Introduction to RNA-Seq Data Analysis

Peter FitzGerald, PhD

Head, Genome Analysis Unit

Director of BTEP

CCR, NCI

Talk Outline

What will be covered?

- General principles of RNA-Seq
- Guidance on best practices for experimental design
- A walk-through of the steps involved in RNA-Seq data analysis
- References to applicable file formats
- References to appropriate software tools and pipelines for RNA-Seq data analysis

What will *NOT* be covered

- How to use individual software tools or pipelines
- How to analyze Single Cell RNA-Seq data

What is RNA-Seq ?

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**

RNA-Seq is only valid within the context of Differential Expression

What is RNA-Seq ?

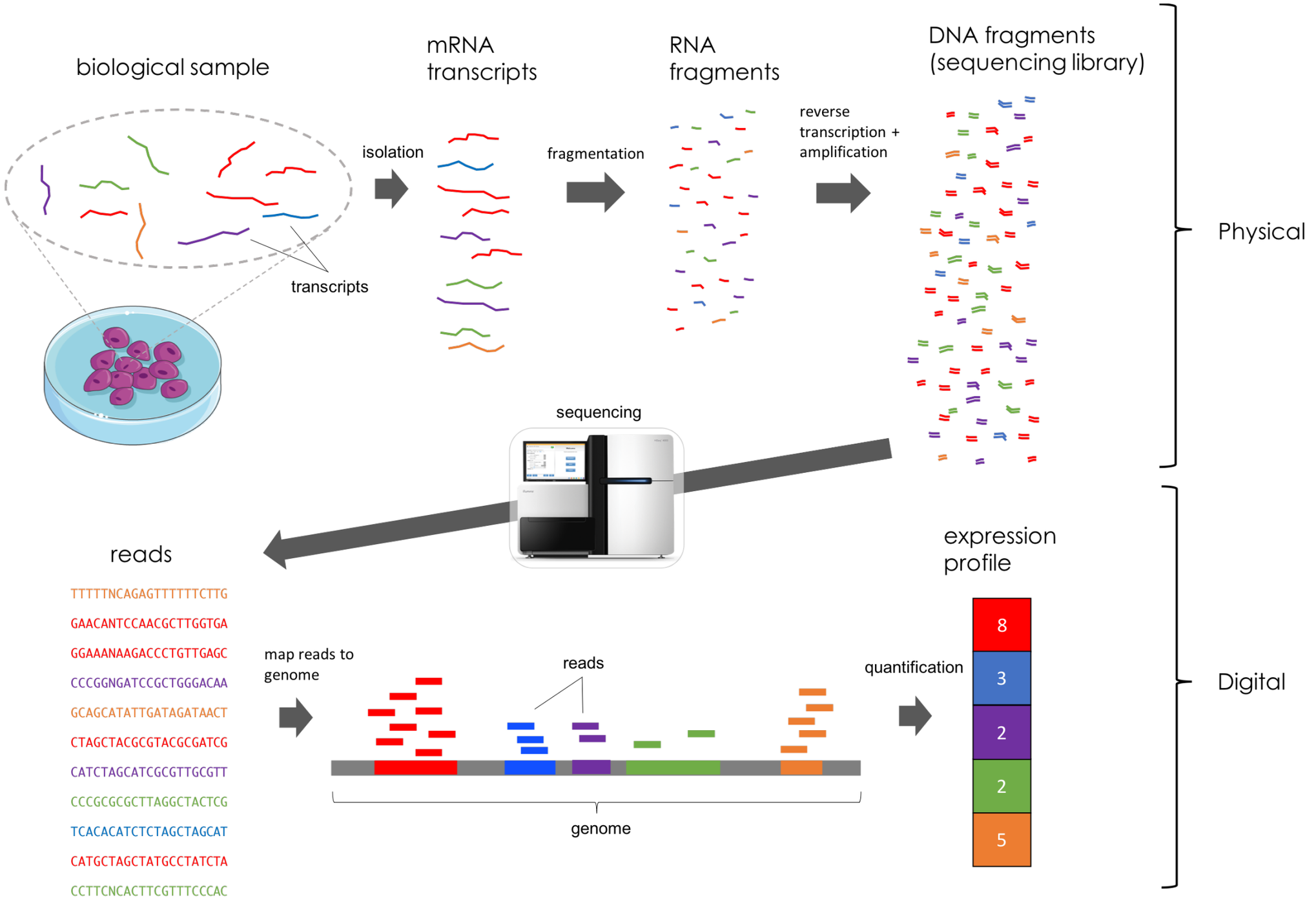


Image from: https://mbernste.github.io/posts/rna_seq_basics/

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]

RNA-Seq - 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

- **Data Analysis (Computation)**

  [*Starting point - fastq (reads) or data-table (counts)***]**

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)

Remember

- RNA-Seq looks at steady state mRNA levels which is the sum of transcription and degradation
- Protein levels are assumed to be driven by mRNA levels
- RNA-Seq can measure relative abundance not absolute abundance
- RNA-Seq is really all about sequencing cDNA

What are the Goals of your Experiment

- What genes are expressed?
- What genes are differentially expressed?
- Are different splicing isoforms expressed?
- Are there novel genes or isoforms expressed?
- Should you be doing targeted long-read sequencing?
- If this a standalone experiment, a pilot, or a “fishing trip” ?

The answers to these questions should guide you in the sequencing technology to use and analytic roadmap to follow.

Read Choices

- **Read Depth**
  - More depth needed for lowly expressed genes
  - Detecting low fold differences need more depth
- **Read Length**
  - Longer reads are 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 RNA-Seq experiments
- **Biological Replicates** (Always useful/necessary)
  - 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

You need replicates

Batch Effects

Variations in samples NOT due to biological effects

- Differences in sample treatment
  - Samples processed on different days/times
  - Samples processed by different people
  - Samples sequenced at different times/lanes/machines
  - Samples are a mixture of different sexes

If all samples cannot be treated the same, never process all treatment or control samples is a single "batch"

Avoid at All Costs !

Data Analysis Questions

- Where will the primary data be stored (fastq)?
  Data Management Environment (DME)
- 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 whom? (what metadata will they require)**
- Are you doing reproducible science?

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

Sample Preparation

Costs (mRNA total)

CCR Sequencing Facility (subsidized pricing**)**

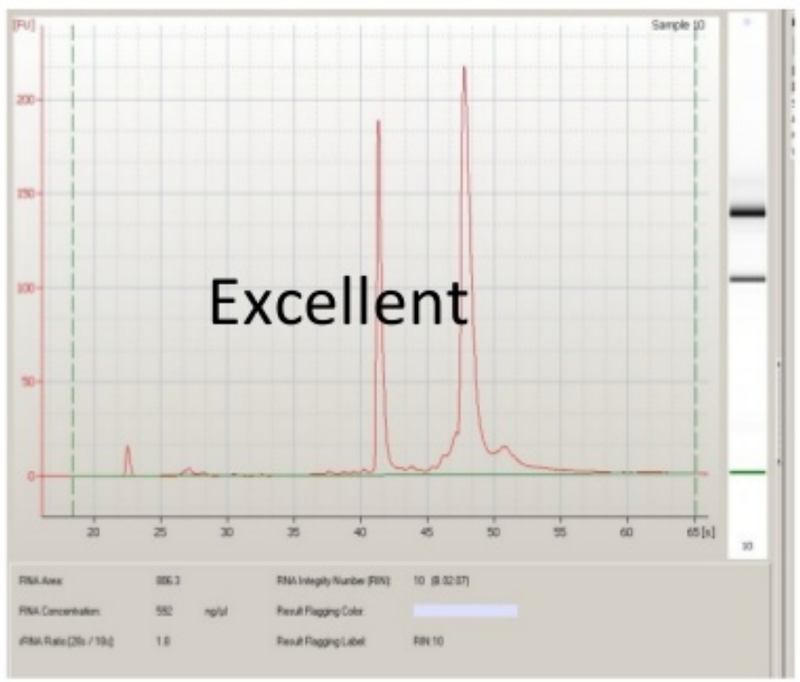
Library Construction	$87	
Illumina HiSeq 4000	$1007/lane	PE 2 x 75 (all 8 lanes)
Illumina NovaSeq	$4382/lane	1 x 100 bp
Illumina NextSeq High Output	$1956	2 x 75 bp (V2)
Illumina MiSeq	$623	PE 2 x 75 bp (V3)

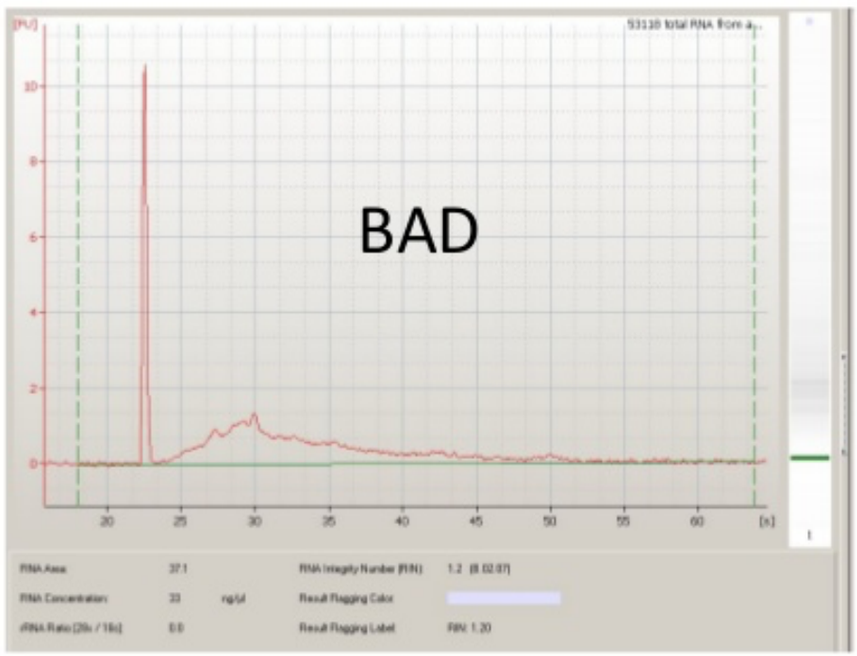
General Rules for Sample Preparation

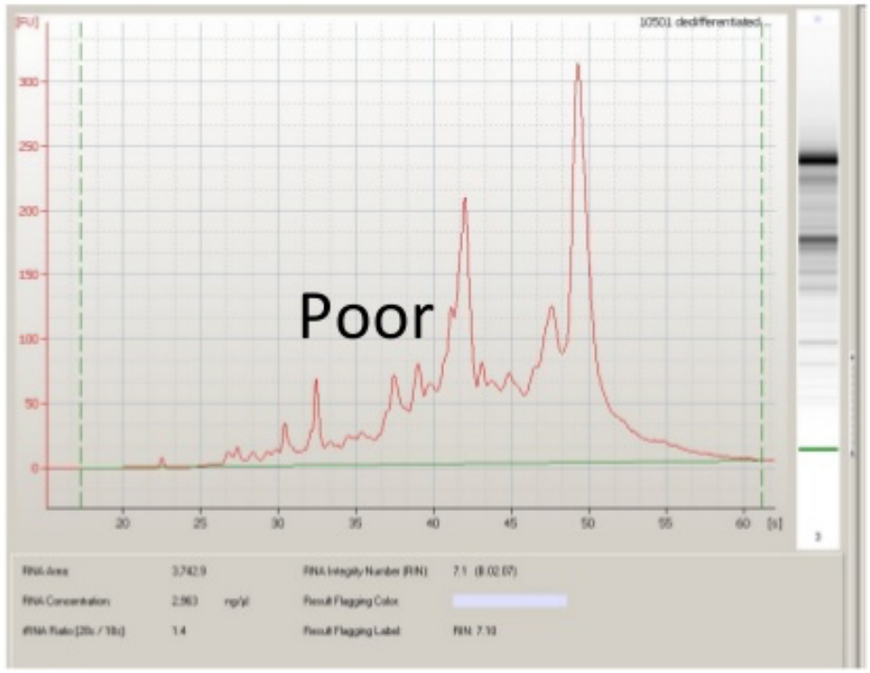
- Prepare all samples at the same time or as close as possible. The same person should prepare all samples
- Do not prepare “experiment” and “control” samples on different days or by different people (Batch effects)
- Use high quality means to determine sample quality (**R**NA **I**ntegrity **N**umber) (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)

Pilot experiments are your friends

Determining Library size distribution







CENTER FOR CANCER COMPUTATIONAL BIOLOGY

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%
0	Coding > 50%	Coding > 35%
Picard RNA-SeqMetrics	Intronic + Intergenic < 25%	Intronic + Intergenic < 40%

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****. For cost and specific information about setting up an RNA-Seq experiment, please visit the Sequencing Facility website or contact Bao Tran*

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.

PacBio Sequel II

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

MinION

GridION

Oxford Nanopore

Data Analysis

RNA-Seq - Data Analysis

What version of the Genome should you align against ?

Sequence and annotation - Same sequence can have different annotations

Factors that determine the answer to this question are:

- Are you trying to match published data or previous experiments?
- Are you interested in a particular type of annotation (GeneID, EnsembleIDs, refseqID, etc.?
- Are you interested in Genes or transcripts?
- If there are no other overriding factors, use the latest genome sequence and annotations (Biowulf has many pre-built)
- Is it desirable to align against the T2T genome
- If no reference genome, you will have to use a different approach

Remember to make note of this choice and advise the core

RNA-Seq Pipeline

RNA-Seq Analysis process can be broken down into two main steps

Primary Analysis

FASTQ -> Count-file

Secondary+ Analysis

Count-file -> Differential Expression, PATHWAY ANALYSIS …

RNA-Seq - 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**

The Sequencing Core may do some or all of this

RNA-Seq - Data Analysis WorkFlow II

- **Quantification**
  - Differential Expression - statistics
- * **Data tables, plots**
- **Visualization**
  - Visual inspection - IGV
  - Data representation - scatter plots, 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

- 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*). You do not 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 built into the whole workflow

Computational Prerequisites

- High performance Linux computer (multicore, 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

OR

- Using pre-built workflows
- Using Cloud/Web resources, with pre-built workflows

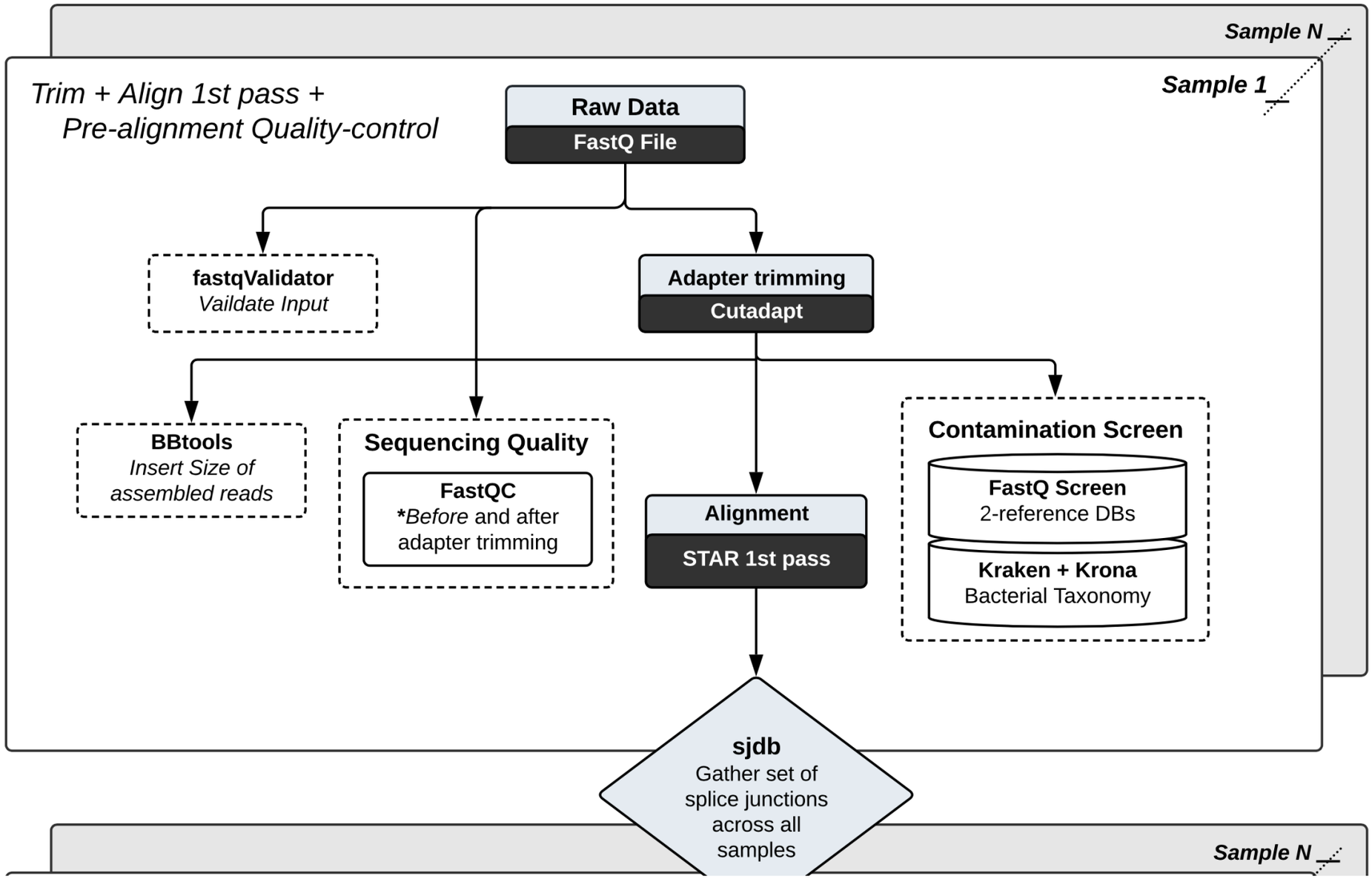
Data Analysis

- Pre-alignment QC & cleanup
- Alignment
- Post-alignment QC & filtering
- Quantification

- Differential Expression
- Biological Interpretation

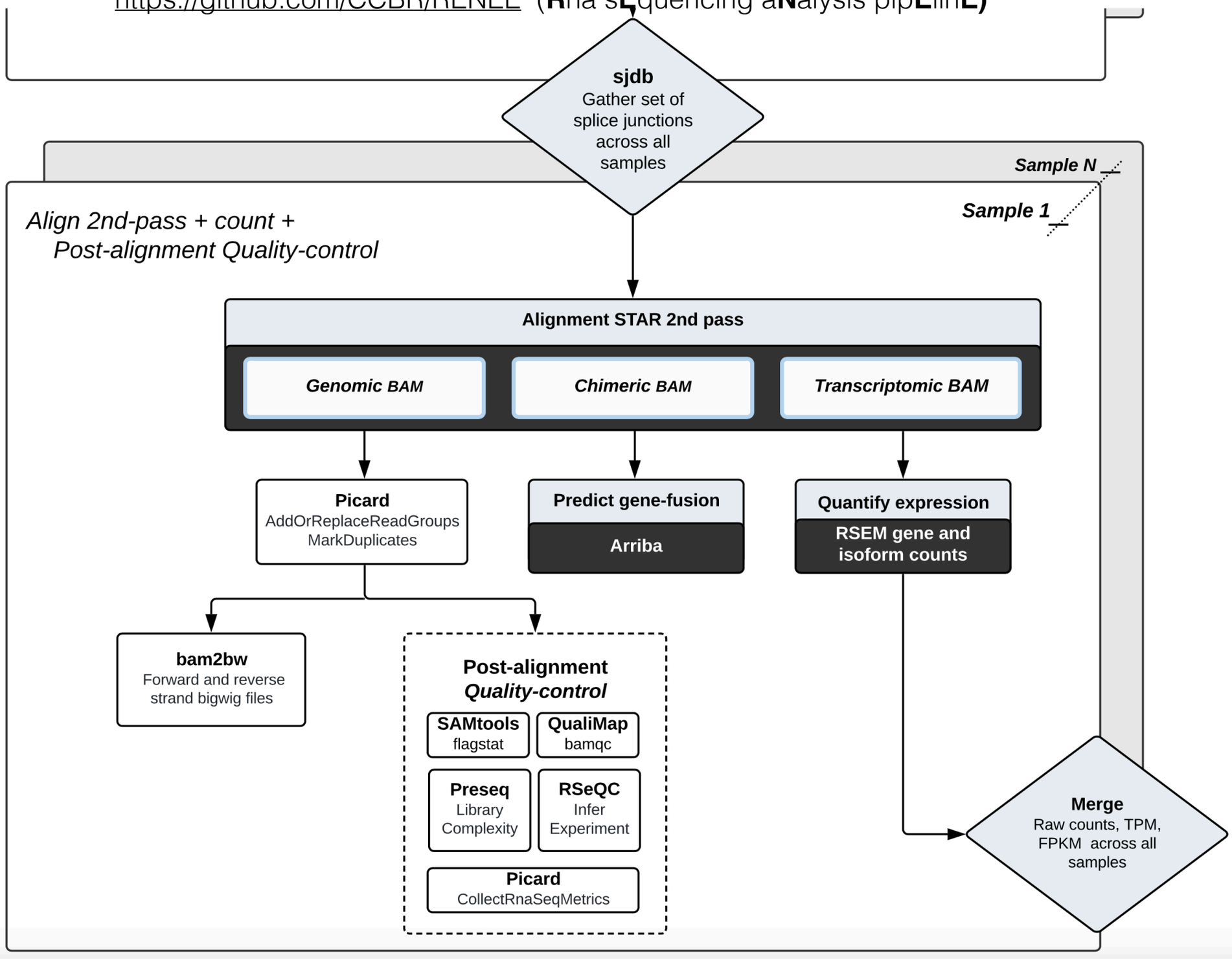
RNA-Seq Pipeline - Primary Analysis

https://github.com/CCBR/RENEE (**R**na s**E**quencing a**N**alysis pip**E**lin**E)**



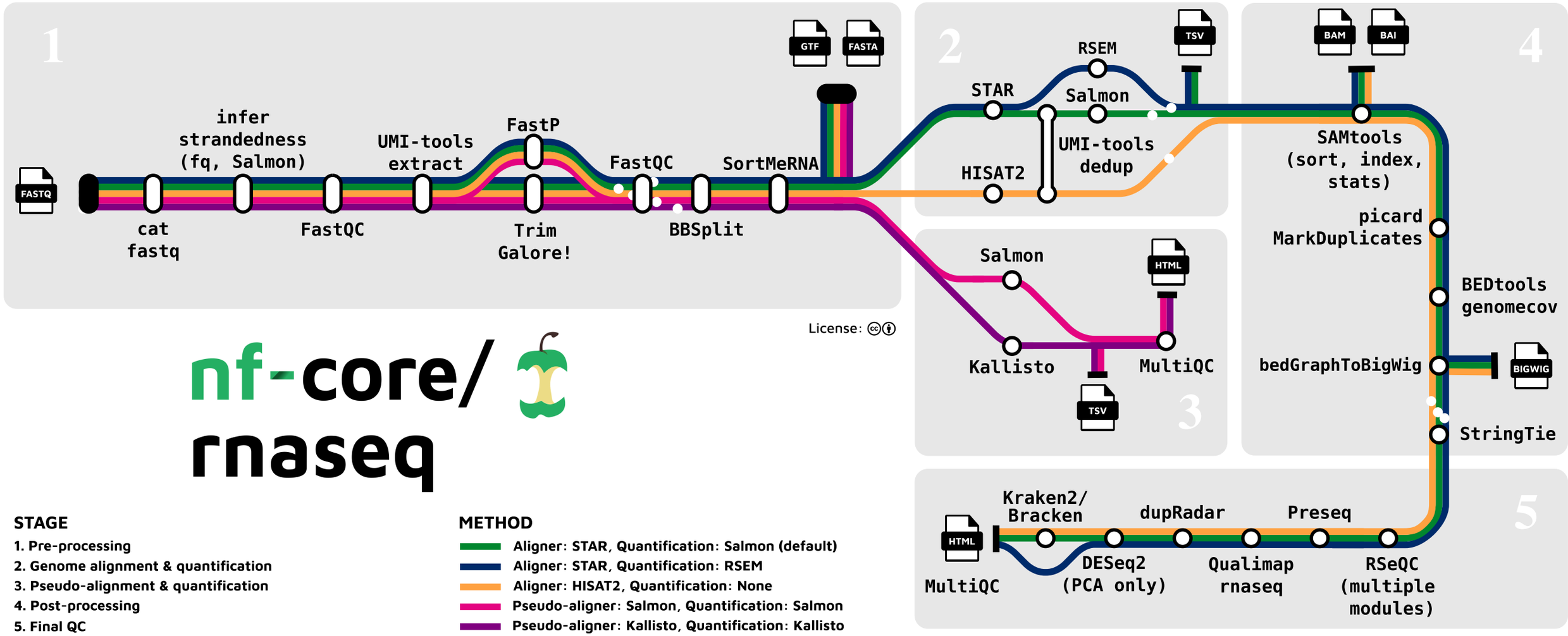
RNA-Seq Pipeline - Primary Analysis

https://github.com/CCBR/RENEE (**R**na s**E**quencing a**N**alysis pip**E**lin**E)**



RNA-Seq Pipeline - Primary Analysis

https://nf-co.re/RNA-Seq



License: CC BY

nf-core/ rnaseq

STAGE

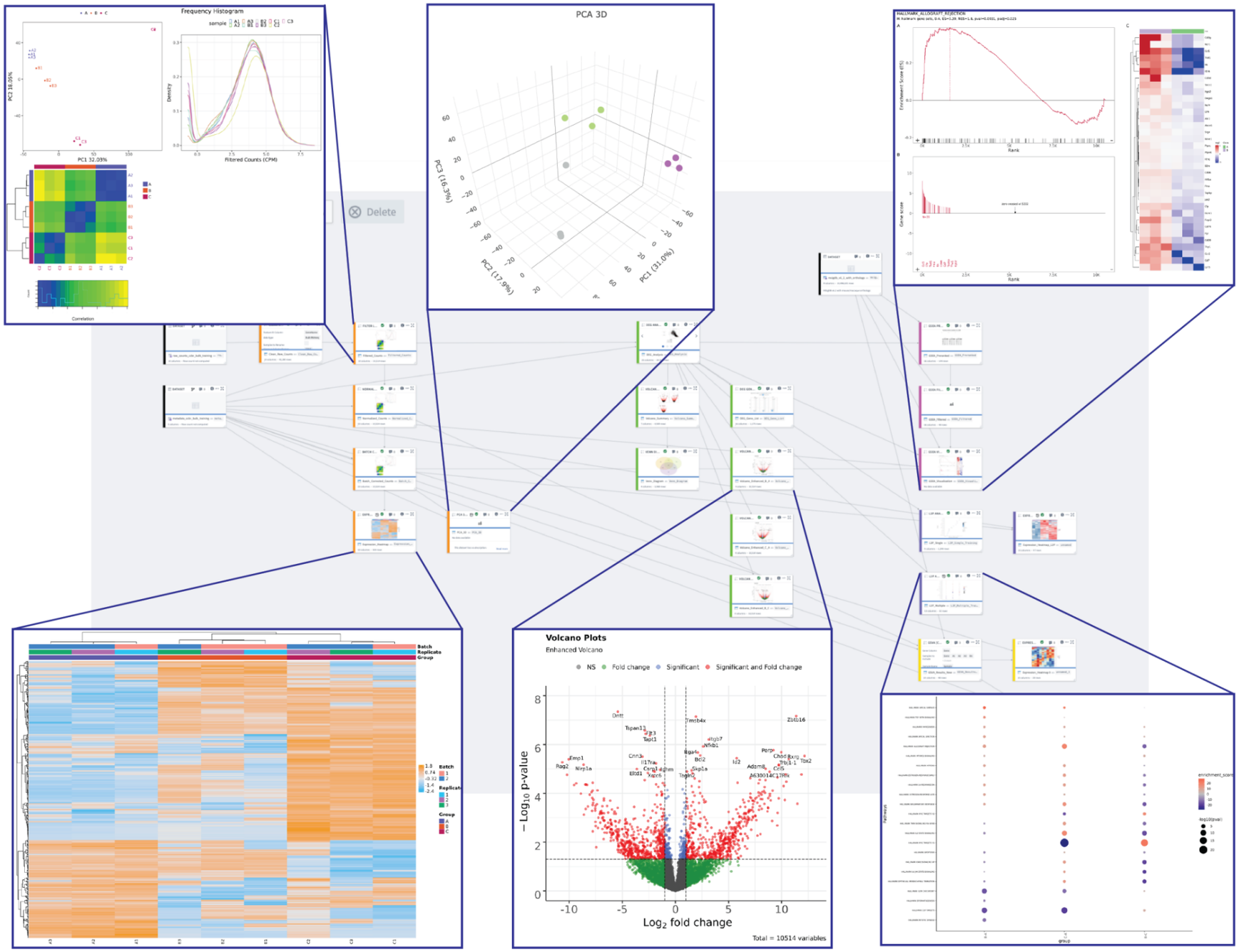
1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

METHOD

- Aligner: STAR, Quantification: Salmon (default)
- Aligner: STAR, Quantification: RSEM
- Aligner: HISAT2, Quantification: None
- Pseudo-aligner: Salmon, Quantification: Salmon
- Pseudo-aligner: Kallisto, Quantification: Kallisto

RNA-Seq Pipeline - Secondary Analysis

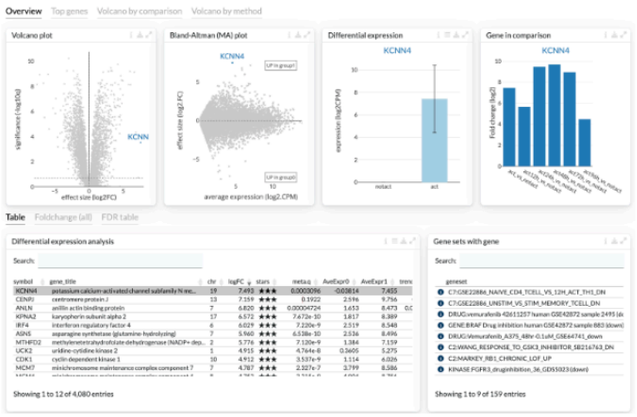
https://nidap.nih.gov/



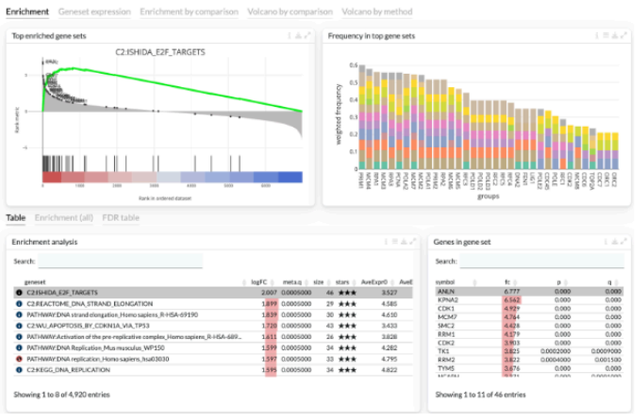
RNA-Seq Pipeline - Secondary Analysis

https://bigomics.ch/rna-seq-data-analysis/

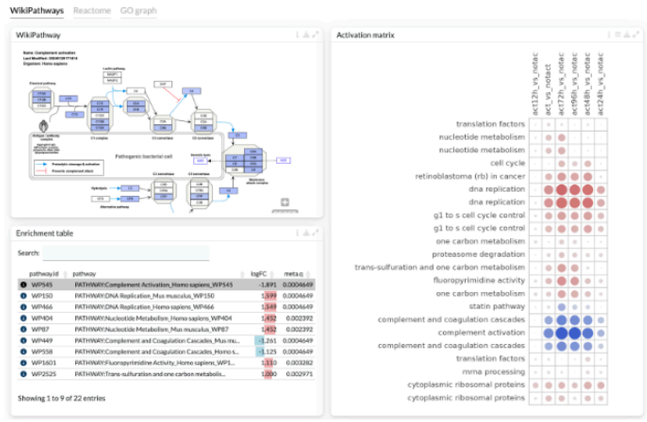
Differential Gene Expression Analysis



Gene Set Enrichment Analysis

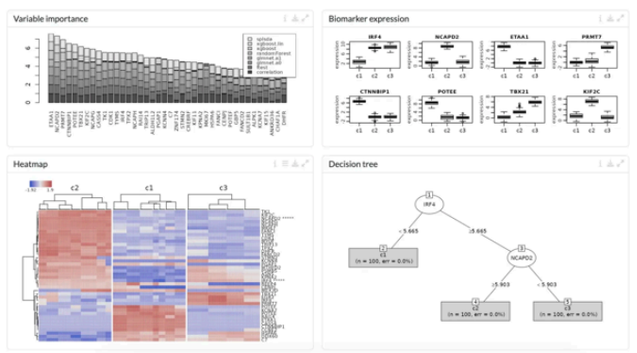


Pathway Analysis

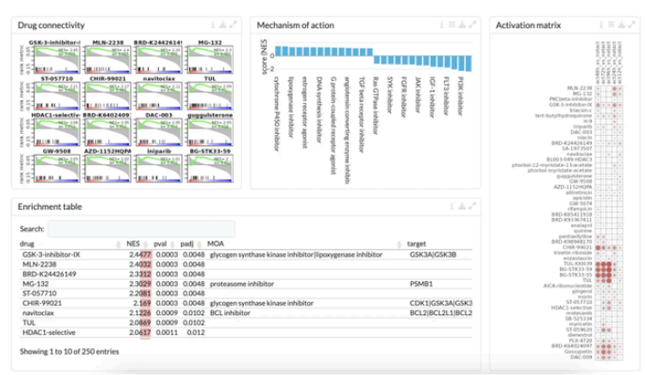


Clustering Analysis

Biomarker Analysis



Drug Connectivity Analysis



Quality Control/Assessment (Pre-Alignment)

Data Quality Assessment

- **Evaluate the read quality to determine**

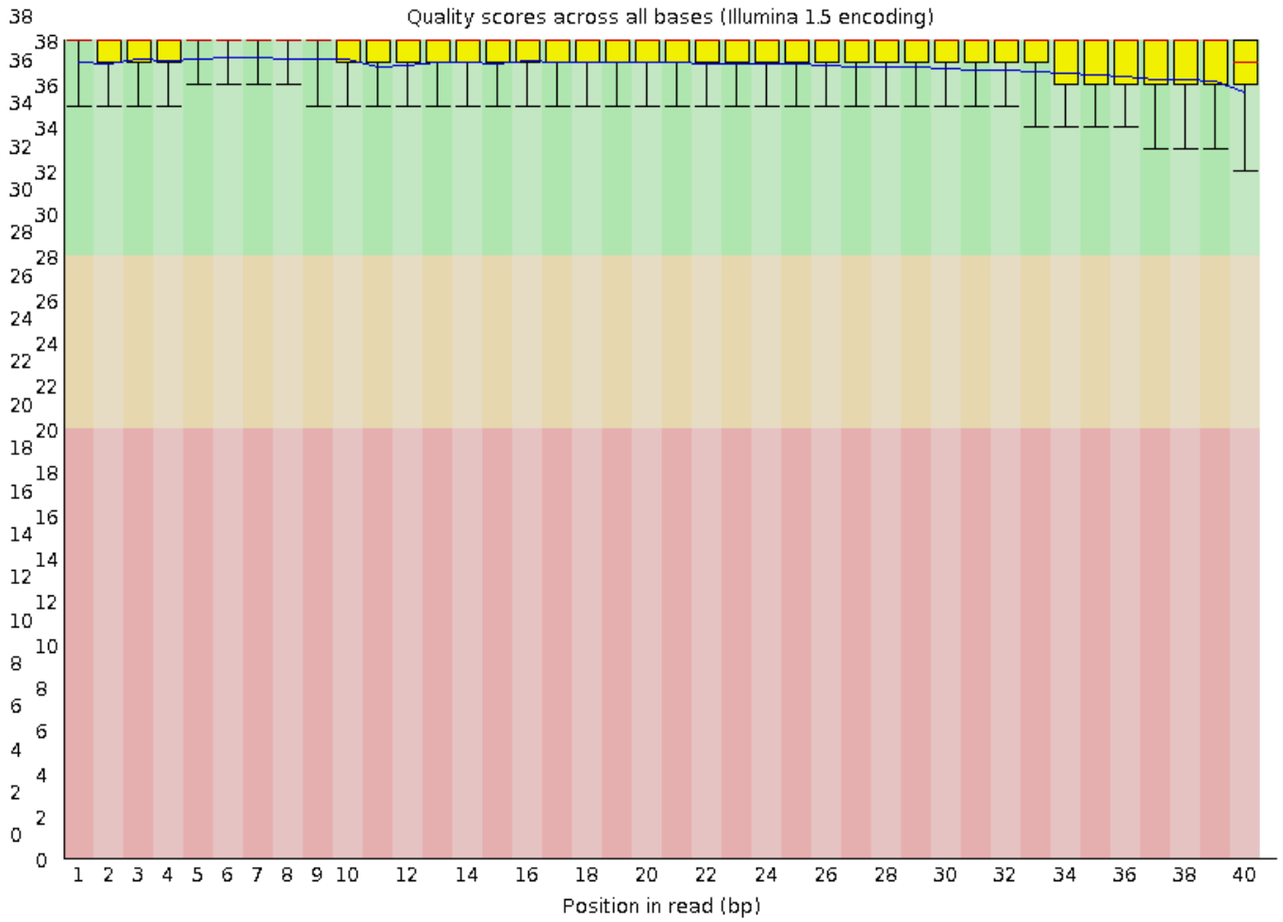
  (Tells us nothing about whether the experiment worked)

  - 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**

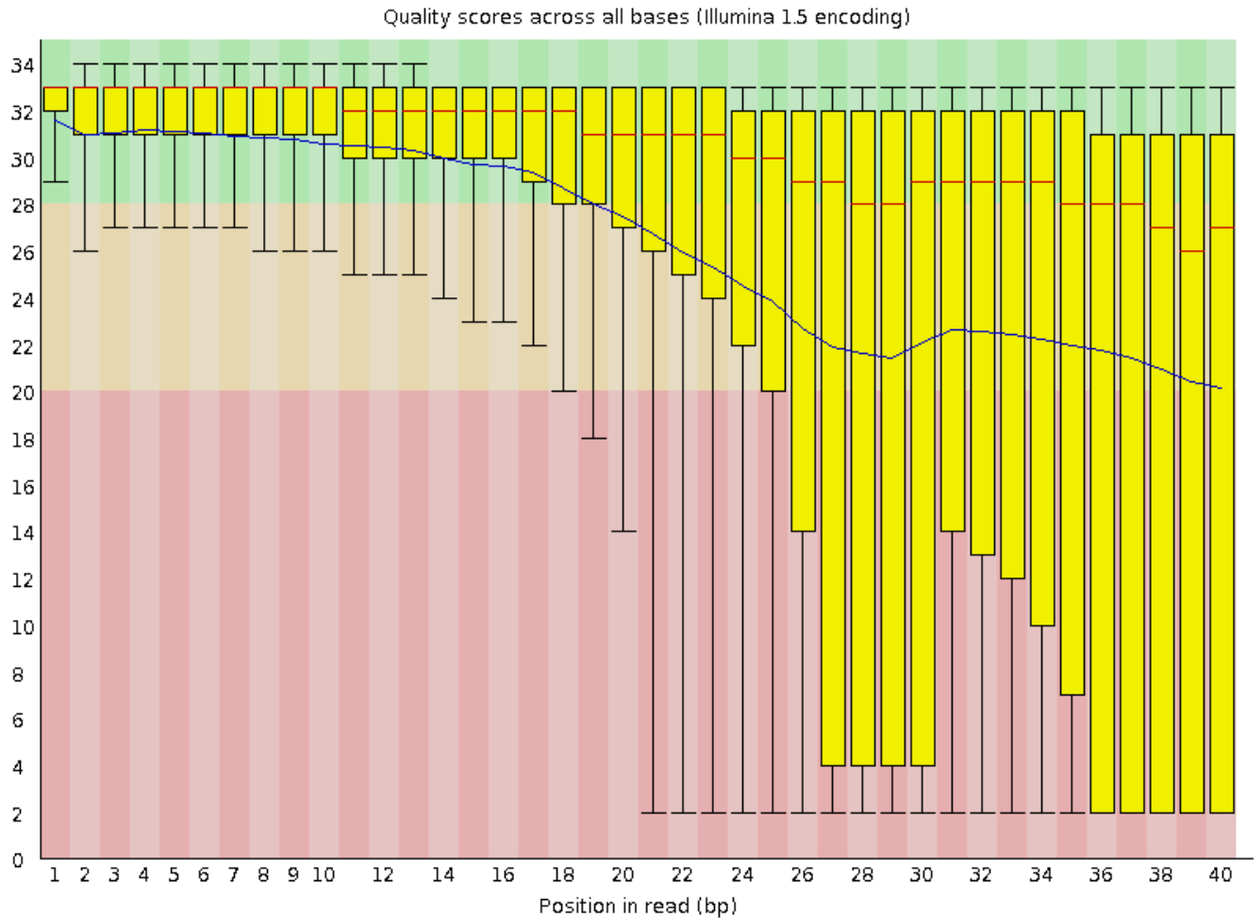
Sequencing cores generally provide some/all of this analysis

FastQC

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



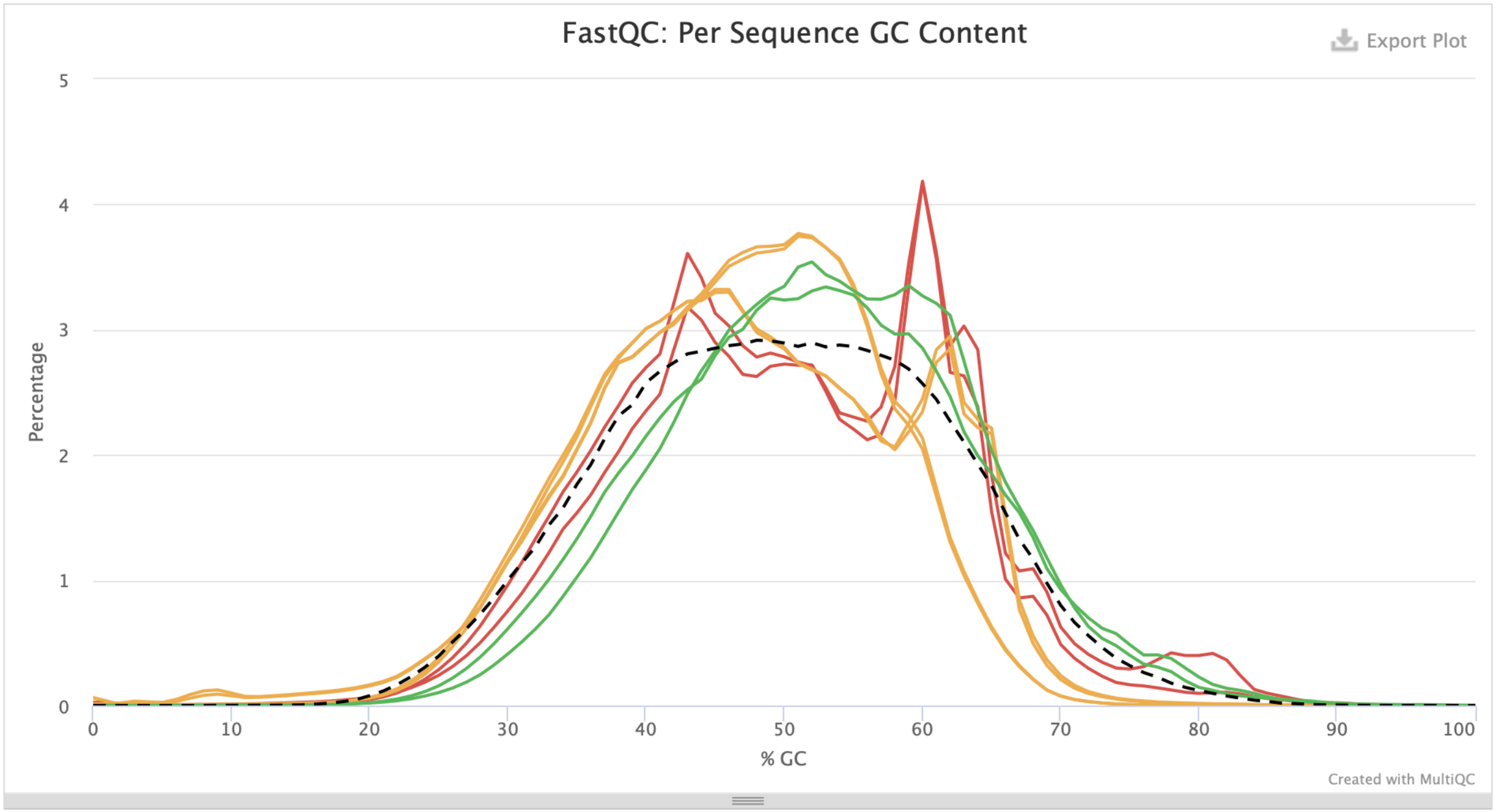
GOOD



BAD

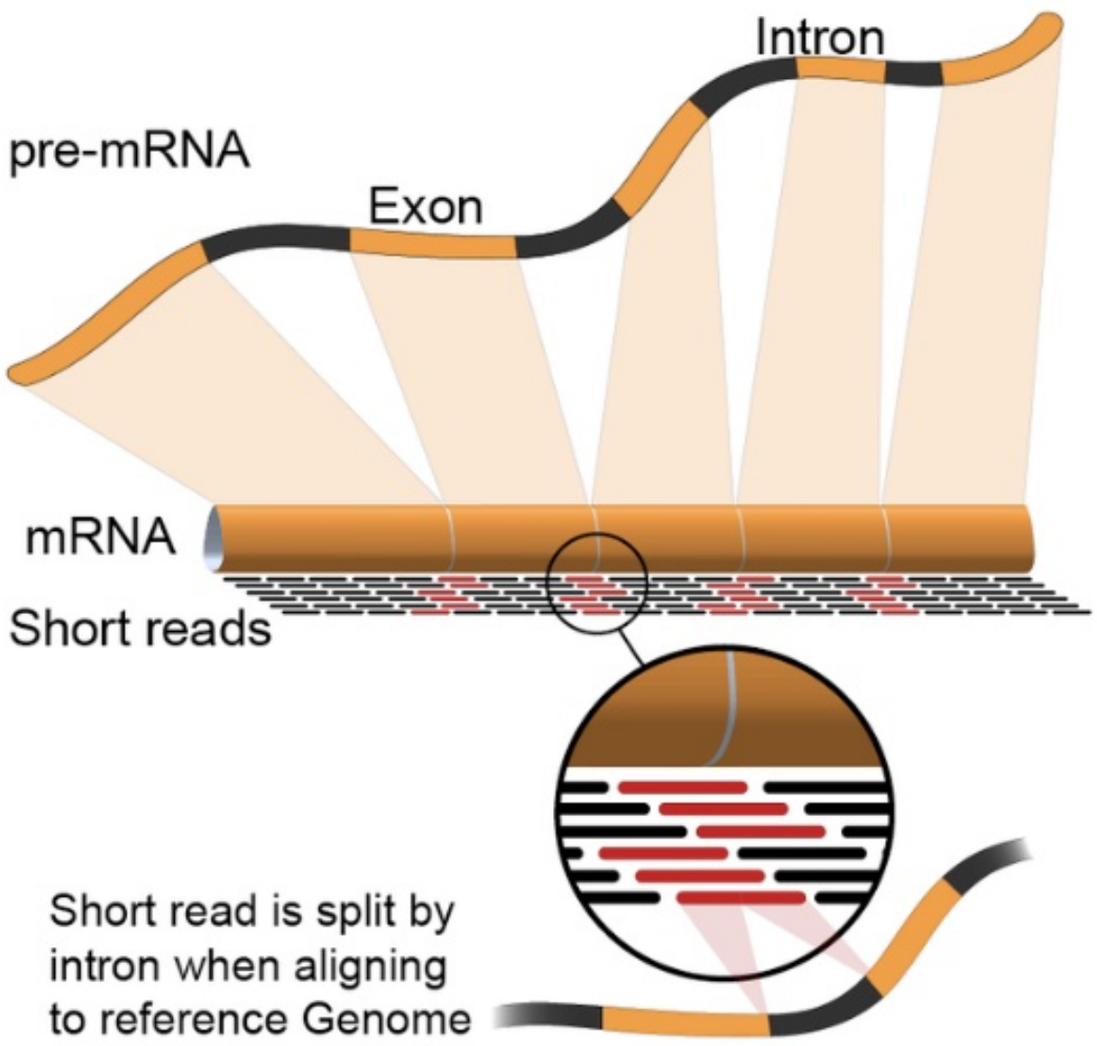
MultiQC

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



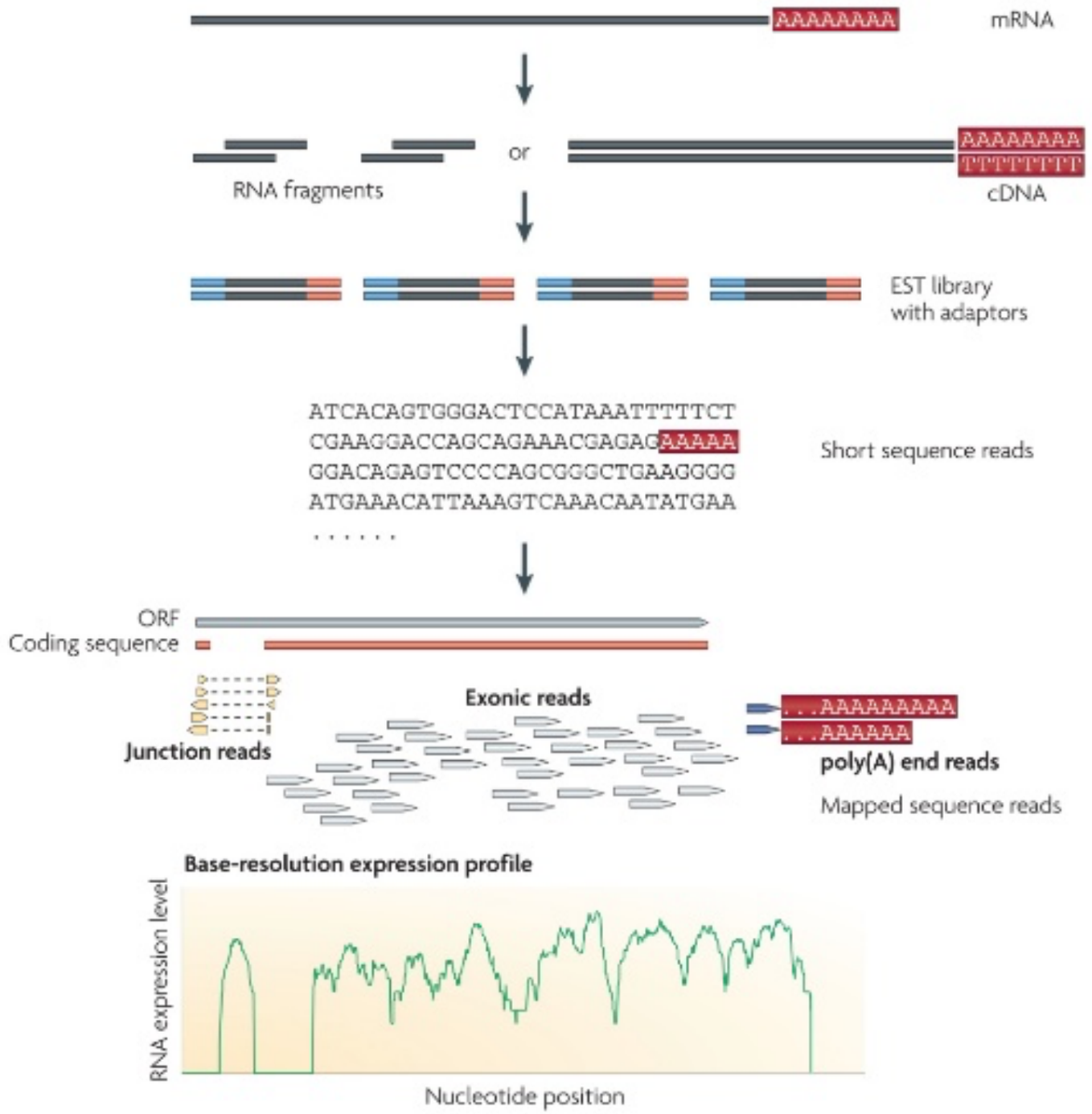
Alignment
(*Computationally Intensive Step*)

RNA-Seq Mapping Challenges



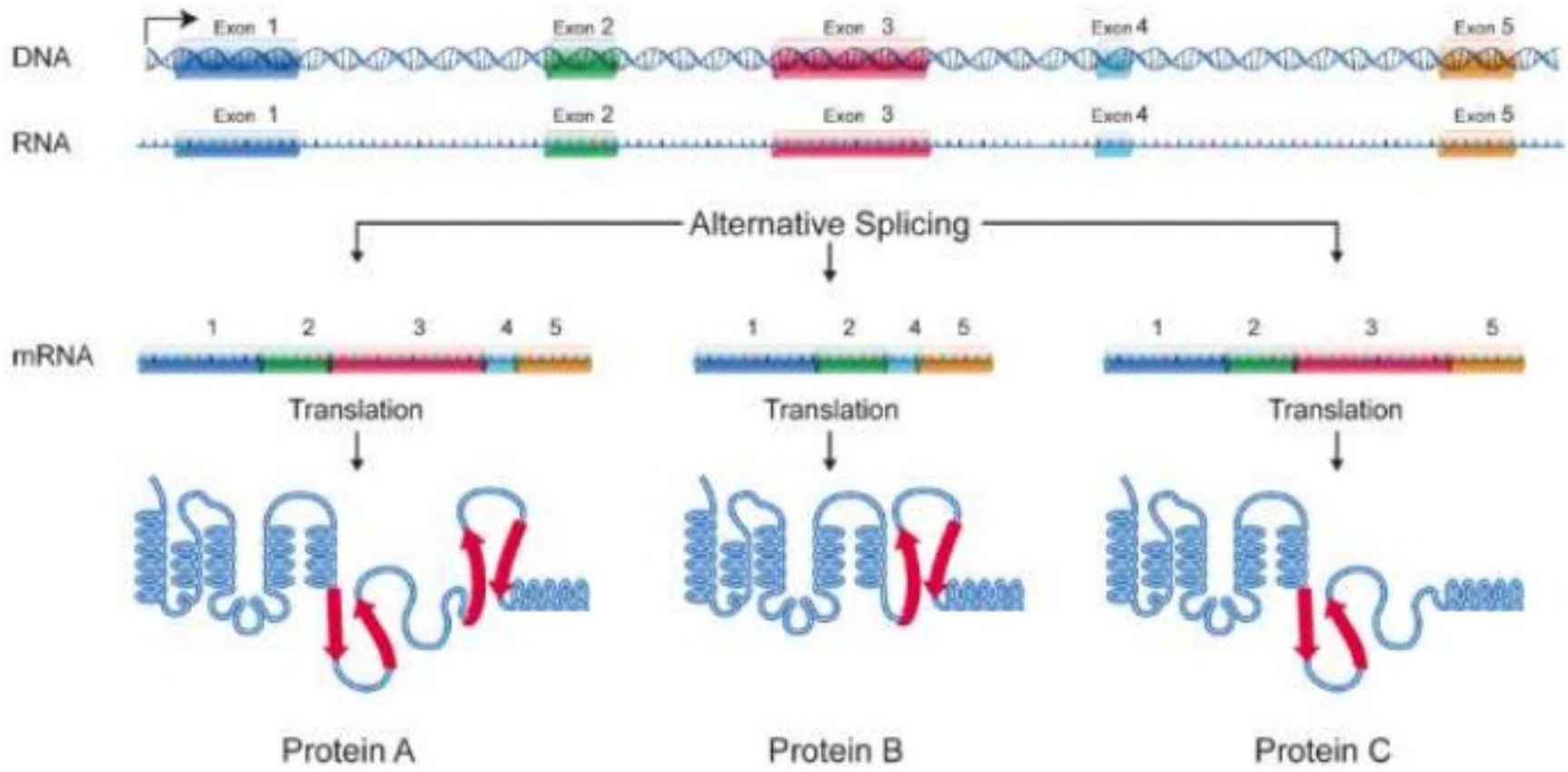
The majority of mRNA derived from eukaryotes is the result of splicing together discontinuous exons

RNA-seq protocol schematic



RNA-Seq: Special Mapping Concerns

Alternate Splicing



Alternate splicing deconvolution is not for the faint of heart

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 provided (genome over transcriptome)

RNA-Seq 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

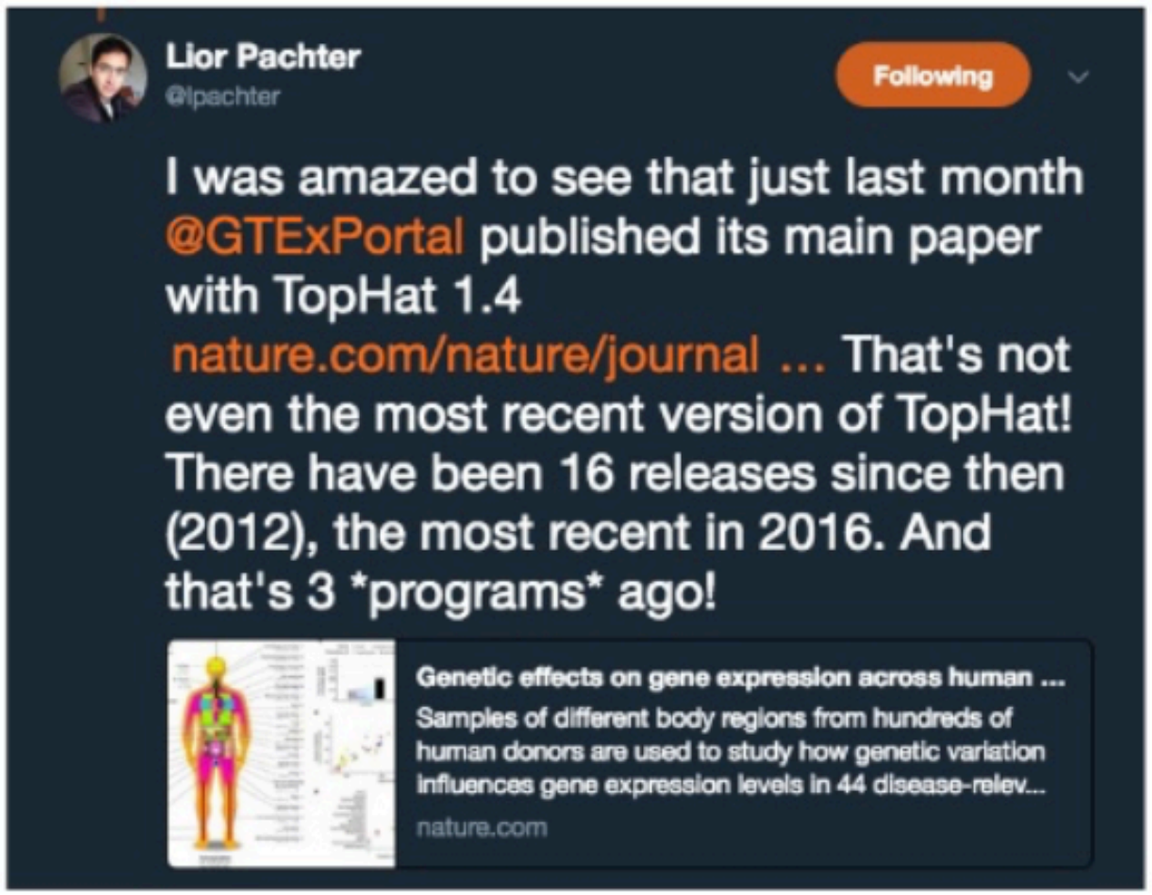
- ***De novo* assembly of RNA-Seq reads**
  - Largely used for uncharacterized genomes

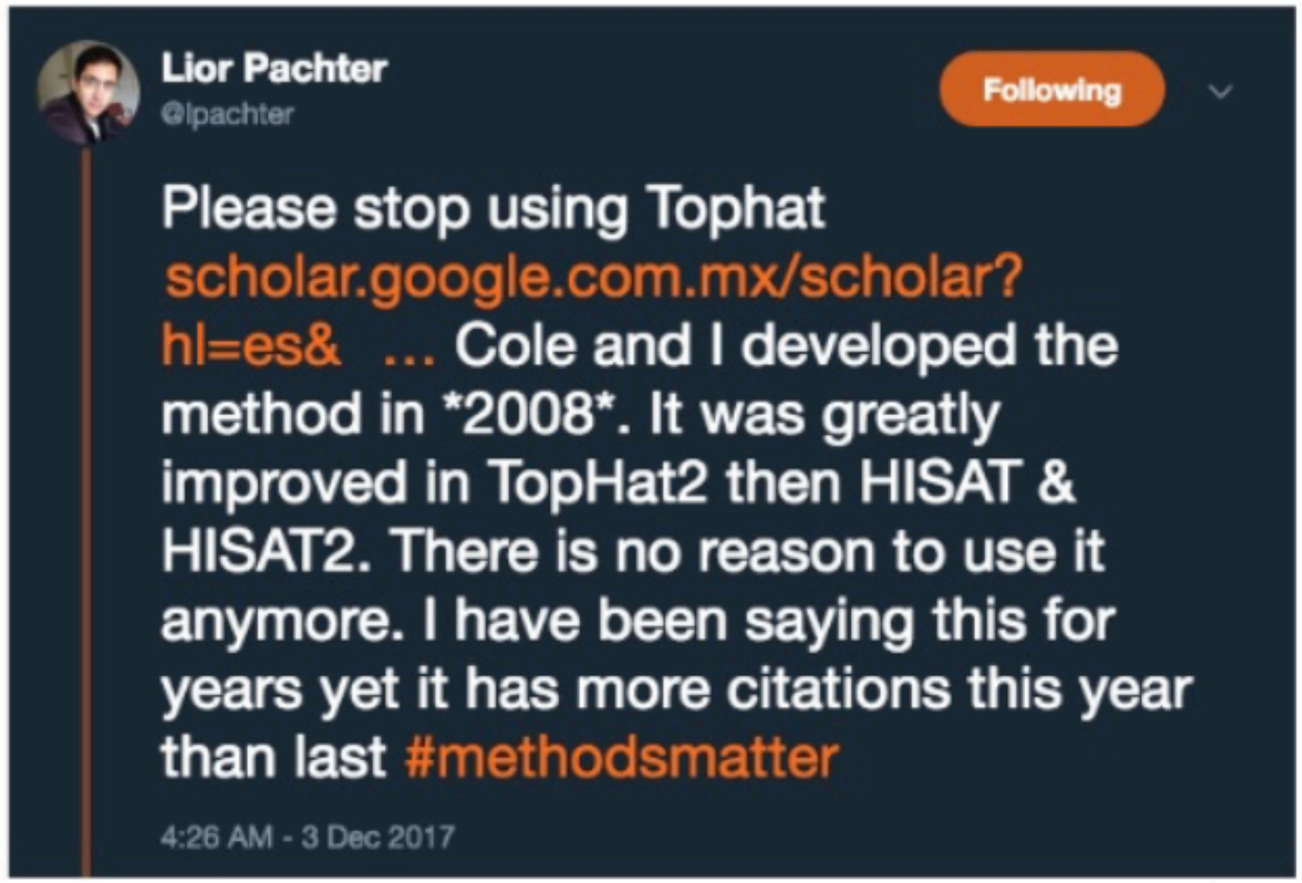
- **Align against the genome using a splice-aware aligner**
  - Most versatile solution

- **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
  - Bootstrapping can give confidence values

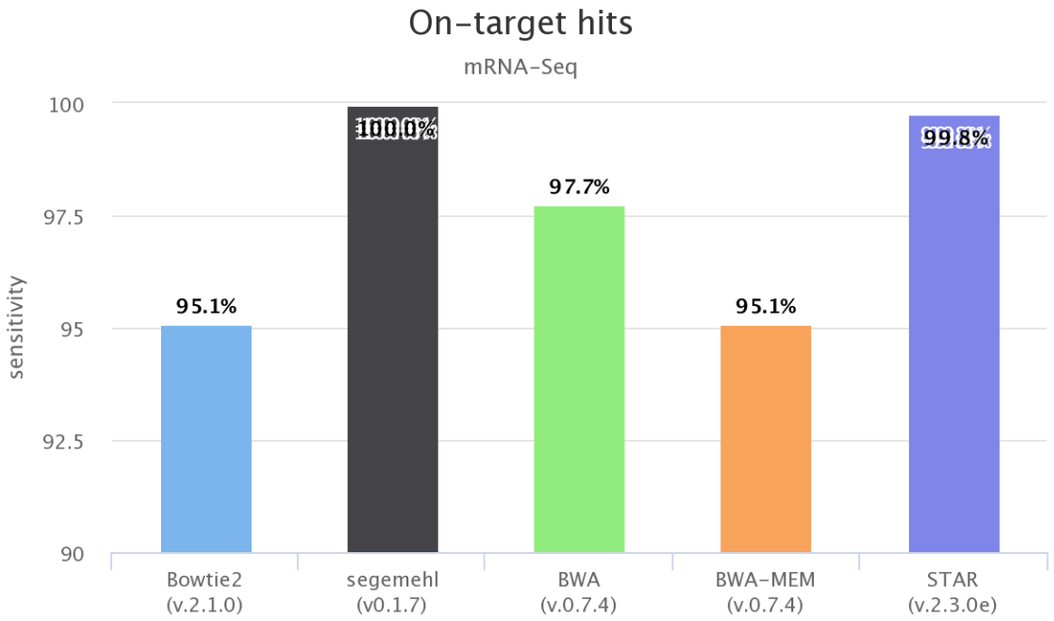
The Times they are a Changin !!

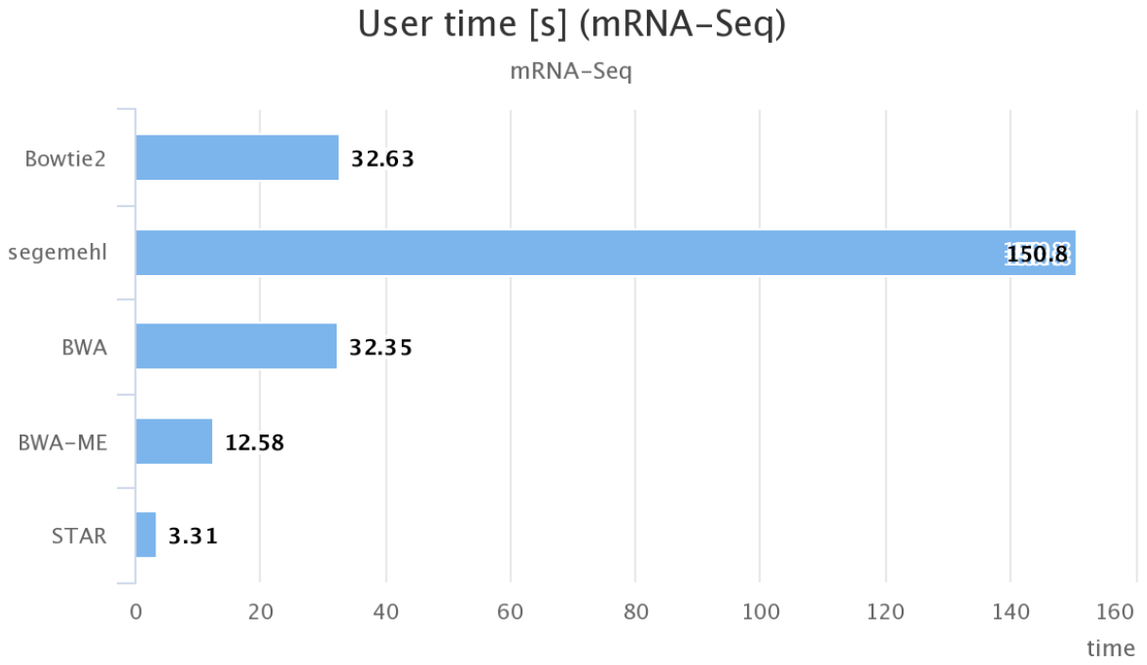
Check or new versions… try new software

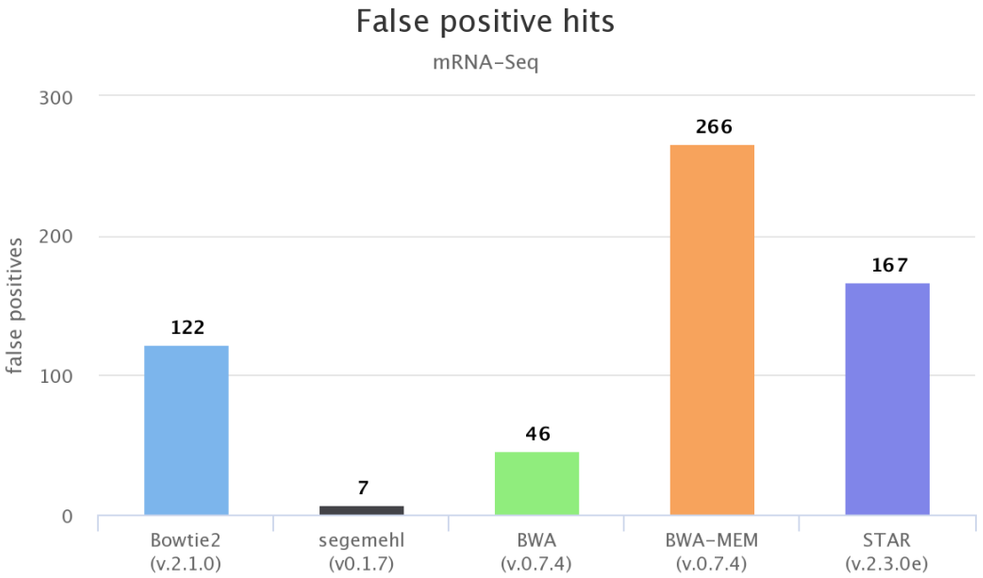


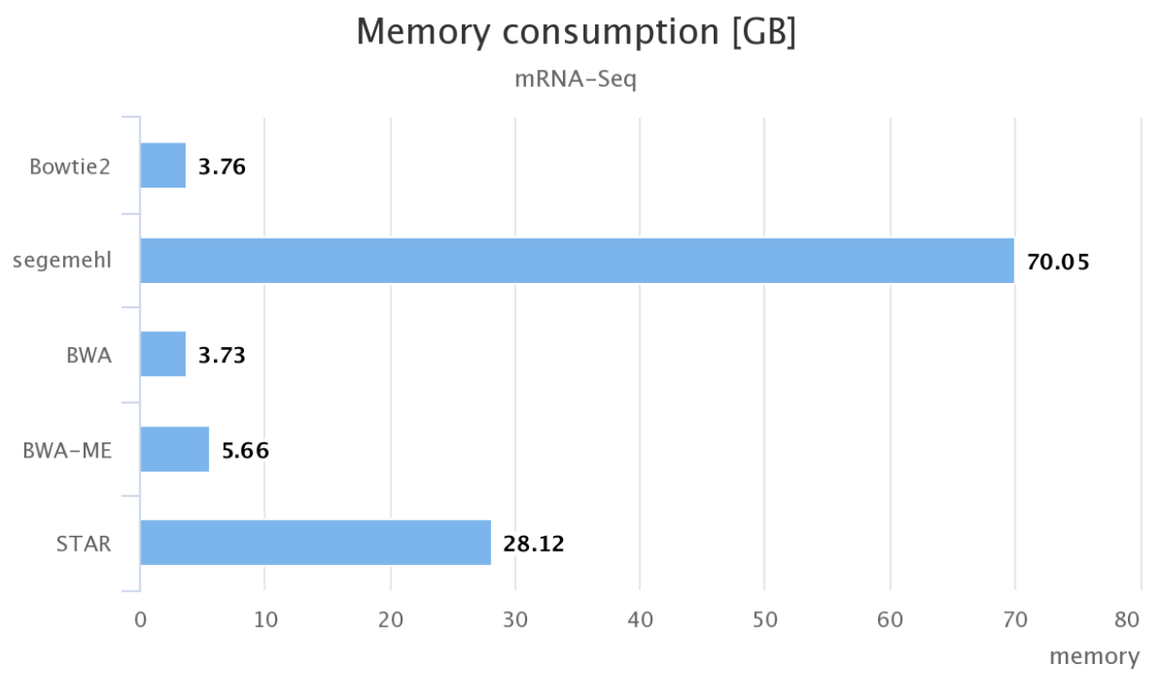


Source: Twitter









ECSEQ BIOINFORMATICS

To Align or not to Align

Aligners typically align against the entire genome and provide an 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 arguably more accurate

Typical Questions about alignment

- What is the best aligner to use?
  ☑STAR - (**Salmon** or Kallisto) - subjective

- What Genome version should I use?
  ☑Depends - most recent or best annotated

- What Genome annotation should I use?
  ☑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?
  - ☑ 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.

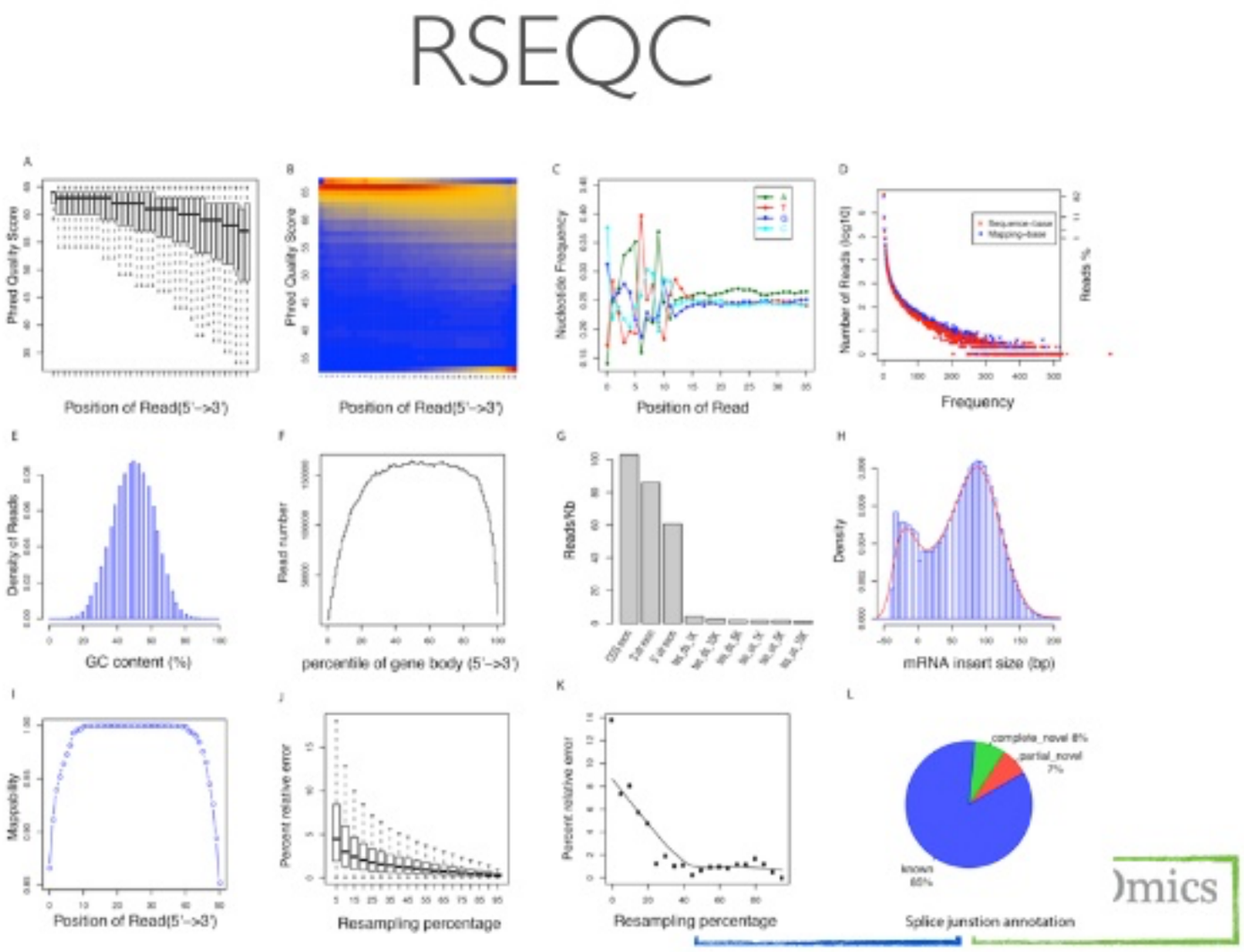
Don’t change parameters that you don’t understand - especially if it produces your preferred result

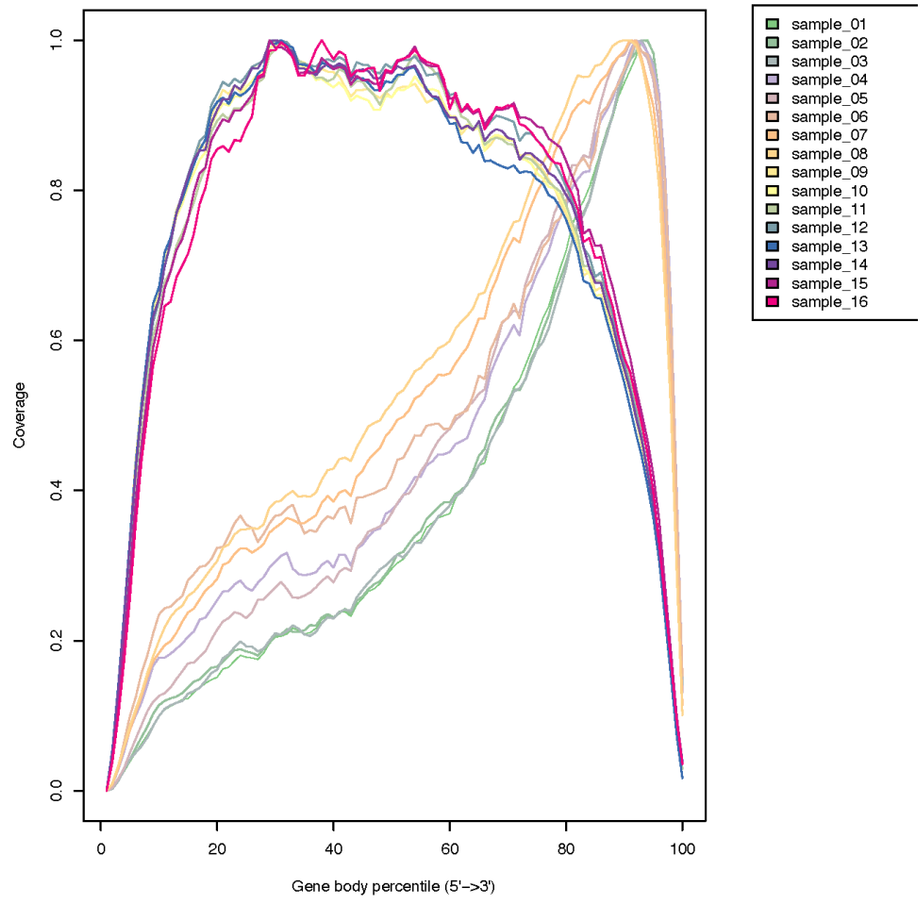
Post-Alignment QC

An important step in accessing the success of the experiment is the post-alignment QC.

- **Important considerations**
  - What is the distribution of the reads across the genome…do they align with know exons.
  - Are the reads distributed across the gene body uniformly
  - Is there a bias in read strand (unstranded protocols)
  - Do the different samples have similar profiles.

RSeQC example of plot types





Quantification

Counting as a measure of Expression

- Most RNA-Seq 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

Normalization

There are three metrics commonly used 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 RNA-Seq, where every read corresponded to a single fragment that was sequenced. FPKM was made for paired-end RNA-seq

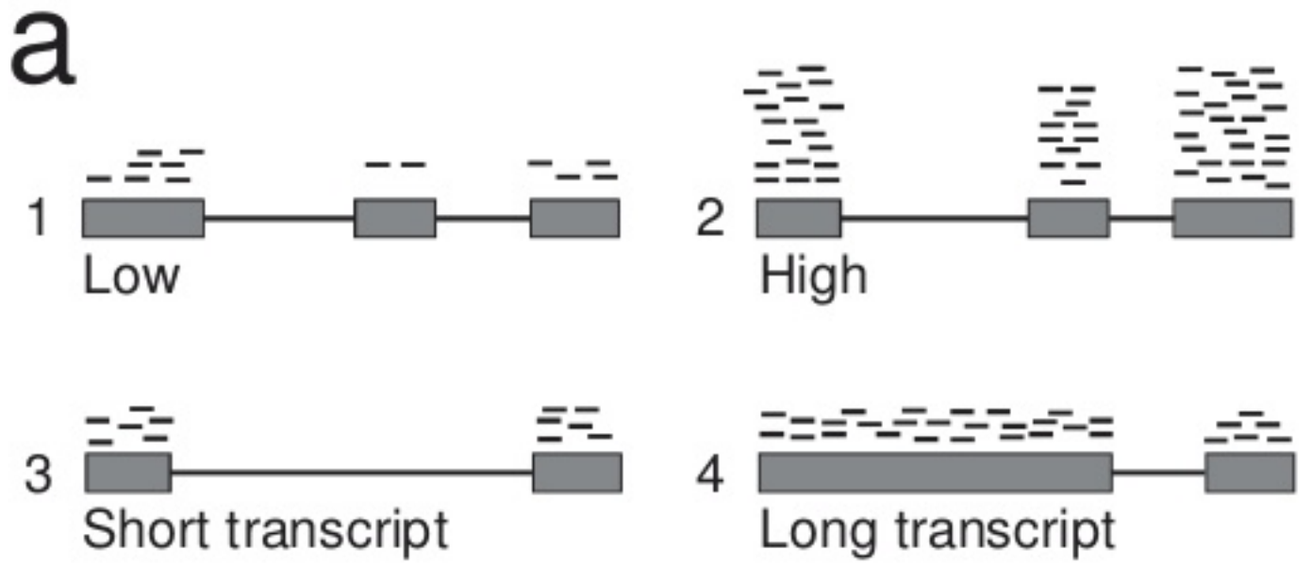
- **TPM = Transcripts Per Million (*Sum of all TPM in samples is the same*)**
  TPM is very similar to RPKM and FPKM. The only difference is the order of operations
  Gene read-count/gene-length (kb) = RPK
  (Sum all RPKs)/1,000,000 = PM
  Gene RPK/PM = TPM

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

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



Most Differential Expression software does its own normalization

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

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

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 two different samples

Differential Expression

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

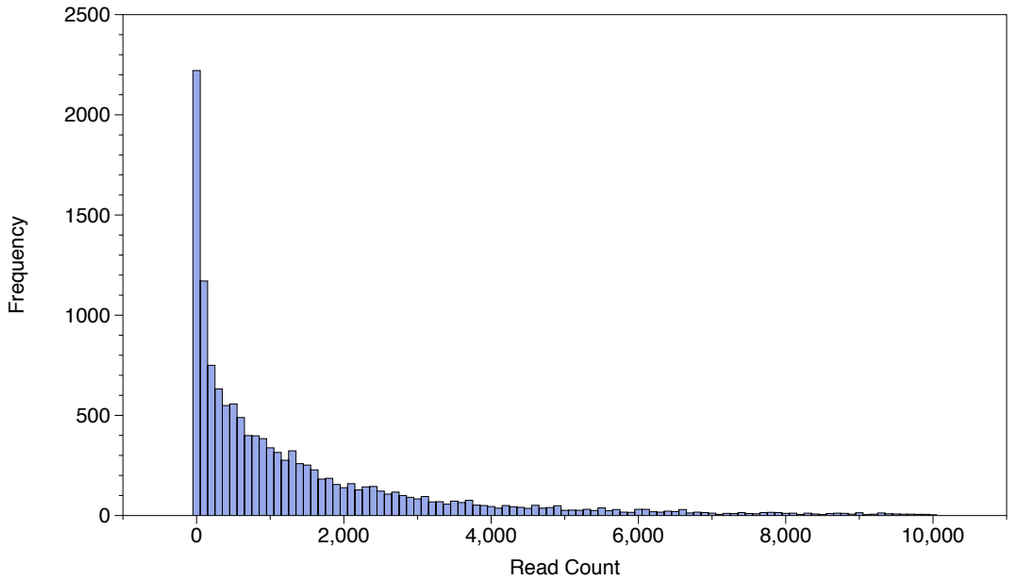
- Normalization of counts - the process of ensuring that values are expressed on the same scale (e.g. RPKM, FPKM, TPM, TMM). Corrects for variable gene length, read depth
- Differential Expression - analysis of the difference in expression of genes under two conditions (pair wise comparison) - *expressed 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*

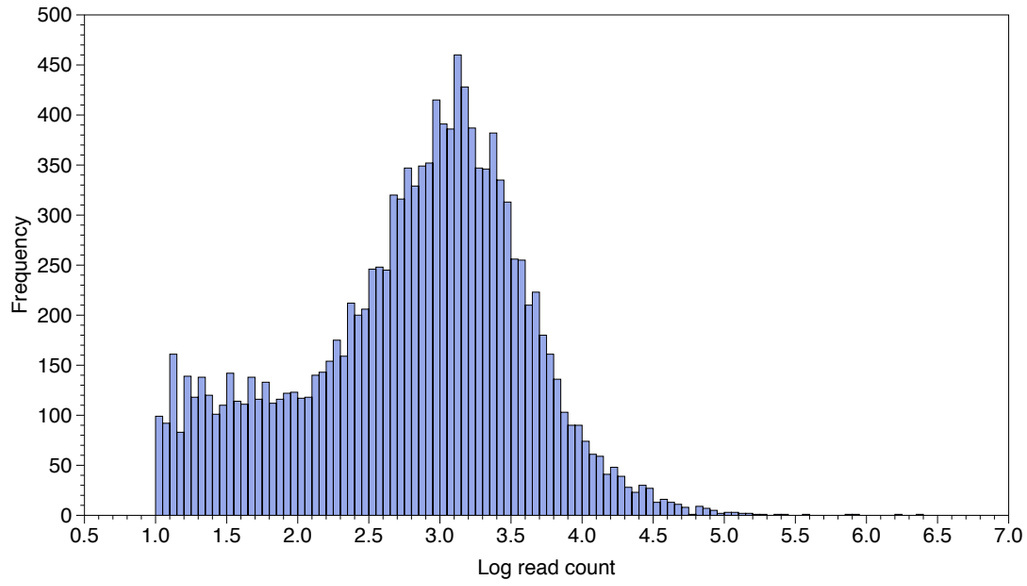
Log Transformed Data

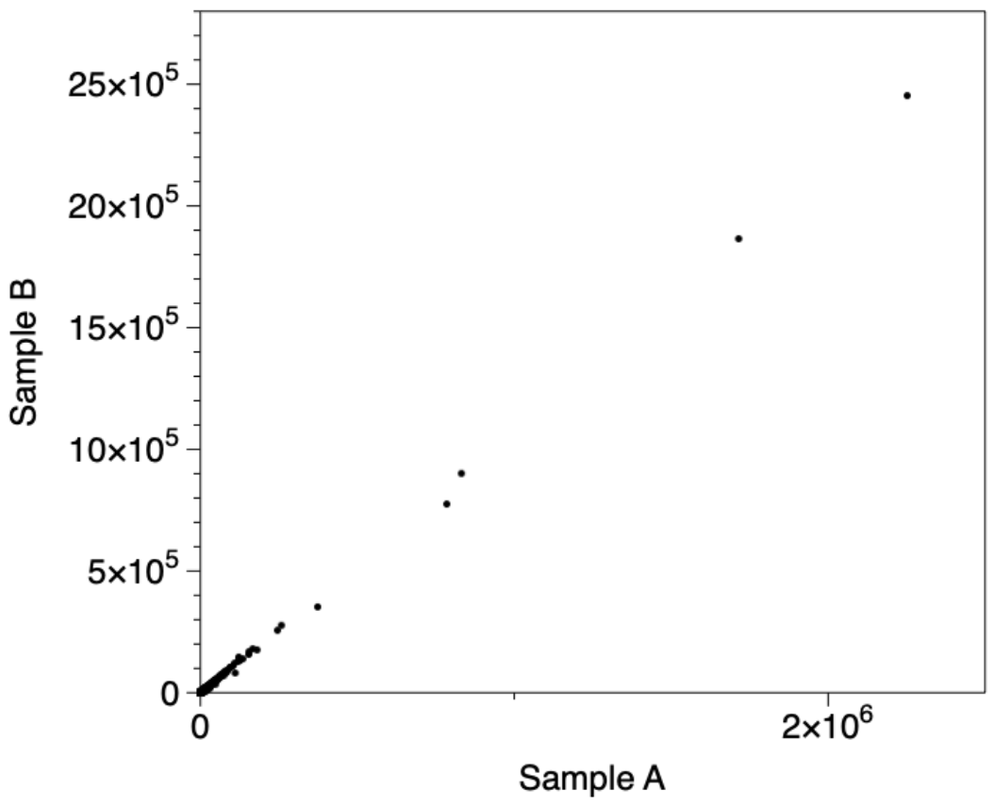
For RNA-Seq data analysis, just like any dataset, choosing the correct data model is essential for getting meaningful results. If the native data doesn't fit a suitable model it is often necessary to transform the data, such that it fits a standard statistical model

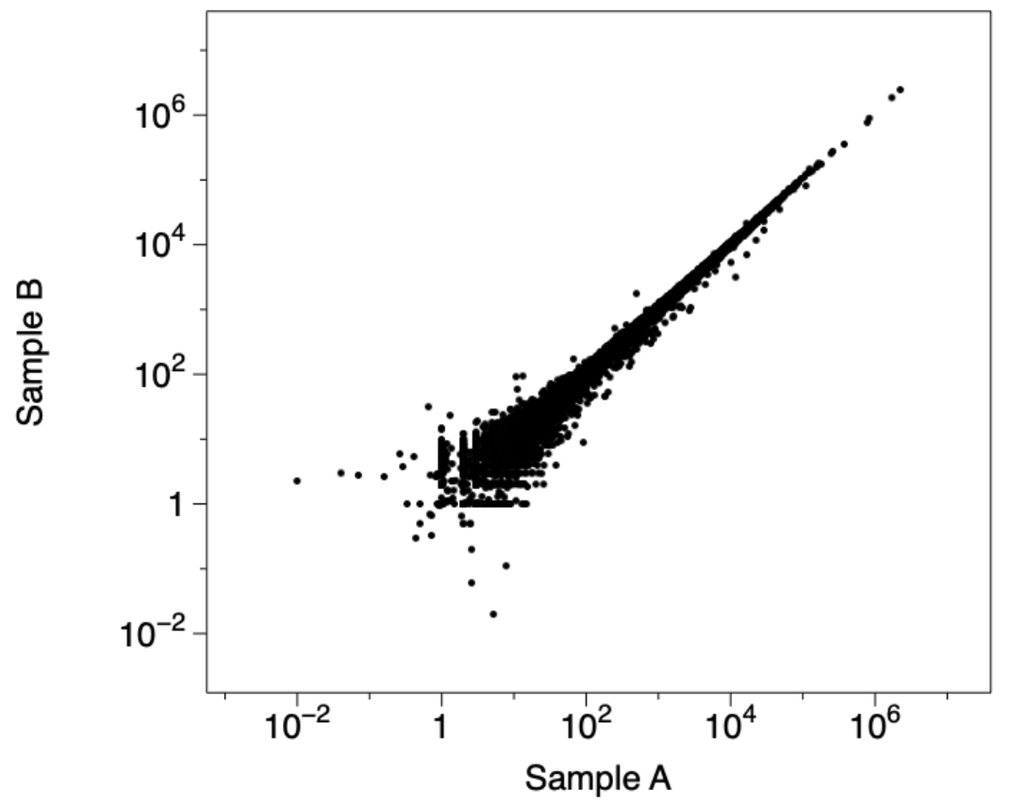
For RNA-Seq data most differential expression software assumes it fits a negative binomial distribution, and this is achieved by taking the log of the raw data The models also make the assumption that the majority of genes have not changed between the two experimental conditions

Log Transformed Data









Differential Expression

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

- Final output is 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 corrections will result in a "corrected p-value", or a qvalue or FDR or padj (adjusted p-value)

Note p-values 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/Meta 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

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

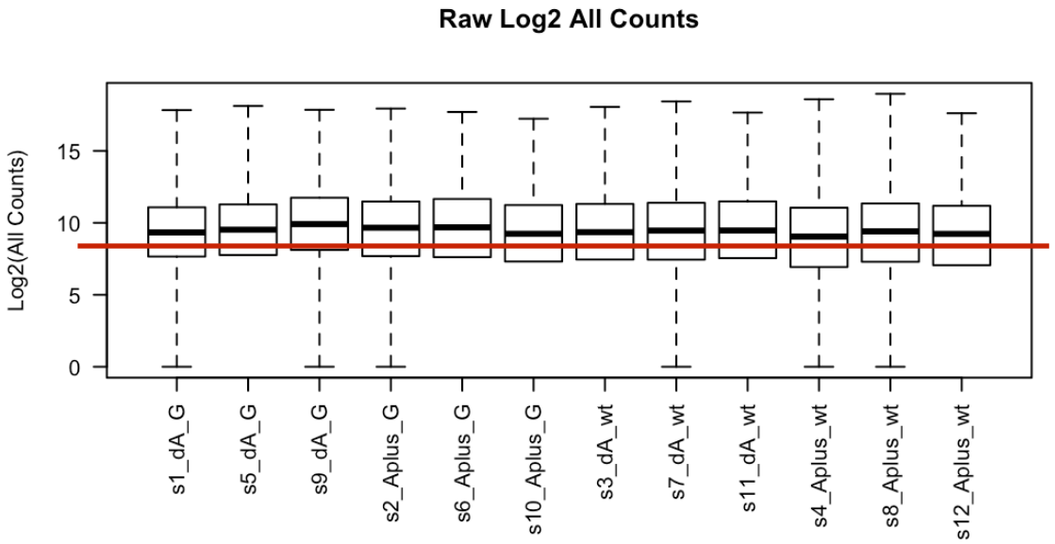
Gene	LogFC	AveExpr	P-Value	FDR
*CA14	-6.72	4.31	1.406716E-10	0.000001
*MCF2L	-10.75	3.25	2.854327E-10	0.000001
*COL5A2	-6.12	4.28	3.678663E-10	0.000001
*TYRP1	-9.31	9.85	4.190114E-10	0.000001
*BCAN	-8.39	5.33	6.384088E-10	0.000001
*CSAG1	10.81	-0.56	7.095577E-10	0.00000

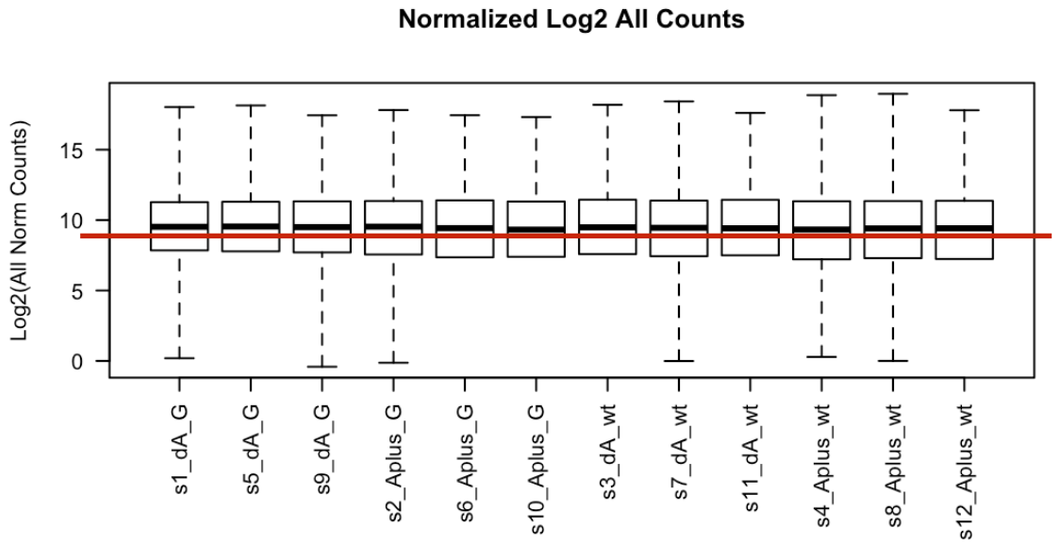
DESEQ2

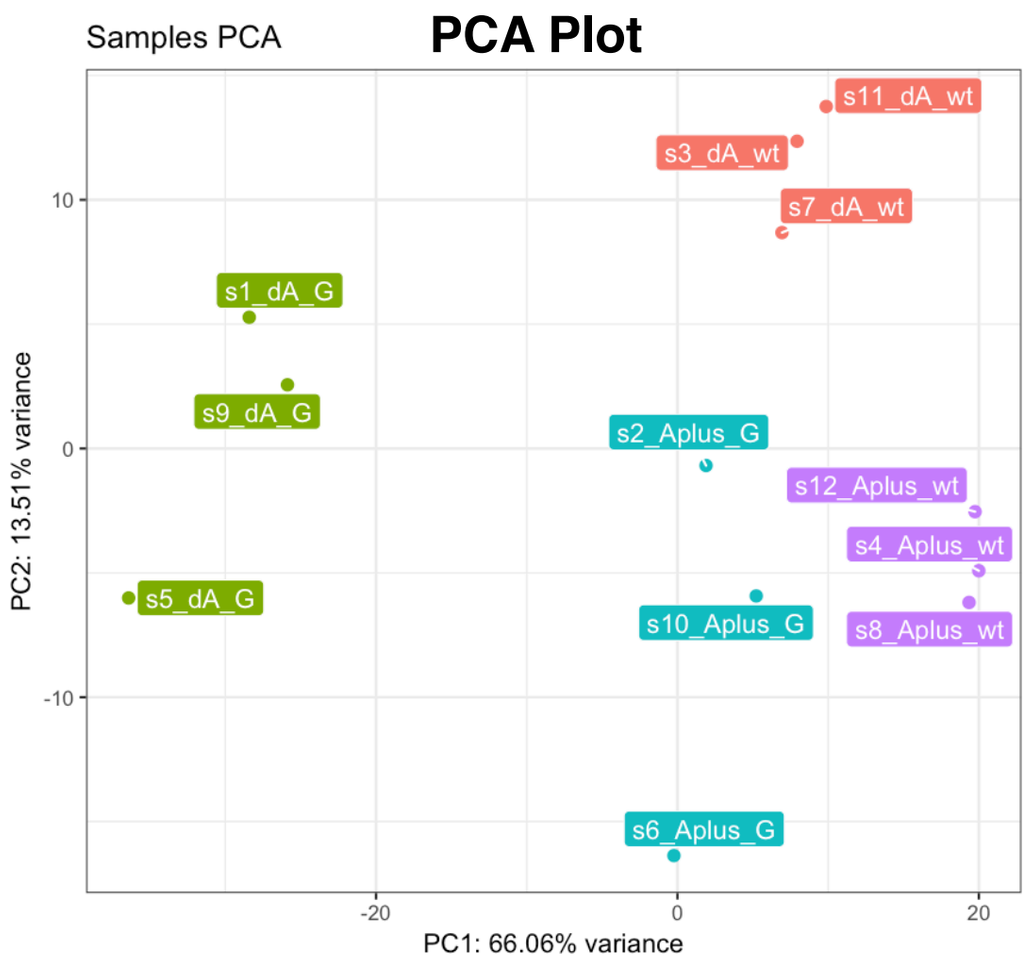
Row-names	Symbol	log2FoldChange	padj	p53_mock_1	p53_mock_2	p53_mock_3	p53_mock_4	p53_IR_1	p53_IR_2	p53_IR_3	p53_IR_4
ENSMUSG00000000001	Gnai3;Gnai3	-0.4763	0.1737	11.584	11.565	11.609	11.621	11.399	11.338	11.997	11.927
ENSMUSG00000000028	Cdc45;Cdc45	-0.4610	0.4125	8.024	7.575	7.668	7.295	7.736	7.675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

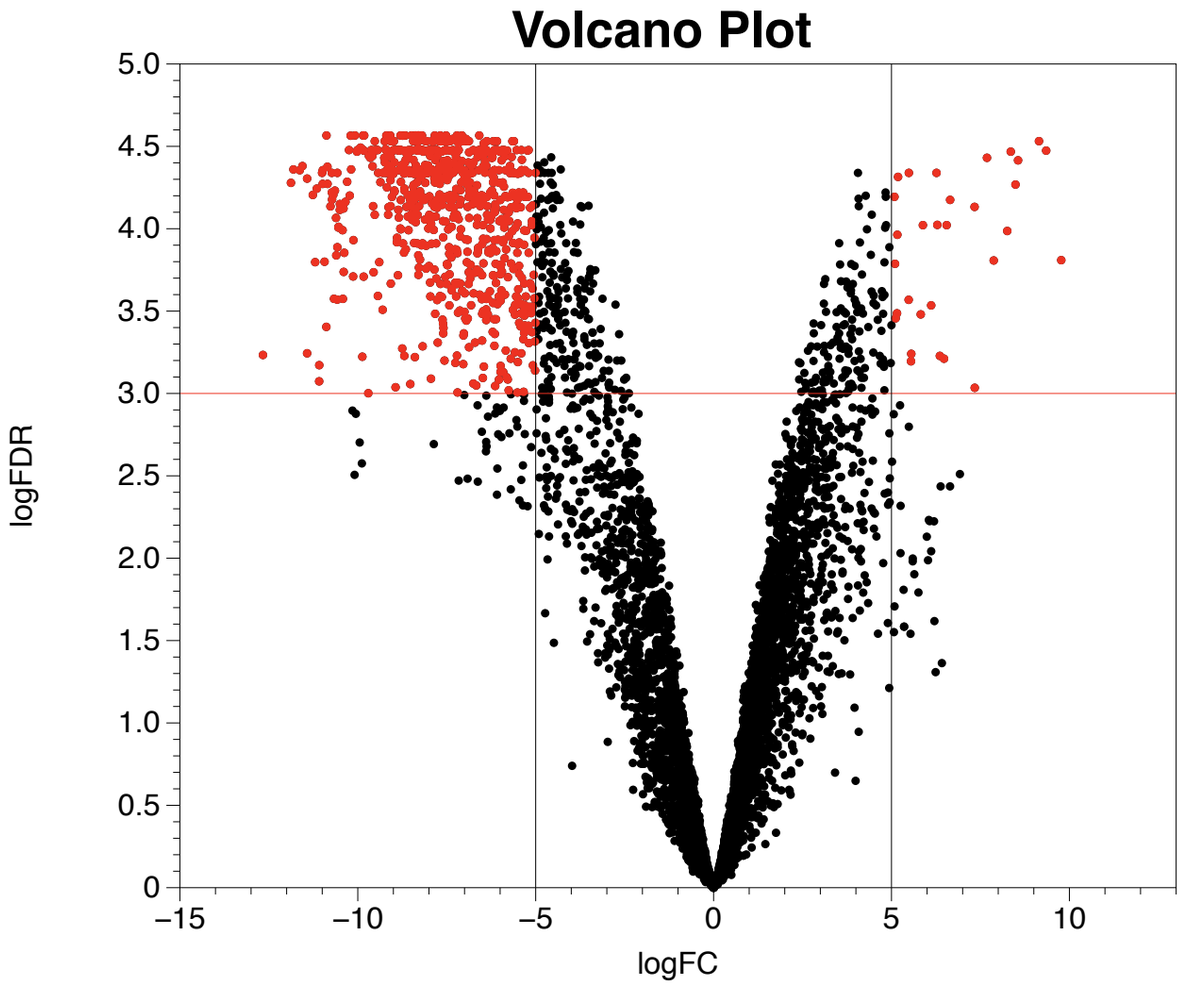
Visualization

Plotting the Data



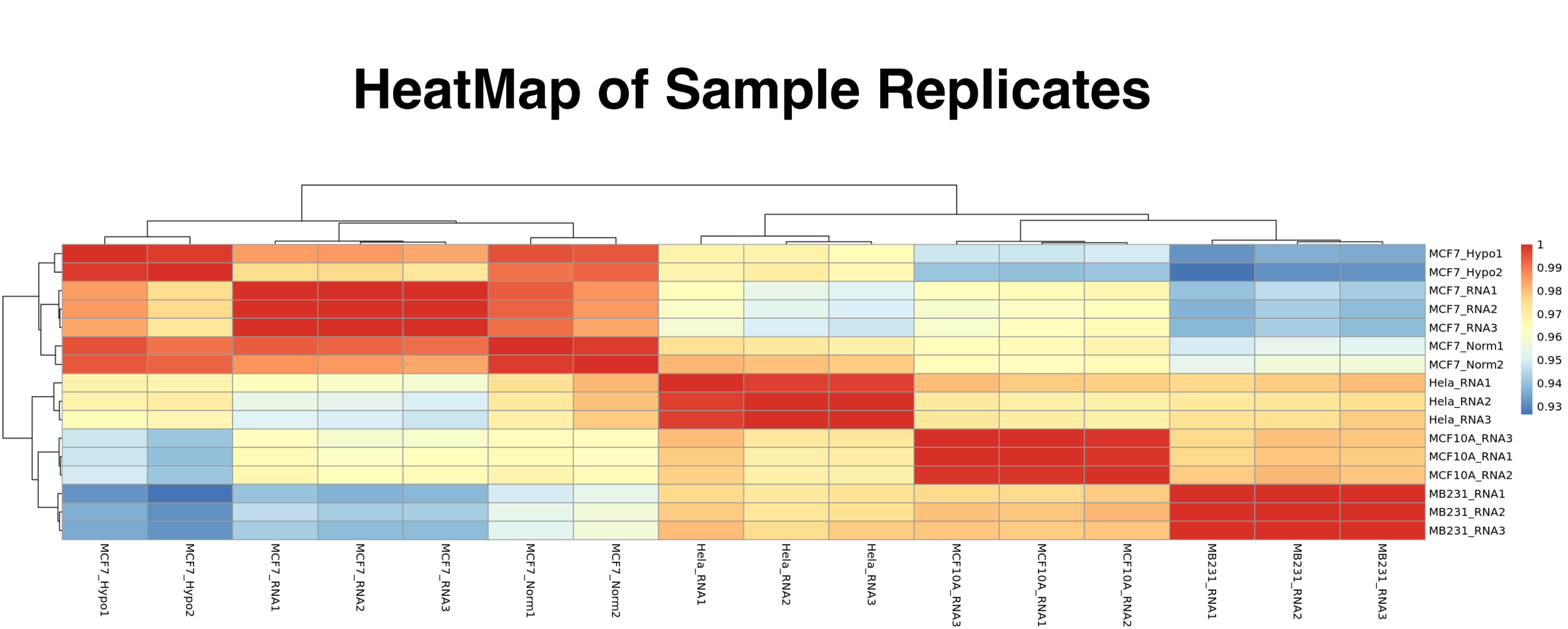






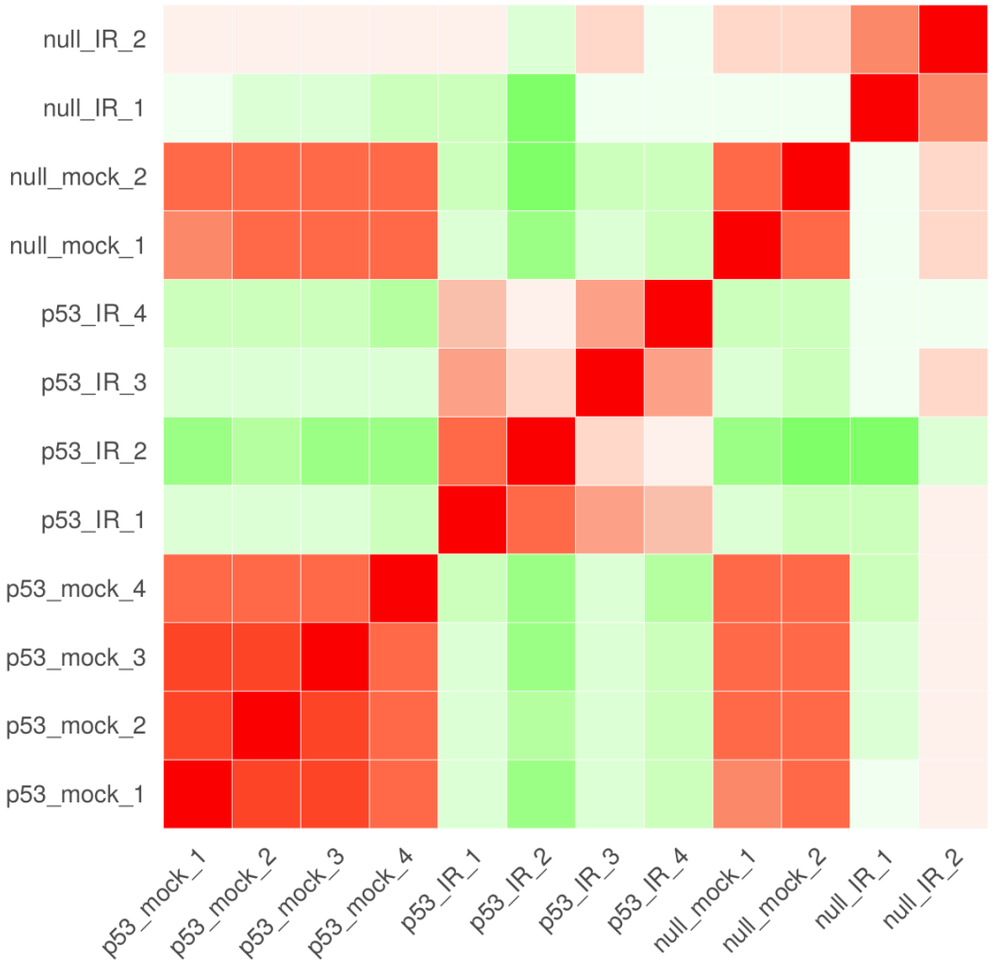
Plotting the Data

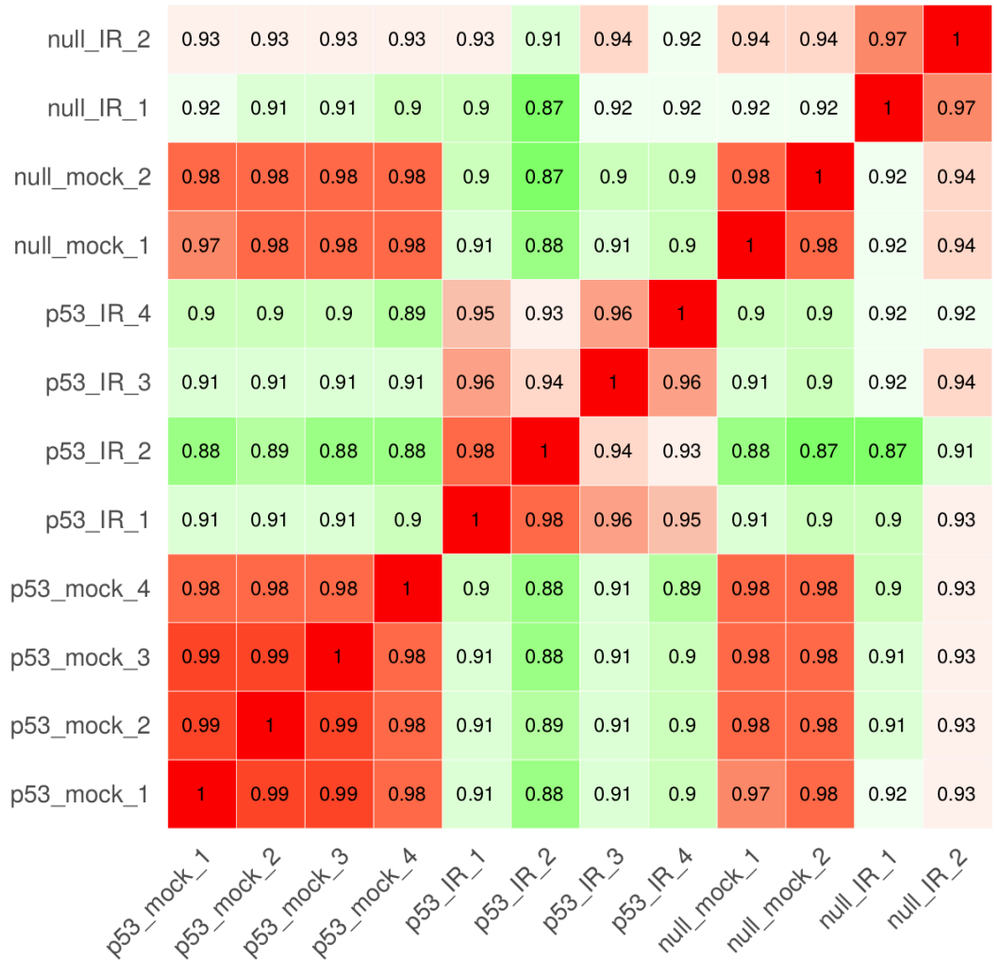
Plotting the Data



Plotting the Data

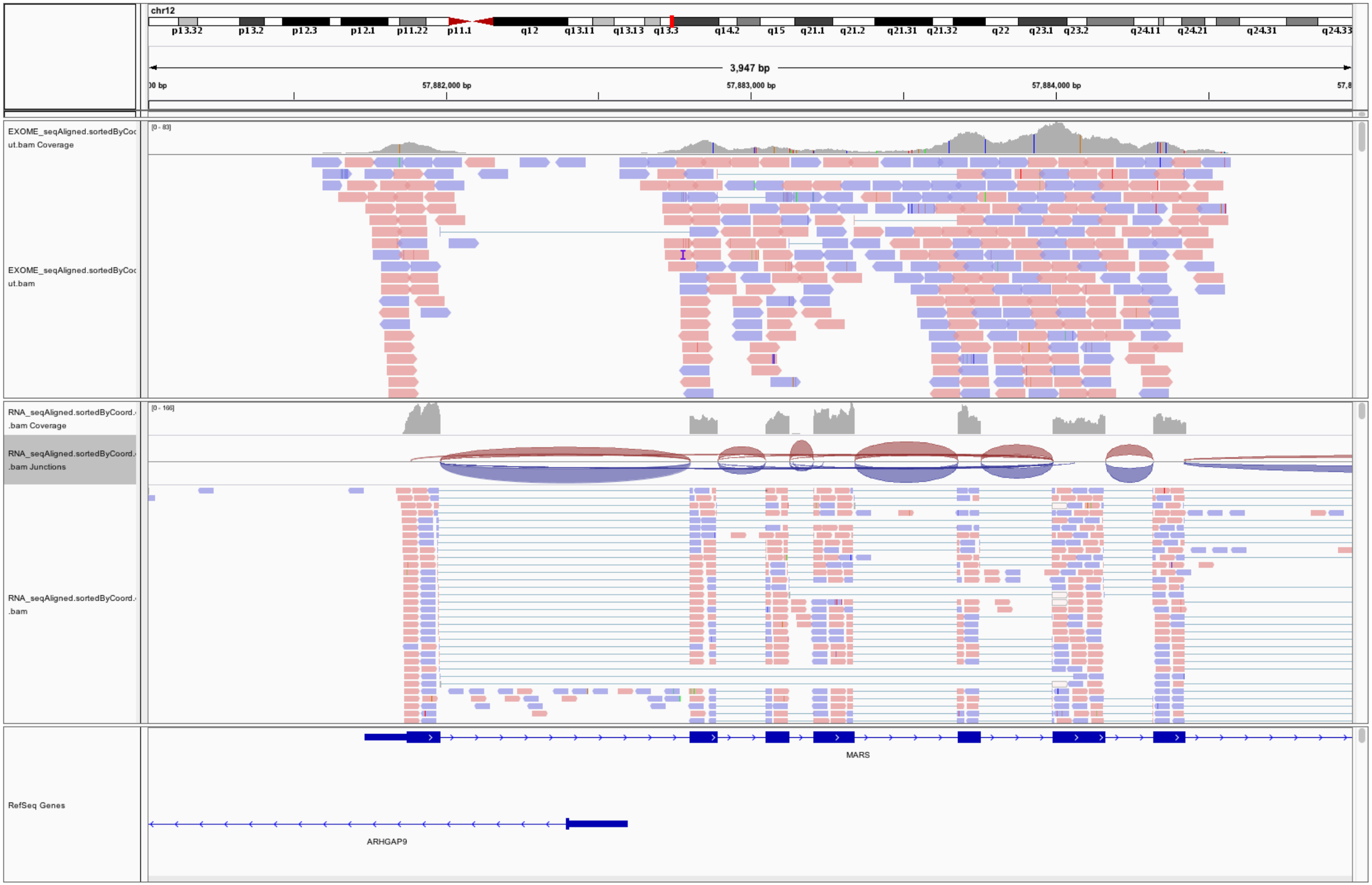
Heat Maps



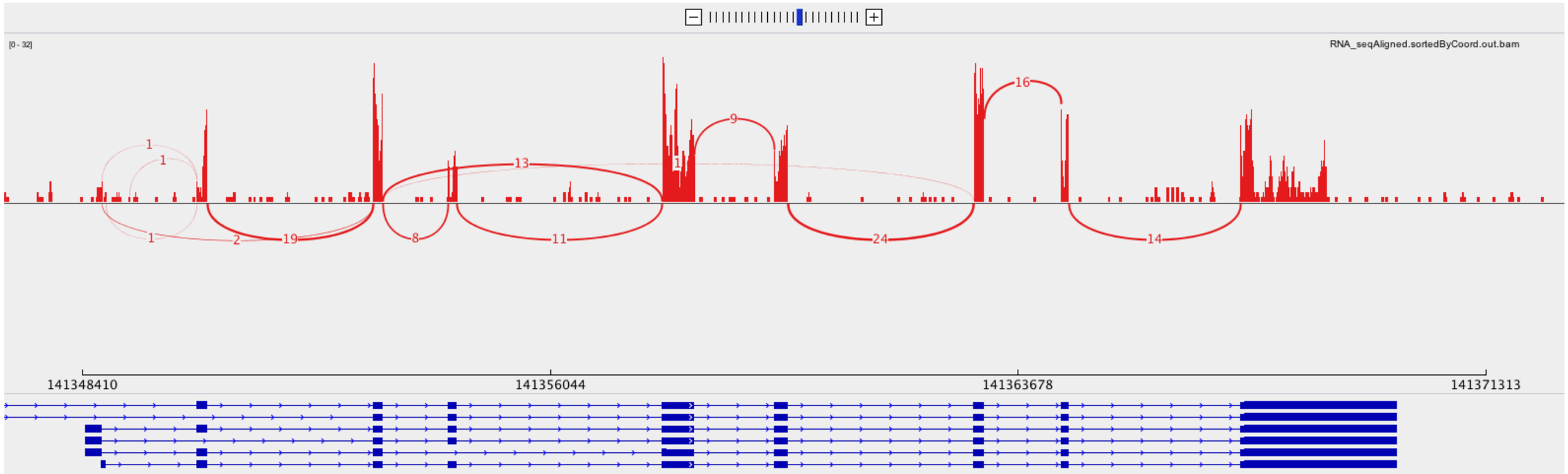


IGV View of RNA-Seq Data

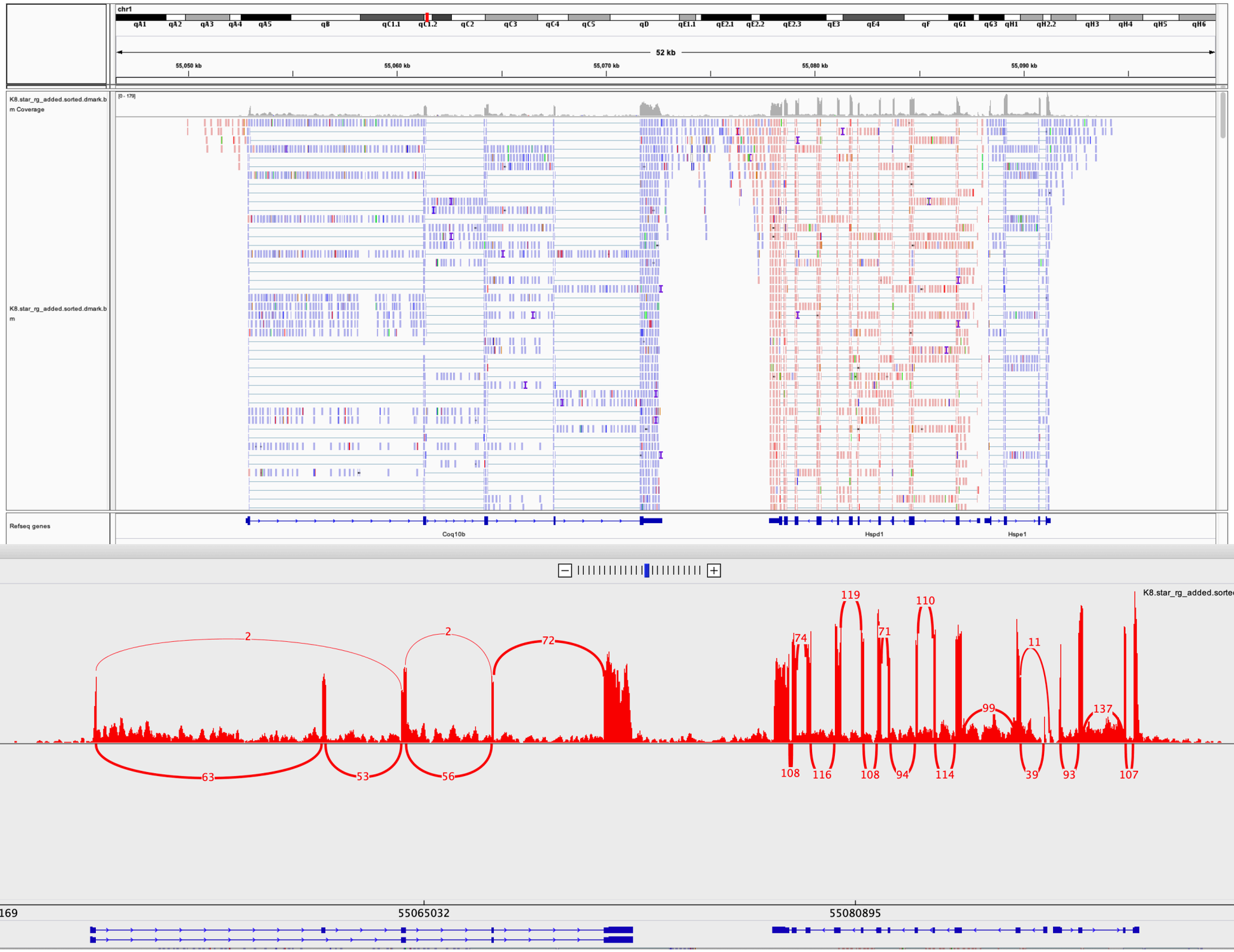
IGV View of RNA-Seq Data



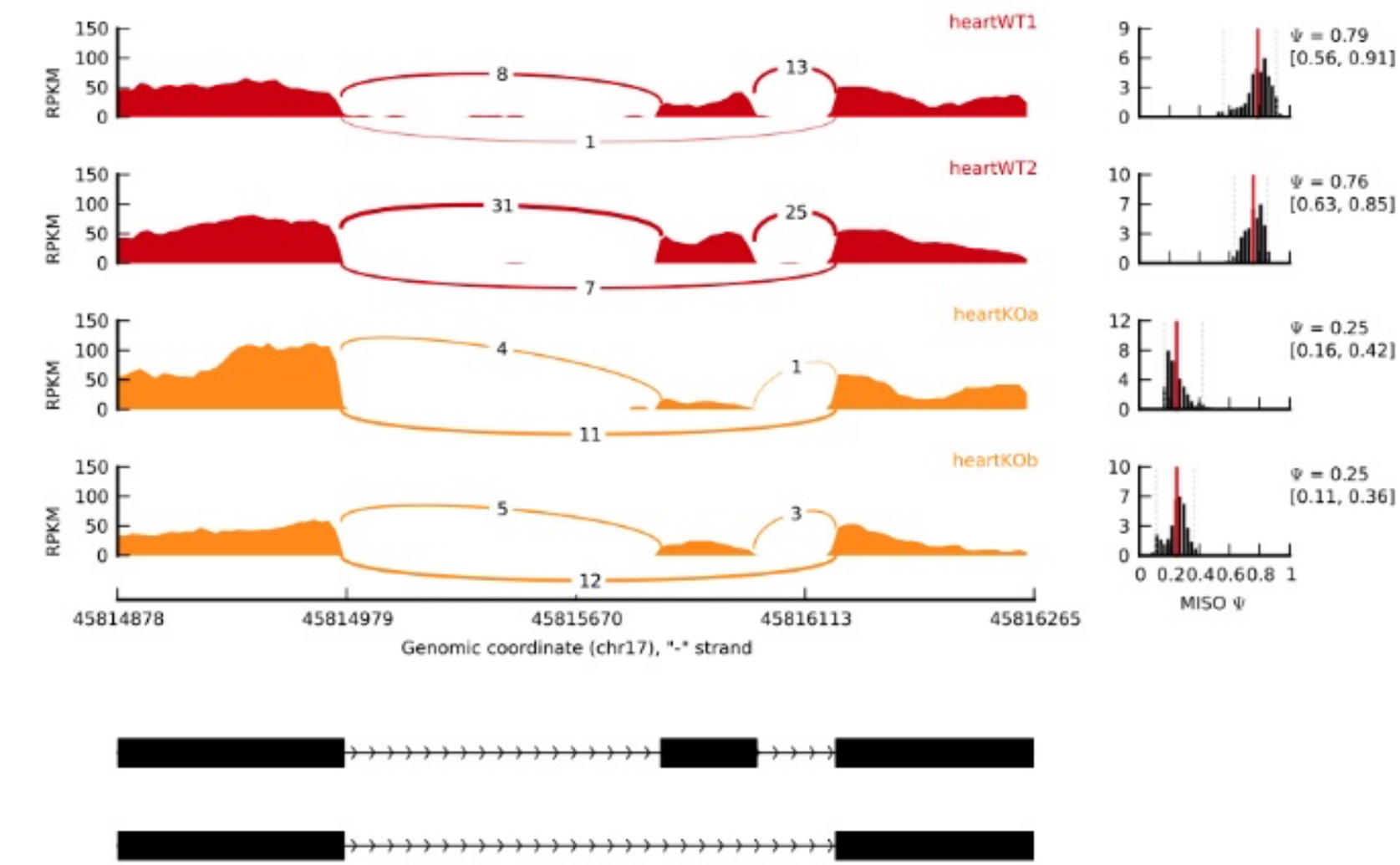
IGV View of RNA-Seq Splicing Data



IGV Stranded RNA-Seq Data



Visualizing Splicing



Tertiary Analysis - Biological Meaning

- **Pathway Analysis**
  - IPA (Qiagen - CCR License) Future talk
  - Reactome (http://www.reactome.org/)
- **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**

Software Solutions

CCR staff have access to a number of resources

- NIH HPC - Biowulf & Helix - CIT maintained large cluster with a huge software library (**Unix command line**)
- CCBR Pipeliner/Renee - RNA-Seq pipeline from CCBR (Biowulf)
- Partek Flow (Local Web Service)
- NIDAP - NIH Integrated Data Analysis Platform (RNA-Seq module)
- Cancer Genomics Cloud CGC
- DNAnexus (Cloud Solution)
- Galaxy -is an open source, web-based platform for data intensive biomedical research

- CLCBio Genomic Workbench (Small genomes - local software)
- Qlucore - (local software)

Web-Based Tools

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

Both allow analysis of your data or many public datasets

NGS File Formats

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

See the **NGS file format document** on the BTEP site

File Transfer

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

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.

- **FastP** a tool designed to provide fast all-in-one preprocessing for FastQ files.

Pseudo-Aligners

Salmon uses new algorithms (specifically, coupling the concept of quasi-mapping with a two-phase inference procedure) to provide accurate expression estimates very quickly while using little memory. Salmon performs its inference using an expressive and realistic model of RNA-seq data that takes into account experimental attributes and biases commonly observed in real RNA-seq data. (*https://doi.org/10.1038/nmeth.4197*)

kallisto is a program for quantifying the abundance of transcripts from RNA-Seq data, or more generally of target sequences using high-throughput sequencing reads. It is based on the novel idea of pseudo-alignment for rapidly determining the compatibility of reads with targets, without the need for alignment. (*https://doi.org/10.1038/nbt.3519*)

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

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

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**
- HISAT2
- 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

Visualization and Pathway Analysis

Visualization

- Integrated Genome Viewer (https://www.broadinstitute.org/igv/)
- UCSC Genome Browser (https://genome.ucsc.edu/)

Pathway Analysis

- QIAGEN Ingenuity Pathway Analysis
- DAVID (http://david.abcc.ncifcrf.gov/tools.jsp)
- ConsensusPathdb (http://cpdb.molgen.mpg.de/)
- Reactome (http://www.reactome.org/)
- Molecular Signatures Database (http://www.netgestalt.org/)
- PANTHER (http://www.pantherdb.org/)
- Cognoscente (http://vanburenlab.medicine.tamhsc.edu/cognoscen
- Pathway Commons (http://www.pathwaycommons.org/)
- PathVisio (http://www.pathvisio.org/)
- Moksiskaan (http://csbi.ltdk.helsinki.fi/moksiskaan/)
- Weighed Gene Co-Expression Network Analysis (WGCNA)s
- More tools in R Bioconductor

Further Reading

RNA-seqlopedia
https://RNA-Seq.uoregon.edu/

RNA-Seq by Example
https://www.biostarhandbook.com/

Reference-based RNA-Seq data analysis

https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html

Further Reading

Introduction to RNA-seq

https://scienceparkstudygroup.github.io/rna-seq-lesson/

RNA-seq: a step-by-step analysis pipeline

https://github.com/CebolaLab/RNA-seq

BTEP - Bioinformatics Resources for CCR Scientists

https://bioinformatics.ccr.cancer.gov/docs/resources-for-bioinformatics/

Questions ?

Contacts:

- **Peter FitzGerald** fitzgepe@nih.gov
- **BTEP** ncibtep@nih.gov