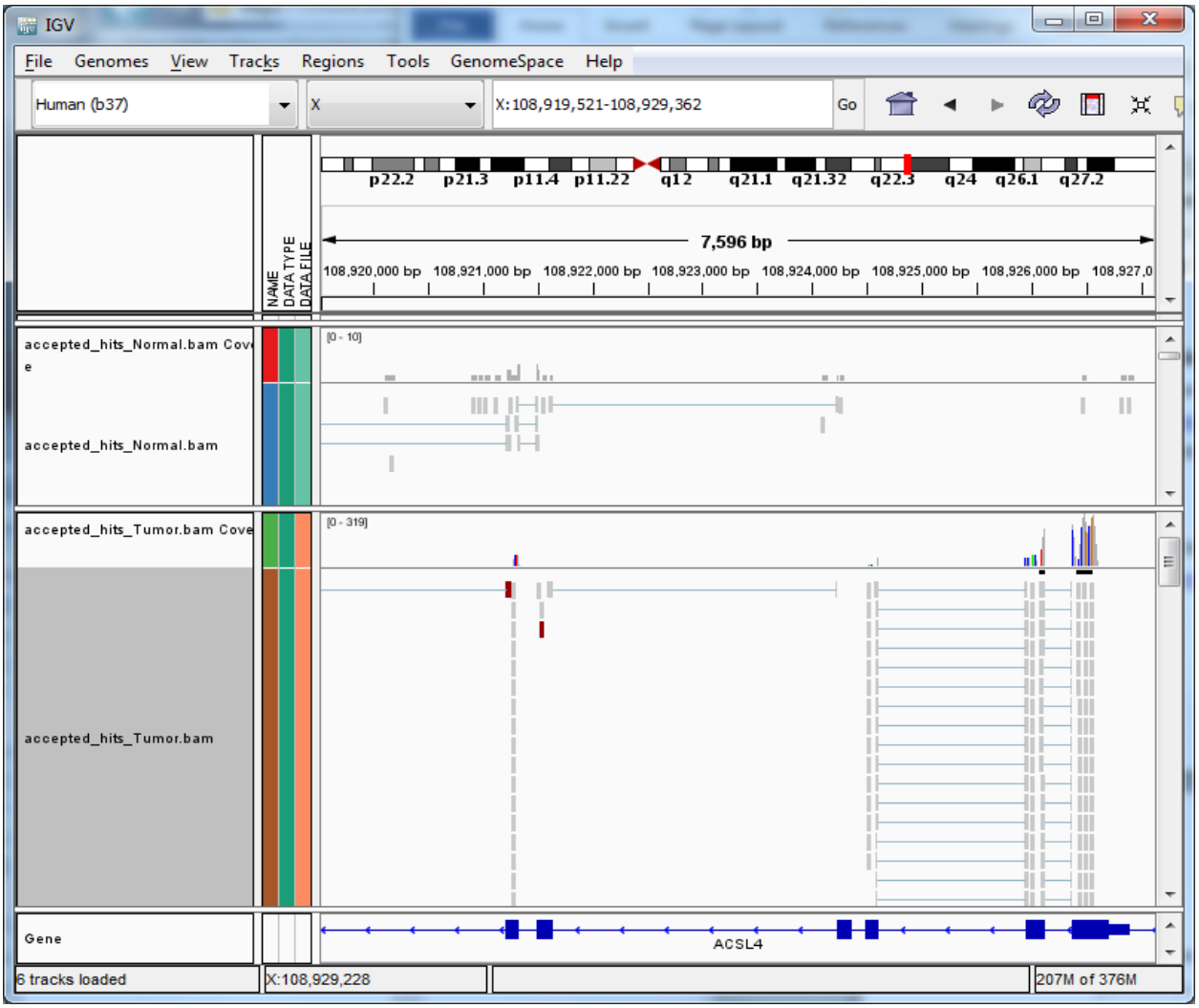
Attribute	Example	Description
gene_id	CUFF.1	Cufflinks gene id

transcript_id	CUFF.1.1	Cufflinks transcript id
FPKM	101.267	Isoform-level relative abundance in **F**ragments **P**er **K**ilobase of exon model per **M**illion mapped fragments
frac	0.7647	Reserved. Please ignore, as this attribute may be deprecated in the future
conf_lo	0.07	Lower bound of the 95% confidence interval of the abundance of this isoform, as a fraction of the isoform abundance. That is, lower bound = FPKM * (1.0 - conf_lo)
conf_hi	0.1102	Upper bound of the 95% confidence interval of the abundance of this isoform, as a fraction of the isoform abundance. That is, upper bound = FPKM * (1.0 + conf_lo)
cov	100.765	Estimate for the absolute depth of read coverage across the whole transcript
full_read_support	yes	When RABT assembly is used, this attribute reports whether or not all introns and internal exons were fully covered by reads from the data.

Visualization of Data Using the IGV Viewer

Instruction of free IGV download and installation can be found at https://www.broadinstitute.org/software/igv/download and the IGV User Guide can be found at https://www.broadinstitute.org/software/igv/UserGuide .

1. Visualization of alignment by opening the file accepted_hits.bam and its index file accepted_hits.bam.bai in the IGV. First, load the human GRCh37 genome if not done yet. Double click your IGV icon to open the viewer, click the Genomes pull-down menu, select Load Genome From Server, and select Human (b37). Then, click the File pull-down menu and select Load from File, navigate to the Visualization folder and select the .bam file of interest. Once the .bam and .bai files are automatically loaded, alignment of a particular gene can be viewed simply by copying and pasting the locus coordinates directly from the Microsoft Excel file DiffExprResults.xlsx into the box near the top (or type the gene name) and clicking **Go**. Use the zoom bar at the upper right corner to zoom in to see alignment.



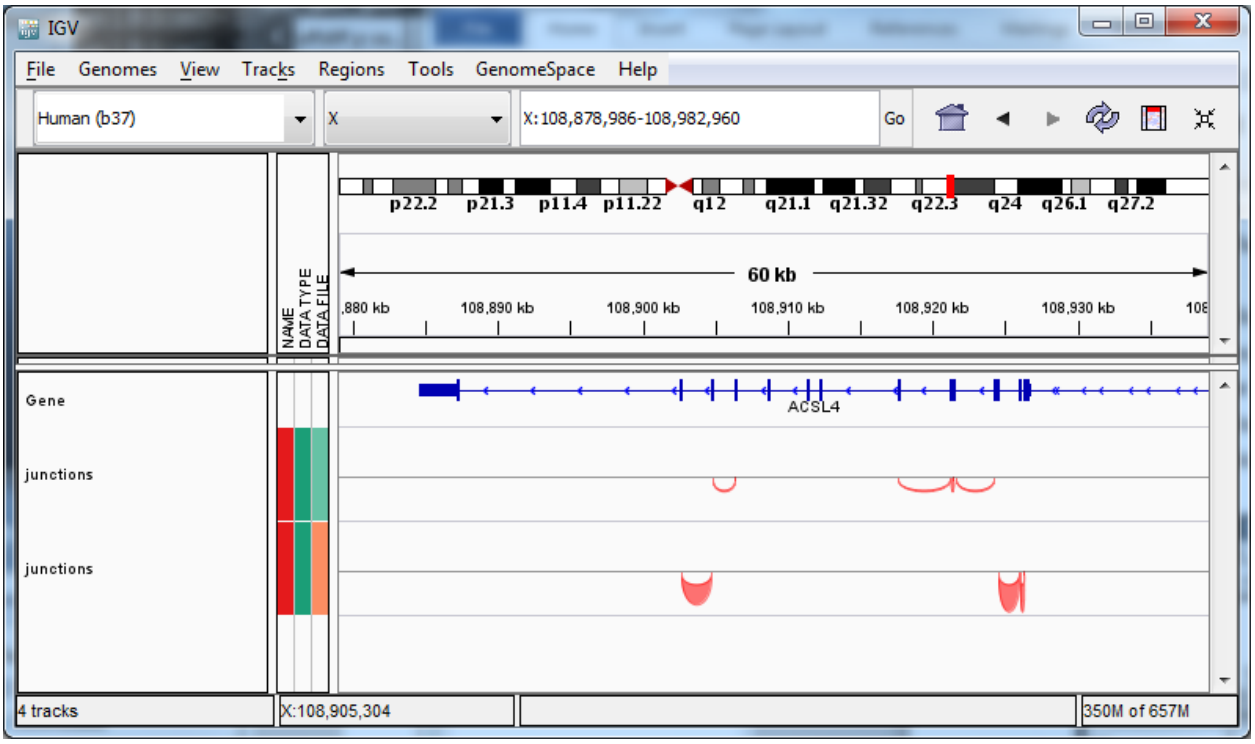
2. Viewing splice junctions: please also see description at https://www.broadinstitute.org/software/igv/splice_junctions

The splice junction view displays an alternative representation of .bed files encoding splice junctions, such as the "junctions.bed" file produced by the TopHat program. This view is enabled by including a track line that specifies either name=junctions or graphType=junctions. TopHat's "junctions.bed" file includes a track line specifying name=junctions by default, so no action is required for these files. Junction files should be in the standard .bed format. The 'score' field is used to indicate depth of coverage.

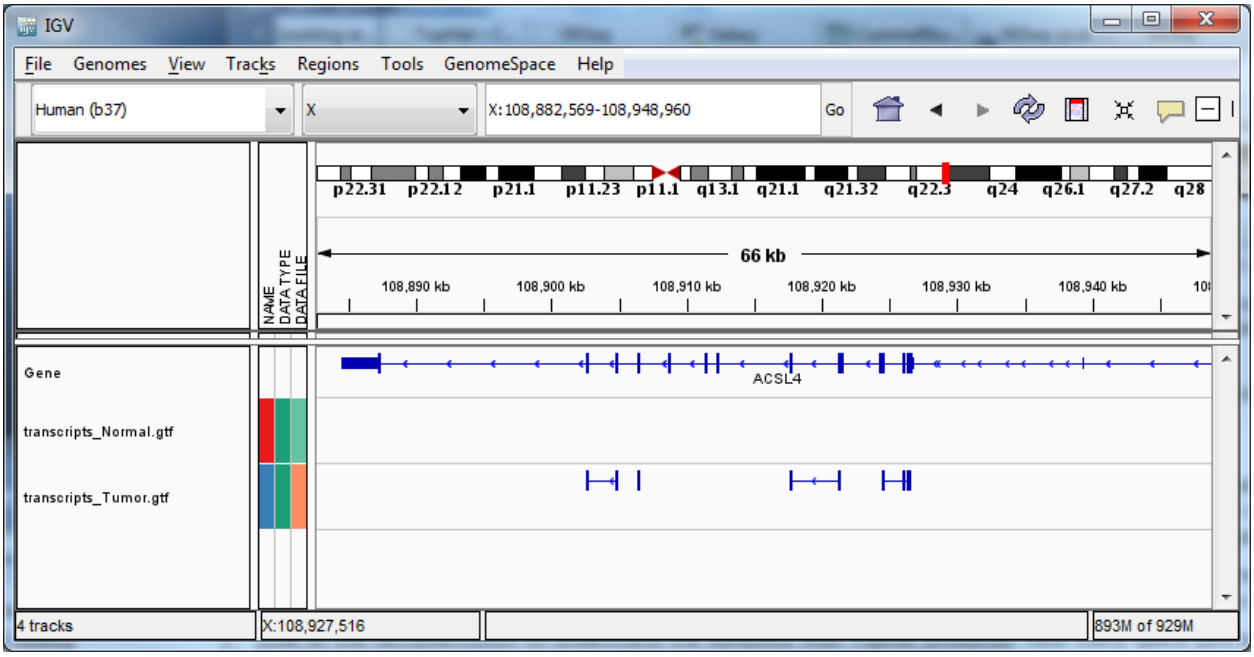
Each splice junction is represented by an arc from the beginning to the end of the junction. Junctions from the '+' strand are colored red and extend upward from the center line. Junctions from the '-' strand are blue and extend downward. The height of the arc, and its

thickness, are proportional to the depth of read coverage. All junctions with more than 50 reads have the same thickness. Hovering the mouse over a junction will display the coverage.

Visualization of splice junctions by opening junctions.bed files in IGV:



3. Visualization of assembled transcripts by opening transcripts.gtf files in IGV:



ZYMO RESEARCH CORP.

17062 Murphy Ave. ▪ Irvine, CA 92614

Phone: 1-888-882-9682 ▪ 1-949-679-1190 ▪ Fax: 1-949-266-9452

info@zymoresearch.com ▪ www.zymoresearch.com