Practical Guide to Interpreting RNA-seq Data

Skyler Kuhn[1,2]
Mayank Tandon[1,2]

1. CCR Collaborative Bioinformatics Resource (CCBR), Center for Cancer Research, NCI
2. Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research

Overview

I. Experimental Design

- Hypothesis-driven
- Overview of Best Practice

II. Quality-control

- Pre- and post- alignment QC metrics
- Interpretation

III. Pipeline

- FastQ Files -> Counts matrix
- Reproducibility

IV. Downstream Analysis

- Principal Components Analysis (PCA)
- Differential Expression
- Pathway Analysis

V. Advanced Visualizations

- Group comparisons
- Alternative Splicing Events
- Pathway Diagrams

I.
Experimental Design

I. Experimental Design: Overview

Hypothesis-driven

Addresses a well thought-out quantifiable question

Considerations:

Library Construction: mRNA versus total RNA

Single-end versus Paired-end Sequencing

Sequencing Depth: quantifying gene-level or transcript-level expression

Number of Replicates: statistical-power and ability drop a *bad* sample

Reducing Batch Effects

I. Experimental Design: Library Construction

Total RNA contains high-levels of ribosomal RNA (rRNA): 80%

mRNA

poly(A) selection ~ standard profiling for gene expression

Low RIN may results in 3’ bias

Total RNA

rRNA depletion

mRNA + non-coding RNA species (lncRNA)

Prokaryotic samples

I. Experimental Design: Sequencing Depth

mRNA: poly(A)-selection

Recommended Sequencing Depth: 10-20M paired-end reads (or 20-40M reads)

RNA must be high quality (RIN > 8)

Total RNA: rRNA depletion

Recommended Sequencing Depth: 25-60M paired-end reads (or 50-120M reads)

RNA must be high quality (RIN > 8)

* ***Differential Isoform regulation or alternative splicing events***: > 100M paired-end reads

I. Experimental Design: Number of Replicates

Recommended

Biological Replicates > Technical Replicates

Number of Replicates: 4

Peace-of-mind: Ability drop a *bad* sample without compromising statistical power

Bare Minimum

Biological Replicates > Technical Replicates

Number of Replicates: 3

I. Experimental Design: Reducing Batch Effects

Unwanted sources of technical variation

- Decrease batch effects by uniform processing

Protocol-driven

- Different Lab Technicians
- Different processing times
- Different Reagent Lots

Sequencing

- Lane effect

Sample Name	Group	Batch	Batch*
Treatment_r1	KO	1	1
Treatment_r2	KO	2	1
Treatment_r3	KO	1	1
Treatment_r4	KO	2	1
Cntrl_r1	WT	1	2
Cntrl_r2	WT	2	2
Cntrl_r3	WT	1	2
Cntrl_r4	WT	2	2

*** Confounded Groups and Batches!**

II.
Quality Control

II. Quality-control: Overview

No need to reinvent the wheel... but there are a lot of wheels!

Pre-alignment Quality-control

- Sequencing Quality
- Contamination Screening

Post-alignment Quality-control

- Alignment Quality

Aggregation and Interpretation

- MultiQC Report
- QC metric guidelines

II. Quality-control: Pre-alignment

Sequencing Quality

FastQC: run twice on *raw* and *trimmed* data

Contamination Screening

FastQ Screen

Kraken

BioBloom

II. Quality-control: Pre-alignment

FastQC (raw) → Adapter Trimming → **FastQC (trimmed)**

FastQC

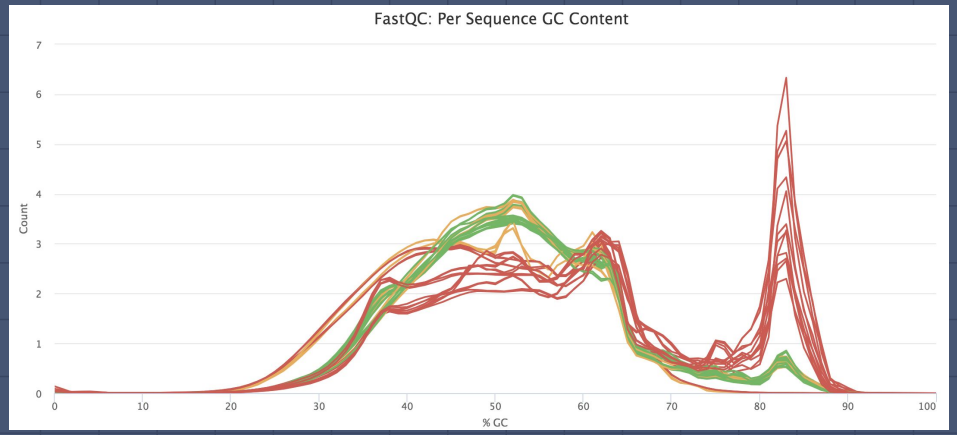
Identify potential problems that can arise during sequencing or library prep

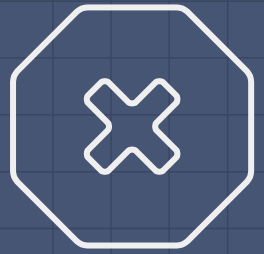
Run on raw reads (pre-adapter removal) and trimmed reads (post-adapter removal)

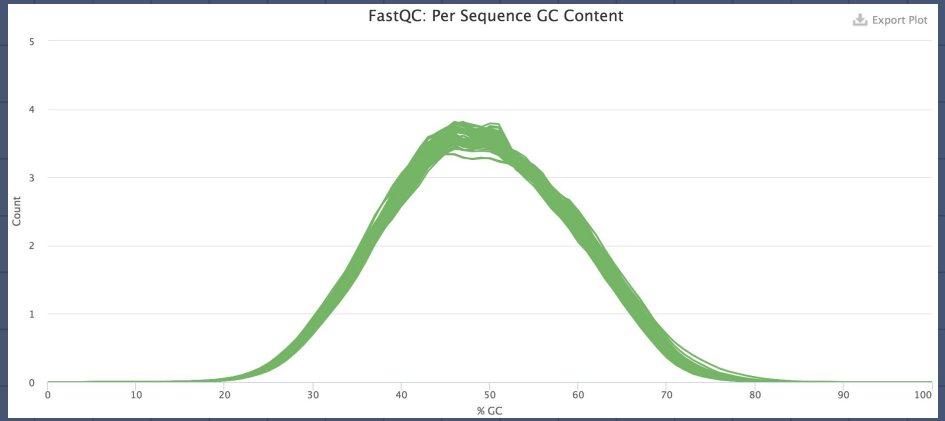
Summarizes:

- Per base and per sequence quality scores
- Per sequence GC content
- Per sequence adapter content
- Per sequence read lengths
- Overrepresented sequences

II. Quality-control: FastQC







II. Quality-control: Pre-alignment

Adapter Trimming | **Contamination Screen** | Alignment

FastQ Screen

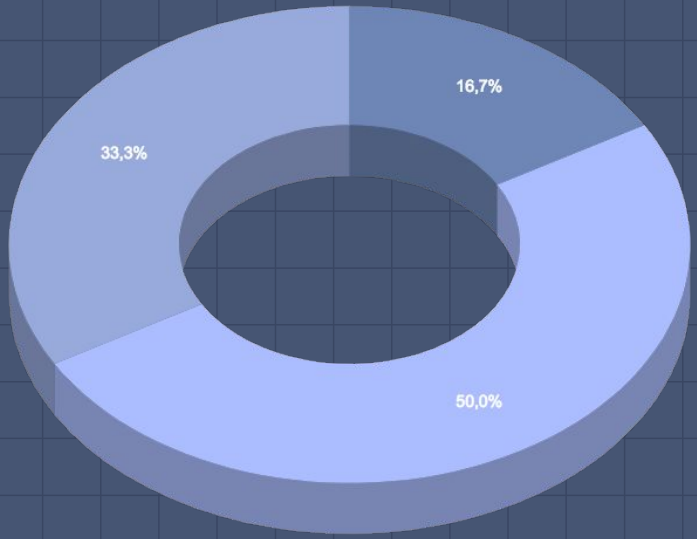
Aligns to Human, Mouse, Fungi, Bacteria, Viral references

Easy to interpret and important QC step

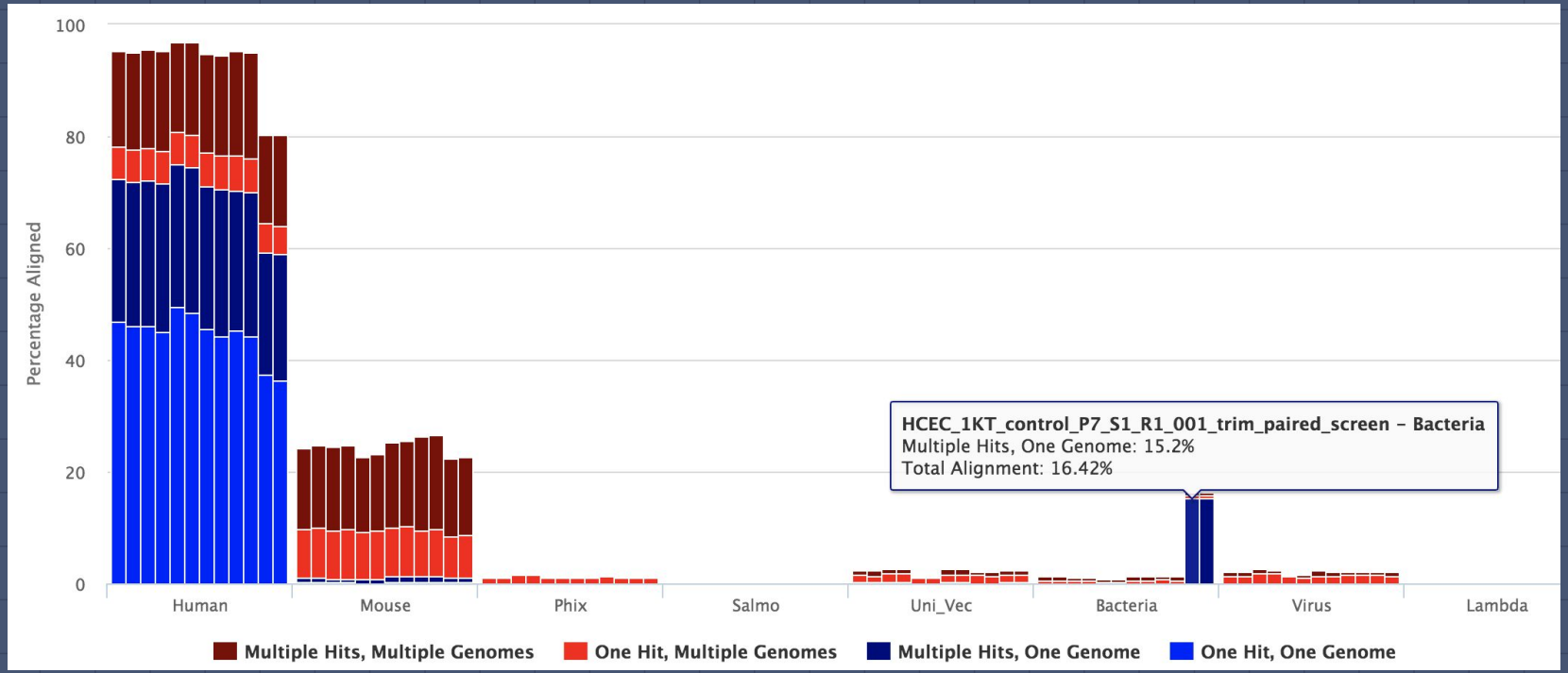
Kraken

Taxonomic composition of microbial contamination

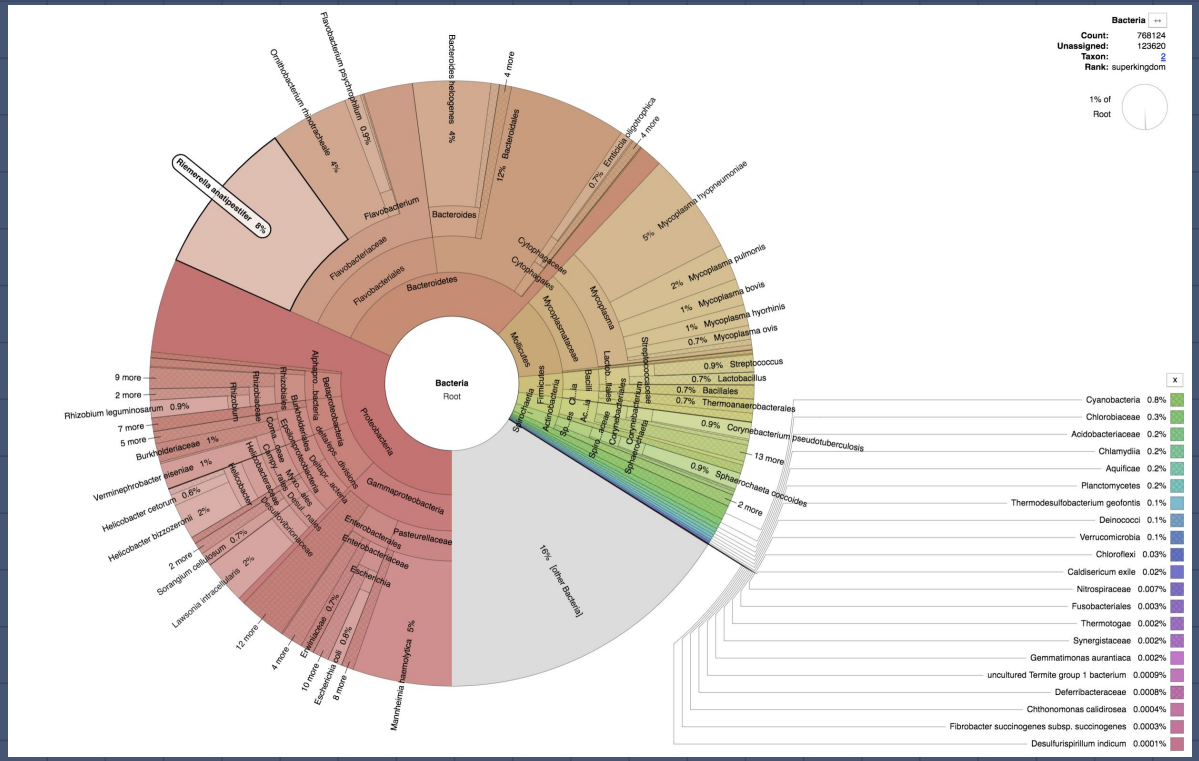
- Archaea
- Bacteria
- Plasmid
- Viral



FastQ Screen **Contamination Screening**



Kraken + Krona **Microbial Taxonomic Composition**



II. Quality-control: Post-alignment

Alignment → **Alignment Quality** → Quantify Counts

Preseq

Estimates library complexity

Picard RNAseqMetrics

Number of reads that align to coding, intronic, UTR, intergenic, ribosomal regions

Normalize gene coverage across a meta-gene body

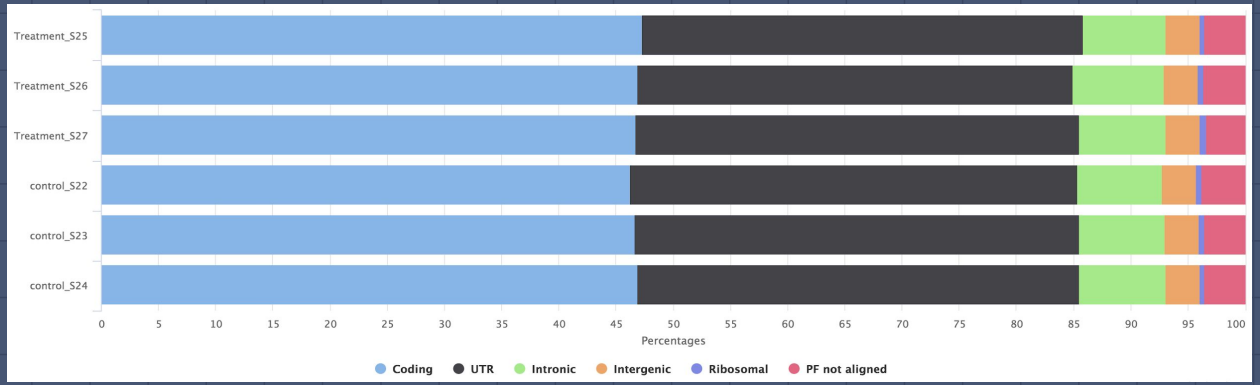
- Identify 5’ or 3’ bias

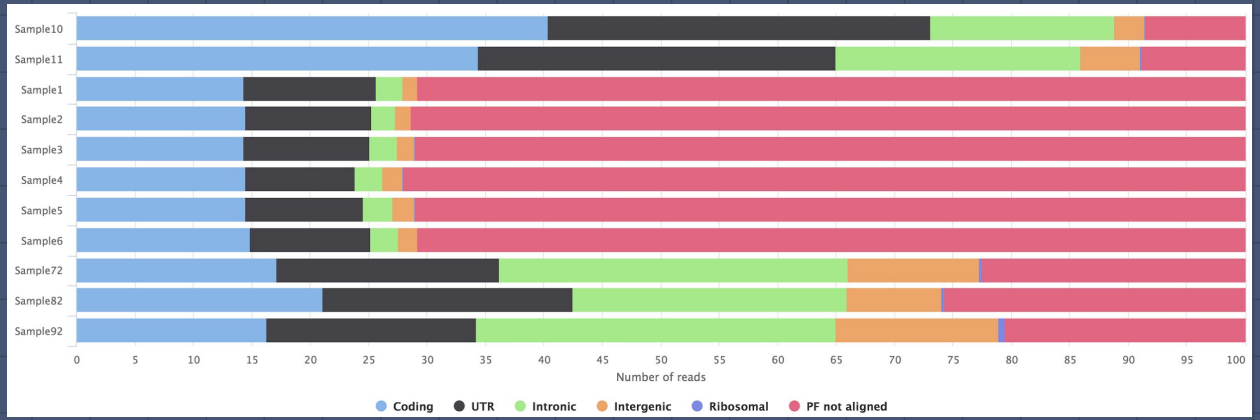
RSeQC

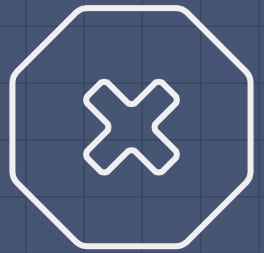
Suite of tools to assess various post-alignment quality

- Calculate distribution of Insert Size
- Junction Annotation (% Known, % Novel read spanning splice junctions)
- BAM to BigWig (Visual Inspection with IGV)

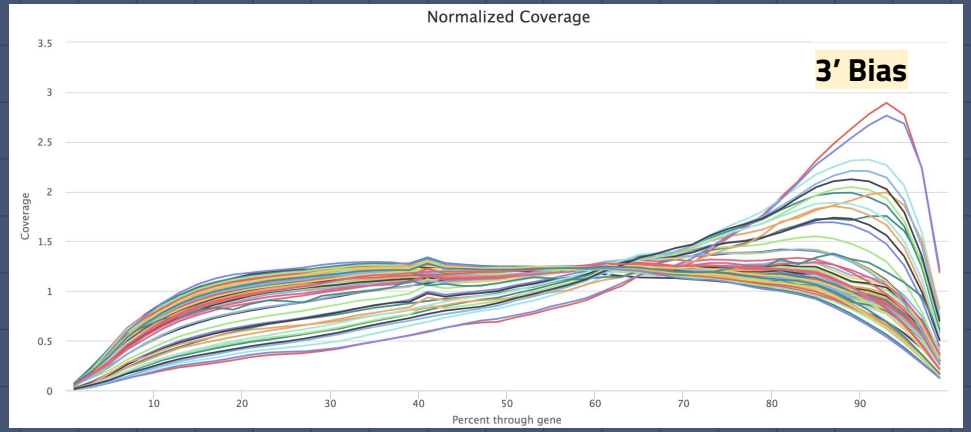
CollectRnaseqMetrics **Alignment Summary**

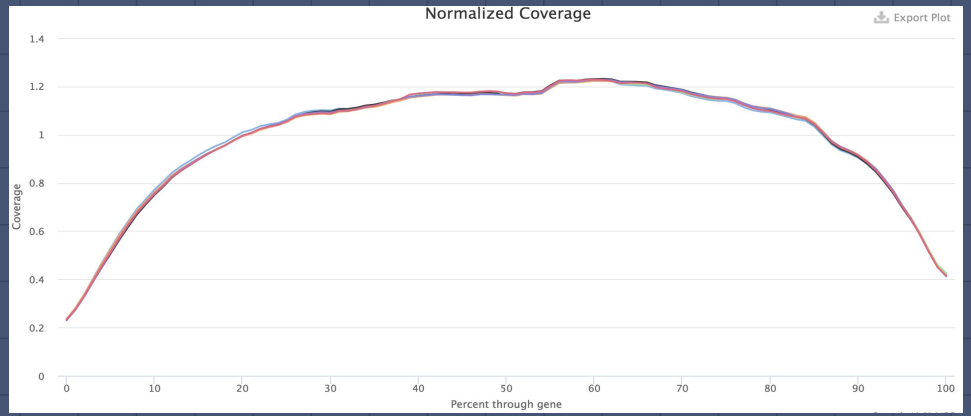






Picard CollectRnaseqMetrics **Normalized Gene Coverage**





II. Quality-control: Aggregation

MultiQC

HTML report that aggregates information across all samples

- Plots, filtering, and highlighting

Highly customizable with great documentation

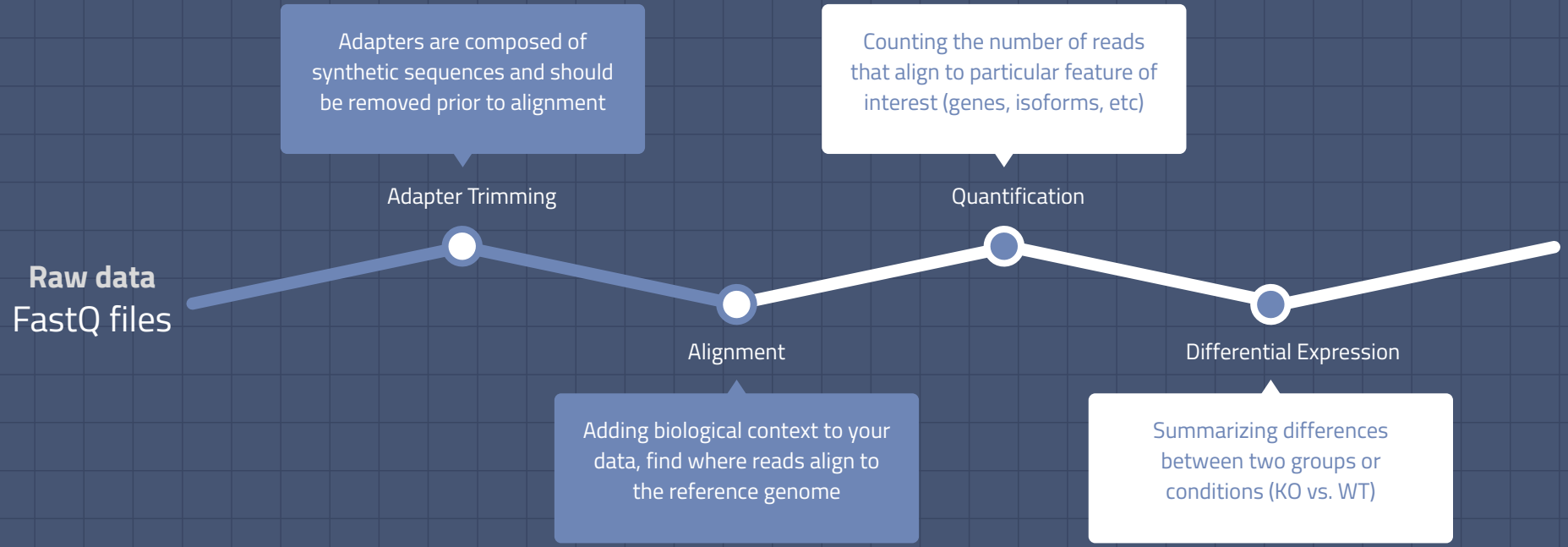
- Add text and embed custom figures
- Create your own module to extend missing functionality

Supports over 73 commonly-used open source bioinformatics tools

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%

III.
Pipeline

III. Processing Pipeline **Conceptual Diagram**



III. Processing Pipeline **Practical Example**

Cutadapt → STAR → RSEM

FastQC: Pre- and post- trimming

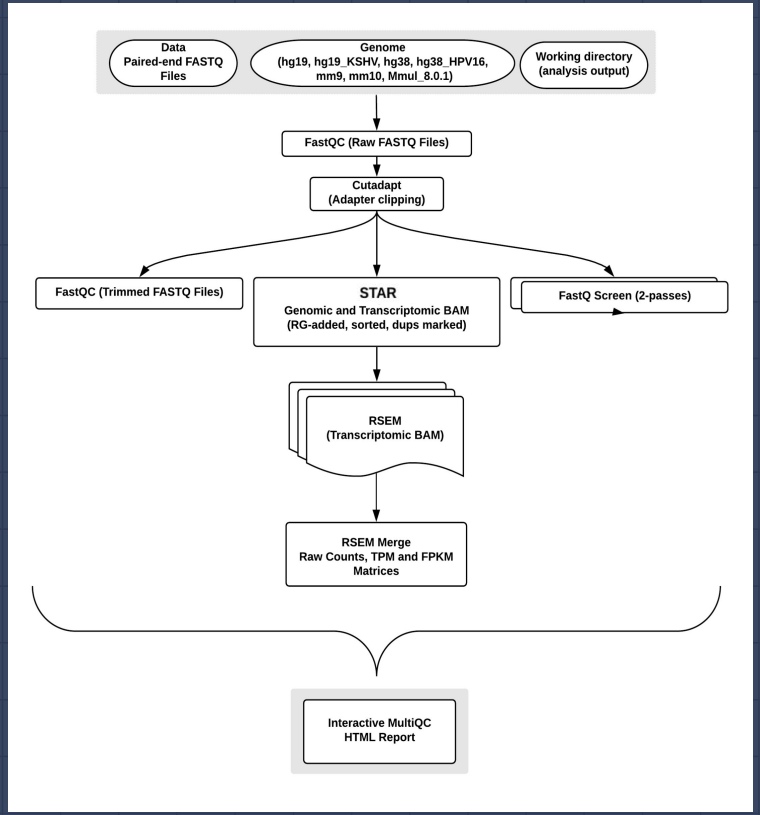
Cutadapt: Remove adapters

FastQ Screen: Run twice on different set of references

STAR: Splice-aware aligner

RSEM: Generates gene and isoform counts

MultiQC: Aggregates everything into an HTML report



FastQ files to raw counts matrix

III. Processing Pipeline: Reproducibility

Workflow management systems

Snakemake, Nextflow

Package management

No active management: rat's nest of interdependencies prone to break

Python: virtual environments

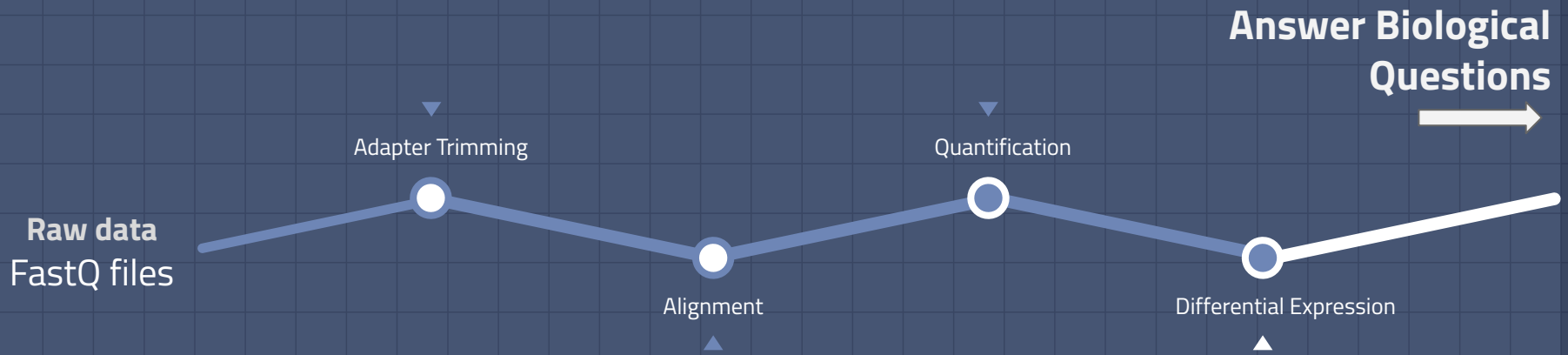
Conda: Python, R, Scala, Java, C/C++, FORTRAN

Docker or Singularity: Portability and high reproducibility

IV.
Downstream Analysis

IV. Downstream Analysis

- ✓ Step 1: Think
- ✓ Step 2: Analyze
- ✓ Step 3: QC
- ☐ ???
- ☐ Step 4: Nobel Prize!



IV. Downstream Analysis

Principal Components Analysis (PCA)

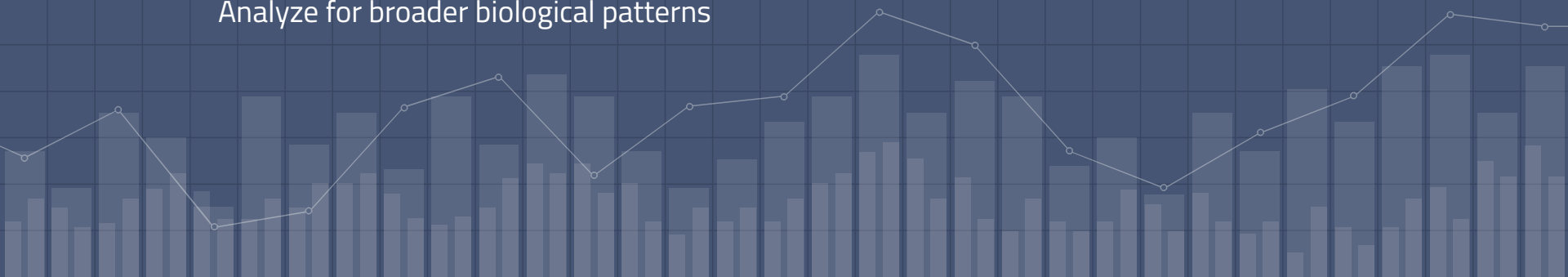
Data summarization, visualization, and QC tool

Differential Expression

Find genes that are different between groups of interest

Pathway Enrichment

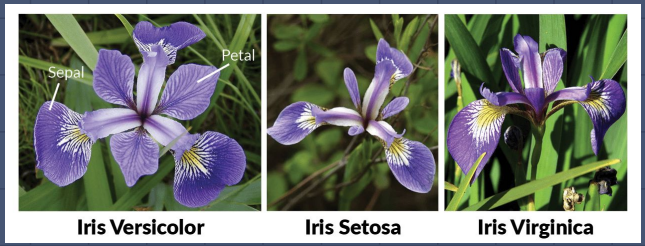
Analyze for broader biological patterns



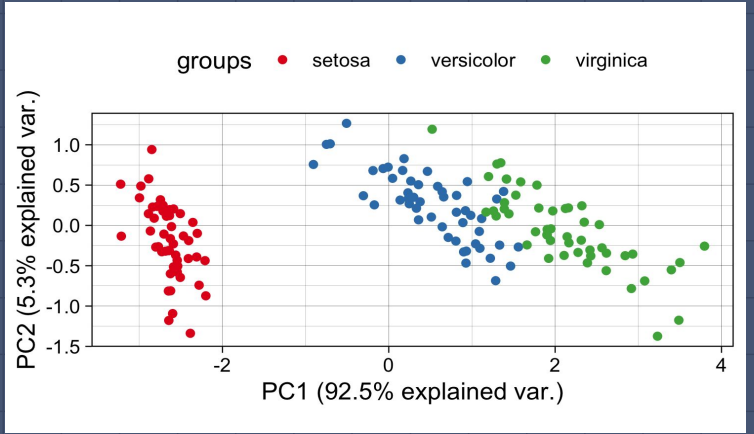
IV. Downstream Analysis: **PCA**

Principal Components Analysis (PCA)

- Dimensionality reduction technique
- Captures patterns of variance into singular values
- Visualizes global transcriptomic patterns



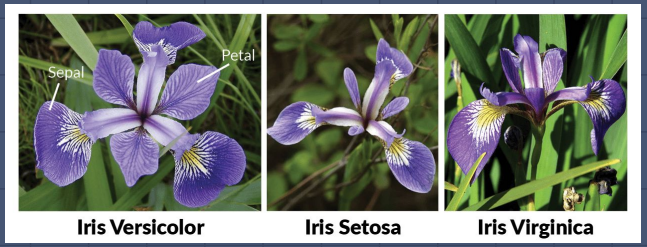
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
4.9	2.5	4.5	1.7	virginica
4.4	3.0	1.3	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.1	3.7	1.5	0.4	setosa
5.7	3.8	1.7	0.3	setosa
6.3	2.5	5.0	1.9	virginica
6.3	3.3	6.0	2.5	virginica
5.4	3.4	1.7	0.2	setosa
6.4	3.1	5.5	1.8	virginica
6.1	3.0	4.6	1.4	versicolor
5.9	3.0	5.1	1.8	virginica

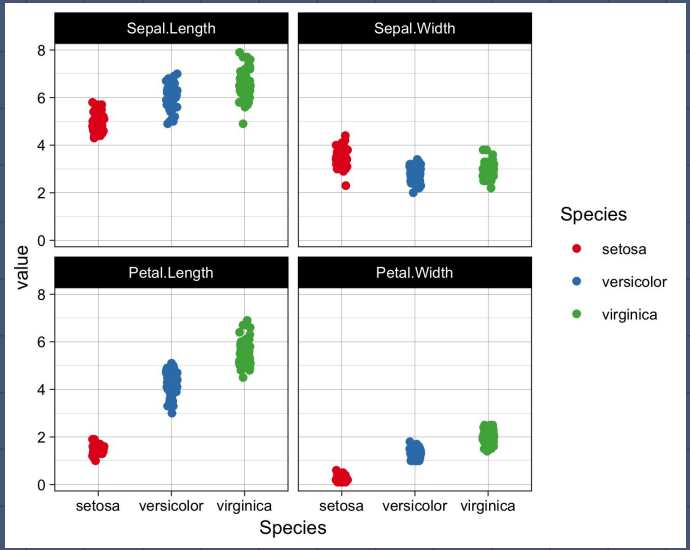


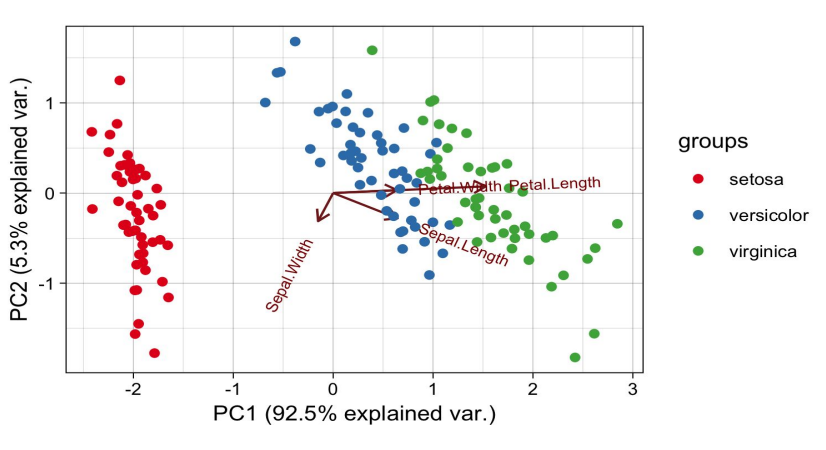
IV. Downstream Analysis: **PCA**

Principal Components Analysis (PCA)

- Dimensionality reduction technique
- Captures patterns of variance into singular values
- Visualizes global transcriptomic patterns

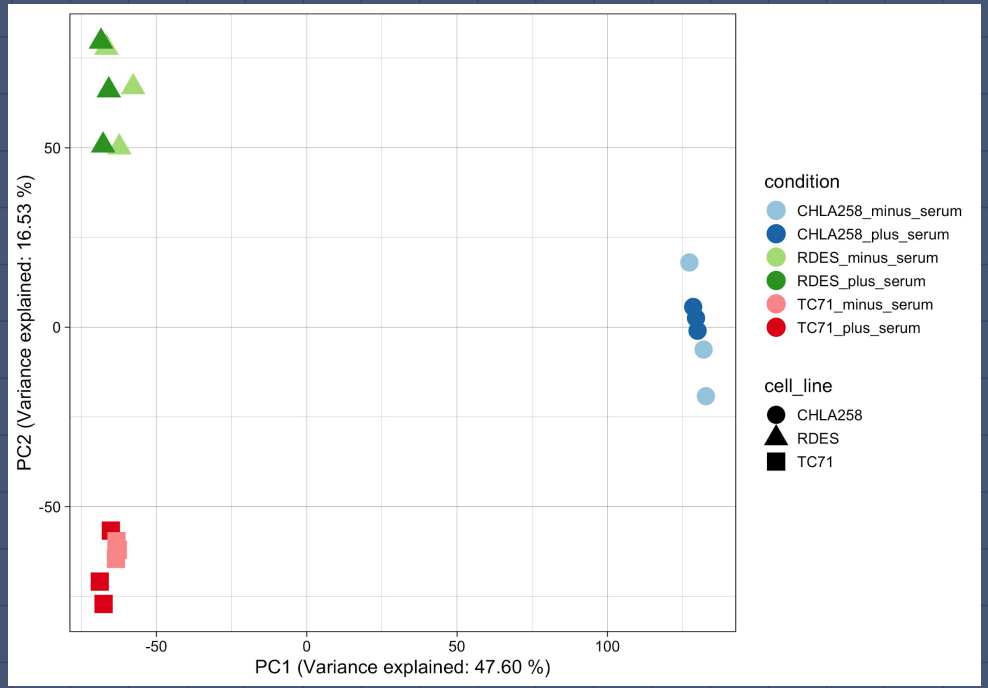






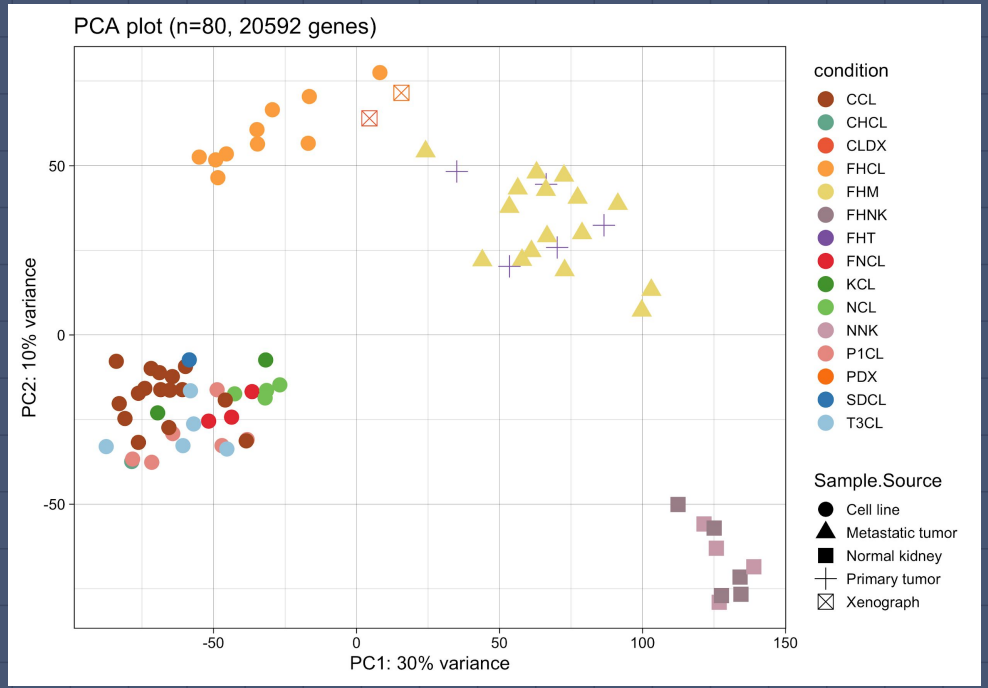
IV. Downstream Analysis: **PCA**

PCA can help drive biological insights...



IV. Downstream Analysis: **PCA**

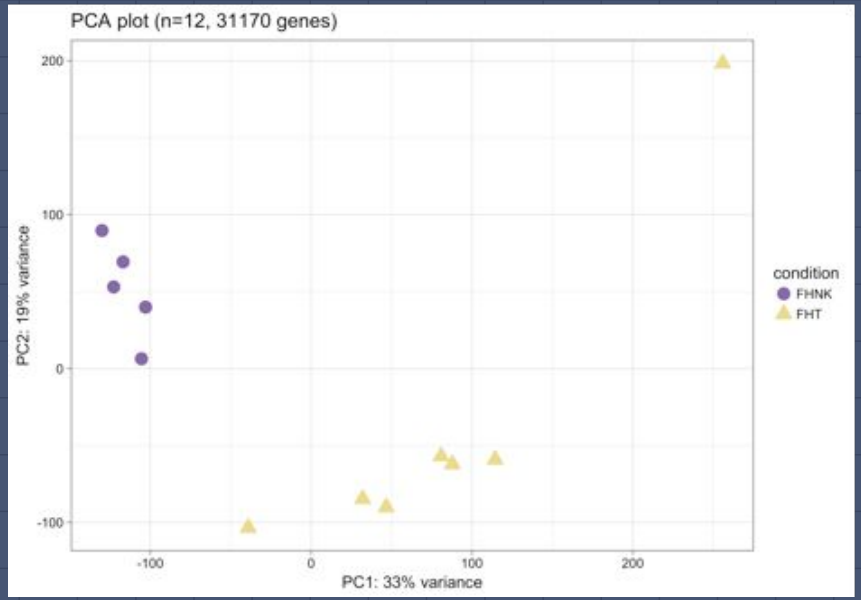
PCA can help drive biological insights...



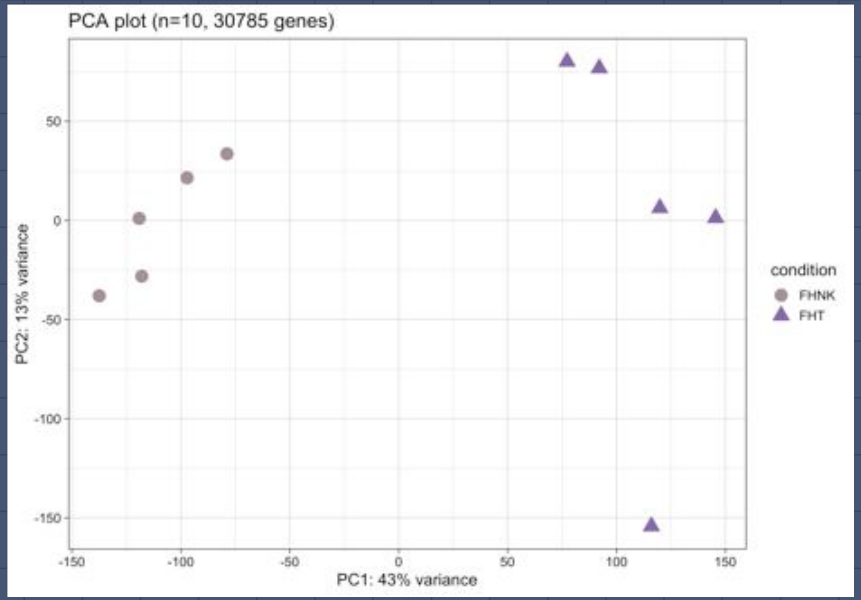
IV. Downstream Analysis: **PCA**

… or be used as a QC tool

Original



Outliers Removed



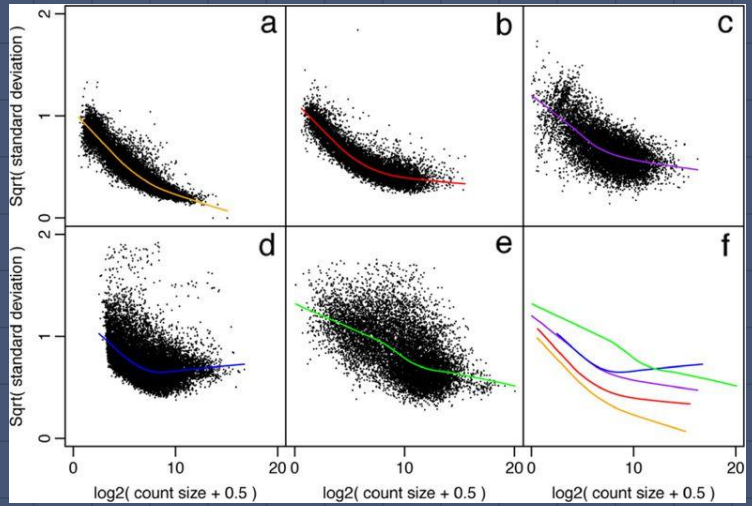
IV. Downstream Analysis: **Differential Expression**

Goal: Identify genes or transcripts that vary due to *biological* effects

Question: Can't I just use a t-test to do that?

Answer: Sure. But data are noisy... bad idea

So we apply normalization and/or employ specialized statistical tests.



Law, C. W., et al. (2014). "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." Genome Biol **15**(2): R29.

IV. Downstream Analysis: **Differential Expression**

Table 1:

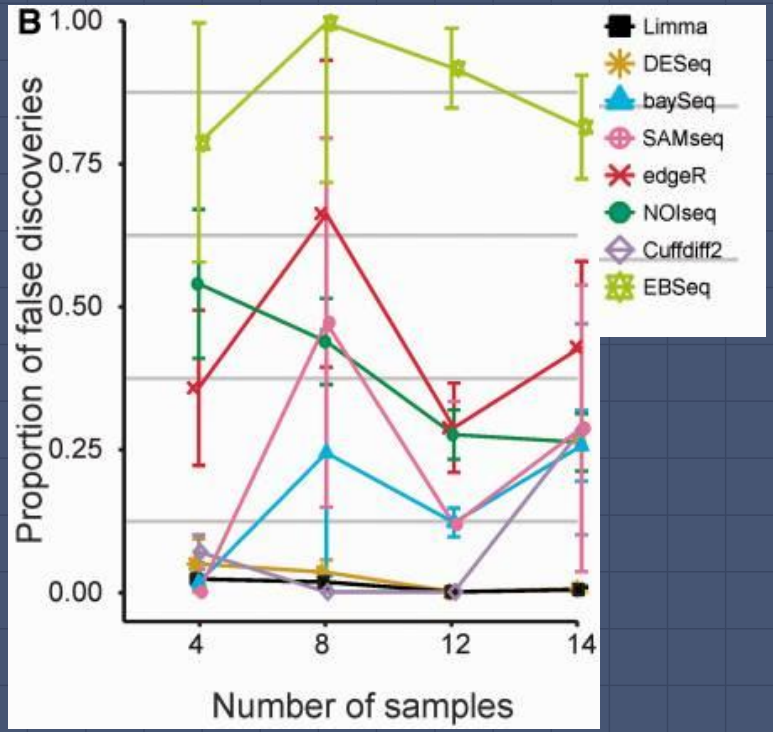
Software packages for detecting differential expression

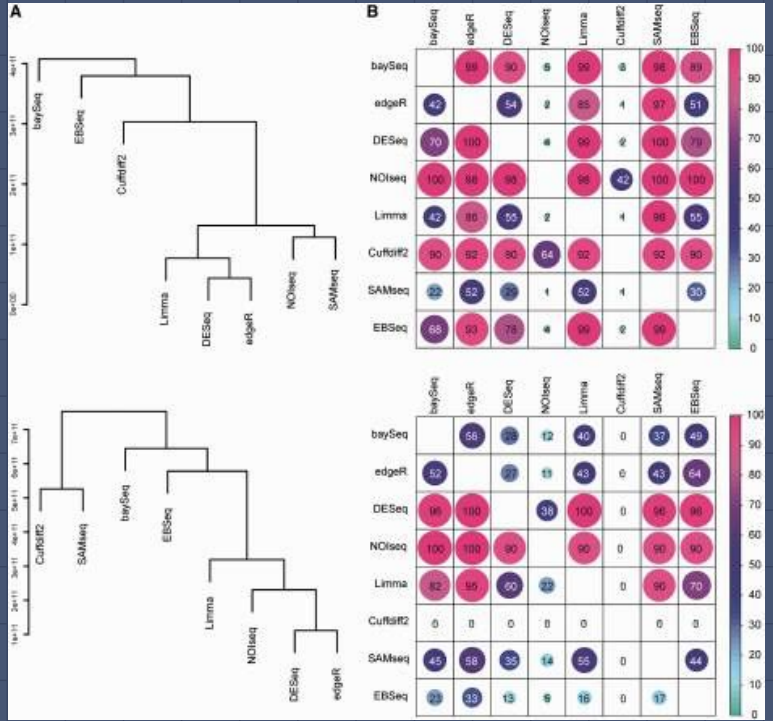
Method	Version	Reference	Normalization[a]	Read count distribution assumption	Differential expression test
edgeR	3.0.8	[4]	TMM/Upper quartile/RLE (DESeq-like)/None (all scaling factors are set to be one)	Negative binomial distribution	Exact test
DESeq	1.10.1	[5]	DESeq sizeFactors	Negative binomial distribution	Exact test
baySeq	1.12.0	[6]	Scaling factors (quantile/TMM/total)	Negative binomial distribution	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods
NOIseq	1.1.4	[7]	RPKM/TMM/Upper quartile	Nonparametric method	Contrasts fold changes and absolute differences within a condition to determine the null distribution and then compares the observed differences to this null
SAMseq (samr)	2.0	[8]	SAMseq specialized method based on the mean read count over the null features of the data set	Nonparametric method	Wilcoxon rank statistic and a resampling strategy
Limma	3.14.4	[9]	TMM	voom transformation of counts	Empirical Bayes method
Cuffdiff 2 (Cufflinks)	2.0.2-beta	[10]	Geometric (DESeq-like)/quartile/classic-fpkm	Beta negative binomial distribution	*t*-test
EBSeq	1.1.7	[11]	DESeq median normalization	Negative binomial distribution	Evaluates the posterior probability of differentially and non-differentially expressed entities (genes or isoforms) via empirical Bayesian methods

[a]In case of availability of several normalization methods, the default one is underlined.

Seyednasrollah, F., et al. (2015). "Comparison of software packages for detecting differential expression in RNA-seq studies." Brief Bioinform **16**(1): 59-70.

IV. Downstream Analysis: **Differential Expression**





Seyednasrollah, F., et al. (2015). "Comparison of software packages for detecting differential expression in RNA-seq studies." Brief Bioinform **16**(1): 59-70.

IV. Downstream Analysis: **Differential Expression**

Practical Rules of Thumb

Limma, DESeq2, and EdgeR will work be **very similarly** in most cases

- Consensus or intersection of the three is sometimes used

Limma works better with **larger** cohorts (7 or more samples per **group**)

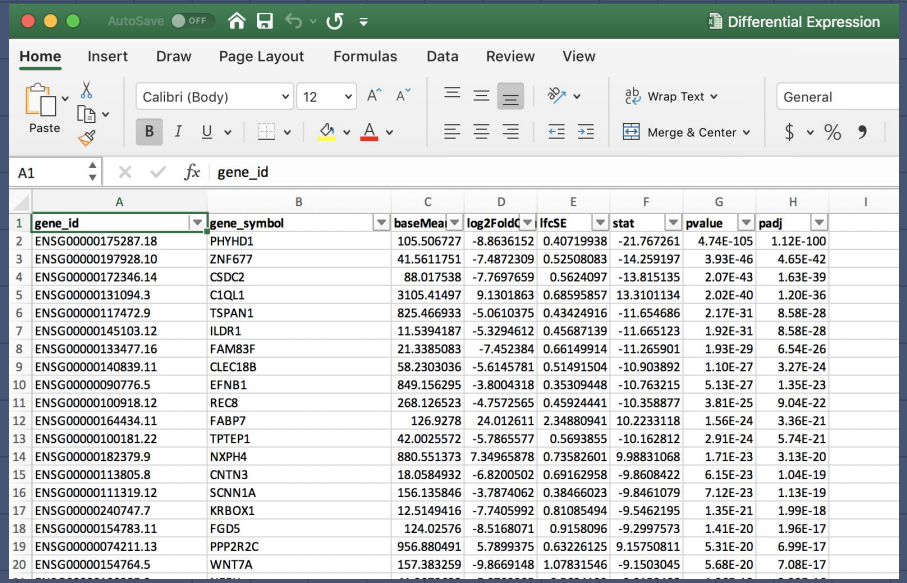
DESeq2 works better with **small** cohorts (3 or less per **group**)

- May also be more sensitive for low depth data

EdgeR provides convenience functions for converting to various normalized values

IV. Downstream Analysis: **Differential Expression**

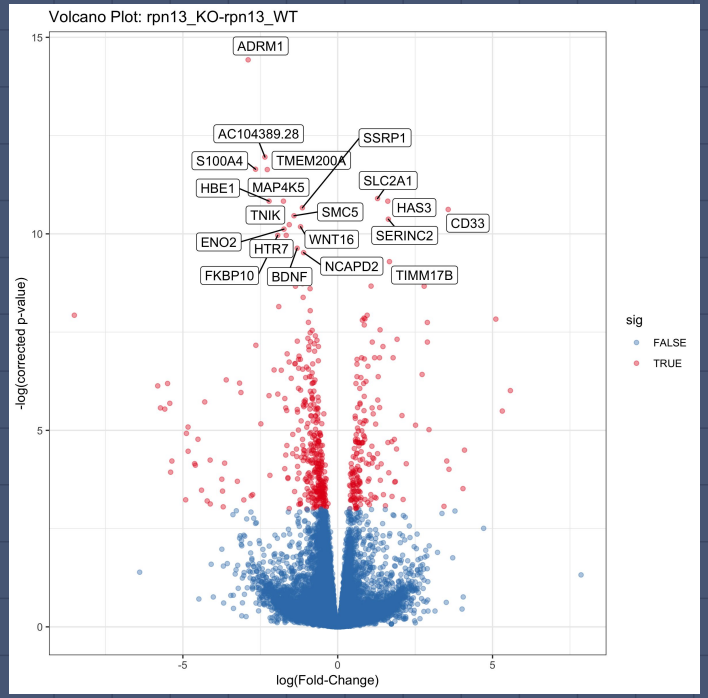
Output

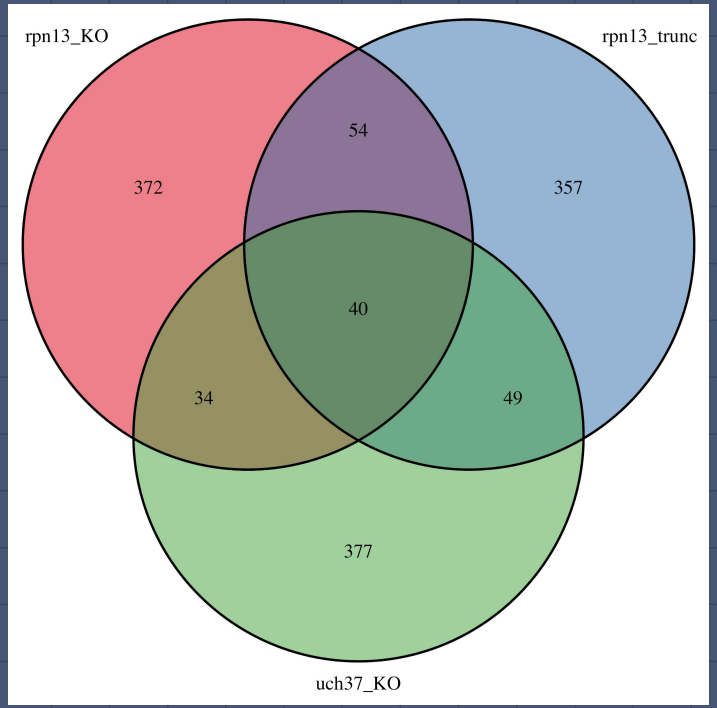


	A	B	C	D	E	F	G	H	I
1	gene_id	gene_symbol	baseMea	log2Fold	lfcSE	stat	pvalue	padj	
2	ENSG00000175287.18	PHYHD1	105.506727	-8.8636152	0.40719938	-21.767261	4.74E-105	1.12E-100	
3	ENSG00000197928.10	ZNF677	41.5611751	-7.4872309	0.52508083	-14.259197	3.93E-46	4.65E-42	
4	ENSG00000172346.14	CSDC2	88.017538	-7.7697659	0.5624097	-13.815135	2.07E-43	1.63E-39	
5	ENSG00000131094.3	C1QL1	3105.41497	9.1301863	0.68595857	13.3101134	2.02E-40	1.20E-36	
6	ENSG00000117472.9	TSPAN1	825.466933	-5.0610375	0.43424916	-11.654686	2.17E-31	8.58E-28	
7	ENSG00000145103.12	ILDR1	11.5394187	-5.3294612	0.45687139	-11.665123	1.92E-31	8.58E-28	
8	ENSG00000133477.16	FAM83F	21.3385083	-7.452384	0.66149914	-11.265901	1.93E-29	6.54E-26	
9	ENSG00000140839.11	CLEC18B	58.2303036	-5.6145781	0.51491504	-10.903892	1.10E-27	3.27E-24	
10	ENSG00000090776.5	EFNB1	849.156295	-3.8004318	0.35309448	-10.763215	5.13E-27	1.35E-23	
11	ENSG00000100918.12	REC8	268.126523	-4.7572565	0.45924441	-10.358877	3.81E-25	9.04E-22	
12	ENSG00000164434.11	FABP7	126.9278	24.012611	2.34880941	10.2233118	1.56E-24	3.36E-21	
13	ENSG00000100181.22	TPTEP1	42.0025572	-5.7865577	0.5693855	-10.162812	2.91E-24	5.74E-21	
14	ENSG00000182379.9	NXPH4	880.551373	7.34965878	0.73582601	9.98831068	1.71E-23	3.13E-20	
15	ENSG00000113805.8	CNTN3	18.0584932	-6.8200502	0.69162958	-9.8608422	6.15E-23	1.04E-19	
16	ENSG00000111319.12	SCNN1A	156.135846	-3.7874062	0.38466023	-9.8461079	7.12E-23	1.13E-19	
17	ENSG00000240747.7	KRBOX1	12.5149416	-7.7405992	0.81085494	-9.5462195	1.35E-21	1.99E-18	
18	ENSG00000154783.11	FGD5	124.02576	-8.5168071	0.9158096	-9.2997573	1.41E-20	1.96E-17	
19	ENSG00000074211.13	PPP2R2C	956.880491	5.7899375	0.63226125	9.15750811	5.31E-20	6.99E-17	
20	ENSG00000154764.5	WNT7A	157.383259	-9.8669148	1.07831546	-9.1503045	5.68E-20	7.08E-17	

IV. Downstream Analysis: **Differential Expression**

Output





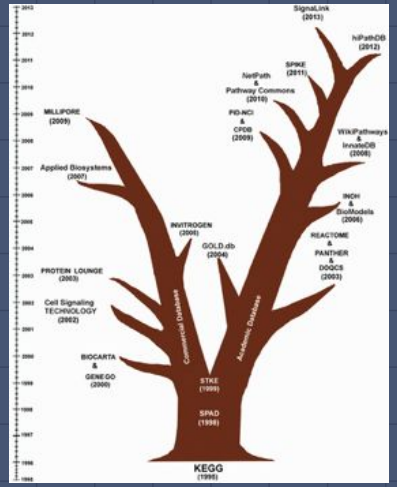
IV. Downstream Analysis: **Pathway Enrichment**

Gene annotation and network databases capture biological meaning

- Manual curation, text mining
- Gene function and/or interactions

Dozens of databases and hundreds of tools

- Depends on how you want to look at gene-pathway relationships



IV. Downstream Analysis: **Pathway Enrichment**

Types of pathway analysis

Simple enrichment test: Qualitative

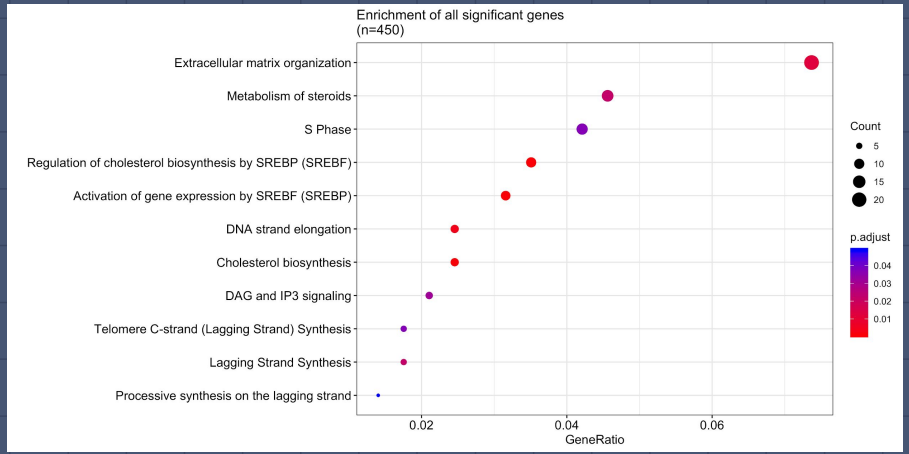
- Fisher's Exact Test
- Hypergeometric test

Enrichment algorithms: Quantitative

- GSEA (Broad Institute)

Network Analysis

Commercial vs. open source



IV. Downstream Analysis: **Pathway Enrichment**

Types of pathway analysis

Simple enrichment test: Qualitative

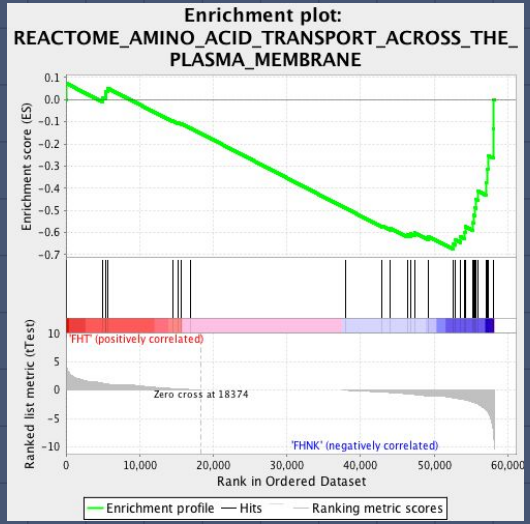
- Fisher's Exact Test
- Hypergeometric test

Enrichment algorithms: Quantitative

- GSEA (Broad Institute)

Network Analysis

Commercial vs. open source



IV. Downstream Analysis: **Pathway Enrichment**

Types of pathway analysis

Simple enrichment test: Qualitative

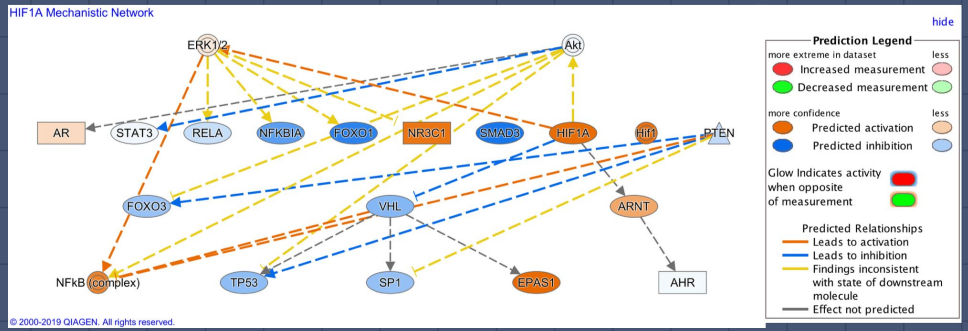
- Fisher's Exact Test
- Hypergeometric test

Enrichment algorithms: Quantitative

- GSEA (Broad Institute)

Network Analysis

Commercial vs. open source



IV. Downstream Analysis: Pathway Enrichment

Types of pathway analysis

	Commercial		Open-source				
	IPA	MetaCore	GSEA	Reactome	KEGG	PANTHER	DAVID
Enrichment Test	x	x	x	x	(x)	x	x
Enrichment Scoring Algorithm	x	x	x		(x)		
Network Analysis	x	x		x	(x)		
Graphical Interface	x	x	x	x	x	x	x

V.
Visualizations

V. Visualizations of RNA-Seq Data

Group comparisons of pathway enrichment

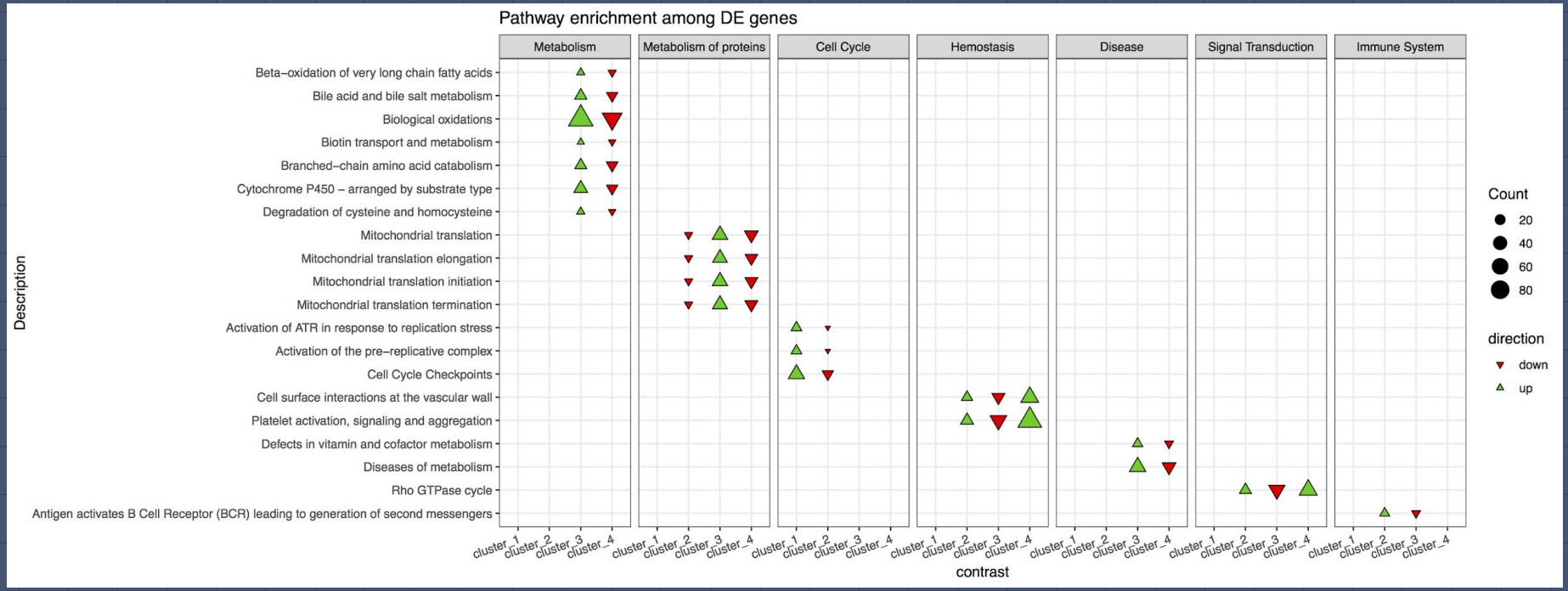
- Heatmaps
- Visualizing Set Overlap
- Dotplots

Sashimi plots

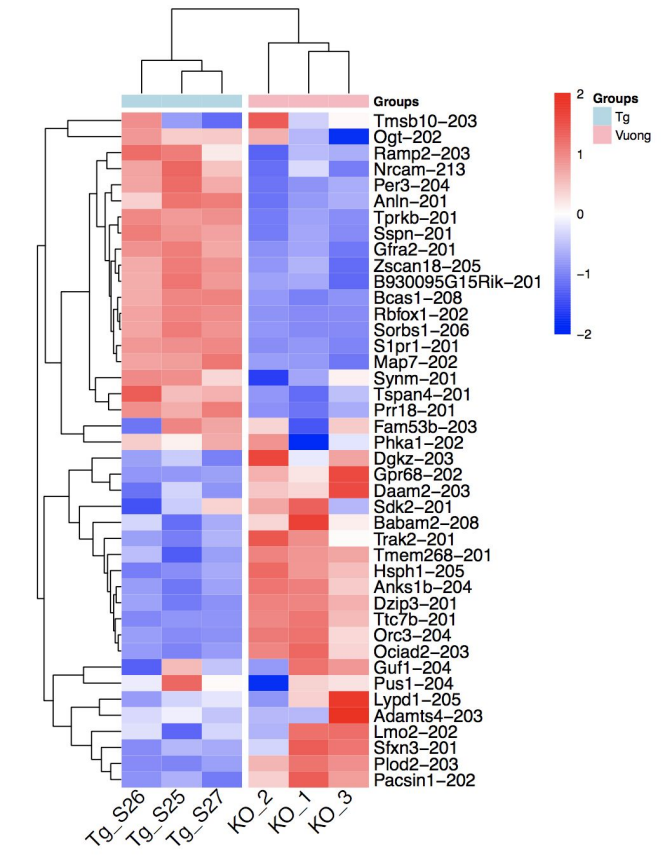
- Alternative Splicing

V. Visualizations: **Group Enrichment**

Group comparison of pathway enrichment: Simple Enrichment Test



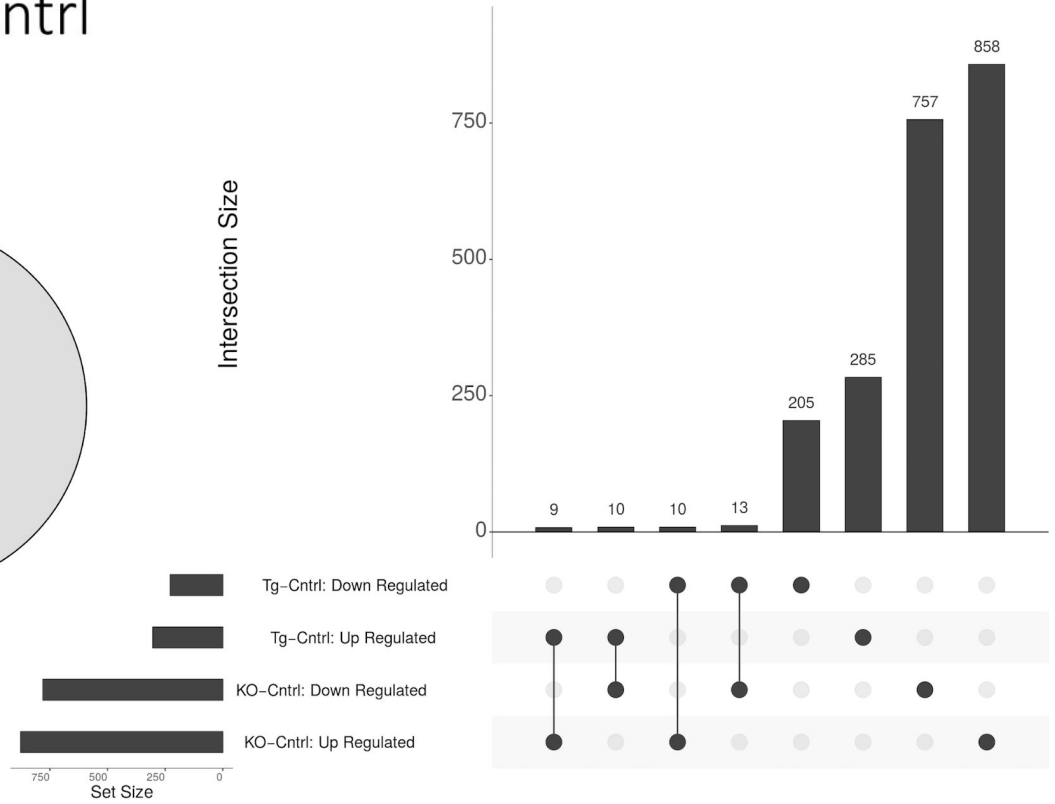
V. Visualizations: **Expression Heatmap**



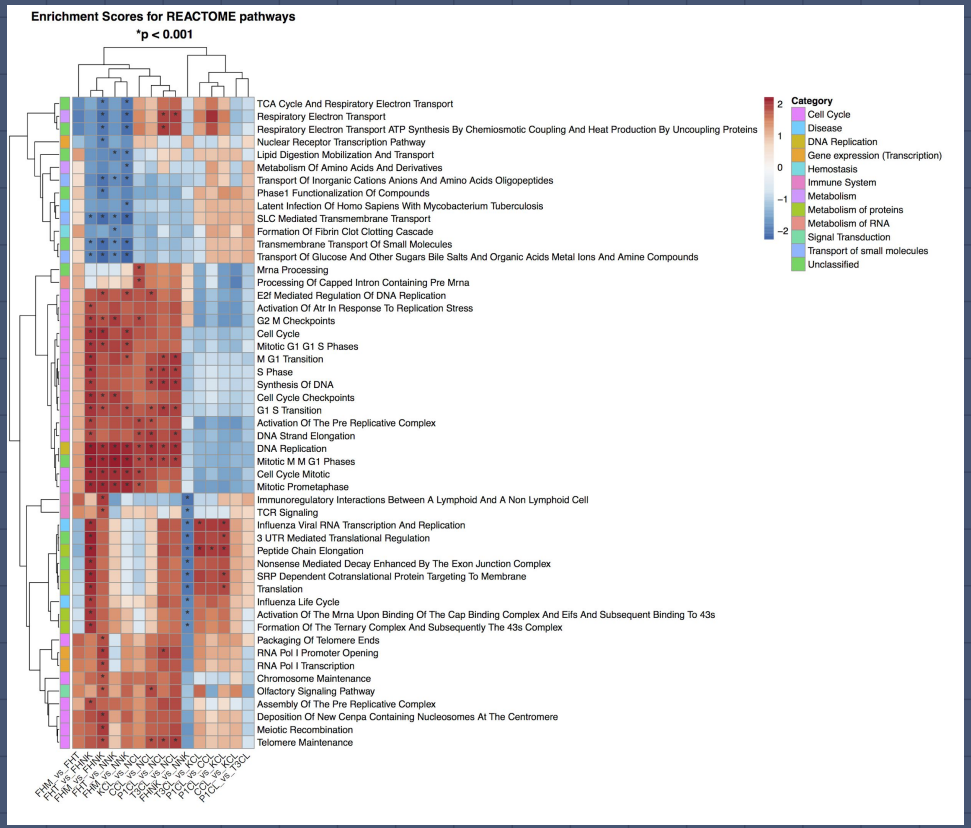
V. Visualizations: **Set Intersection**

Tg-Cntrl ∩ KO-Cntrl

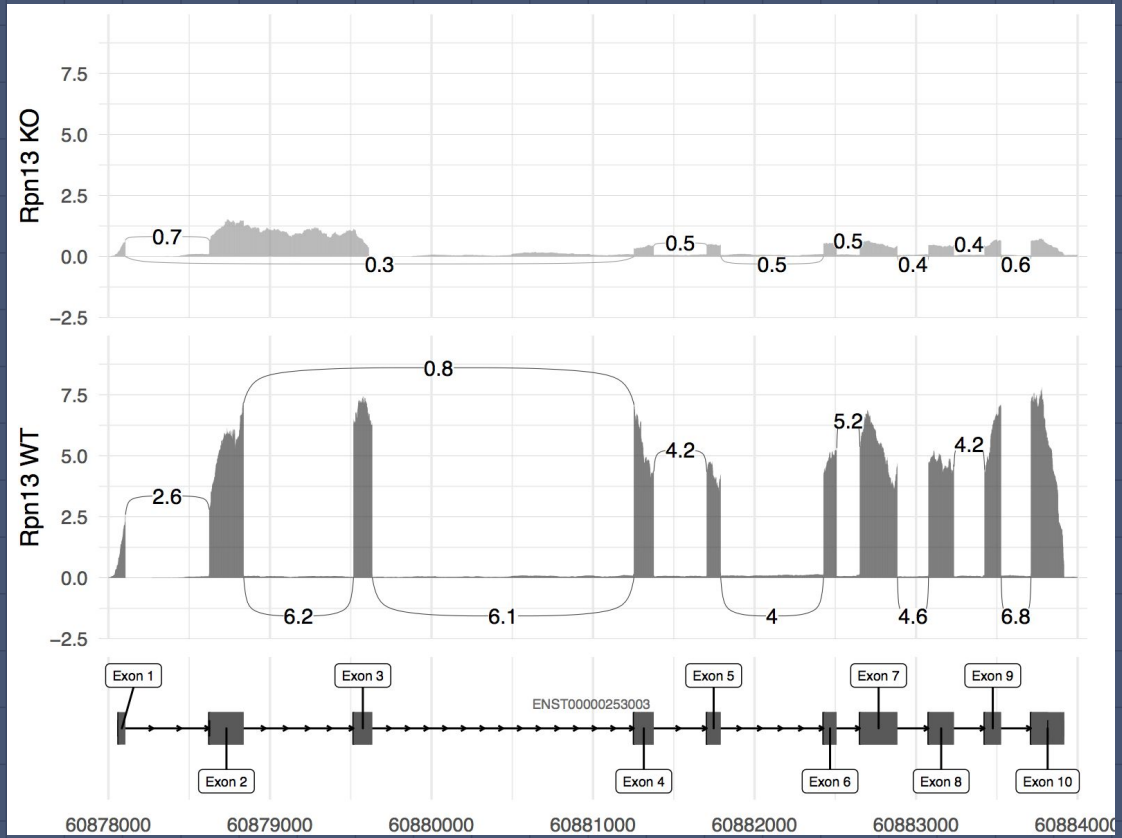
Isoform Regulation



V. Visualizations: **Pathway enrichment**



V. Visualizations: **Sashimi Plot**



Conclusions

Think BEFORE you sequence!

This is a three-way partnership: bench → sequencing → analysis

- Everyone should agree on experimental design, platform, approach

QC is extremely important!

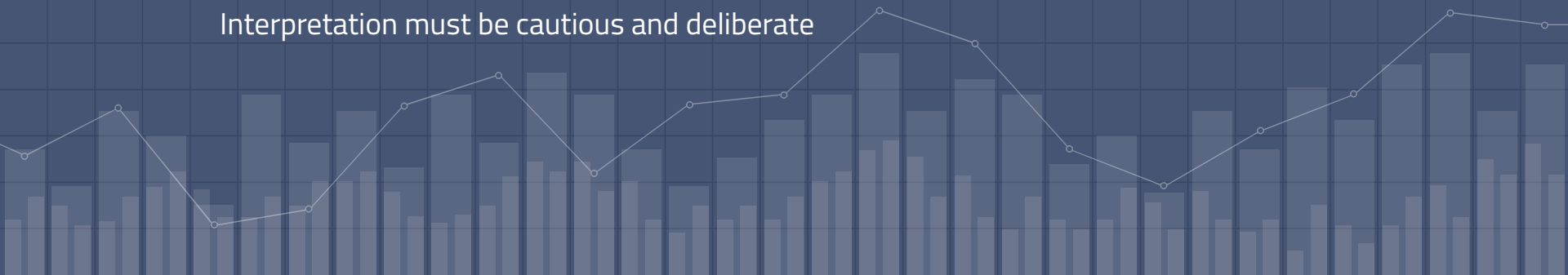
There is no need to reinvent the wheel… *but there are a lot of wheels*

Garbage in, Garbage out!

- Only some problems can be fixed bioinformatically

There will always be significant changes detected

Interpretation must be cautious and deliberate

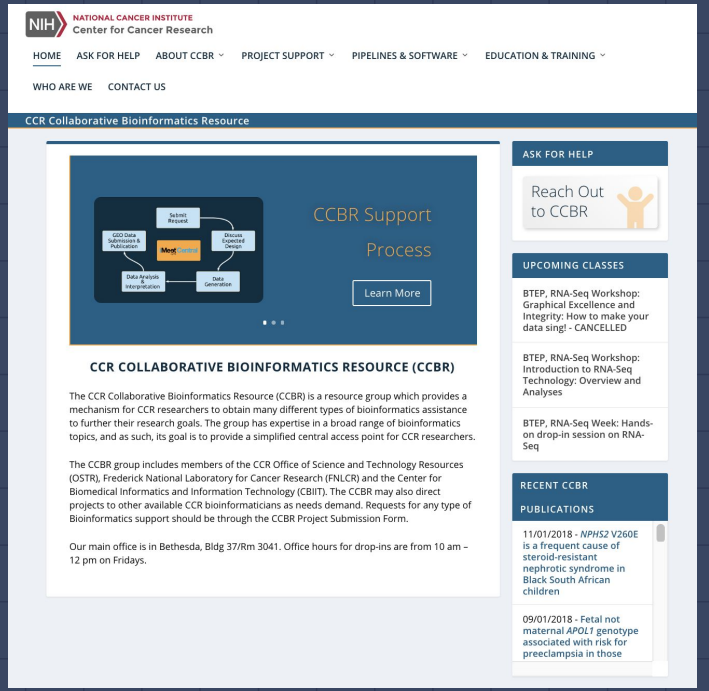


THANKS!

Acknowledgements

CCBR, NCBR, and GAU members

Any questions?



Cost-Benefit Considerations

Caveats:

Expected reads/sample based on **maximum possible yield**

Typical runs likely yield 80% of max

Different platforms may have different turnaround times depending on queue length and popularity

Library Prep cost is not included here:
$50-84 depending on type of kit

	MiSeq	NextSeq	HiSeq 4000	Novaseq
Run Time	4–55 hours	12–30 hours	< 1–3.5 days	~13 - 44 hours
Max Output	15 Gb	120 Gb	1500 Gb	6000 Gb
Max Reads Per Run	25 million	400 million	5 billion	20 billion
Lanes	1	1	8	4
Maximum Read Length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 x 250**
Cost from SF	$623	$1956	$1007/lane	$4382/lane
Max Coverage (12 samples)	**2 million reads**	**33 million reads**	**52 million reads**	**416 million reads**
$ per sample (12 samples)	**$51.91**	**$163.92**	**$83.92**	**$365.16**

https://emea.illumina.com/systems/sequencing-platforms.html?langsel=/ch/