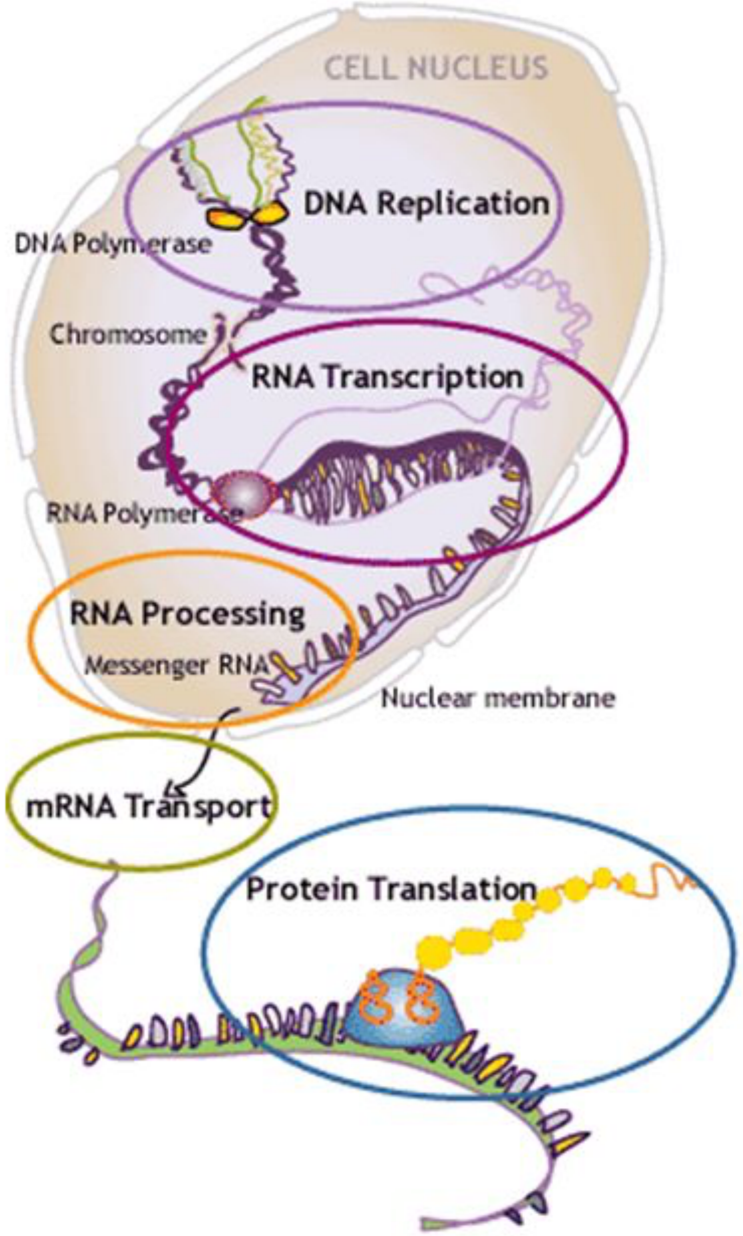
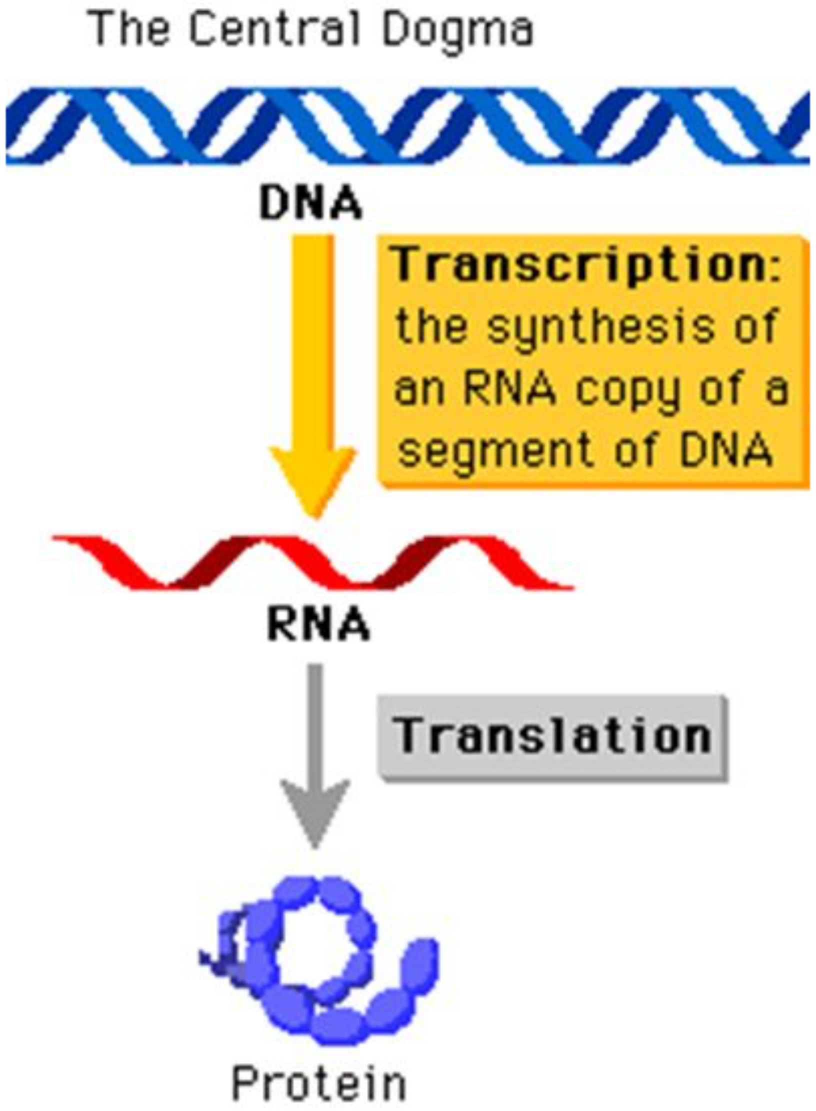
Introduction to RNASeq Data Analysis

Peter FitzGerald, PhD

Head Genome Analysis Unit
Director of BTEP
CCR, NCI

DNA → RNA → Protein




The Central Dogma
DNA
Transcription: the synthesis of an RNA copy of a segment of DNA
RNA
Translation
Protein

What is RNASEQ ?

RNA-Seq (**RNA sequencing**), uses next-generation sequencing (NGS) to reveal the presence and quantity of **RNA** in a biological sample at a given moment. (*Wikipedia*)

- Strictly speaking this could be any type of RNA (mRNA, rRNA, tRNA, snoRNA, miRNA) from any type of biological sample.
- For the purpose of this talk we will be limiting ourselves to **mRNA**.
- Technically, with a few exceptions, we are not actually sequencing **mRNA** but rather **cDNA**.

(RNASeq is only valid within the context of Differential Expression)

RNASEQ - WorkFlow

- **Experimental Design**
  - What question am I asking
  - How should I do it (*does it need to be done*)
- **Sample Preparation**
  - Sample Prep
  - Library Prep
  - Quality Assurance
- **Sequencing**
  - Technology/Platform
  - Detail Choices
- **Data Analysis (Computation)**

Experimental Design

Only Sequence the RNA of interest

- Remember ~90% of RNA is ribosomal RNA
- Therefore enrich your total RNA sample by:
- polyA selection (oligodT affinity) of mRNA (eukaryote)
- rRNA depletion - RiboZero is typically used (costs extra)

What are the Goals of your Experiment

- Which gene are expressed?
- Which genes are differentially expressed?
- Are different splicing isoforms expressed?
- Are there novel genes or isoforms expressed?
- Are you interested in structural variants or SNPs, indels
- Are you interested in non-coding RNAs
- Does your interest lie in micro RNAs
- If this a standalone experiment, a pilot, or a “fishing trip”

Read Choices

- **Read Depth**
  - More depth needed for lowly expressed genes
  - Detecting low fold differences need more depth
- **Read Length**
  - The longer the length the more likely to map uniquely
  - Paired read help in mapping and junctions
- **Stranded Protocols**
  - Give clearer results
- **Replicates**
  - Detecting subtle differences in expression needs more replicates
  - Detecting novel genes or alternate iso-forms need more replicates

Increasing depth, length, and/or replicates increase costs

Replicates

- **Technical Replicates**
  - It's generally accepted that they are not necessary because of the low technical variation in RNASeq experiments
- **Biological Replicates** (Always useful)
  - Not strictly needed for the identification of novel transcripts and transcriptome assembly.
  - Essential for differential expression analysis - must have 3+ for statistical analysis
  - Minimum number of replicates needed is variable and difficult to determine:

    3+ for cell lines

    5+ for inbred samples

    20+ for human samples (rarely possible)
  - More is always better

Data Analysis Questions

- Where will the primary data be stored (fastq)?
- Where will the processed data be stored (bam)?
- Who will do the primary analysis?
- Who will do the secondary analysis?
- **Where will the published data be deposited and by who? (what metadata will they require)**
- Are you doing reproducible science?

Talk *to the people who will be analyzing your data* ***BEFORE*** *doing the experiment*

Best Practice Guidelines from Bioinformatic Core (CCBR):

1. Factor in at least 3 replicates (absolute minimum), but 4 if possible (optimum minimum). Biological replicates are recommended rather than technical replicates.

2. Always process your RNA extractions at the same time. Extractions done at different times lead to unwanted batch effects.

3. There are 2 major considerations for RNA-Seq libraries:

- If you are interested in coding mRNA, you can select to use the mRNA library prep. The recommended sequencing depth is between 10-20M paired-end (PE) reads. Your RNA has to be high quality (RIN > 8).
- If you are interested in long noncoding RNA as well, you can select the total RNA method, with sequencing depth ~25-60M PE reads. This is also an option if your RNA is degraded.

4. Ideally to avoid lane batch effects, all samples would need to be multiplexed together and run on the same lane. This may require an initial MiSeq run for library balancing. Additional lanes can be run if more sequencing depth is needed.

5. If you are unable to process all your RNA samples together and need to process them in batches, make sure that replicates for each condition are in each batch so that the batch effects can be measured and removed bioinformatically.

6. For sequence depth and machine requirements, visit Illumina Sequencing Coverage website

For cost estimates, visit Sequencing Facility pricing for NGS
For further assistance in planning your RNA-Seq experiment or to discuss specifics of your project, please contact us by email: ***CCBR@mail.nih.gov*** *OR visit us during office hours on Fridays 10am to noon (Bldg37/Room3041). For cost and specific information about setting up an RNA-Seq experiment, please visit the Sequencing Facility website or contact Bao Tran*

Sample Preparation

General Rules for Sample Preparation

- Prepare all samples at the same time or as close as possible. The same person should prepare all samples
- Do not prepare “experiment” and “control” samples on different days or by different people. (Batch effects).
- Use high quality means to determine sample quality (**R**NA **I**ntegrity **N**umber) (RIN >0.8) and quantity, and size (Tapestation, Qibit, Bioanalyzer)
- Don’t assume everything will work the first time (do pilot experiments) or every time (prepare extra samples)

Sample Amounts

Type of Library	Minimum DNA/RNA Requirement for Library Construction	Recommended DNA/RNA for Optimal Library Construction	Maximum Sample Volume Requirement for Library Construction	Additional Requirements
mRNA Sequencing	100 ng	1 µg	50 µL	RIN should be at least 8.0, DNase treated
mRNA ultralow Clonetech	100 pg	10 ng	10 µL	RIN should be at least 8.0, DNase treated
microRNA Sequencing	100 ng	1 µg	6 µL	
Total RNA Sequencing	100 ng	1 µg	10 µL	DNase treated, FFPE and degraded RNA can be used
Total RNA ultralow	10 ng	1 µg	10 µL	DNase treated, FFPE and degraded RNA can be used

RNA-Seq Sample Recommendations (CCBR)

QC Metric Guidelines	mRNA	total RNA
RNA Type(s)	Coding	Coding + non-coding
RIN	8 [low RIN = 3' bias]	> 8
Single-end vs Paired-end	Paired-end	Paired-end
Recommended Sequencing Depth	10-20M PE reads	25-60M PE reads
FastQC	Q30 > 70%	Q30 > 70%
Percent Aligned to Reference	70%	> 65%
Million Reads Aligned Reference	7M PE reads (or > 14M reads)	16.5M PE reads (or > 33M reads)
Percent Aligned to rRNA	< 5%	< 15%
Picard RNAseqMetrics	Coding > 50%	Coding > 35%
Picard RNAseqMetrics	Intronic + Intergenic < 25%	Intronic + Intergenic < 40%

Sequencing

Illumina Sequencing Platforms

Illumina
Sequencing by Synthesis (SbS)
/NovaSeq/HiSeq/NextSeq/MiSeq
Short read length (50 to 300 bp)

Selection driven by cost, precision, speed, number of samples and number of reads required

Consult with the Sequencing Core

Illumina
NovaSeq

Illumina
NextSeq

Illumina
MiSeq

Long Read Sequencing Platforms

PacBio
120,000 bases per molecule, with maximum read lengths > 200,000 bases. Good for repetitive regions and isomers, modified bases.

Oxford Nanopore
Direct DNA or RNA sequencing (Max length 2 Mb) Good for modified bases, repetitive regions, isomers, small genomes.

Consult with the Sequencing Cores

PacBio Sequel II

MinION

GridION

Oxford Nanopore

Data Analysis

RNASEQ - Data Analysis WorkFlow I

- **Quality Control**
  - Sample quality and consistency
  - Is Trimming appropriate - quality/adaptors
- **Reports**
- **Alignment/Mapping**
  - Reference Target (Sequence and annotation)
  - Alignment Program & parameters
  - Mark Duplicates
  - Post-Alignment Quality Assurance
- **BAM, WIG, files and reports**
- **Quantification**
  - Counting Method and Parameters
- **BED files, count matrices**

RNASEQ - Data Analysis WorkFlow II

- **Quantification**
  - Differential Expression - statistics
- * **Data tables, plots**
- **Visualization**
  - Visual inspection - IGV
  - Data representation - scatter, violin plots, heat-maps
- * **Images and Graphs**
- **Biological Meaning**
  - Gene Set Enrichment
  - Pathway Analysis
- * **Data tables, network maps**

Computational Considerations
THE GOOD NEWS

For the most part the computational aspects have been taken care of for you.
(no need to develop new algorithms or code).

There are pre-built workflows that can automate many of the processes involved, and facilitate reproducibility

Computational Considerations
THE BAD NEWS

Like most of NGS data analysis, the complexity of RNA-Seq data analysis revolves around data and information management and the dealing with "unexpected" issues.

Consider the simplest experiment

(*Two conditions three replicates*)

6-12 fastq starting files

6-12 quality control files

6-12 fastq files post trimming of adaptors

6 bam file, and 6 bam index files

6 gene count files

36-48 files minimum (big files)

Computational Considerations
The Challenges

There is no single **best method** for RNA-Seq data analysis - it depends on your definition of best, and even then it varies over time and with the particular goals and specifics of a given experiment

It’s for this reason that you should learn enough about the process to make “sensible choices” and to know when the results are reasonable and correct.

Treating an RNA-Seq (or any NGS) analysis as a black box is a “recipe for disaster” (*or at least bad science*). That’s not to say that you need to know the particulars of every algorithm involved in a workflow, but you should know the steps involved and what assumptions and/or limitations are build into the whole workflow

Computational Prerequisites

- High performance Linux computer (multi core, high memory, and plenty of storage) for the alignment phase
- Familiarity with the “command line” and at least one programming/scripting language.
- Basic knowledge of how to install software
- Basic knowledge of R and/or statistical programming
- Basic knowledge of Statistics and model building

Data Analysis

Pre-alignment QC & cleanup

Alignment

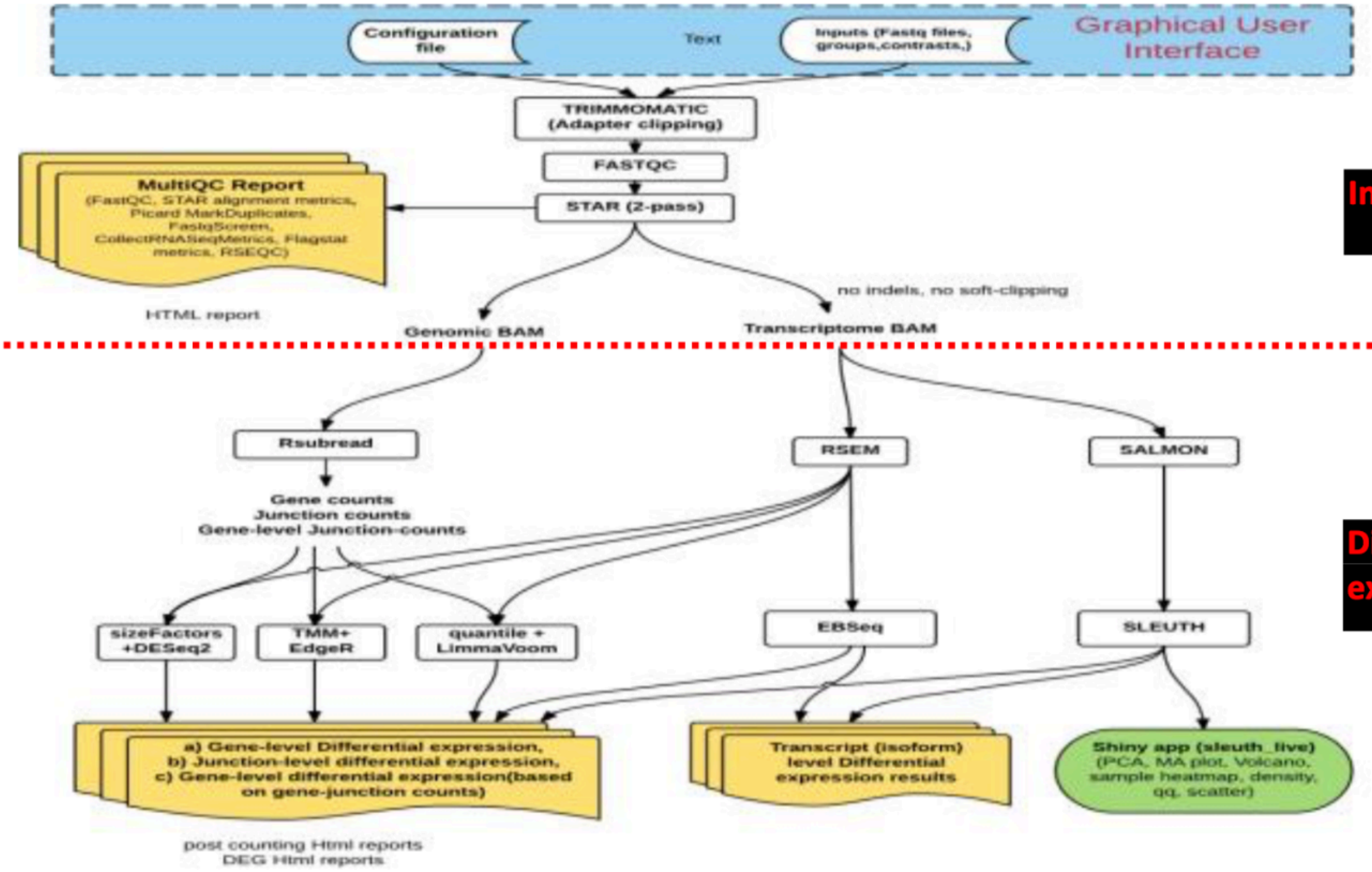
Post-alignment QC & filtering

Quantification

Differential Expression

RNASEQ Pipeline

https://github.com/CCBR/Pipeliner/blob/master/RNASeqDocumentation.pdf

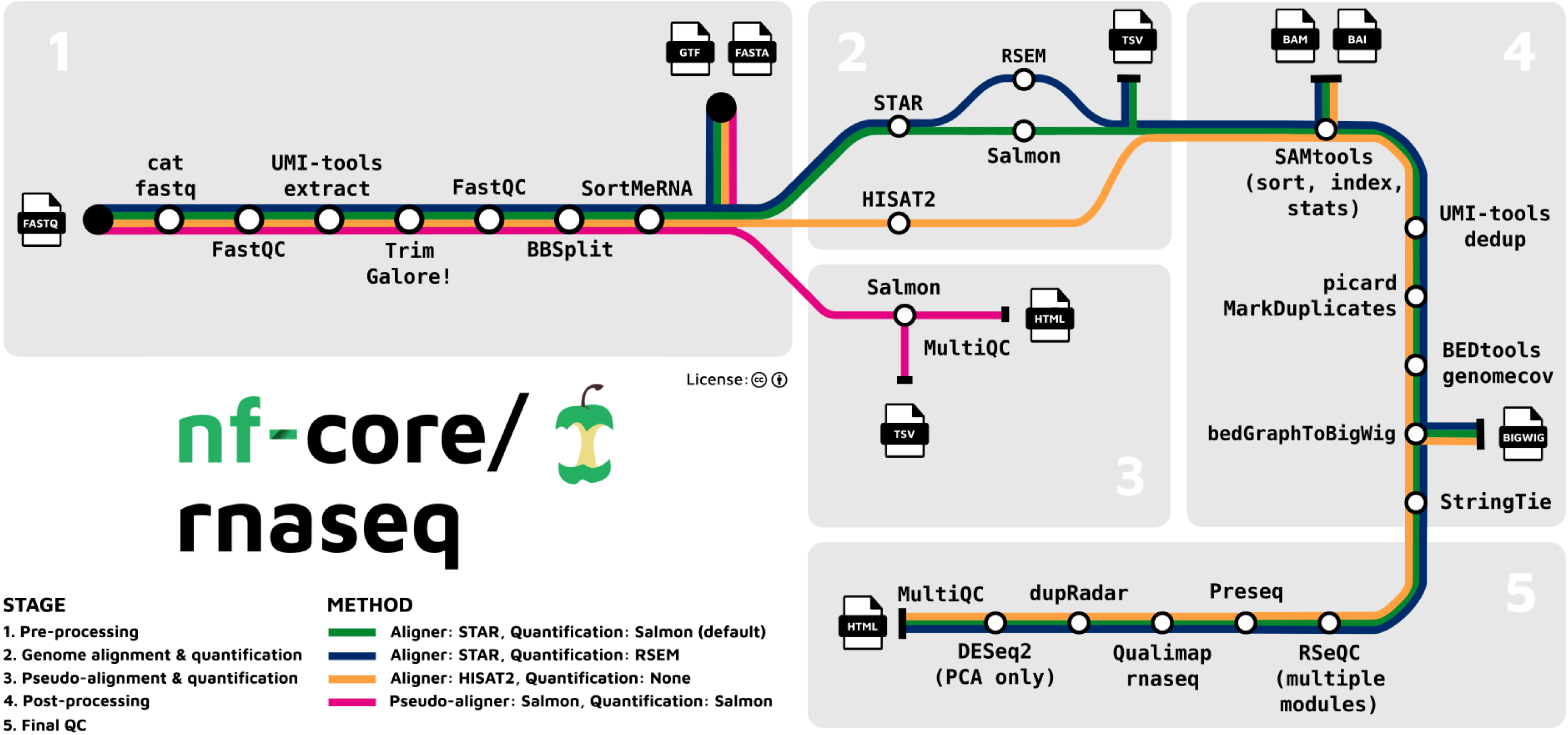

Configuration file
Text
Inputs (Fastq files, groups,contrasts,)
Graphical User Interface
TRIMMOMATIC (Adapter clipping)
FASTQC
STAR (2-pass)
MultiQC Report (FastQC, STAR alignment metrics, Picard MarkDuplicates, FastqScreen, CollectRNASeqMetrics, Flagstat metrics, RSEQC)
HTML report
no indels, no soft-clipping
Genomic BAM
Transcriptome BAM
Rsubread
RSEM
SALMON
Gene counts
Junction counts
Gene-level Junction-counts
sizeFactors +DESeq2
TMM+ EdgeR
quantile + LimmaVoom
EBSeq
SLEUTH
a) Gene-level Differential expression, b) Junction-level differential expression, c) Gene-level differential expression(based on gene-junction counts)
Transcript (isoform) level Differential expression results
Shiny app (sleuth_live) (PCA, MA plot, Volcano, sample heatmap, density, qq, scatter)
post counting Html reports
DEG Html reports

Initial QC

Differential expression

RNASEQ Pipeline

https://nf-co.re/rnaseq



STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

METHOD

- 🟩 Aligner: STAR, Quantification: Salmon (default)
- 🟦 Aligner: STAR, Quantification: RSEM
- 🟧 Aligner: HISAT2, Quantification: None
- 🟪 Pseudo-aligner: Salmon, Quantification: Salmon

Quality Control/Assesment (Pre-Alignment)

Data Quality Assessment

- **Evaluate the read quality to determine**

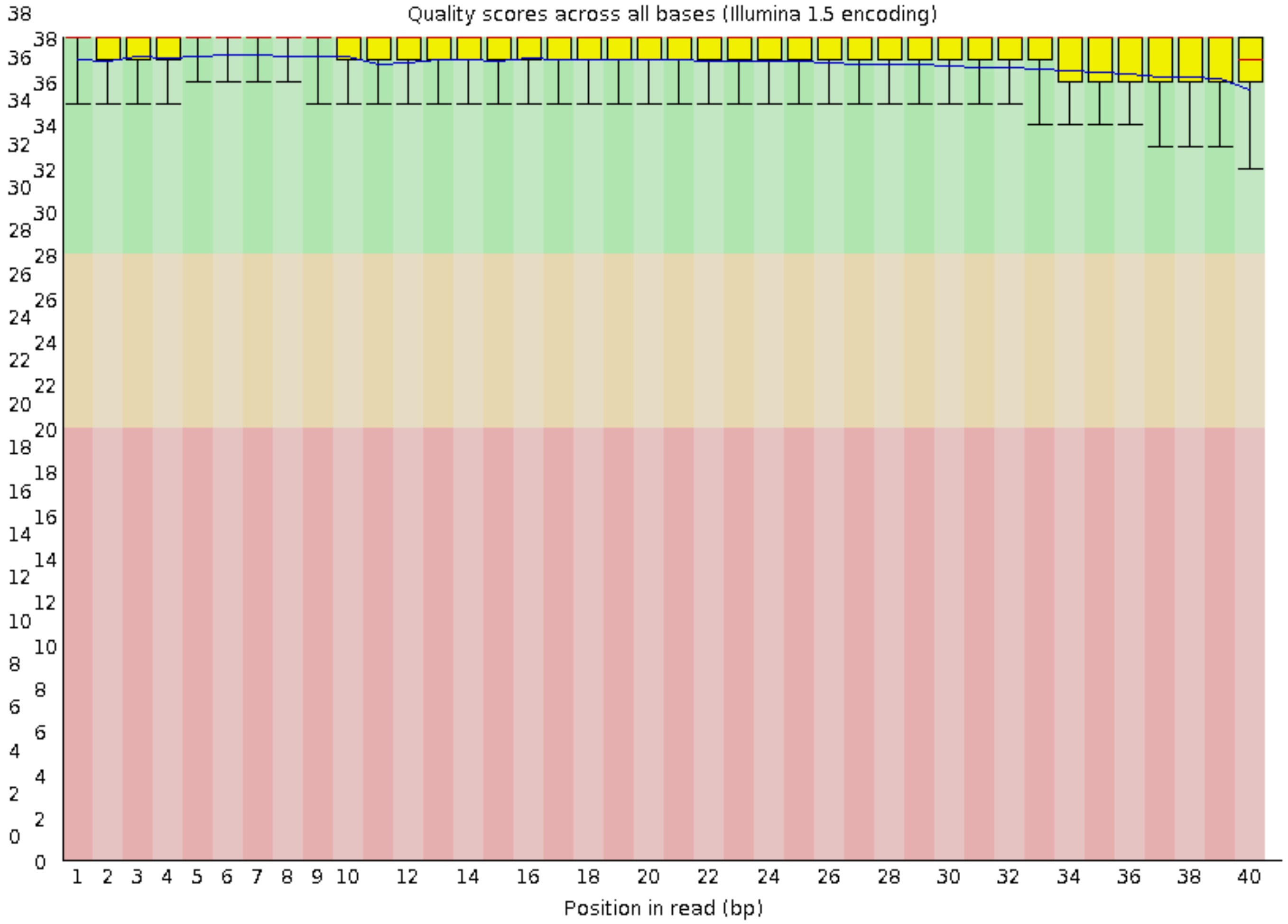
  (Tells us nothing about whether the experiment worked)

  - Is the data of sufficiently high quality to be analyzed?
  - Are there technical artifacts?
  - Are there poor quality samples?
- **Evaluate the following features**
  - Overall sequencing quality scores and distributions
  - GC content distribution
  - Presence of adapter or contamination
  - Sequence duplication levels
- **Data should be filtered, trimmed, or rejected as appropriate**

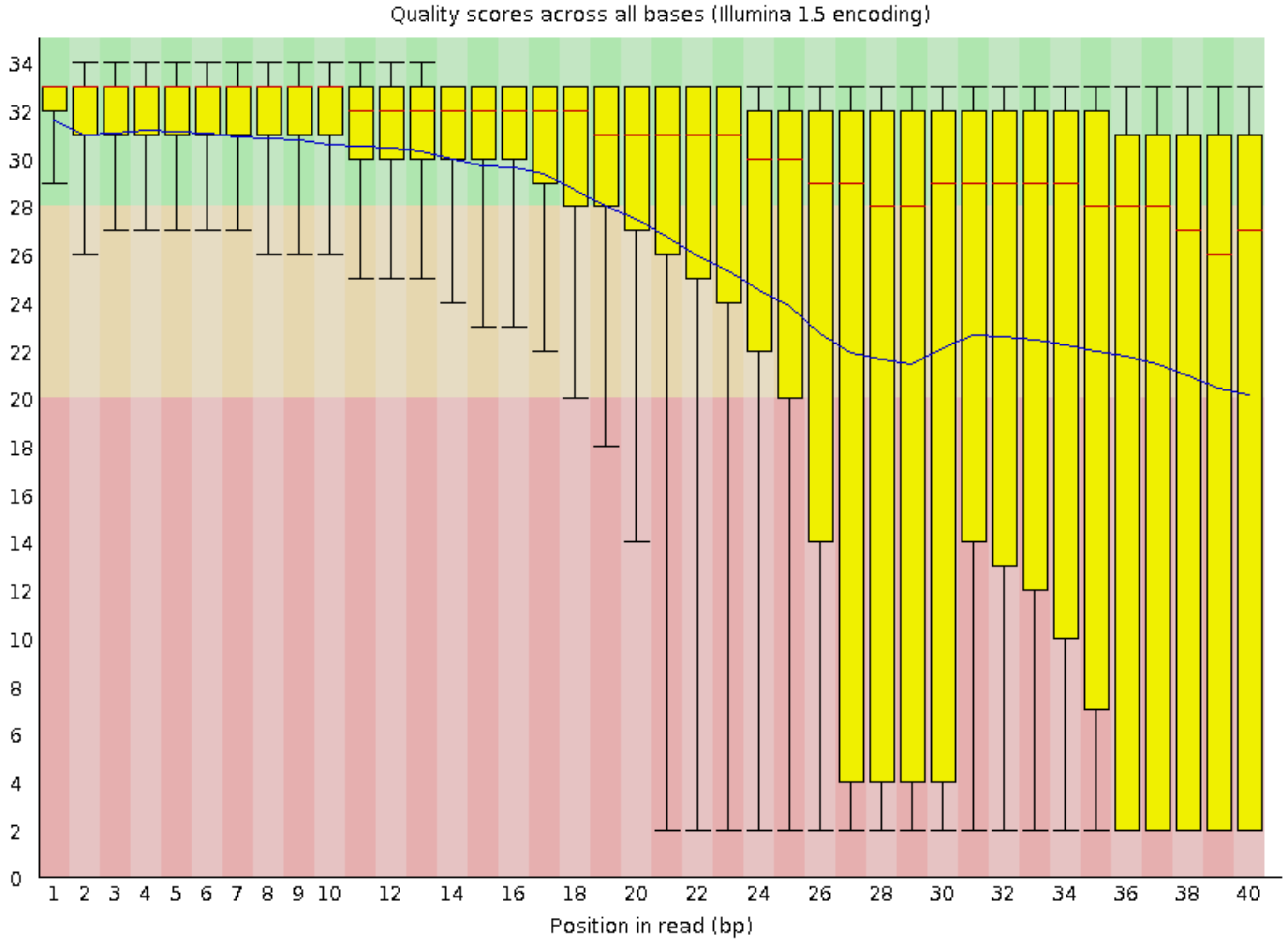
Sequencing cores generally provide some/all of this analysis

FastQC

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
good_sequence_short_fastqc.html



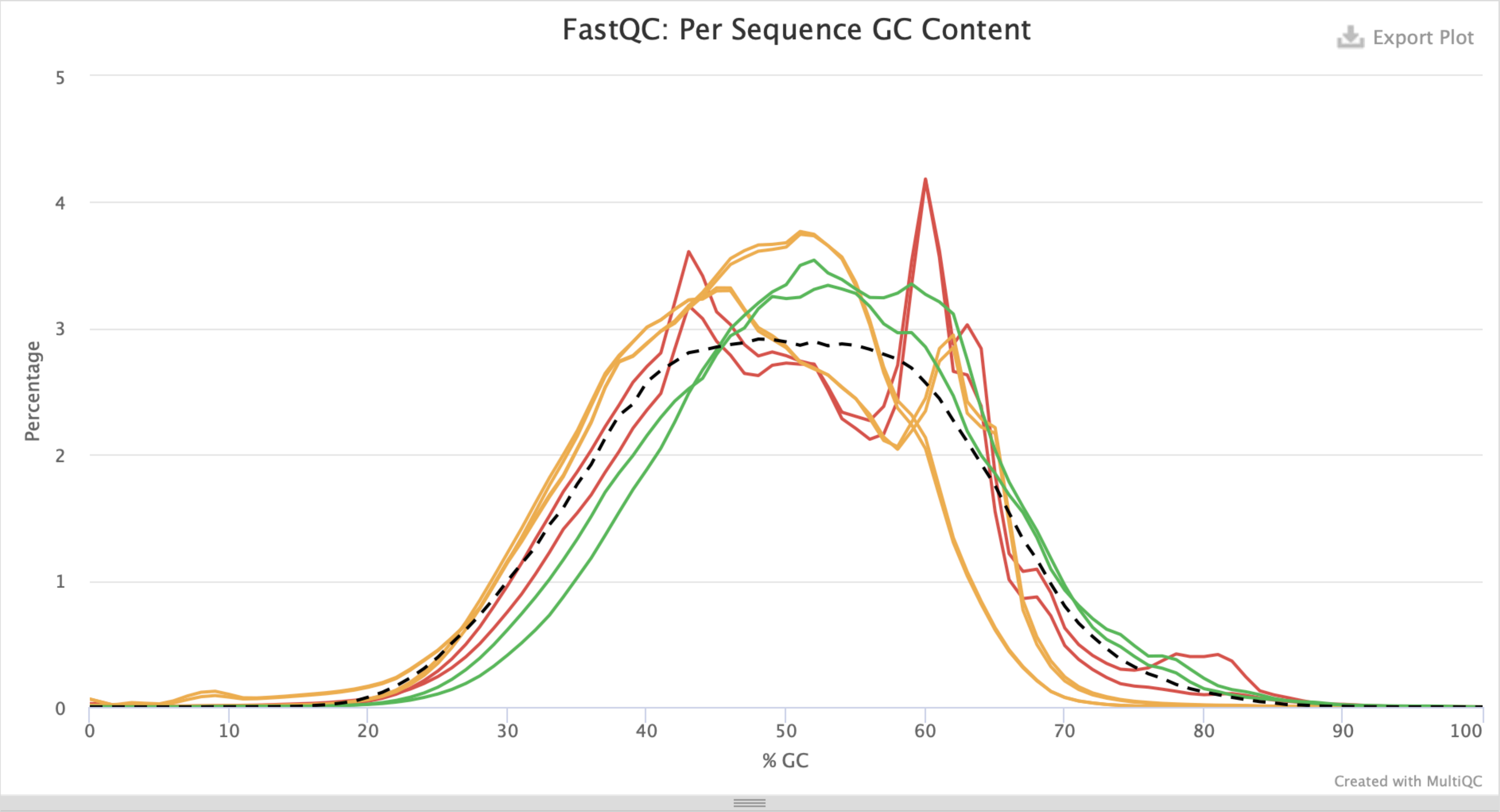
GOOD



BAD

MultiQC

https://multiqc.info/examples/rna-seq/multiqc_report.html



Raw Sequence Cleanup

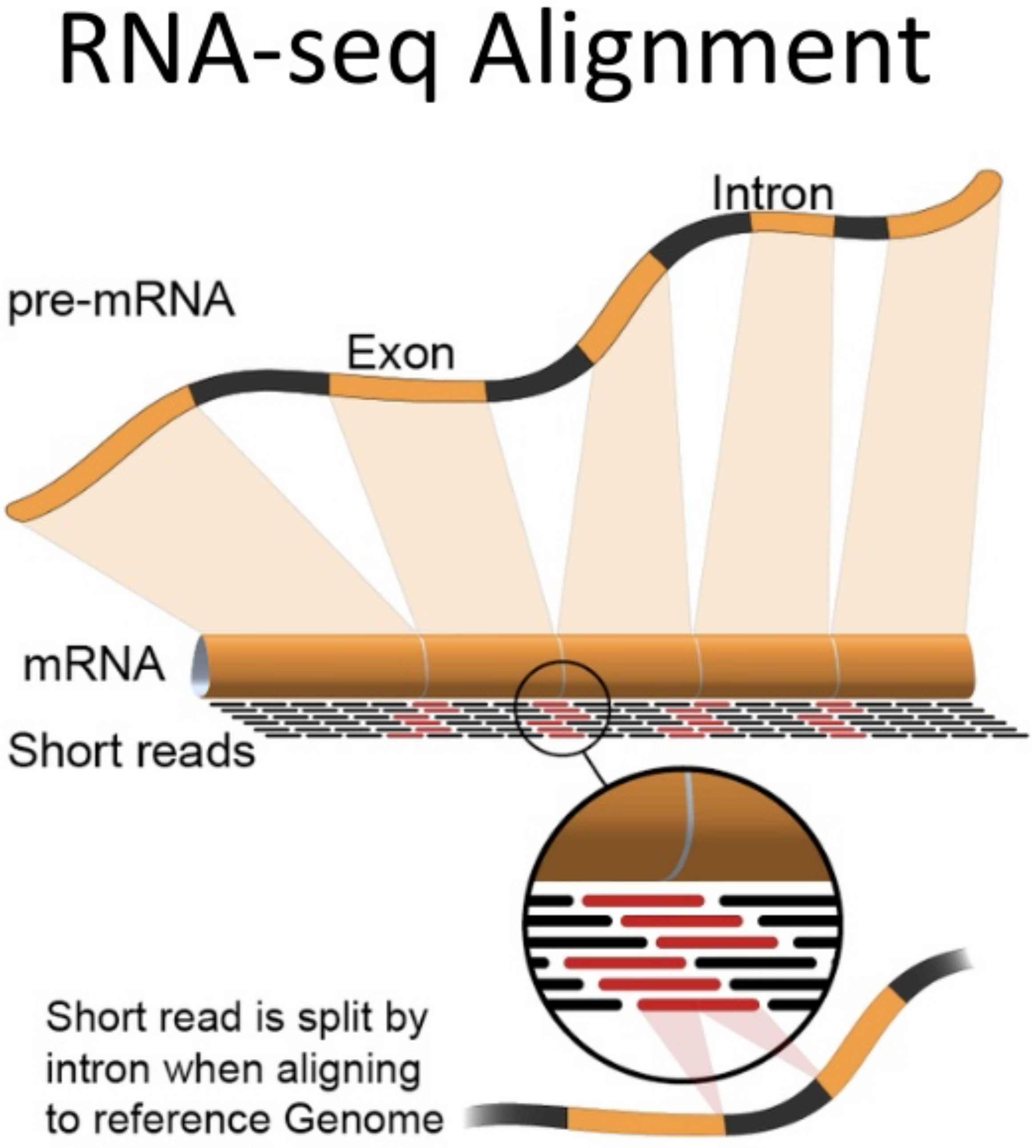
Trim and/or filter sequence to remove sequencing primers/adaptor and poor quality reads. Example programs:

- **FASTX-Toolkit** is a collection of command line tools for Short-Reads FASTA/ FASTQ files preprocessing.
- **SeqKit** is an ultrafast comprehensive toolkit for FASTA/Q processing.
- **Trimmomatic** is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters.
- **TrimGalore** is a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries.
- **Cutadapt** finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads.

Alignment

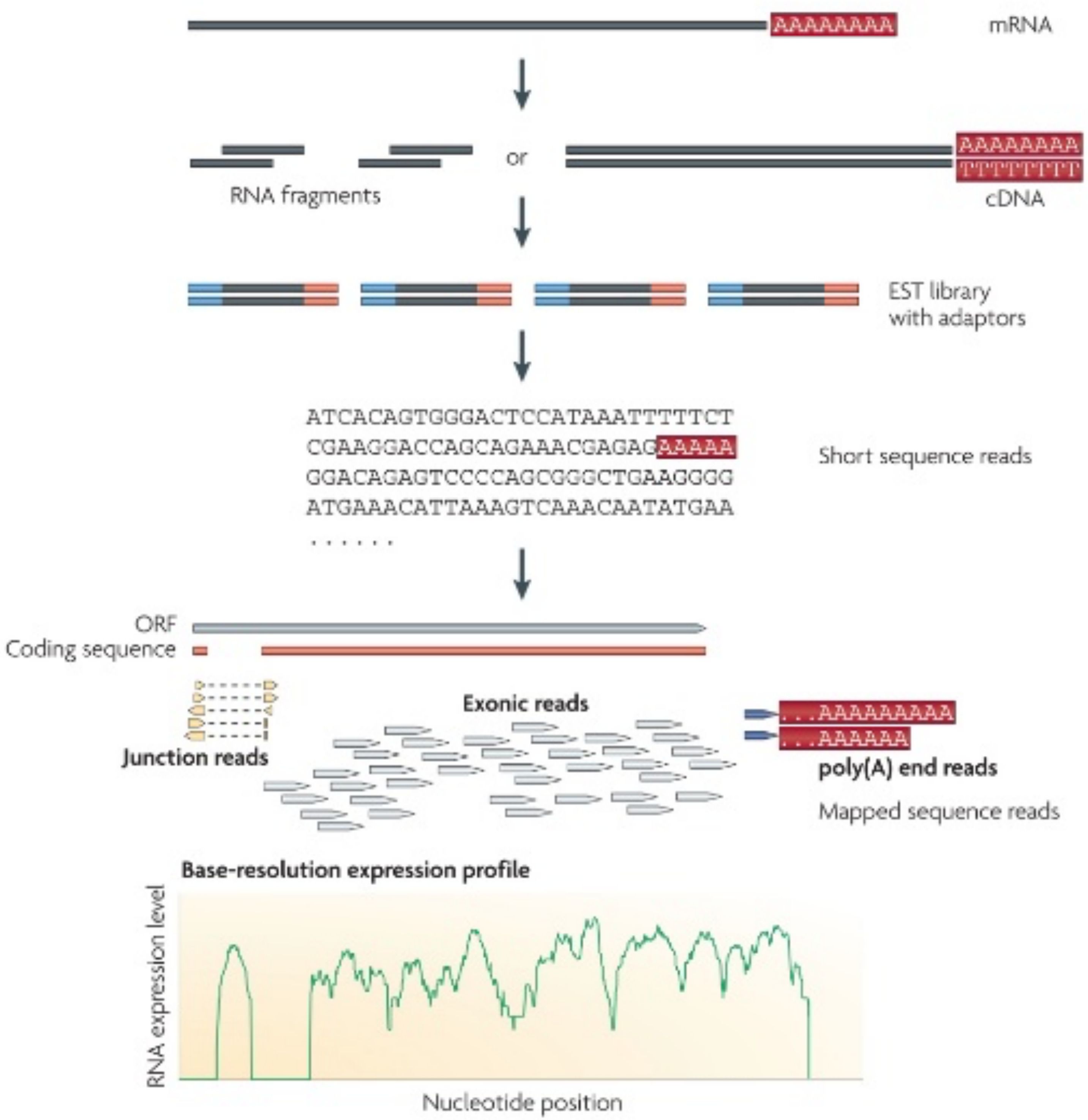
(Computationally Intensive Step)

RNASeq Mapping Challenges


RNA-seq Alignment
pre-mRNA
Exon
Intron
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

The majority of mRNA derived from eukaryotes is the result of splicing together discontinuous exons.

RNA-seq protocol schematic



Mapping Challenges

- Reads not perfect
- Duplicate molecules (PCR artifacts skew quantitation)
- Multimapped reads - Some regions of the genome are thus classified as unmappable
- Aligners try **very** hard to align **all** reads, therefore fewest artifacts occur when all possible genomic locations are provides (genome over transcriptome)

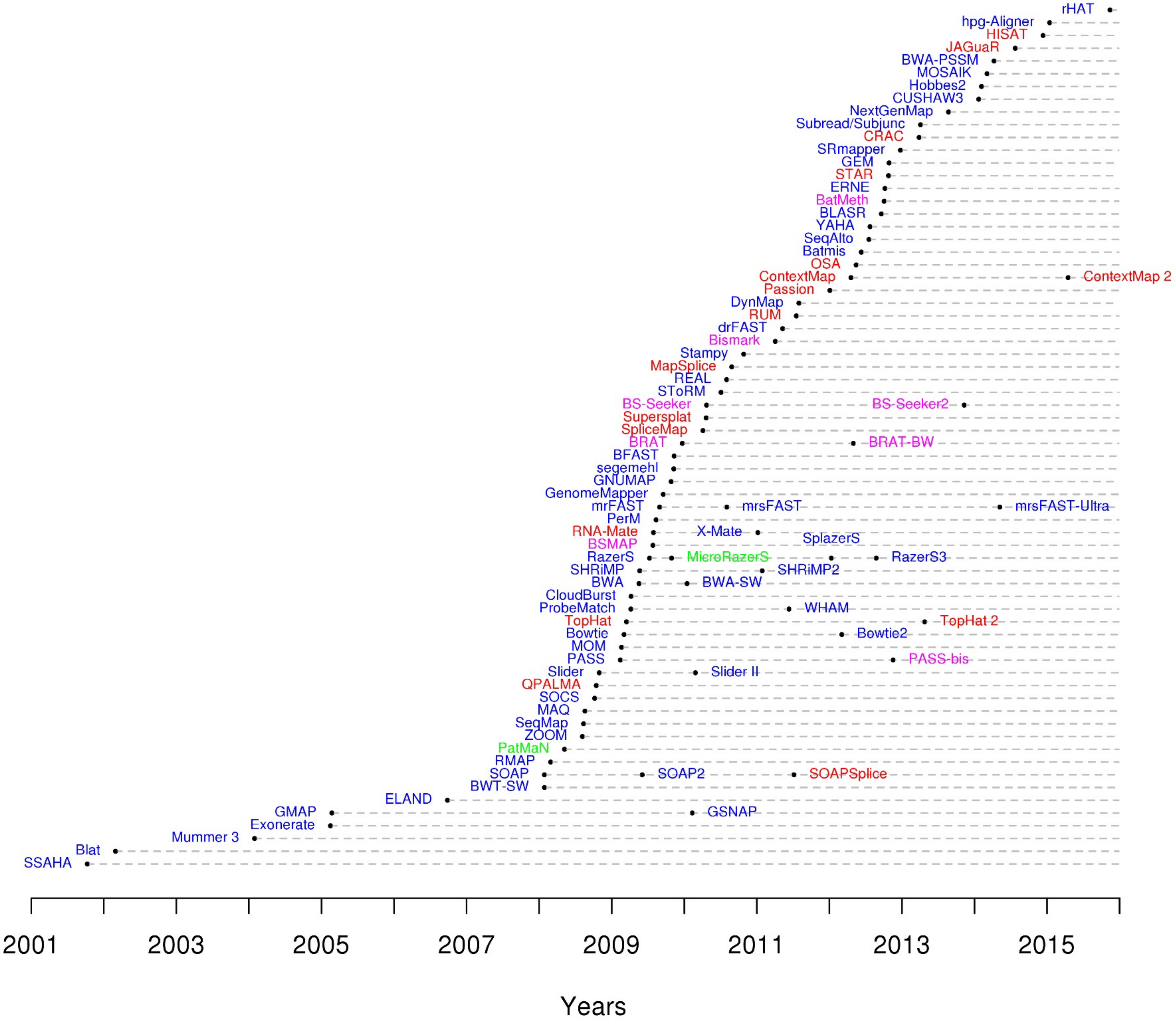
RNASeq Mapping Solutions

- **Align against the transcriptome**
  - Many/All transcriptomes are incomplete
  - Can only measure known *genes*
  - Won’t detect non-coding RNAs
  - Can’t look at splicing variants
  - Can’t detect fusion genes or structure variants

- **De novo assembly of RNASeq reads**
  - Largely used for uncharacterized genomes

- **Align against the genome using a splice-aware aligner**
  - Most versatile solution

- **Pseudo-Aligner - quasi mappers (Salmon and Kalisto)**
  - New class of programs - blazingly fast
  - Map to transcriptome (not genome) and does quantitation
  - Surprisingly accurate except for very low abundance signals
  - With bootstrapping can give confidence values


rHAT
hpg-Aligner
HISAT
JAGuaR
BWA-PSSM
MOSAIK
Hobbes2
CUSHAW3
NextGenMap
Subread/Subjunc
CRAC
SRmapper
GEM
STAR
ERNE
BatMeth
BLASR
YAHA
SeqAlto
Batmis
OSA
ContextMap
ContextMap 2
Passion
DynMap
RUM
drFAST
Bismark
Stampy
MapSplice
REAL
SToRM
BS-Seeker
BS-Seeker2
Supersplat
SpliceMap
BRAT
BRAT-BW
BFAST
segemehl
GNUMAP
GenomeMapper
mrFAST
mrsFAST
mrsFAST-Ultra
PerM
RNA-Mate
X-Mate
SplazerS
BSMAP
RazerS
MicroRazerS
RazerS3
SHRiMP
SHRiMP2
BWA
BWA-SW
CloudBurst
ProbeMatch
WHAM
TopHat
TopHat 2
Bowtie
Bowtie2
MOM
PASS
PASS-bis
Slider
Slider II
QPALMA
SOCS
MAQ
SeqMap
ZOOM
PatMaN
RMAP
SOAP
SOAP2
SOAPSplice
BWT-SW
ELAND
GMAP
GSNAP
Exonerate
Mummer 3
Blat
SSAHA
2001
2003
2005
2007
2009
2011
2013
2015
Years

Common Aligners

Most alignment algorithms rely on the construction of auxiliary data structures, called indices, which are made for the sequence reads, the reference genome sequence, or both. Mapping algorithms can largely be grouped into two categories based on properties of their indices: algorithms based on hash tables, and algorithms based on the Burrows-Wheeler transform

- Bowtie2
- BWA/BWA-mem
- **STAR**
- HISAT
- HISAT2
- TopHat
- TopHat2

Tools for mapping high-throughput sequencing data

Nuno A. Fonseca Johan Rung Alvis Brazma John C. Marioni Author Notes

Bioinformatics, Volume 28, Issue 24, 1 December 2012, Pages 3169–3177, https://doi.org/10.1093/bioinformatics/bts605

To Align or not to Align

Aligners typically align against the entire genome and provide a output where the results can be **visibly inspected** (bam file via IGV). They must be used for detecting novel genes/transcripts. Quantitation of aligned reads to specific genes is typically done by a separate program

PseudoAligners assign reads to the most appropriate transcript… can't find novel genes/transcripts or other anomalies. Generally much faster than aligners and are likely more accurate (Recent improvements in salmon have increased its accuracy, at the expense of being somewhat slower than the original)

Typical Questions about alignment

- What is the best aligner to use?
- What Genome version should I use?
- What Genome annotation should I use?

Answers

- STAR - (**Salmon** or Kallisto) - *subjective*
- Depends ! most recent or best annotated
- GeneCode with caveats - know what is being annotated and what is not and how it effects your results

Questions not asked

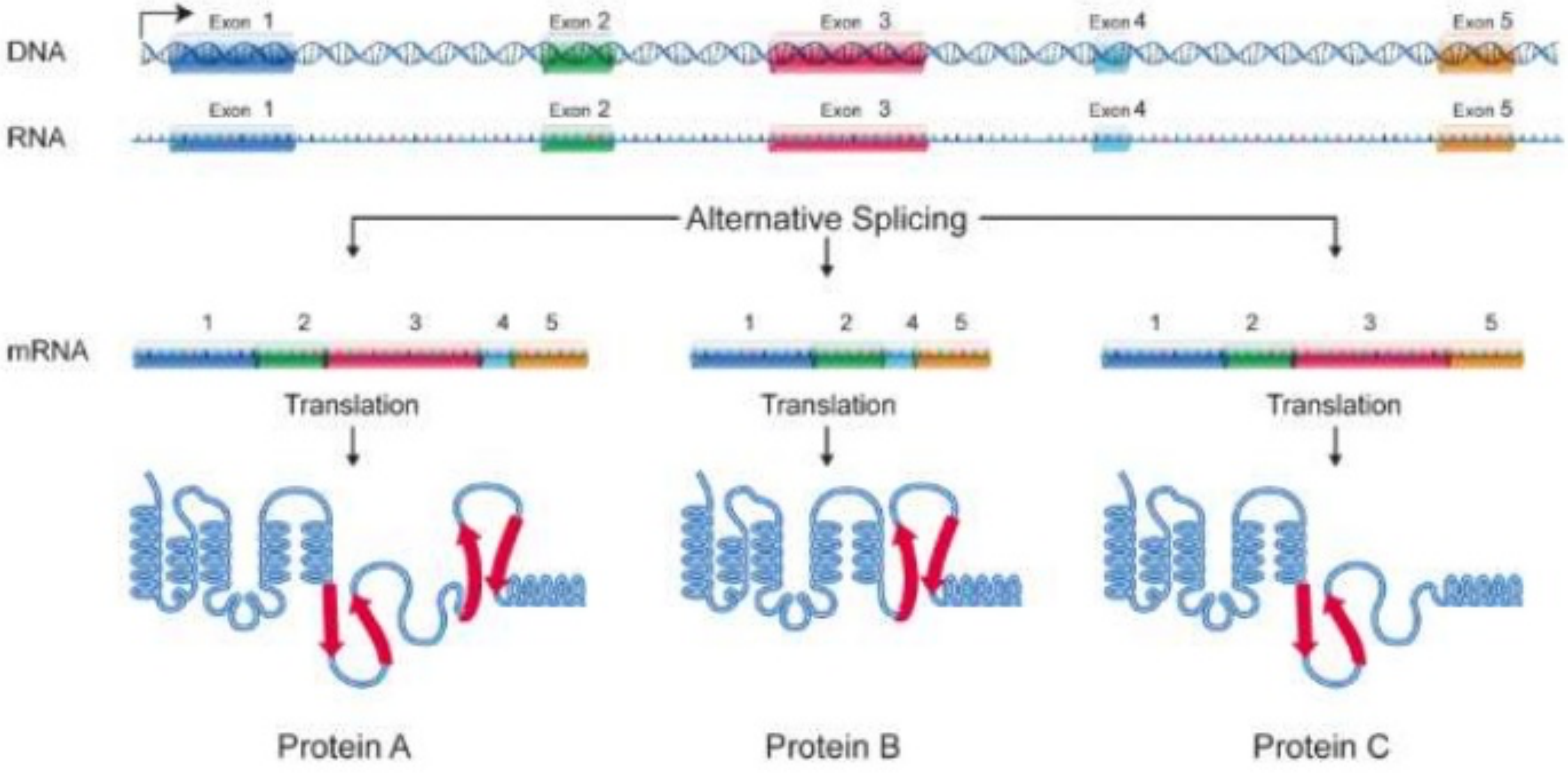
- What parameters should I use?

Answers

- Most programs have lots of optional parameters that can tweak the results, but most are set to defaults that should work in most common situations.
(*Don't touch what you don't understand -* ***especially*** *if it gets you, your favorite answer*)

RNA-Seq: Special Mapping Concerns

Alternate Splicing



genome.gov

Post Alignment QC

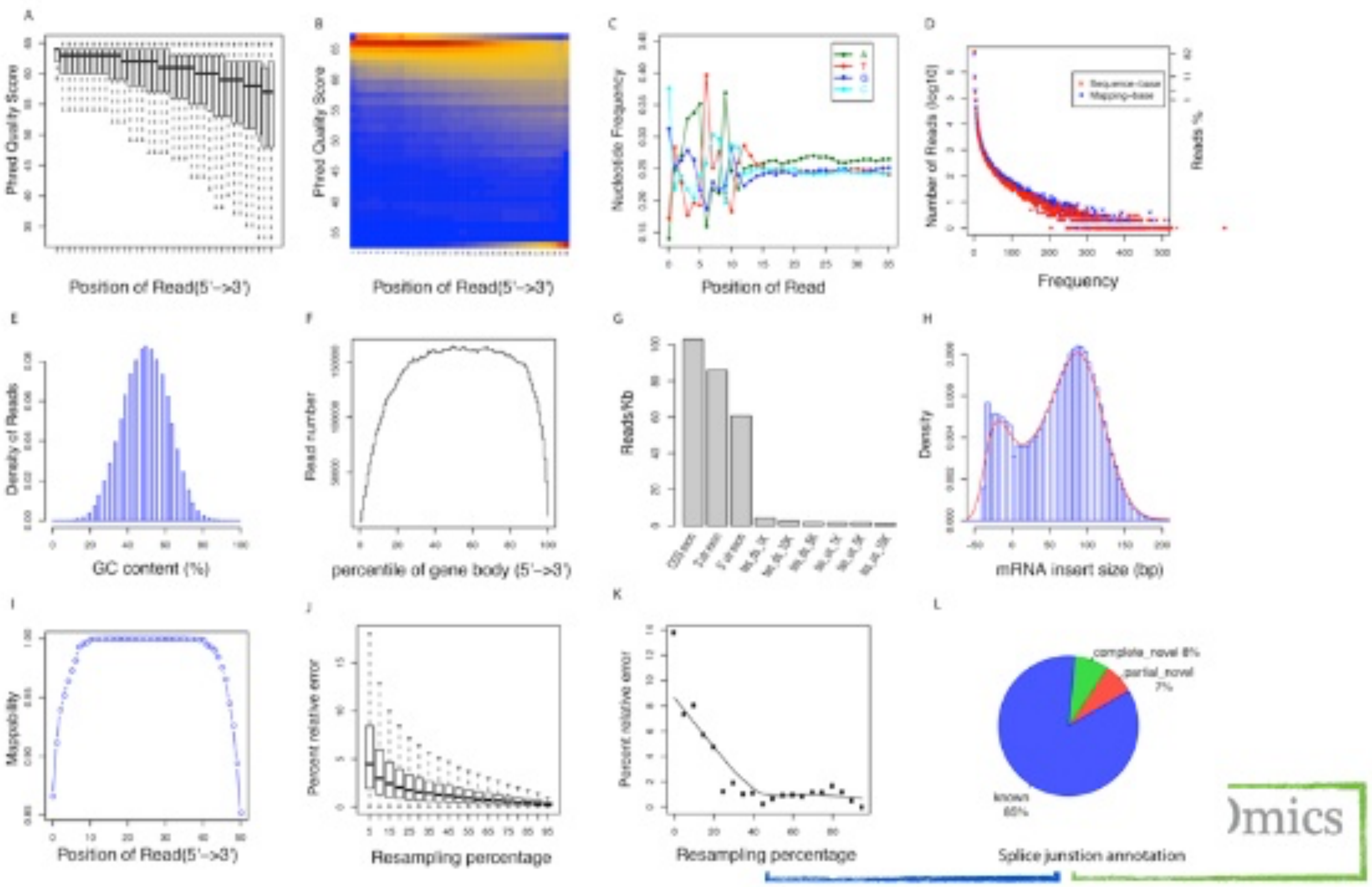
RSeQC package provides a number of useful modules that can comprehensively evaluate high throughput sequence data especially RNA-seq data. “Basic modules” quickly inspect sequence quality, nucleotide composition bias, PCR bias and GC bias, while “RNA-seq specific modules” investigate sequencing saturation status of both splicing junction detection and expression estimation, mapped reads clipping profile, mapped reads distribution, coverage uniformity over gene body, reproducibility, strand specificity and splice junction annotation.

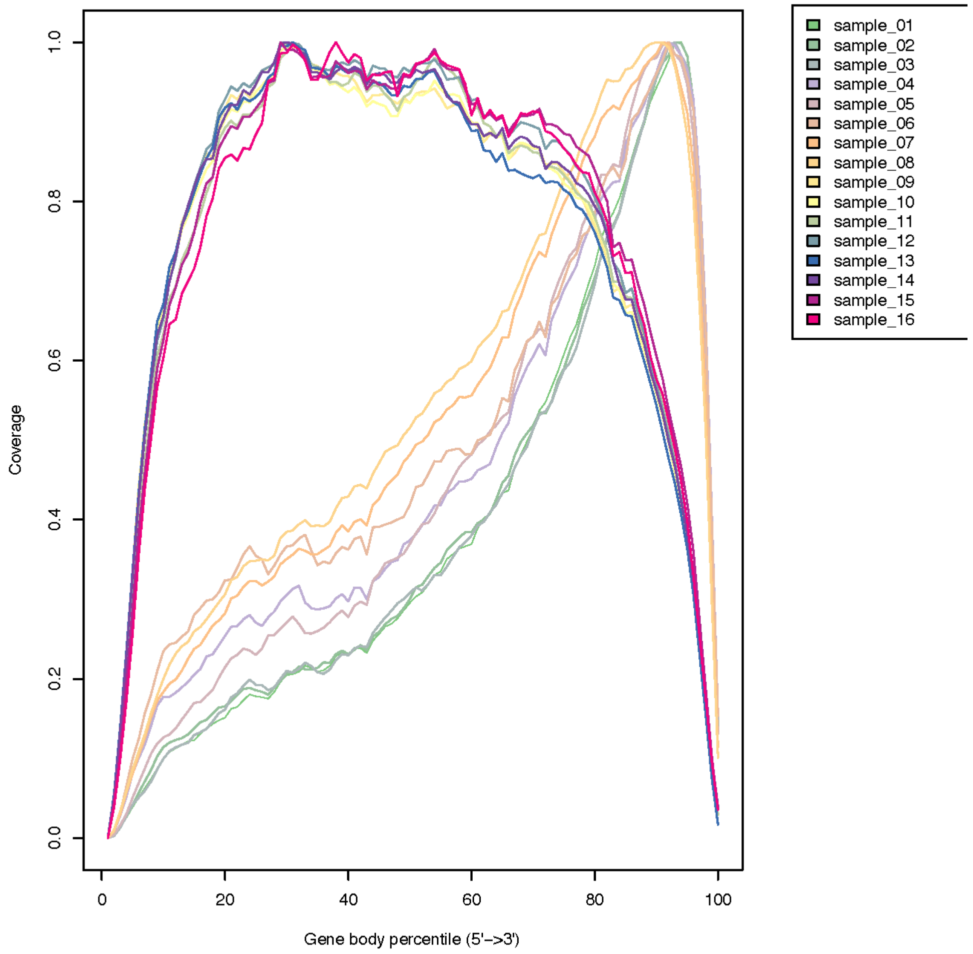
MultiQC is a modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Picard Tools - RNAseqMetrics is a module that produces metrics about the alignment of RNA-seq reads within a SAM file to genes

RSeQC example of plot types

RSEQC





Post Alignment Cleanup

Picard is a set of command line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. (mark pcr duplicates)

Samtools provide various utilities for manipulating alignments in the SAM/BAM format, including sorting, merging, indexing and generating alignments in a per-position format.

BamTools is a command-line toolkit for reading, writing, and manipulating BAM (genome alignment) files.

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

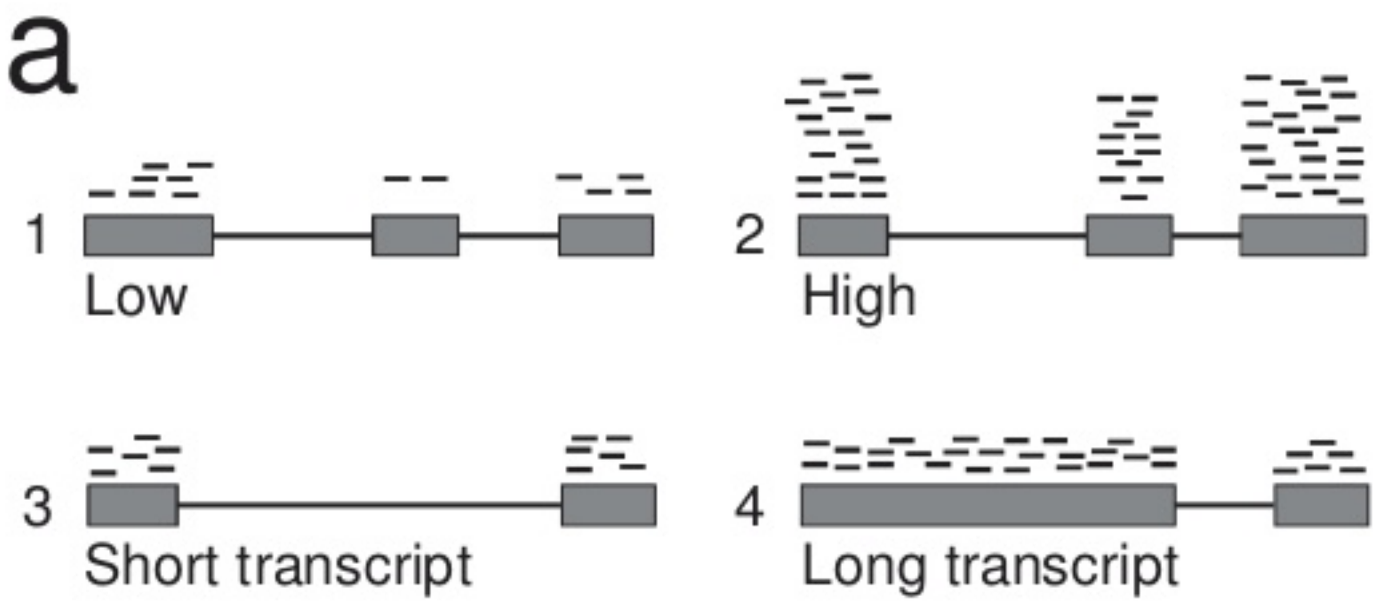
Counting as a measure of Expression

- 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

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/

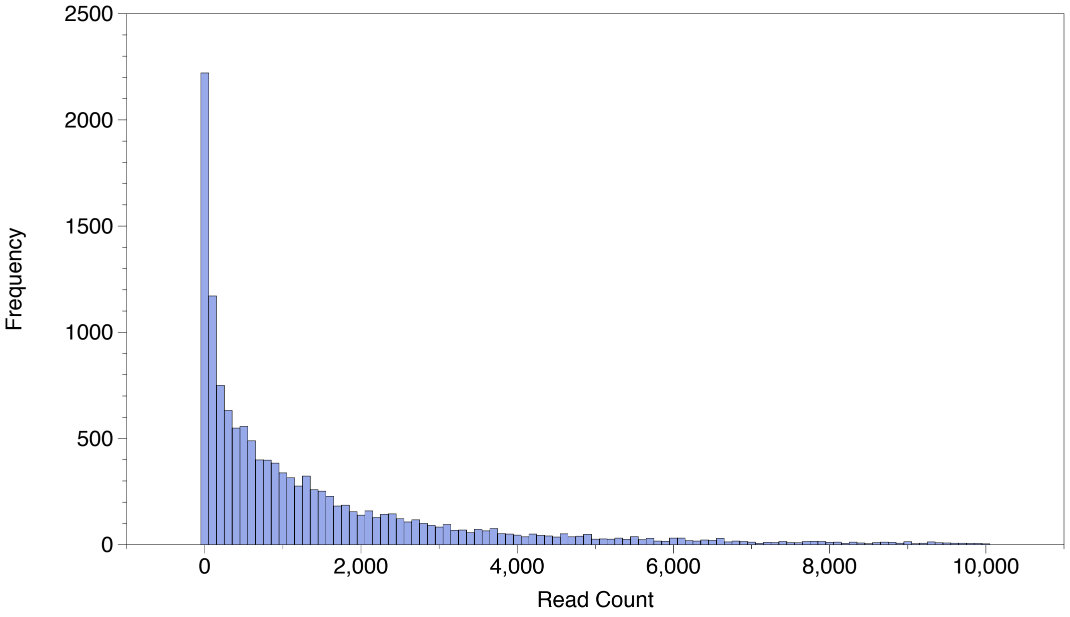
Counting as a measure of Expression

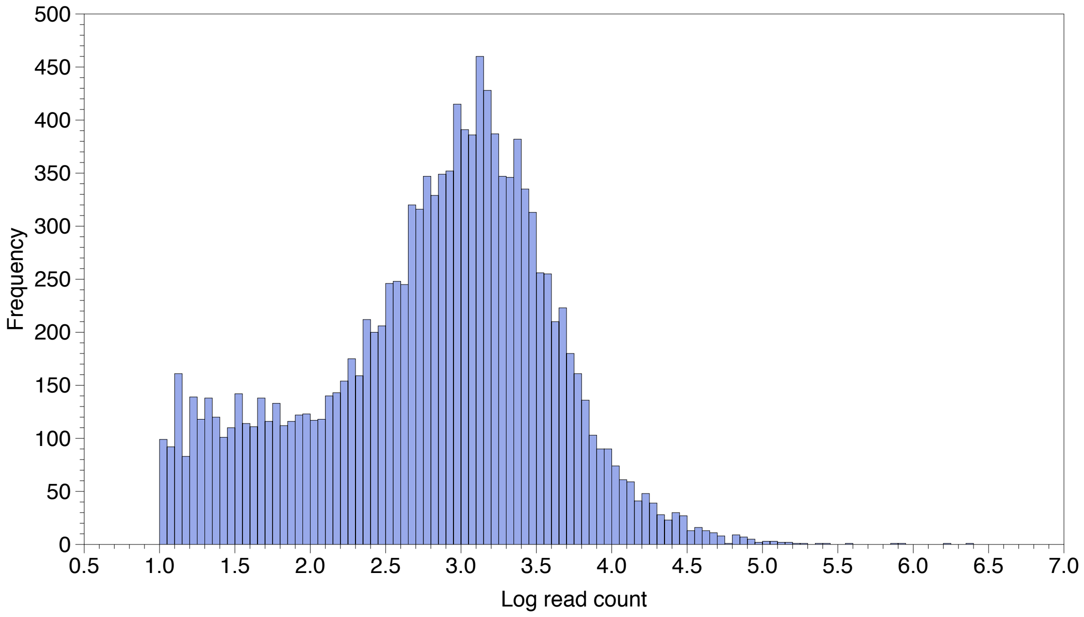
Name	Length	EffectiveLength	TPM	NumReads
ENSG00000121410.12_4	509.732	325.991	3.22494	322.674
ENSG00000268895.6_6	1823.71	1633.86	0.9255	464.119
ENSG00000148584.15_4	5354.1	5164.27	0	0
ENSG00000175899.14_4	4544.77	4354.95	0.039651	53
A2M-AS1	2592.39	2402.54	0.008136	5.999
A2ML1	1749	1561.55	0	0
SLC7A2	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

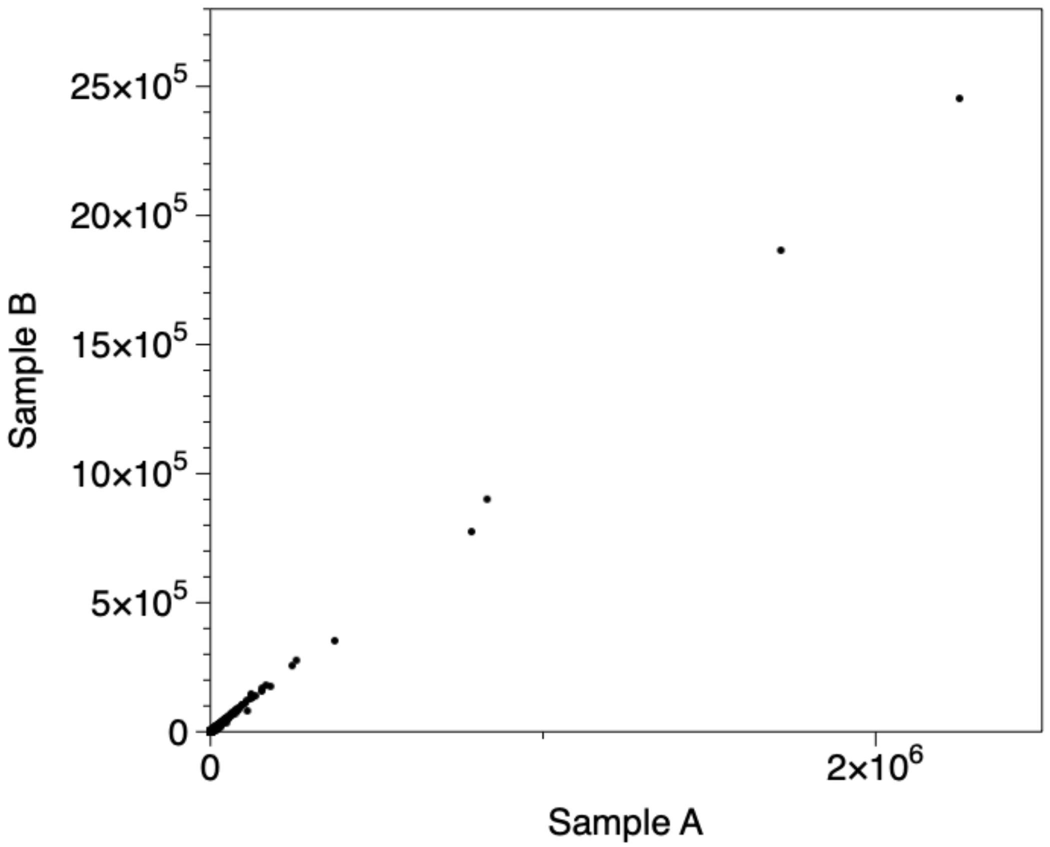
Different ways of annotating the genes

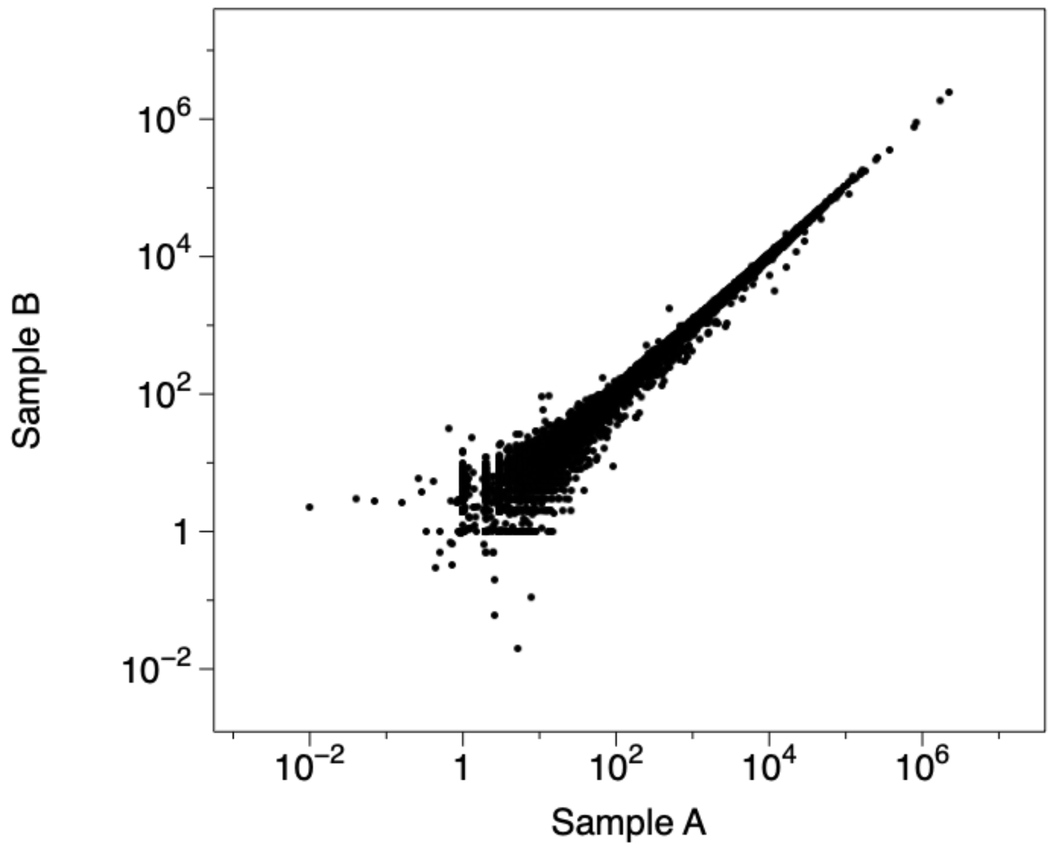
Not always integers -
Decimal values are not acceptable
to some programs

Log Transformed Data









Counting as a measure of Expression

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

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 two different samples.

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). Corrects for variable gene length, read depth.

- Differential Expression - analysis of the difference in expression of genes under two conditions (pair wise comparison) - *expressed as fold difference*.
  A statistical test determines whether the observed difference is statistically significant (i.e. the likelihood of the observation is greater than that expected from random biological variability). Such analyses are typically based on a negative binomial distribution - *expressed as P or corrected P value*.

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.

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.

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

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

Differential Expression Output

EDGER

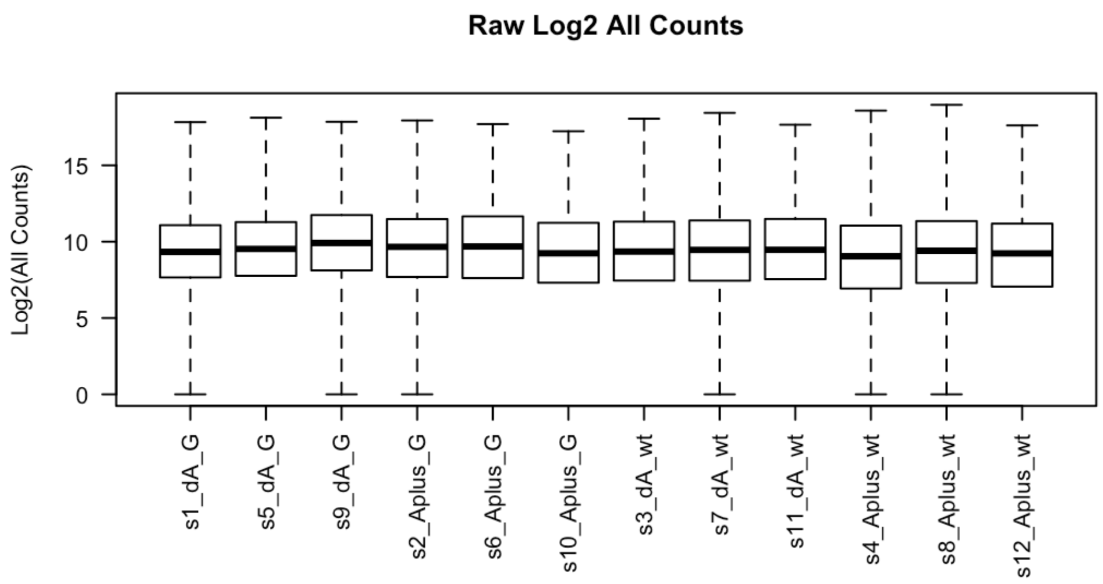
Gene	LogFC	AveExpr	P-Value	FDR
*CA14	-6.72	4.31	1.406716E-10	0.000001
*MCF2L	-10.75	3.25	2.854327E-10	0.000001
*COL5A2	-6.12	4.28	3.678663E-10	0.000001
*TYRP1	-9.31	9.85	4.190114E-10	0.000001
*BCAN	-8.39	5.33	6.384088E-10	0.000001
*CSAG1	10.81	-0.56	7.095577E-10	0.00000

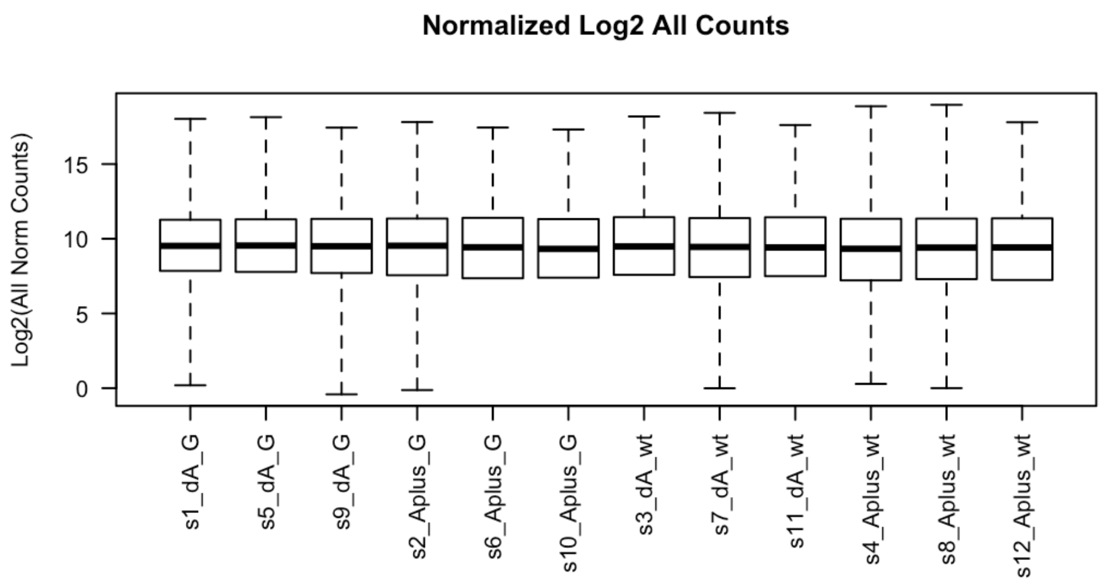
DESEQ2

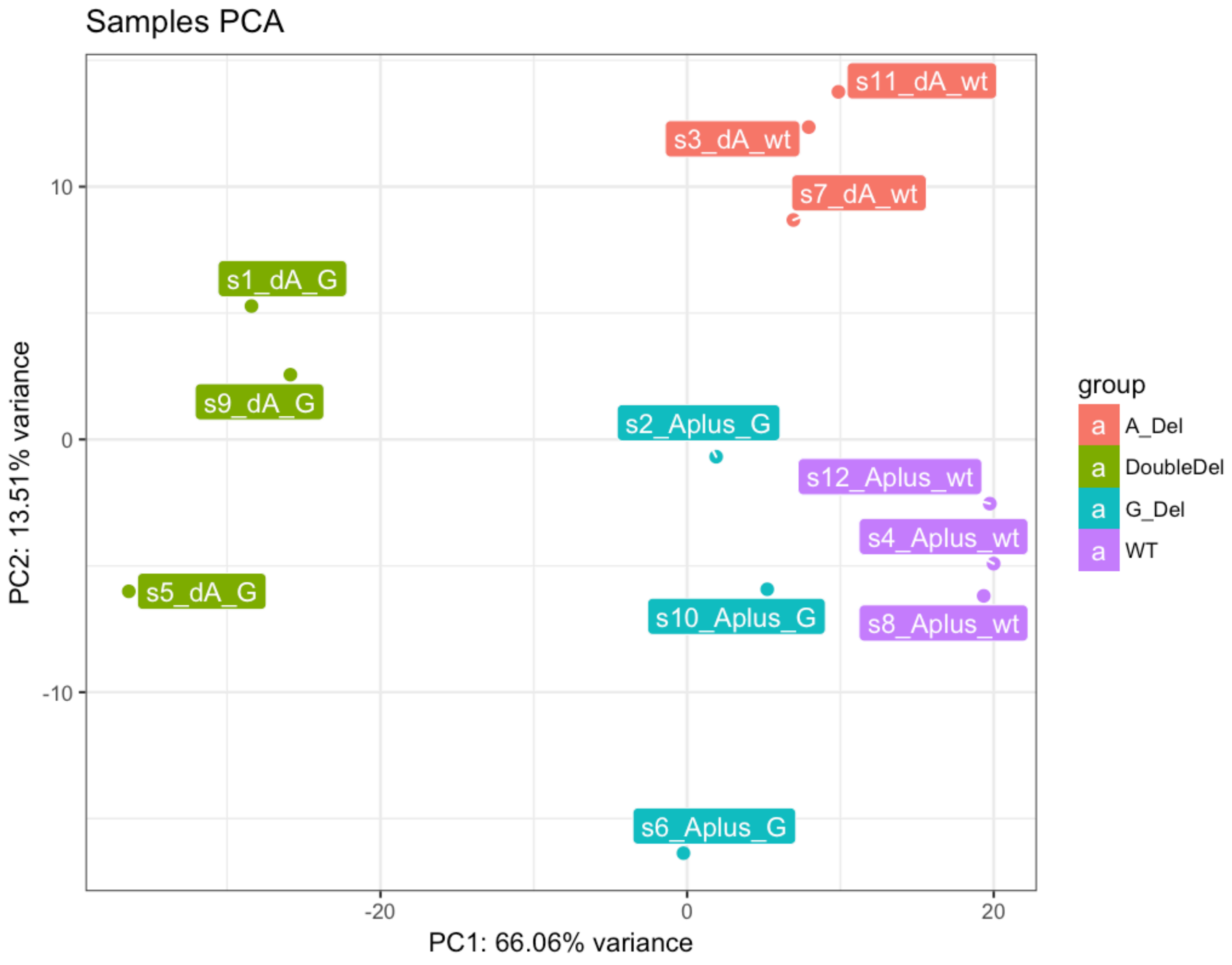
Row-names	Symbol	log2FoldChange	padj	p53_mock_1	p53_mock_2	p53_mock_3	p53_mock_4	p53_IR_1	p53_IR_2	p53_IR_3	p53_IR_4
ENSMUSG00000000001	Gnai3;Gnai3	-0.4763	0.1737	11.584	11.565	11.609	11.621	11.399	11.338	11.997	11.927
ENSMUSG00000000028	Cdc45;Cdc45	-0.4610	0.4125	8.024	7.575	7.668	7.295	7.736	7.675	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

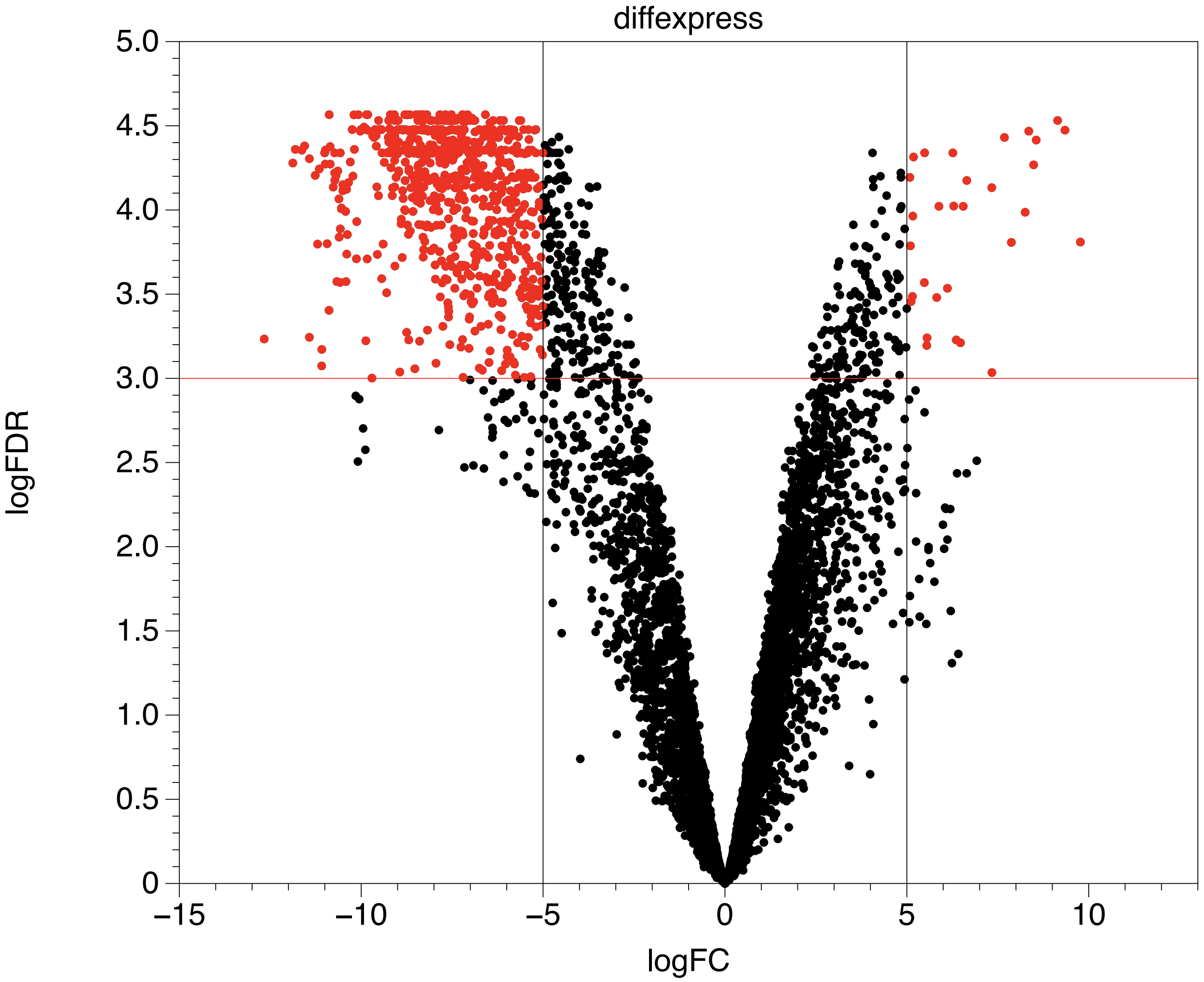
Visualization

Plotting the Data

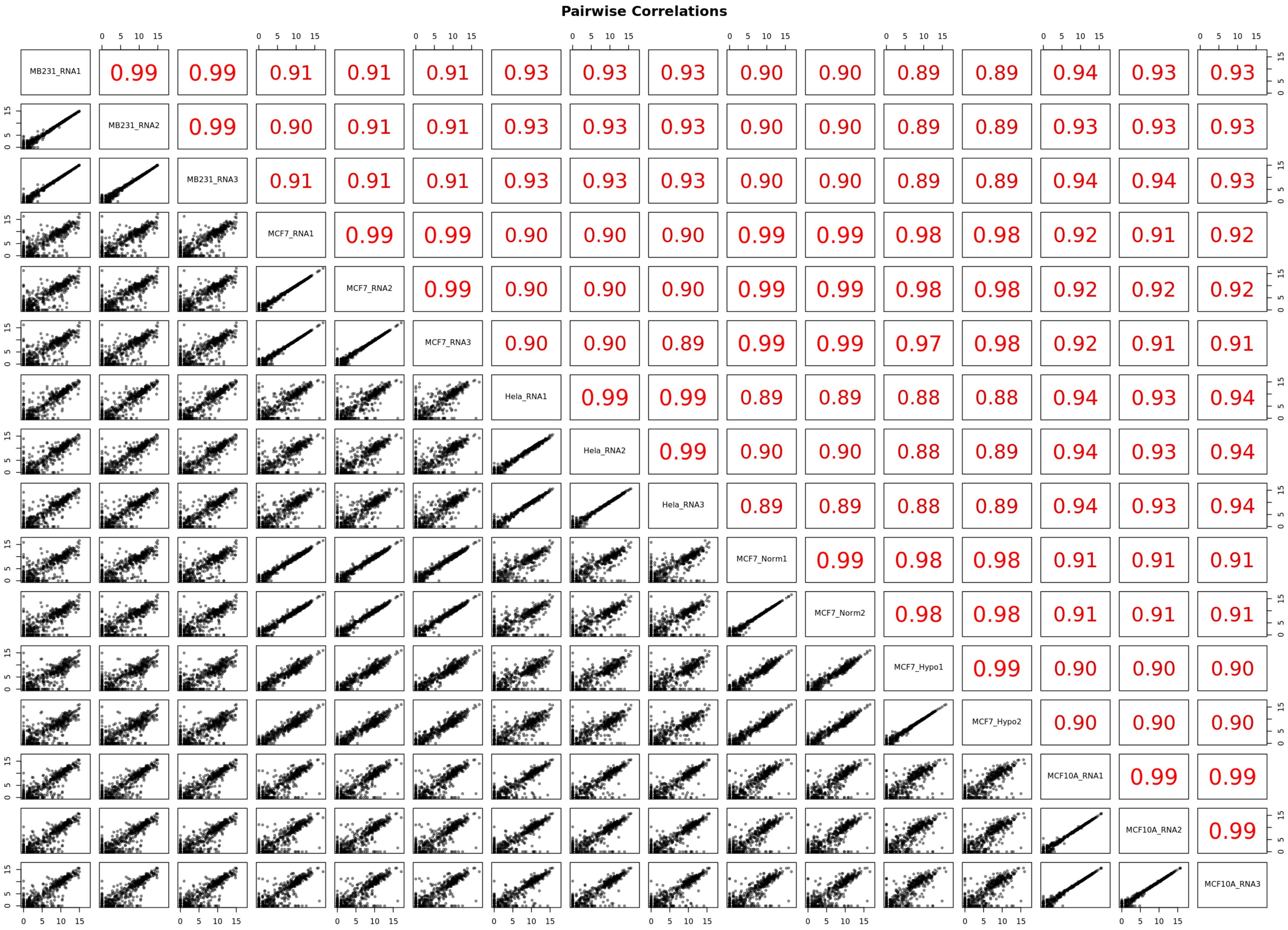

Raw Log2 All Counts


Normalized Log2 All Counts

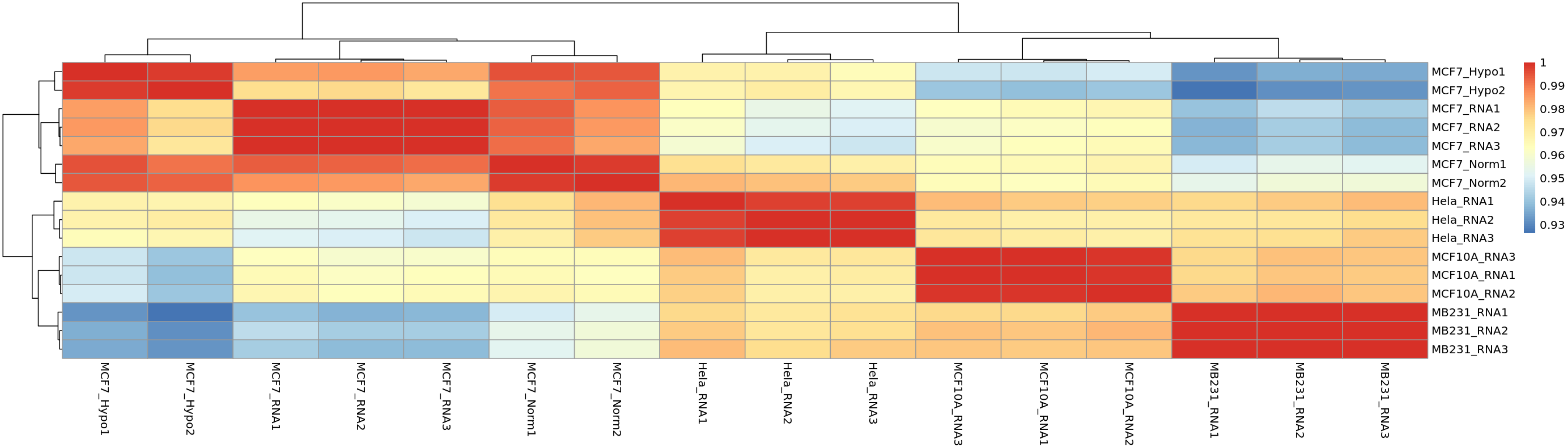

Samples PCA


diffexpress

Plotting the Data

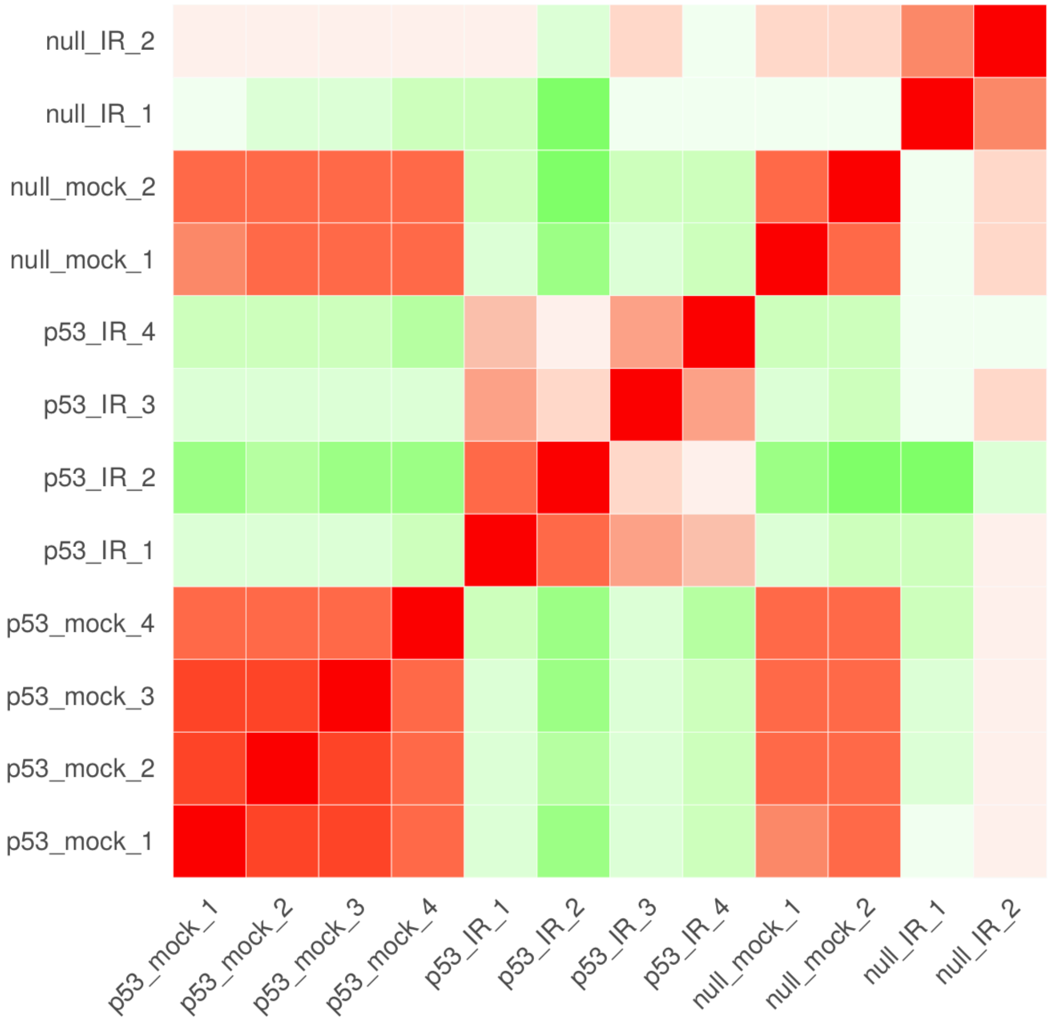


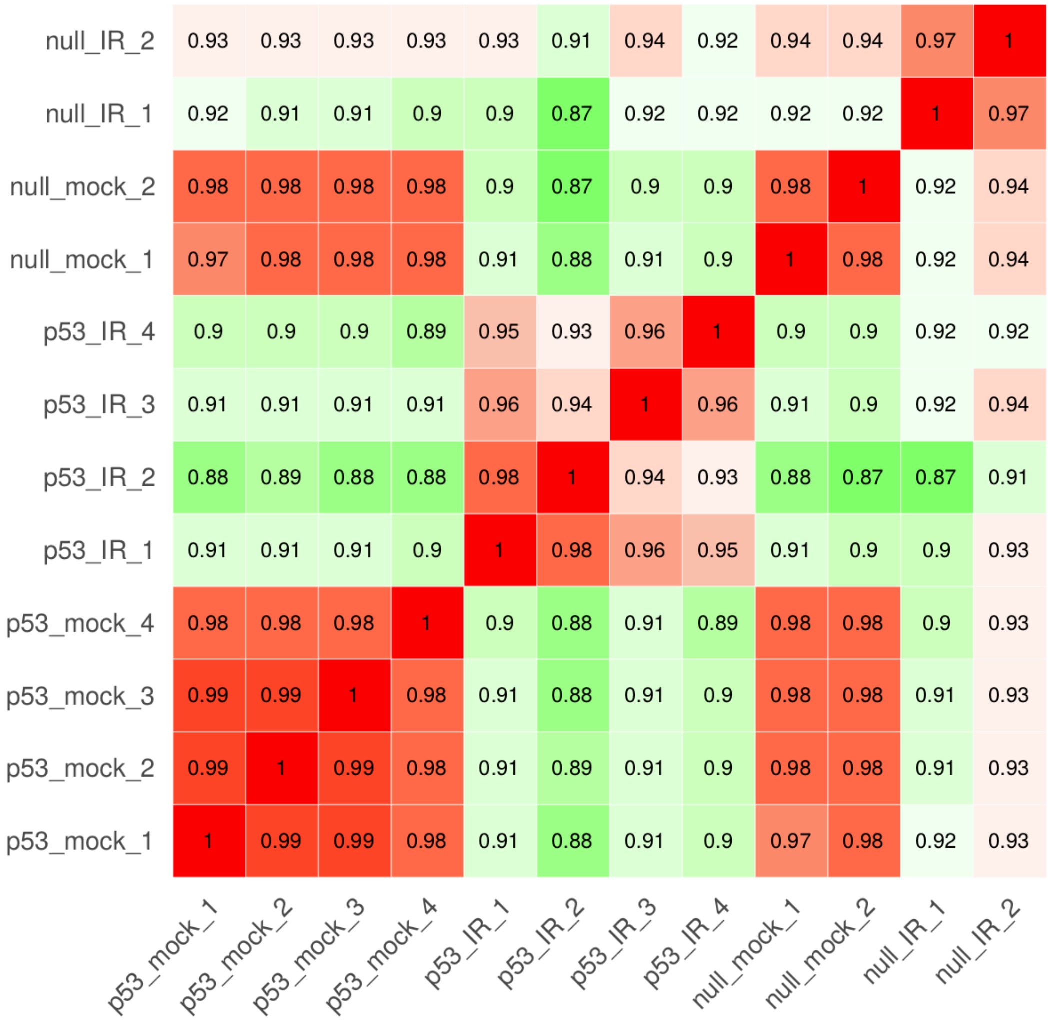
Plotting the Data



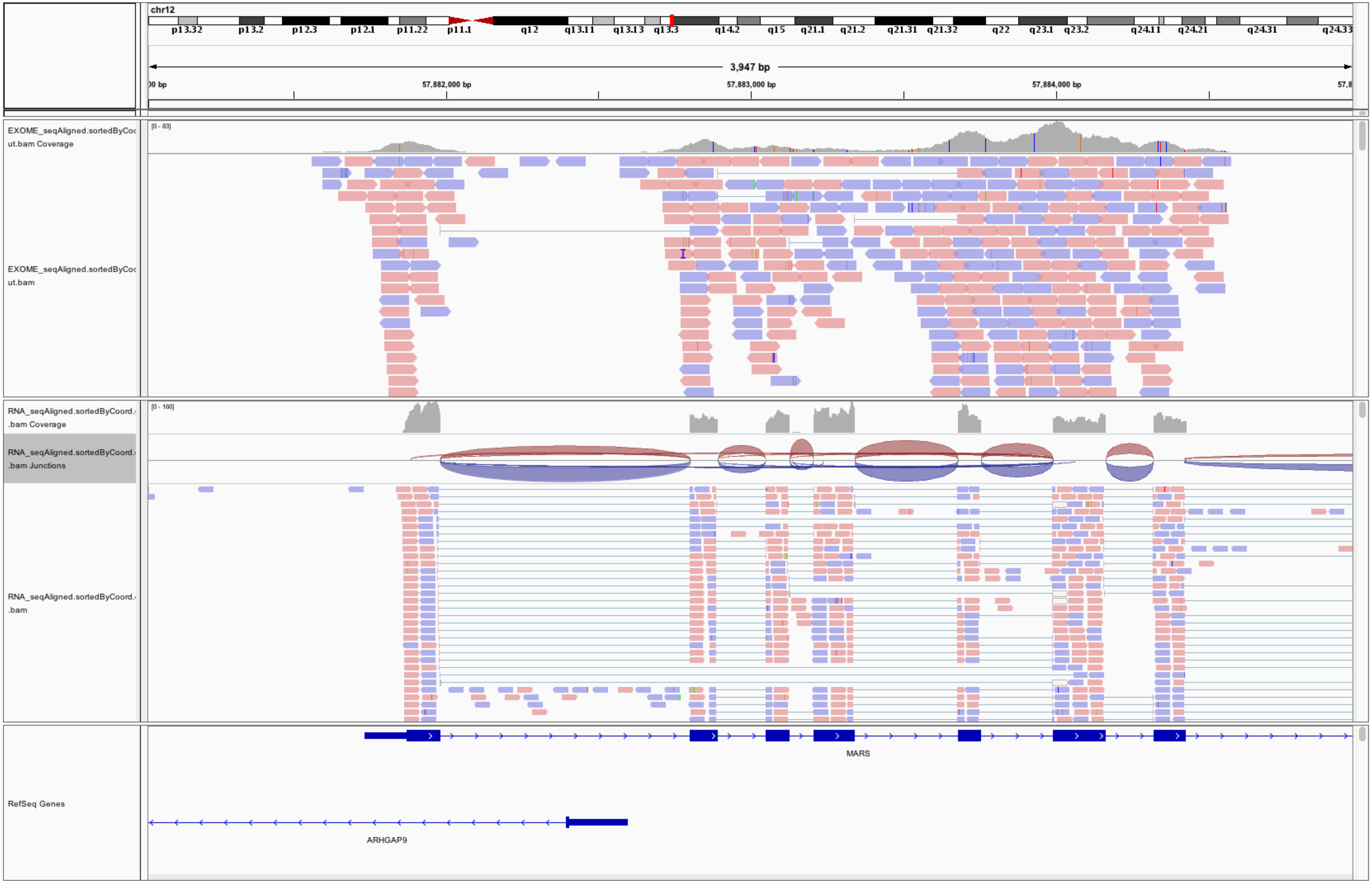
Plotting the Data

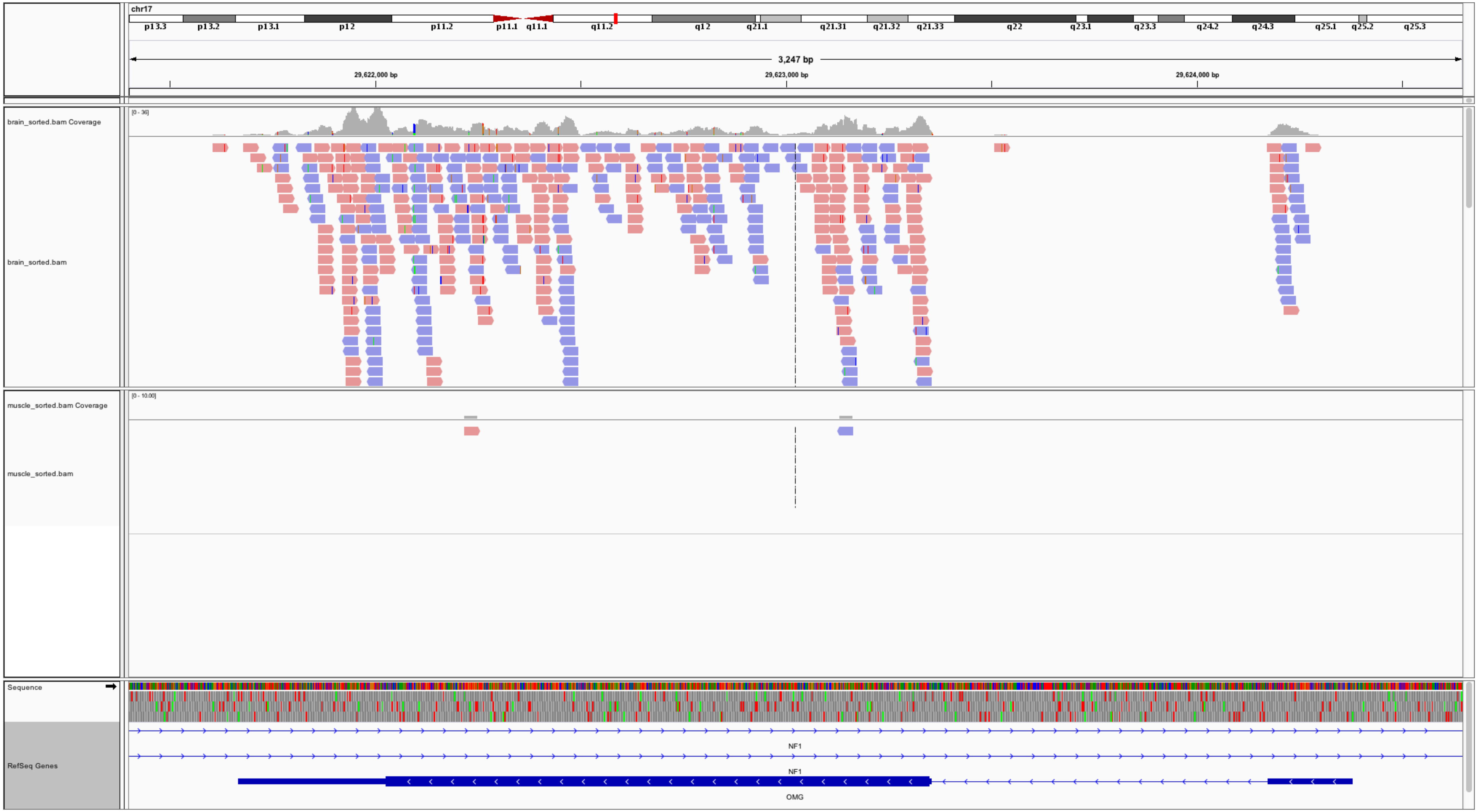
Heat Maps

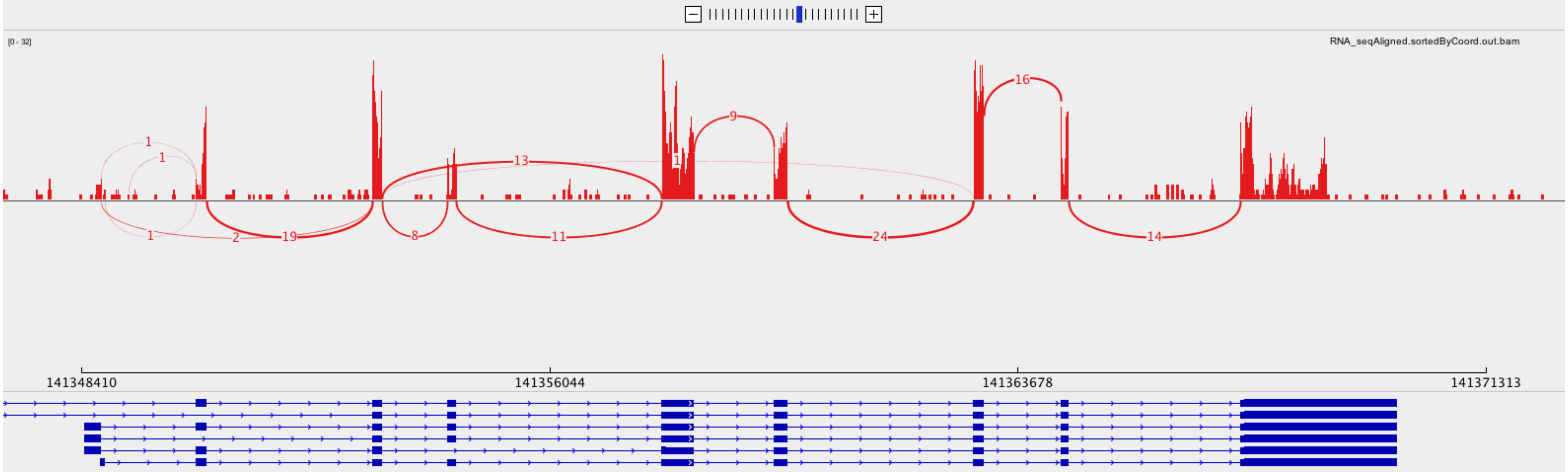




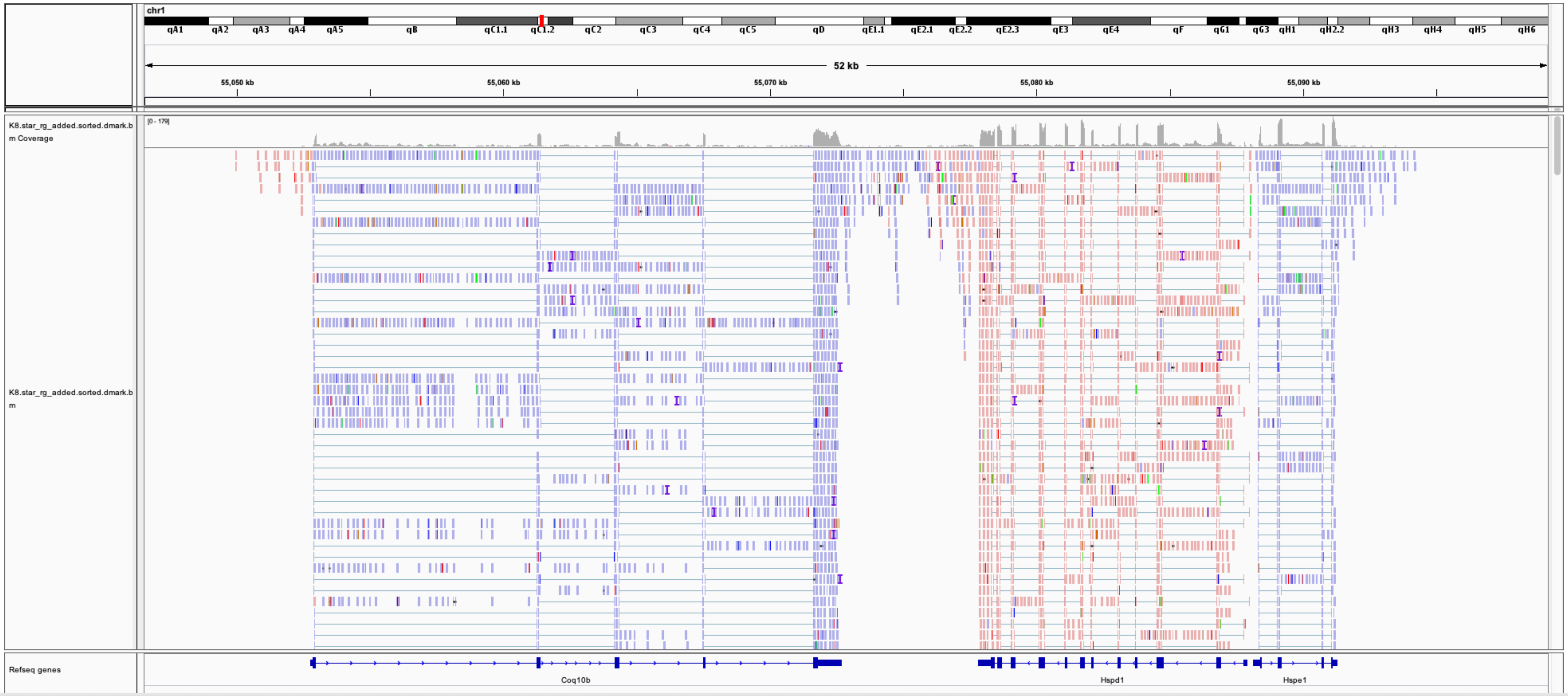
	p53_mock_1	p53_mock_2	p53_mock_3	p53_mock_4	p53_IR_1	p53_IR_2	p53_IR_3	p53_IR_4	null_mock_1	null_mock_2	null_IR_1	null_IR_2
null_IR_2	0.93	0.93	0.93	0.93	0.93	0.91	0.94	0.92	0.94	0.94	0.97	1
null_IR_1	0.92	0.91	0.91	0.9	0.9	0.87	0.92	0.92	0.92	0.92	1	0.97
null_mock_2	0.98	0.98	0.98	0.98	0.9	0.87	0.9	0.9	0.98	1	0.92	0.94
null_mock_1	0.97	0.98	0.98	0.98	0.91	0.88	0.91	0.9	1	0.98	0.92	0.94
p53_IR_4	0.9	0.9	0.9	0.89	0.95	0.93	0.96	1	0.9	0.9	0.92	0.92
p53_IR_3	0.91	0.91	0.91	0.91	0.96	0.94	1	0.96	0.91	0.9	0.92	0.94
p53_IR_2	0.88	0.89	0.88	0.88	0.98	1	0.94	0.93	0.88	0.87	0.87	0.91
p53_IR_1	0.91	0.91	0.91	0.9	1	0.98	0.96	0.95	0.91	0.9	0.9	0.93
p53_mock_4	0.98	0.98	0.98	1	0.9	0.88	0.91	0.89	0.98	0.98	0.9	0.93
p53_mock_3	0.99	0.99	1	0.98	0.91	0.88	0.91	0.9	0.98	0.98	0.91	0.93
p53_mock_2	0.99	1	0.99	0.98	0.91	0.89	0.91	0.9	0.98	0.98	0.91	0.93
p53_mock_1	1	0.99	0.99	0.98	0.91	0.88	0.91	0.9	0.97	0.98	0.92	0.93

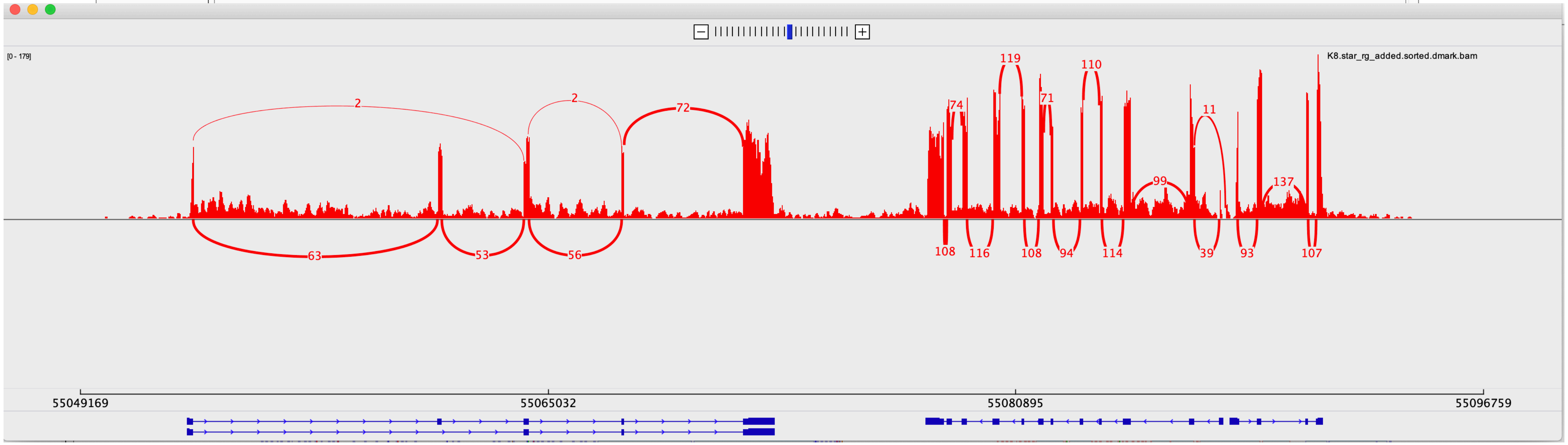

chr12
p13.32
p13.2
p12.3
p12.1
p11.22
p11.1
q12
q13.11
q13.13
q13.3
q14.2
q15
q21.1
q21.2
q21.31
q21.32
q22
q23.1
q23.2
q24.11
q24.21
q24.31
q24.33
3,947 bp
00 bp
57,882,000 bp
57,883,000 bp
57,884,000 bp
57,8
EXOME_seqAligned.sortedByCoo ut.bam Coverage
[0 - 83]
EXOME_seqAligned.sortedByCoo ut.bam
RNA_seqAligned.sortedByCoord. .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
11
9
1
24
16
14
141348410
141356044
141363678
141371313

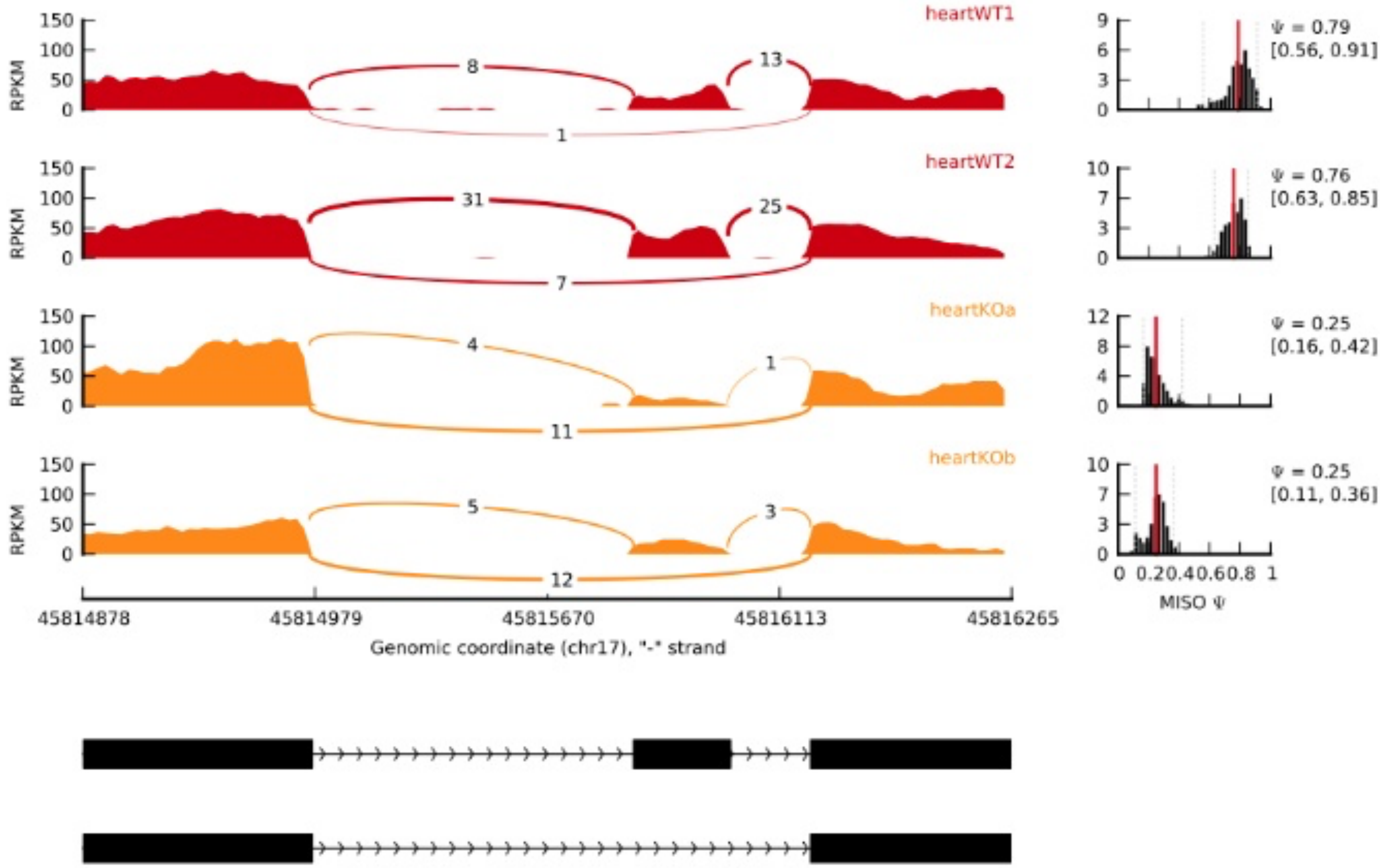
Stranded RNA-Seq Data





Visualizing Splicing

chr17:45816186:45816265:-@chr17:45815912:45815950:-@chr17:45814875:45814965:-



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**

Software Solutions

CCR staff have access to a number of resources

- Biowulf (Helix) - CIT maintained large cluster with a huge software library (Unix command line)
- CCBR Pipeliner (Biowulf)
- Partek Flow (Local Web Service)
- DNAnexus (Cloud Solution)
- CLCBio Genomic Workbench (Small genomes)

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
BTEP	ncibtep@nih.gov