Introduction to RNASeq Data Analysis (Part 2)

Peter FitzGerald, PhD

Head Genome Analysis Unit

Director of BTEP

CCR, NCI

RNA-SEQ WEEKS

- Registration, Introduction to RNA-Seq Analysis, Part 1
  Tuesday, Nov 5 @ 1 PM
- Registration, Introduction to RNA-Seq Analysis, Part 2
  NOW
- Partek Flow*-
  Thursday, Nov 12 @ 1 PM
- Qiagen Ingenuity Pathway Analysis (IPA)
  Thursday Nov 19 @ 1 PM
- RNA-Seq Analysis on the DNAnexus platform*
  Thursday, Dec 3 @ 1 PM
- CCBR-Pipeliner for analysis of RNA-Seq data
  Thursday, Dec 10 @ 1 PM
- Bulk RNA-Seq Analysis on the NIDAP platform*
  Thursday, Dec 17 @ 1 PM

* Web-based tools

BTEP - Calendar **https://btep.ccr.cancer.gov**

RNA-SEQ WEEKS **https://btep.ccr.cancer.gov/rna-seq-weeks-coming-in-october/**

RNA-SEQ WEEKS

- ~~Experimental Design~~
- ~~Sample Preparation~~
- ~~Sequencing~~
- **Data Analysis (Computation)**
- **Quantitation** - how to get an expression value for each gene
- **Differential Expression** - relative expression of each gene under different conditions
- **Visualization -** visual examination and confirmation of results
- **Next Step**s - tertiary analysis - deriving biological meaning
- **File Formats -** pointers to info on different file formats
- **Utilities -** short list of essential programs

Quantitation

Counting as a measure of Expression

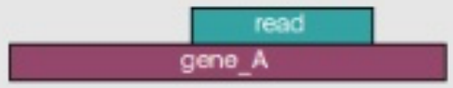
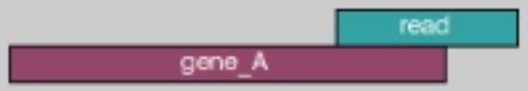
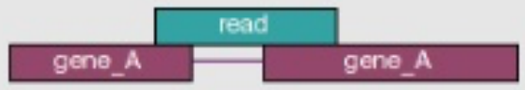
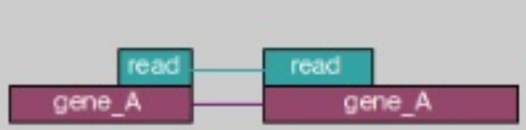
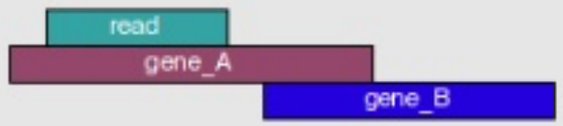
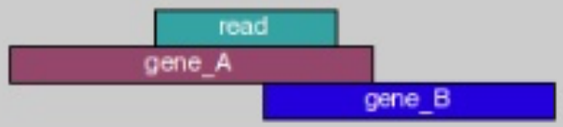
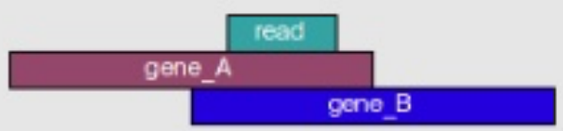
- Most RNASEQ techniques deal with count data.
  The reads are mapped to a reference and the number of reads mapped to each gene/transcript is counted
- Read counts are roughly proportional to gene-length and abundance
- The more reads the better

  - Artifacts occur because of:
    - Sequencing Bias
    - Positional bias along the length of the gene
    - Gene annotations (overlapping genes)
    - Alternate splicing
    - Non-unique genes
    - Mapping errors

Counting as a measure of Expression

- Count mapped **reads**
- Count each read once (deduplicate)
- Discard reads that:
  - have poor quality alignment scores
  - are not uniquely mapped
  - overlap several genes
  - Have paired reads do not map together
- Document what was done

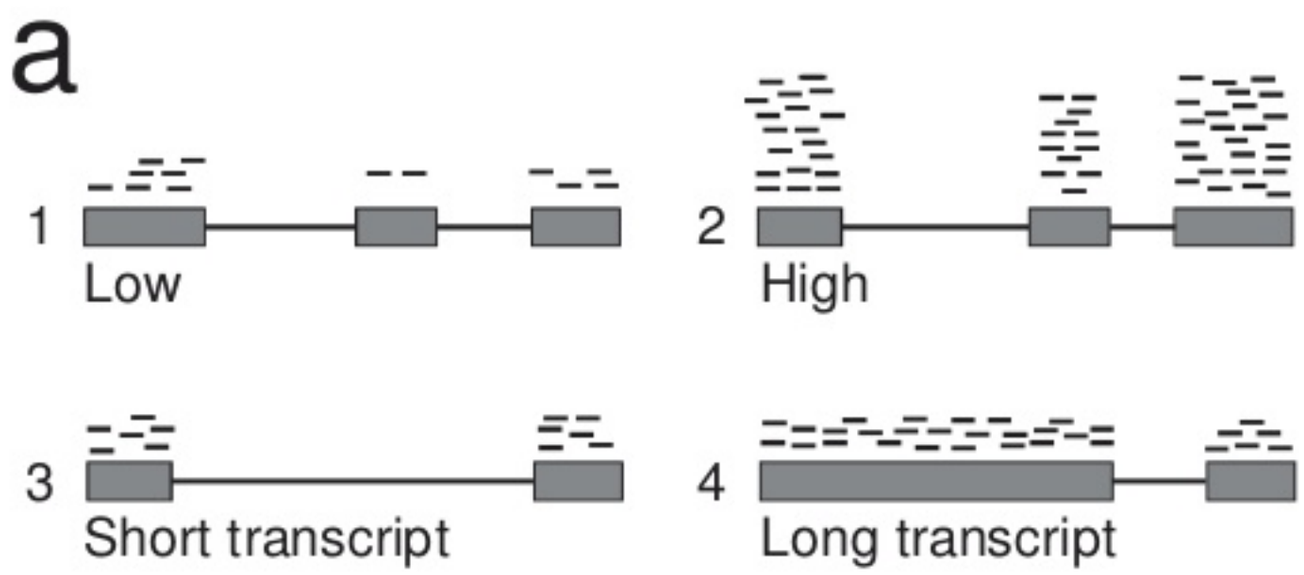
Read Counting

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Count Normalization

- Number of reads aligned to a gene gives a measure of its level of expression

- Normalization of the count data
  - Sequencing depth
  - Length bias



Nature Methods 8, 469–477 (2011), Doi:10.1038/nmeth.1613

Normalization

There are three metrics commonly used to attempt to normalize for sequencing depth and gene length.

- **RPKM = Reads Per Kilobase Million**

  Total Reads/1,000,000 = PM
  Gene read-count/PM = RPKM
  RPM/gene-length (kb) = RPKM

- **FPKM = Fragments Per Kilobase Million**

FPKM is very similar to RPKM. RPKM was made for single-end RNASEQ, where every read corresponded to a single fragment that was sequenced. FPKM was made for paired-end RNA-seq.

- **TPM = Transcripts Per Million (*Sum of all TPM in samples is the same*)**

TPM is very similar to RPKM and FPKM. The only difference is the order of operations

Gene read-count/gene-length (kb) = RPK
(Sum all RPKs)/1,000,000 = PM
Gene RPK/PM = TPM

https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

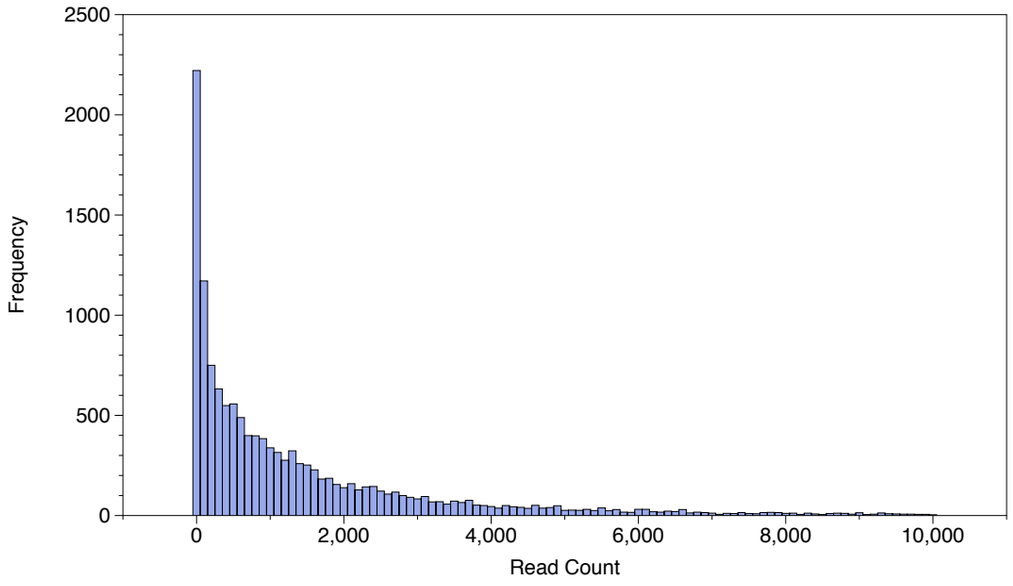
Counting as a measure of Expression

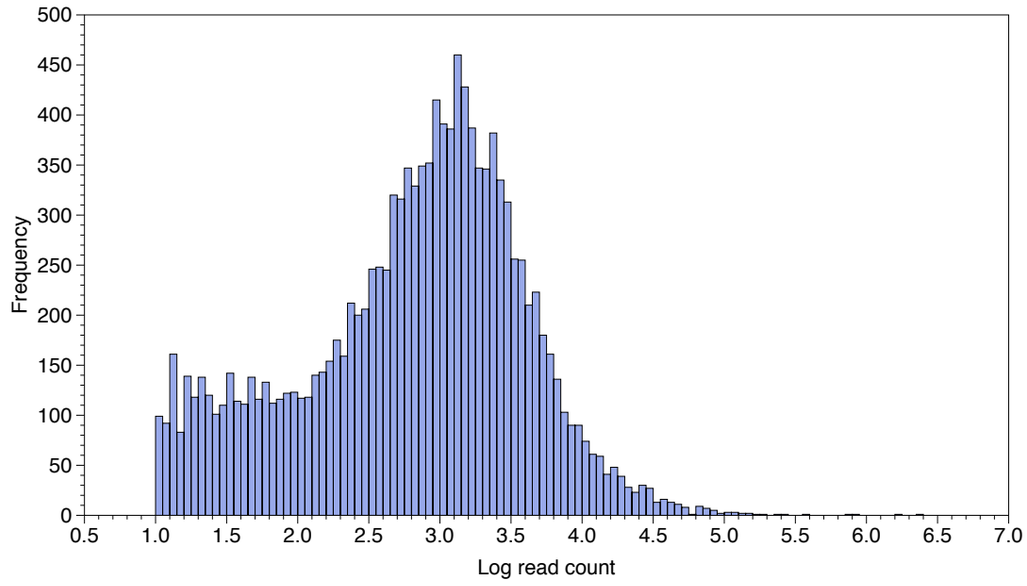
Name	Length	EffectiveLength	TPM	NumReads
ENSG00000121410.12_4	509.732	325.991	3.22494	322.674
ENSG00000268895.6_6	1823.71	1633.86	0.9255	464.119
ENSG00000148584.15_4	5354.1	5164.27	0	0
ENSG00000175899.14_4	4544.77	4354.95	0.039651	53
A2M-AS1	2592.39	2402.54	0.008136	5.999
A2ML1	1749	1561.55	0	0
SLC7A2	452	269.66	0	0
ENSG00000001461.12_NIPAL3	386	208.766	0	0
ENSG00000001497.12_LAS1	1715	1526.05	0	0
ENSG00000001617.7_SEMA3F	1023	833.15	0	0
ENSG00000003096.9_KLHL13	1457.48	1269.51	3.23046	1258.74

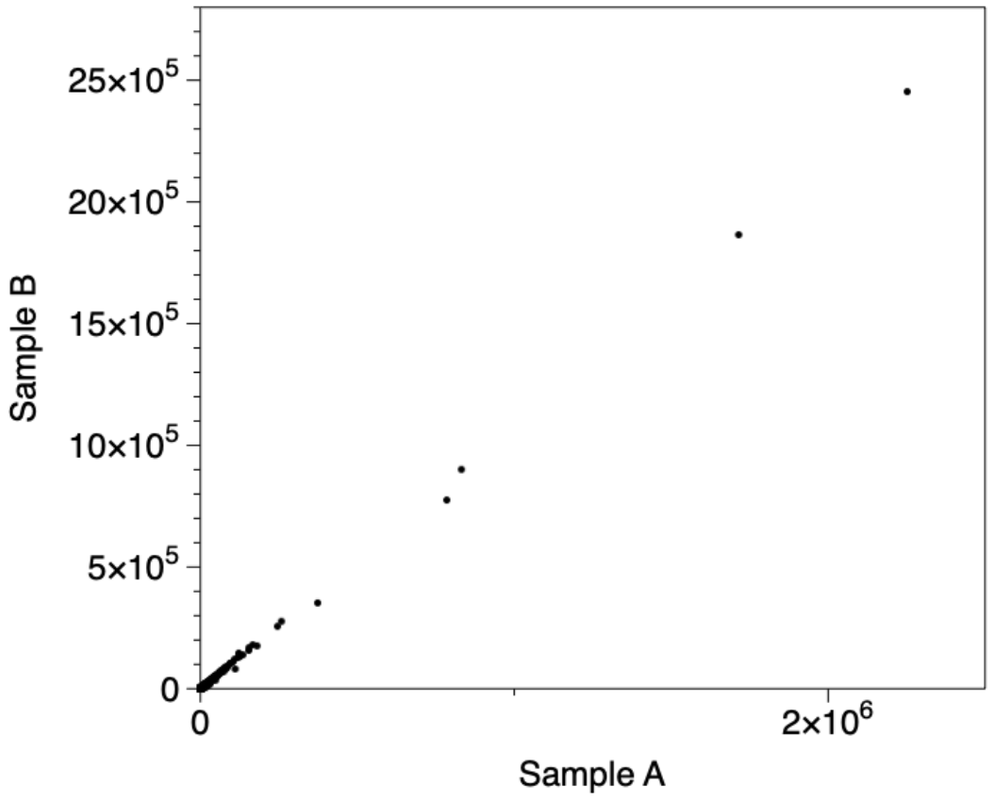
Different ways of annotating the genes

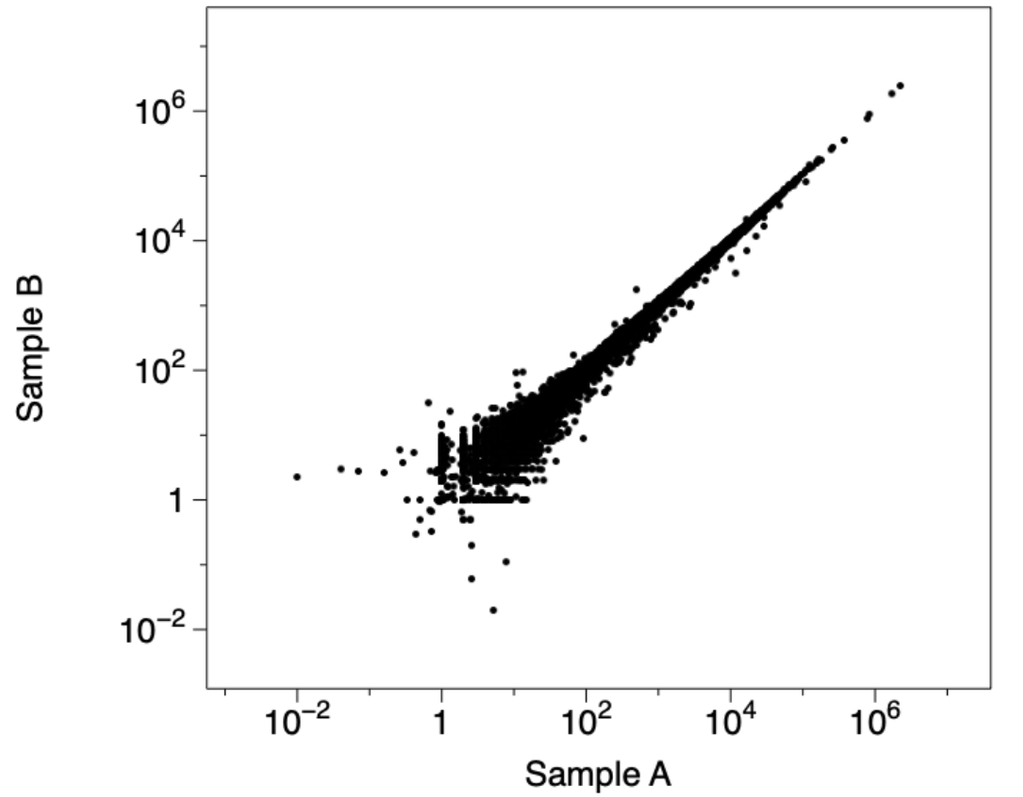
Not always integers -
may not be acceptable to some programs

Log Transformed Data









Counting as a measure of Expression

- Subread (featureCount)
- STAR (quantmode)
- HTseq (counts)
- **RSEM** (RNA-Seq by Expectation Maximization)
- **Salmon**, Kallisto - pseudoaligners

Spike in Controls

The goal of the **spike-in control** is to determine how well we can measure and reproduce data with known (expected) properties. ERCC ExFold Spike-In Mixes are commercially available, pre-formulated blends of 92 transcripts, derived and traceable from NIST-certified DNA plasmids. The transcripts are designed to be 250 to 2,000 nt in length, which mimic natural eukaryotic mRNAs.

Differential Expression

Differential Expression

Differential expression involves the comparison of **normalized** expression counts of different samples and the application of **statistical measures** to identify quantitative changes in gene expression between the different samples.

The two most important step are:

- The normalization of the data to ensure all samples are comparable (variable gene length, read depth)
- The statistical test that determines whether an observed difference is statistically significant (i.e. the likelyhood of the observation is greater than that expected from random biological variability).

Count Matrix

Data_matrix

Data_matrix	p53_rock_1	p53_rock_2	p53_rock_3	p53_rock_4	p53_IR_1	p53_IR_2	p53_IR_3	p53_IR_4	null_rock_1	null_rock_2	null_IR_1	null_IR_2
C330021F23RIK	83	67	52	117	52	43	38	38	96	71	54	71
CPS1	0	0	0	0	4	8	0	0	0	0	0	1
FAM171B	11	11	6	11	13	10	4	8	14	6	10	10
OLFR910	0	1	0	0	0	1	0	0	0	0	0	0
DYNLL2	462	413	294	529	330	206	317	293	312	275	409	663
NPEPL1	2361	1794	1563	1612	2296	1565	2969	3758	1904	1657	3200	3516
TRAJ2	4	6	6	4	9	13	5	4	7	4	5	2
SLC2A4	9	11	3	3	15	10	13	21	2	7	0	0
ZFP655	2874	2474	2006	2517	1640	1276	1881	1948	2666	2412	3157	3315
SLC8A1	1074	839	941	921	657	340	469	320	852	770	337	803
CYB5R4	7431	6425	4866	6215	4502	3800	4170	4656	6602	5619	6059	6843
GM31123	0	0	0	0	0	0	0	0	0	0	0	0
CTDNEP1	1210	1105	869	1323	833	493	951	1094	1063	999	2069	2039
ETS1	44445	38606	27356	39522	10423	7905	8481	10543	42254	41214	20881	27334

Contrast File

Study_design

Study_Design	p53_rock_1	p53_rock_2	p53_rock_3	p53_rock_4	p53_IR_1	p53_IR_2	p53_IR_3	p53_IR_4	null_rock_1	null_rock_2	null_IR_1	null_IR_2
p53	wt	wt	wt	wt	wt	wt	wt	wt	null	null	null	null
Treatment	rock	rock	rock	rock	IR	IR	IR	IR	rock	rock	IR	IR

Study_design-1

Study_Design	p53	Treatment
p53_rock_1	wt	rock
p53_rock_2	wt	rock
p53_rock_3	wt	rock
p53_rock_4	wt	rock
p53_IR_1	wt	IR
p53_IR_2	wt	IR
p53_IR_3	wt	IR
p53_IR_4	wt	IR
null_rock_1	null	rock
null_rock_2	null	rock
null_IR_1	null	IR
null_IR_2	null	IR

Different programs require this file to be organized in different ways

Differential Expression

Biological replicates are essential to derive a meaningful result. Don't mistake the high precision of the technique for the need for biological replicates.

Final output id typically a rank order list of differentially expressed (DE) genes with expression values and associated p-values.

If technical or biological variability exceeds that of the experimental perturbation you will get zero DEs.

Remember not all DE may be directly due to the experimental perturbation, but could be do to cascading effects of other genes.

Differential Expression

Two Statistical Components:(*All statistical methods rely on various assumptions regarding the characteristics of the data)*

- Normalization of counts - the process of ensuring that values are expressed on the same scale (e.g. RPKM, FPKM, TPM, TMM)
- Differential Expression - statistical analysis of the difference in expression of genes under two conditions (pair wise comparison) typically based on a negative binomial distribution.

Inferring Differential Expression (DE)

Method	Normalization	Needs replicas	Input	Statistics for DE	Availability
edgeR	Library size	Yes	Raw counts	Empirical Bayesian estimation based on Negative binomial distribution	R/Bioconductor
DESeq	Library size	No	Raw counts	Negative binomial distribution	R/Bioconductor
baySeq	Library size	Yes	Raw counts	Empirical Bayesian estimation based on Negative binomial distribution	R/Bioconductor
LIMMA	Library size	Yes	Raw counts	Empirical Bayesian estimation	R/Bioconductor
CuffDiff	RPKM	No	RPKM	Log ratio	Standalone

Multiple Testing Correction

Differential Expression data **must** be corrected for multiple testing. Two common methods are the “Bonferroni procedure” and “Benjamini–Hochberg procedure”. These forms or statistical correction will result in a “corrected pvalue”, or a qvalue or FDR or padj (adjusted p value).

Note pvalues refer to the each gene, whereas an FDR (or qvalue) is a statement about a list. So using FDR cuff of 0.05 indicates that you can expect 5% false positives in the list of genes with an FDR of 0.05 or less.

Differential Expression Output

1. **name** - the feature identity. It must be unique within the column. It may be a gene name, a transcript name, an exon
 (i.e. whatever the feature that we chose to quantify… can impact later steps.
2. **baseMean** - the average normalized expression level across all samples. It measure how much total signal is present across both conditions.
3. **baseMeanA** - the average normalized expression level across the first condition.
4. **baseMeanB** - the average normalized expression level across the first condition.
5. **foldChange** - the ratio of baseMeanB/baseMeanA. Very important to always be aware that in the fold change means B/A (second condition/first condition)
6. **log2FoldChange** - the second logarithm of foldChange. Log 2 transformations are convenient as they transform the changes onto a uniform scale. A four-fold increase after transformation is 2 . A four-fold decrease (1/4) after log 2 transform is -2. This property makes it much easier to compare the magnitude of up/down changes.
7. **PValue** - the uncorrected p-value of the likelihood of observing the effect of the size foldChange (or larger) by chance alone. This p-value is not corrected for multiple comparisons.
8. ***PAdj*** - the multiple comparison corrected PValue (via the Hochberg method). This probability is that of having at least one false positive when accounting for all comparisons that were made. This value is usually overly conservative in genomics.
9. ***FDR/q-values*** - the False Discovery Rate - this column represents the fraction of false discoveries for all the rows above the row where the value is listed. For example, if in row number 300 the FDR is 0.05, it means that if you were cut the table at this row and accept all genes at and above it as differentially expressed then, 300 * 0.05 = 15 genes out of the 300 are likely to be false positives.

The normalized matrix of the original count data is rarely given by default but can be very useful.

Differential Expression Output

EDGER

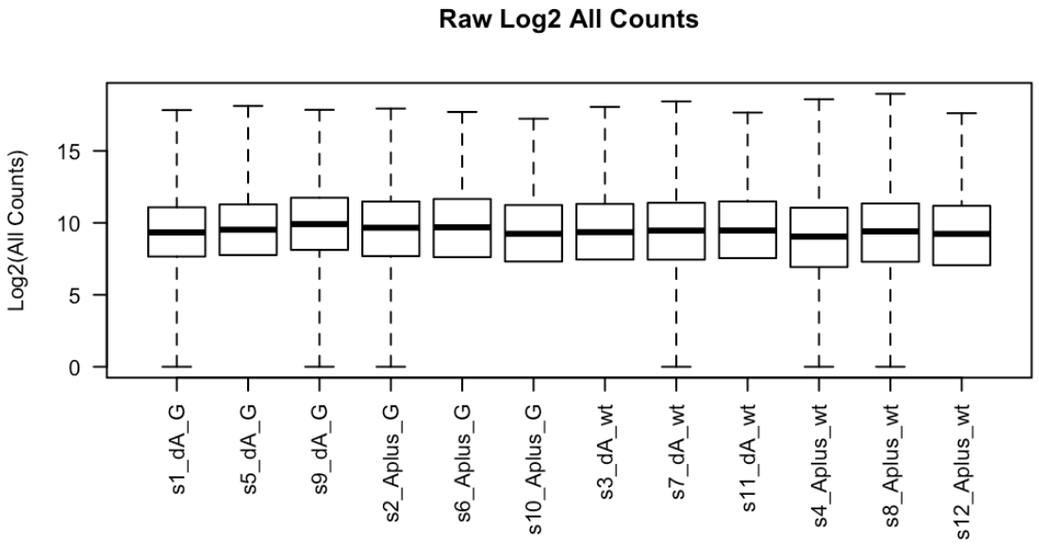
Gene	LogFC	AveExpr	P-Value	FDR
*CA14	-6.72	4.31	1.406716E-10	0.000001
*MCF2L	-10.75	3.25	2.854327E-10	0.000001
*COL5A2	-6.12	4.28	3.678663E-10	0.000001
*TYRP1	-9.31	9.85	4.190114E-10	0.000001
*BCAN	-8.39	5.33	6.384088E-10	0.000001
*CSAG1	10.81	-0.56	7.095577E-10	0.00000

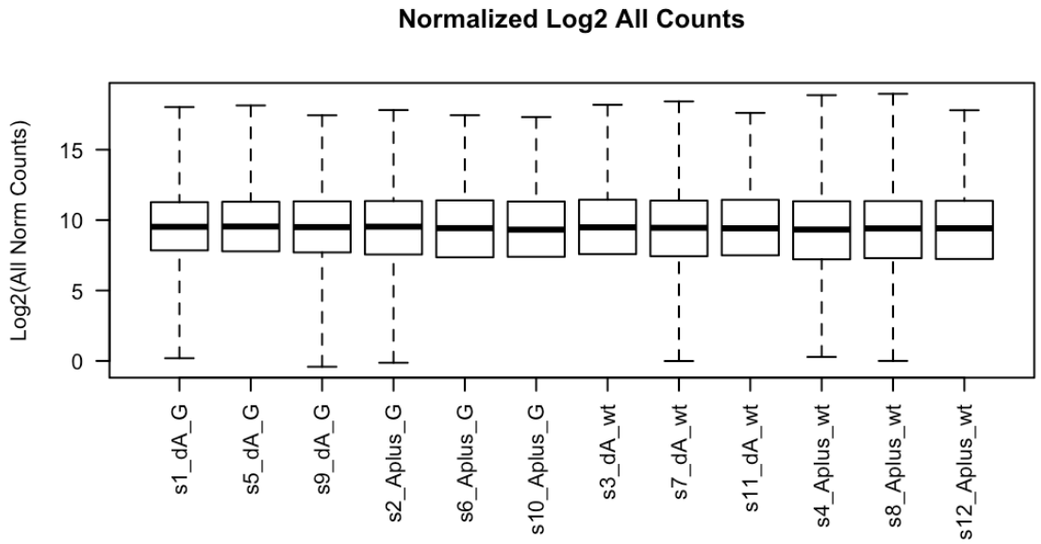
DESEQ2

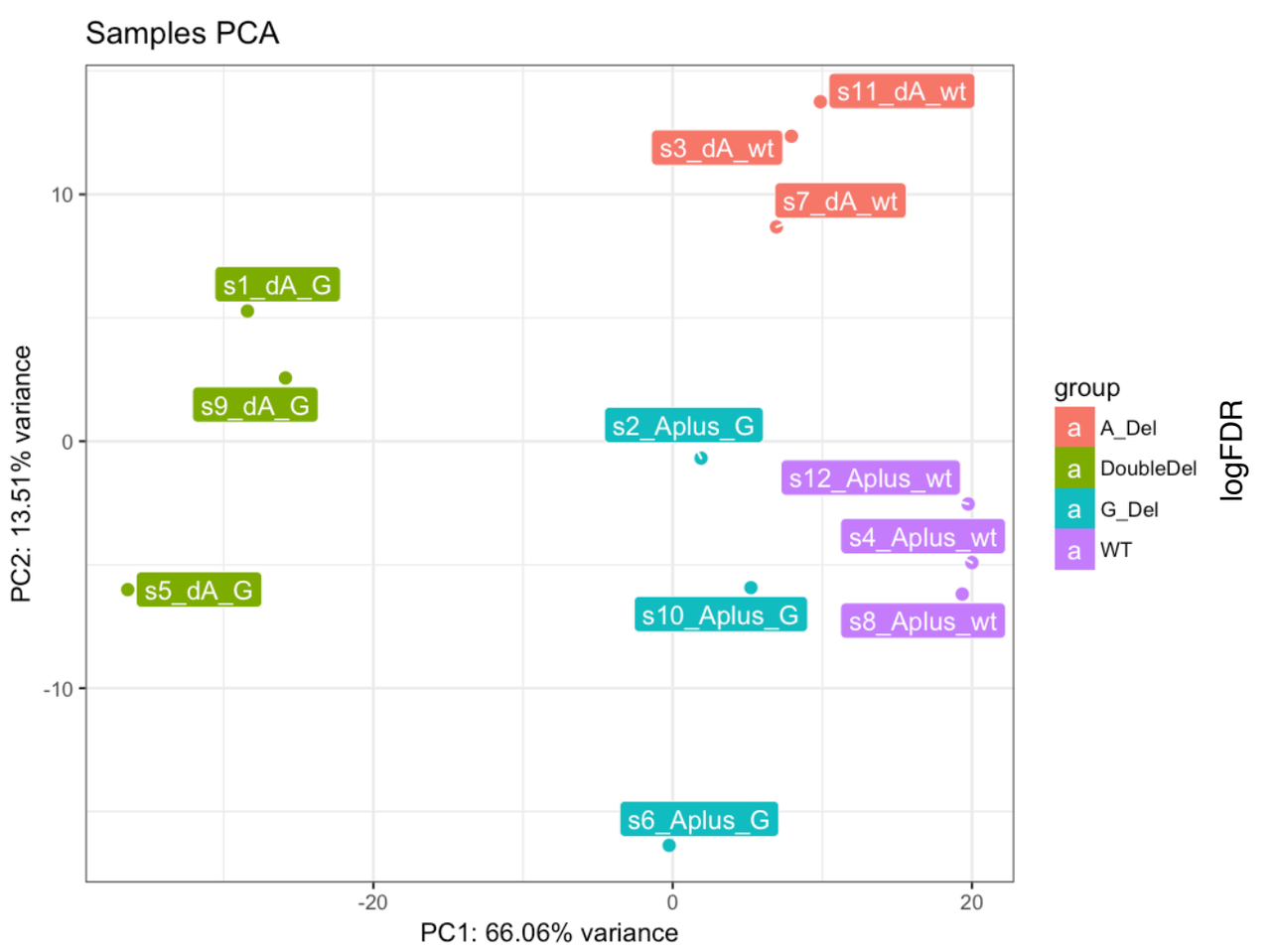
Row-names	Symbol	log2FoldChange	padj	p53_mock_1	p53_mock_2	p53_mock_3	p53_mock_4	p53_IR_1	p53_IR_2	p53_IR_3	p53_IR_4
ENSMUSG00000000001	Gnai3;Gnai3	-0.4763	0.1737	11.584	11.565	11.609	11.621	11.399	11.338	11.997	11.927
ENSMUSG00000000028	Cdc45;Cdc45	-0.4610	0.4125	8.024	7.575	7.668	7.295	7.736	7.675	7.906	7.873
ENSMUSG00000000037	Scml2;Scml2	1.3780	0.1889	3.196	3.554	3.563	3.296	4.592	5.249	4.765	5.262
ENSMUSG00000000056	Narf;Narf	-0.1732	0.8053	10.644	10.609	10.634	10.754	9.640	9.516	10.036	10.127
ENSMUSG00000000058	Cav2;Cav2	-0.3945	0.6751	4.377	4.546	5.292	5.120	4.122	3.531	4.835	4.269
ENSMUSG00000000088	Cox5a;Cox5a	-0.5847	0.2738	9.887	9.754	9.964	9.851	9.692	9.501	10.530	10.467
ENSMUSG00000000120	Ngfr;Ngfr	0.7409	0.2168	7.519	7.746	7.625	8.458	8.053	8.149	7.435	7.406
ENSMUSG00000000127	Fer;Fer	0.1804	0.7480	7.324	7.381	7.368	7.008	7.389	6.650	6.534	6.235
ENSMUSG00000000142	Axin2;Axin2	0.0927	0.9124	5.542	5.920	5.396	5.510	6.008	6.281	5.351	5.484

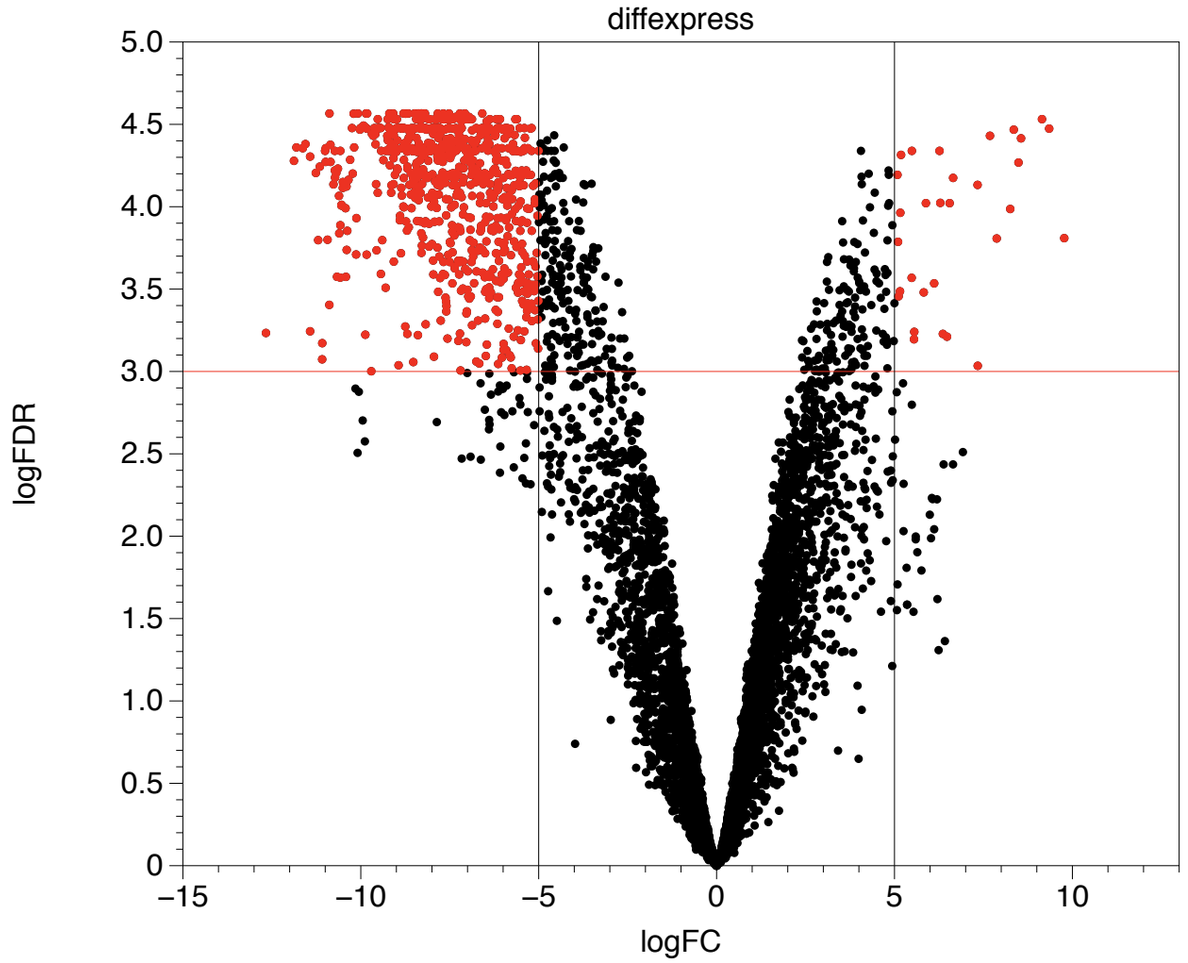
Visualization

Plotting the Data



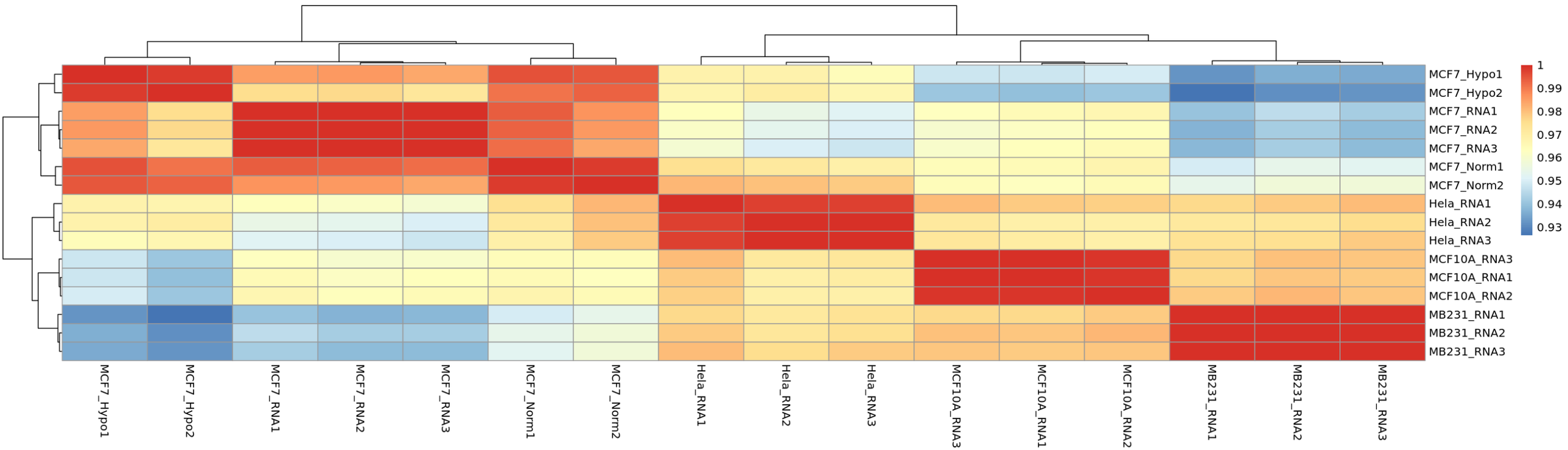






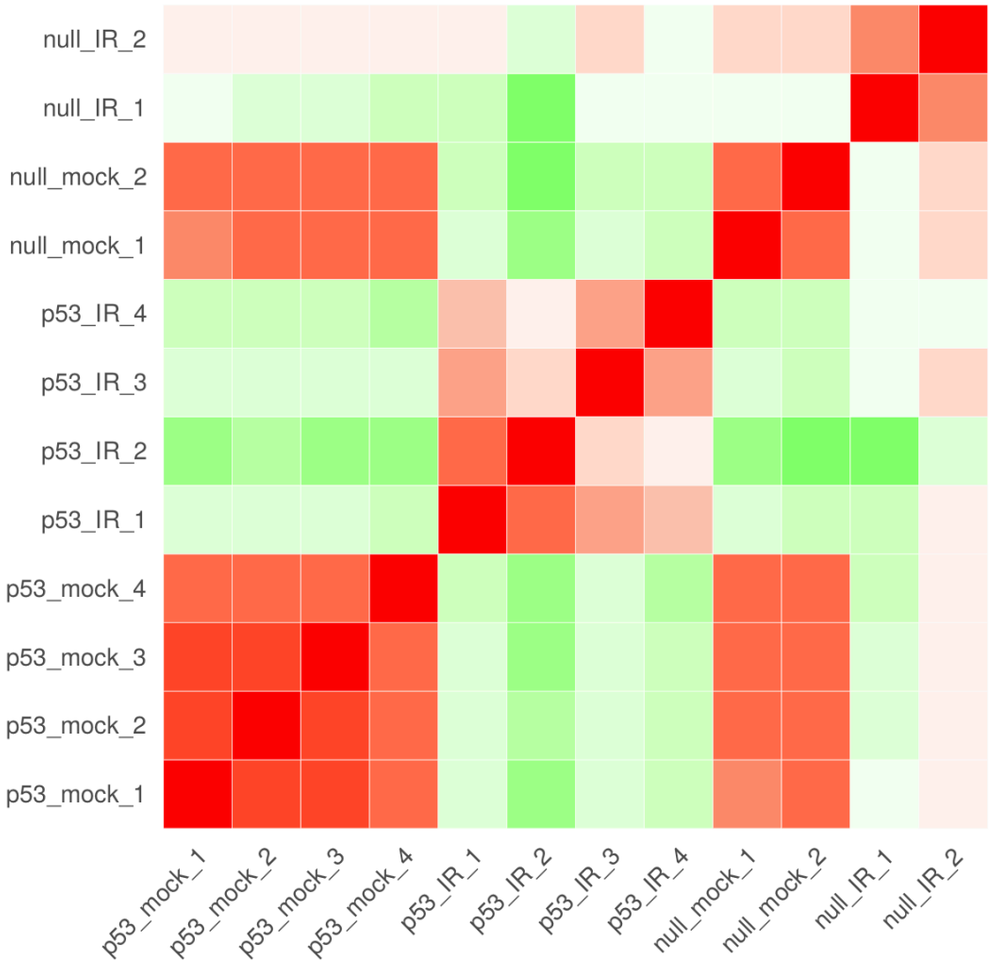
Plotting the Data

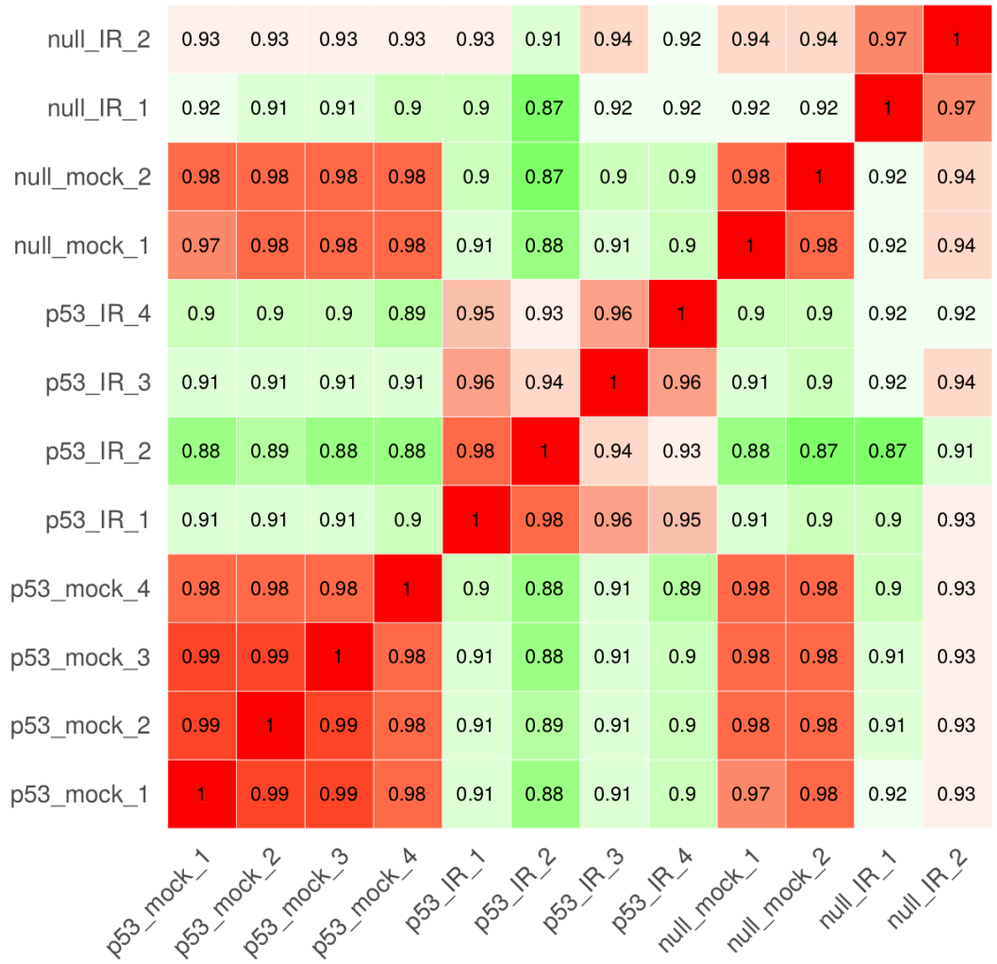
Plotting the Data

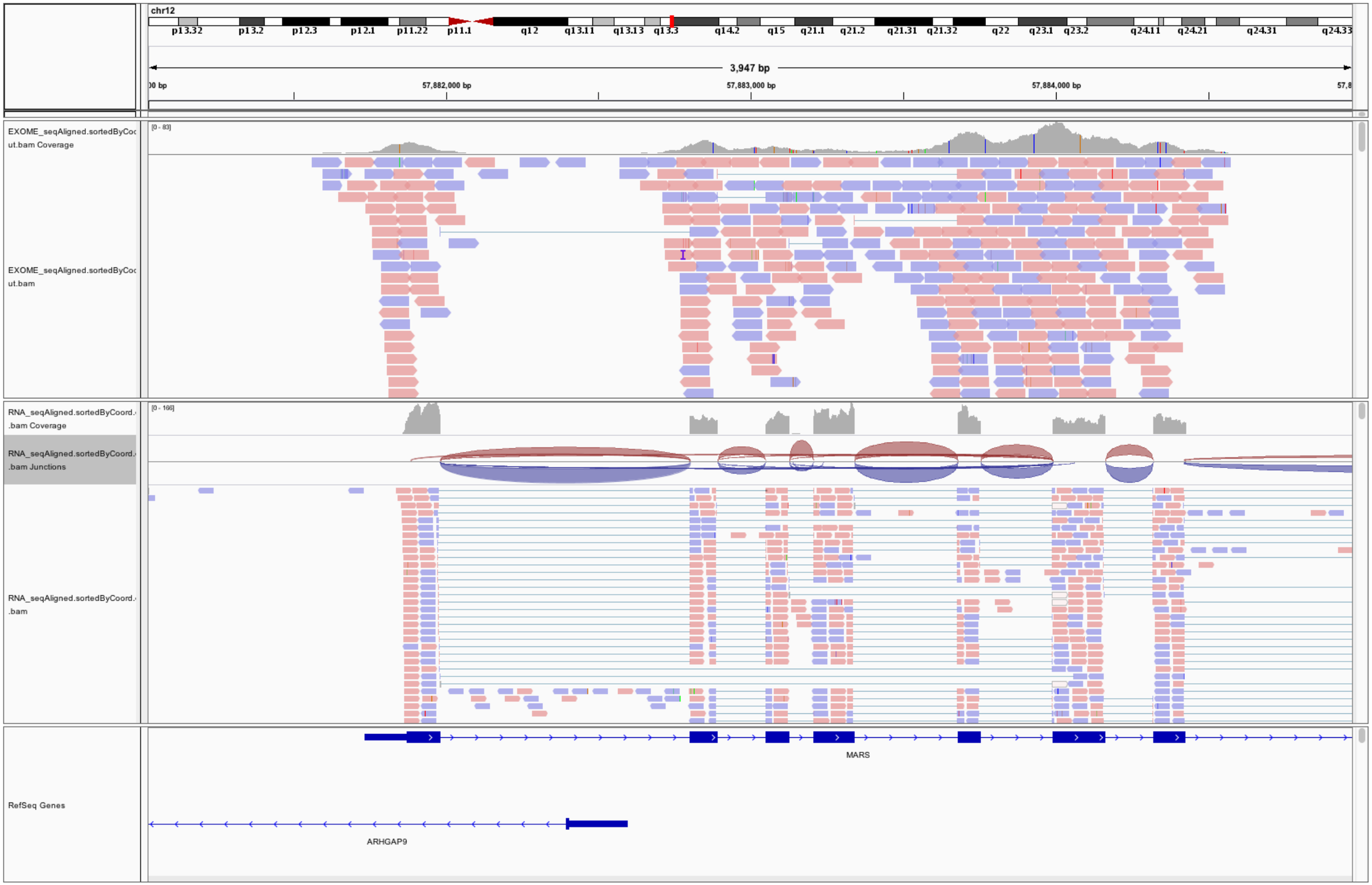


Plotting the Data

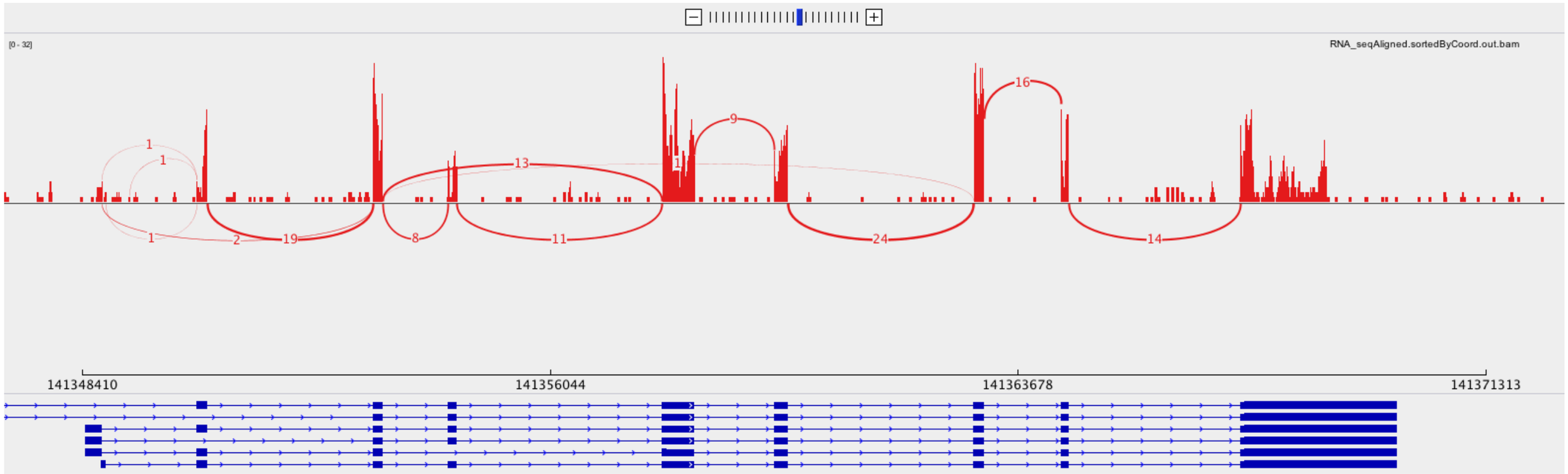
Heat Maps



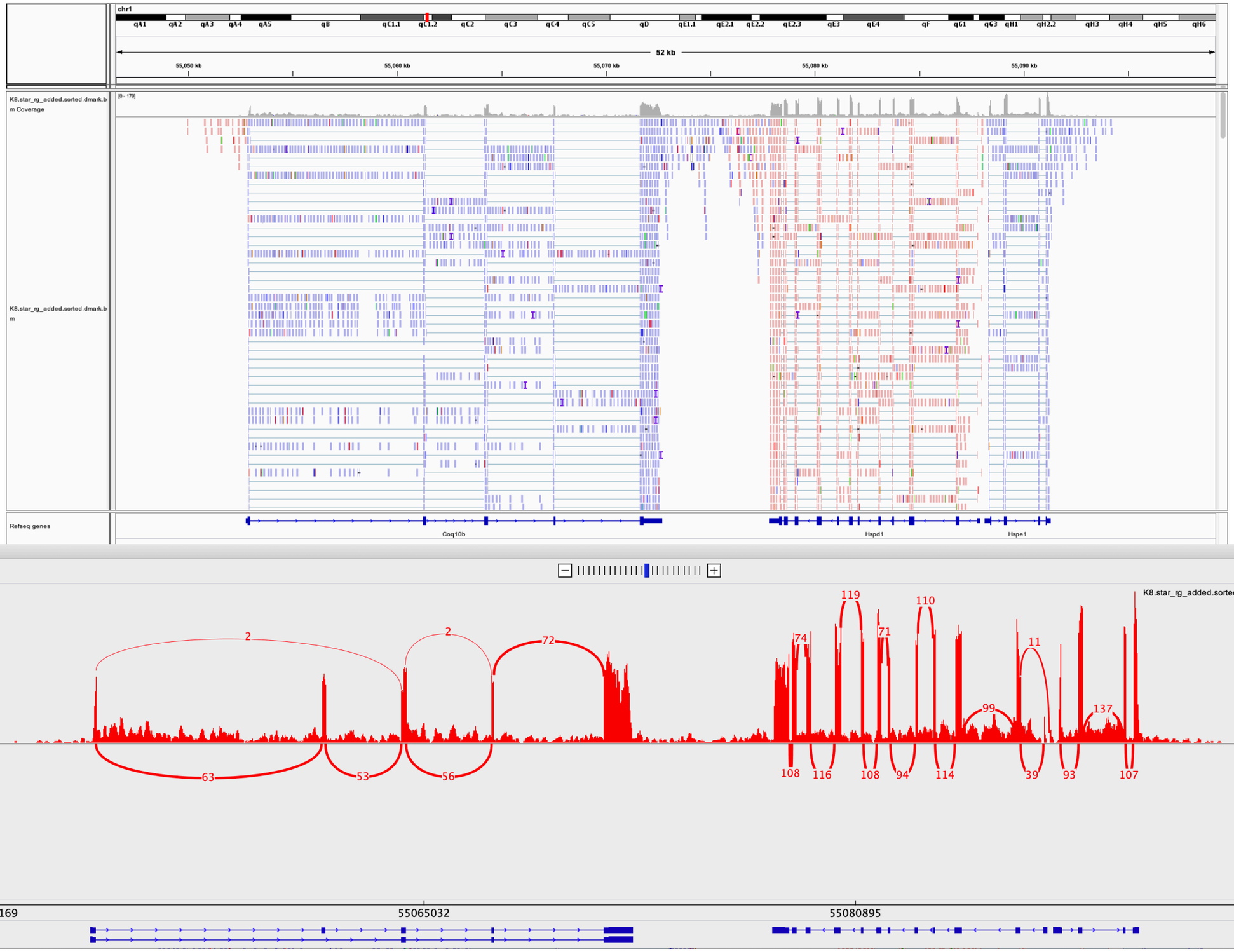



chr12
p13.32 p13.2 p12.3 p12.1 p11.22 p11.1 q12 q13.11 q13.13 q13.3 q14.2 q15 q21.1 q21.2 q21.31 q21.32 q22 q23.1 q23.2 q24.11 q24.21 q24.31 q24.33
3,947 bp
57,882,000 bp
57,883,000 bp
57,884,000 bp
EXOME_seqAligned.sortedByCoo
ut.bam Coverage
[0 - 83]
EXOME_seqAligned.sortedByCoo
ut.bam
RNA_seqAligned.sortedByCoord.
.bam Coverage
[0 - 166]
RNA_seqAligned.sortedByCoord.
.bam Junctions
RNA_seqAligned.sortedByCoord.
.bam
RefSeq Genes
MARS
ARHGAP9

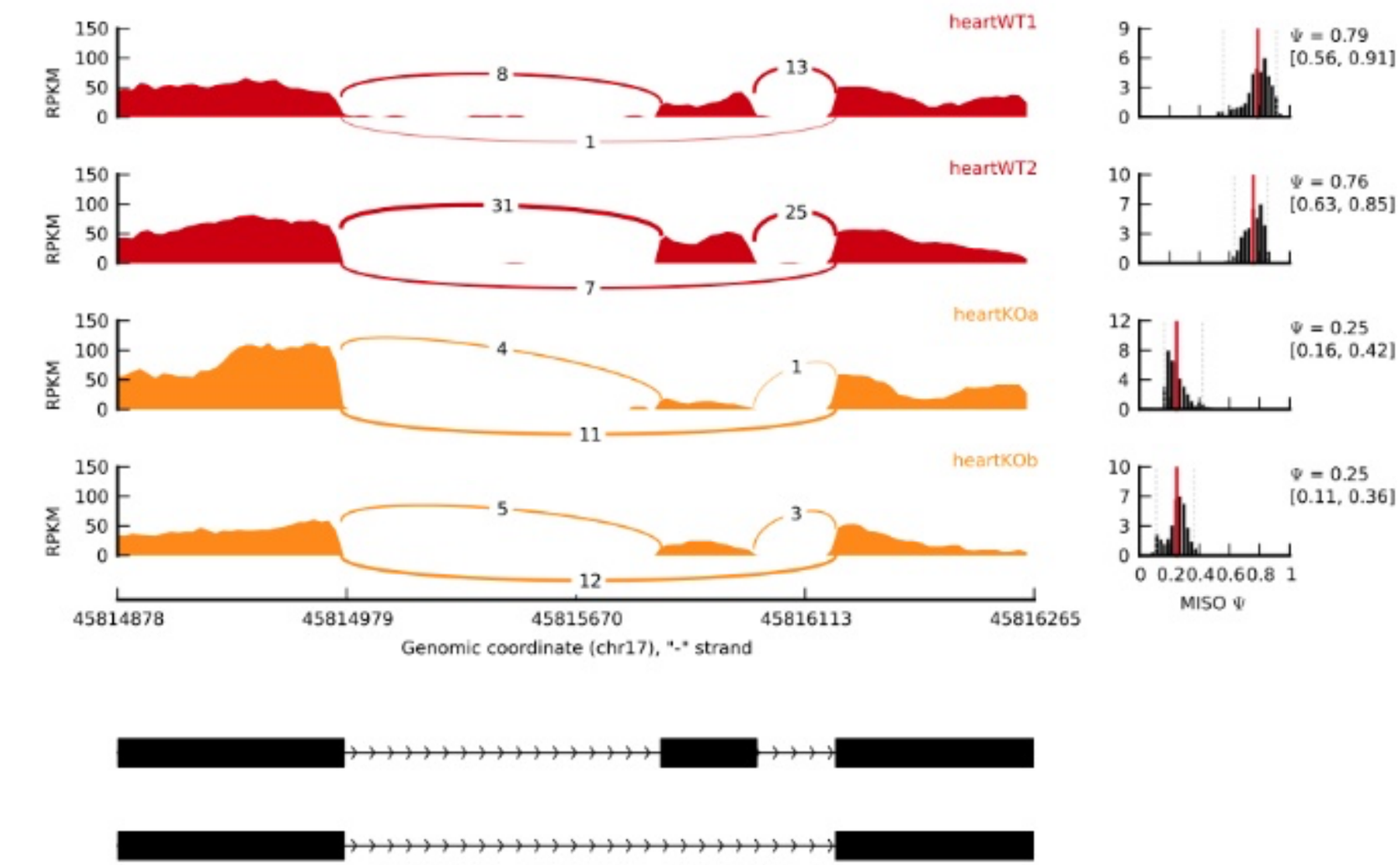

chr17
p13.3
p13.2
p13.1
p12
p11.2
p11.1
q11.1
q11.2
q12
q21.1
q21.31
q21.32
q21.33
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.2
q25.3
3,247 bp
29,622,000 bp
29,623,000 bp
29,624,000 bp
brain_sorted.bam Coverage
[0 - 36]
brain_sorted.bam
muscle_sorted.bam Coverage
[0 - 10.00]
muscle_sorted.bam
Sequence
RefSeq Genes
NF1
NF1
OMG


[0 - 32]
RNA_seqAligned.sortedByCoord.out.bam
1
1
1
2
19
8
13
1
11
9
24
16
14
141348410
141356044
141363678
141371313

Stranded RNA-Seq Data



Visualizing Splicing



UCDAVIS Bioinformatics Core

Visualization and Next step tools

Visualization

1. Integrated Genome Viewer (https://www.broadinstitute.org/igv/)

Further Annotation of Genes

1. DAVID (http://david.abcc.ncifcrf.gov/tools.jsp)
2. ConsensusPathdb (http://cpdb.molgen.mpg.de/)
3. NetGestalt (http://www.netgestalt.org/)
4. Molecular Signatures Database (http://www.netgestalt.org/)
5. PANTHER (http://www.pantherdb.org/)
6. Cognoscente (http://vanburenlab.medicine.tamhsc.edu/cognoscente.shtml)
7. Pathway Commons (http://www.pathwaycommons.org/)
8. Readctome (http://www.reactome.org/)
9. PathVisio (http://www.pathvisio.org/)
10. Moksiskaan (http://csbi.ltdk.helsinki.fi/moksiskaan/)
11. Weighed Gene Co-Expression Network Analysis (WGCNA)s
12. More tools in R Bioconductor

Tertiary Analysis - Biological Meaning

- **Pathway Analysis**
  - IPA (Qiagen - CCR License) Future talk
- **Functional Analysis**
  - Gene Set Enrichment Analysis (GSEA)
    https://www.gsea-msigdb.org/gsea/index.jsp
  - DAVID
    https://david.ncifcrf.gov/
  - Enrichr
    https://maayanlab.cloud/Enrichr/
- **Genomic Location**
- **Transcription Factor Enrichment Analysis**
- **miRNA Enrichment Analysis**

Public sources of RNA-Seq data

- **Gene Expression Omnibus** (GEO) (http://www.ncbi.nlm.nih.gov/geo/)
  - Both microarray and sequencing data
- **Sequence Read Archive (SRA)** (http://www.ncbi.nlm.nih.gov/sra)
  - All sequencing data (not necessarily RNA-Seq)
- **ArrayExpress** (https://www.ebi.ac.uk/arrayexpress/)
  - European version of GEO
- **Homogenized data**: MetaSRA, Toil, recount2, ARCHS[4]

NGS File Formats

- **Sequence**
  - FASTA, FastQ
- **Alignment**
  - SAM, BAM, CRAM
- **Annotation**
  - GTF, GFF, BED (BIGBED)
- **Graphing**
  - WIG (BIGWIG), BEDGRAPH

Utility Programs

- SeqKit
- FastQC, RSeQC, MultiQC
- Cutadapt, Fastp, Trimmomatic, TrimGalore
- STAR,Bowtie, Salmon
- Samtools, Picard, bedrolls, bamtools
- R, Python
- IGV

Web-Based Tools

- BioJupies - Many analysis functions - generates Jupyter Notebook of results (*https://amp.pharm.mssm.edu/biojupies/*)
- IDEP92 - an integrated web application for differential expression and pathway analysis of RNA-Seq data (*http://bioinformatics.sdstate.edu/idep92/*)

Both allow analysis of many public datasets

File Transfer

- Globus (https://hpc.nih.gov/storage/globus.html)
- HPCDME
- BOX
- OneDrive
- (s)FTP
- Network Drives
- Flash Drives

Further Reading

RNA-seqlopedia

https://rnaseq.uoregon.edu/

RNA-Seq by Example

https://www.biostarhandbook.com/

Questions ?

Contacts:

Peter Fitzgerald fitzgepe@nih.gov

Amy Stonelake amy.stonelake@nih.gov

BTEP ncibtep@nih.gov