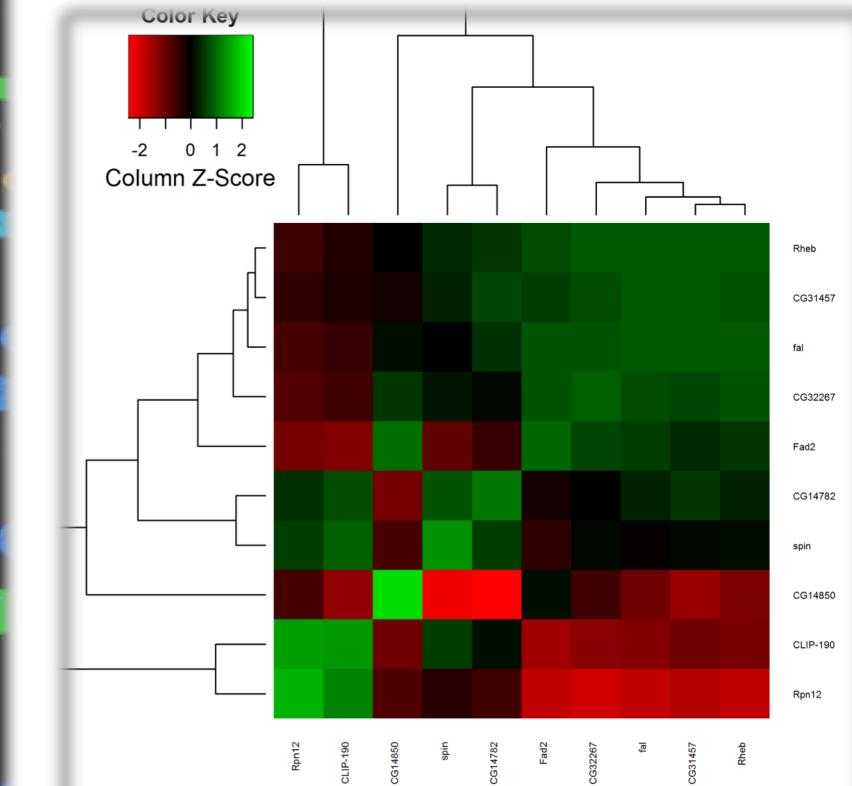
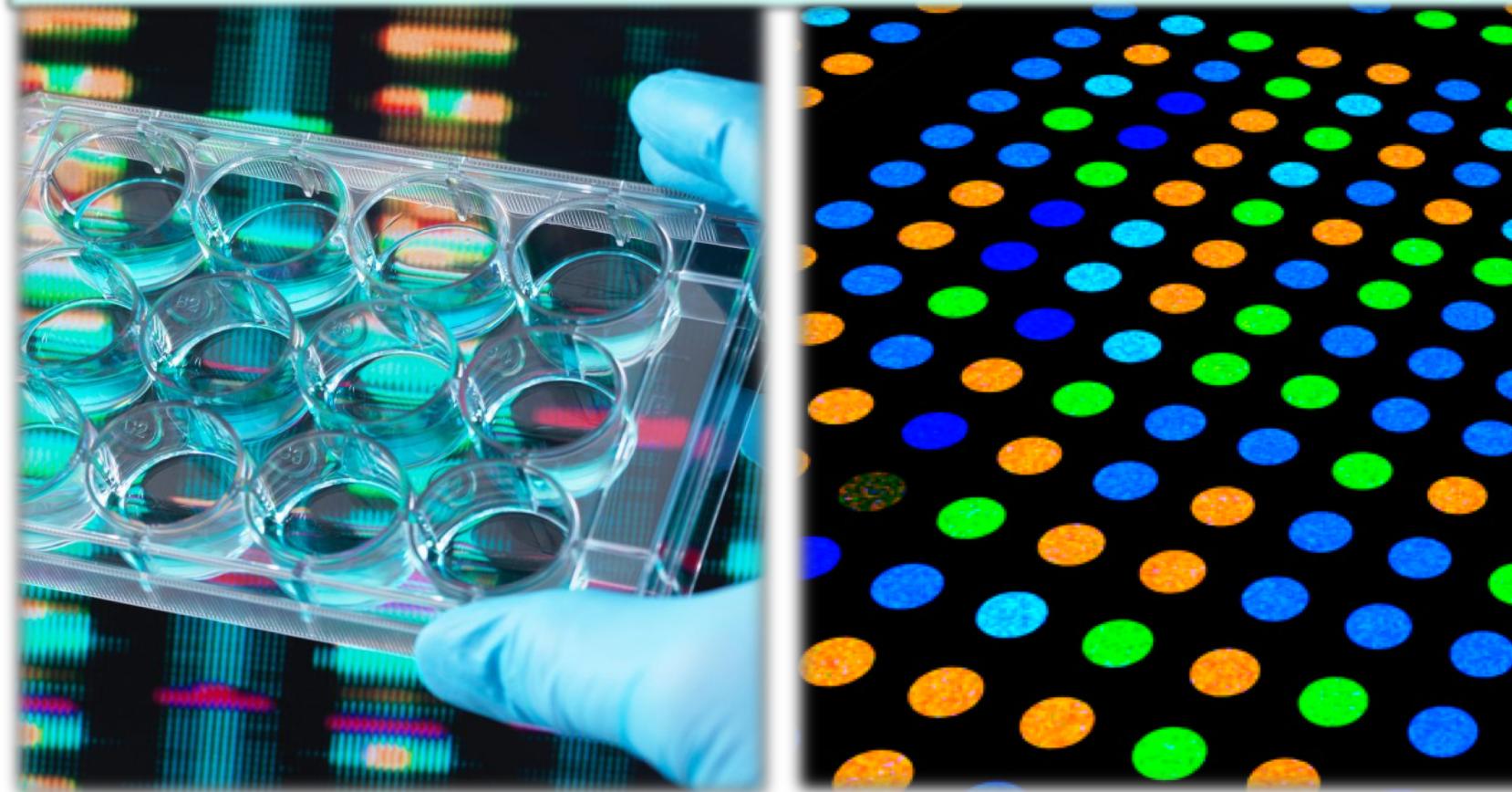


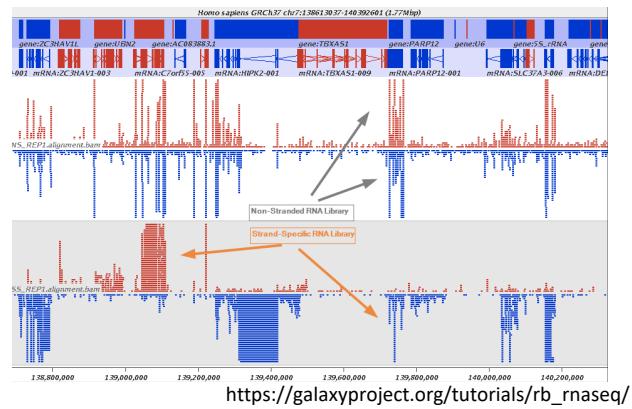
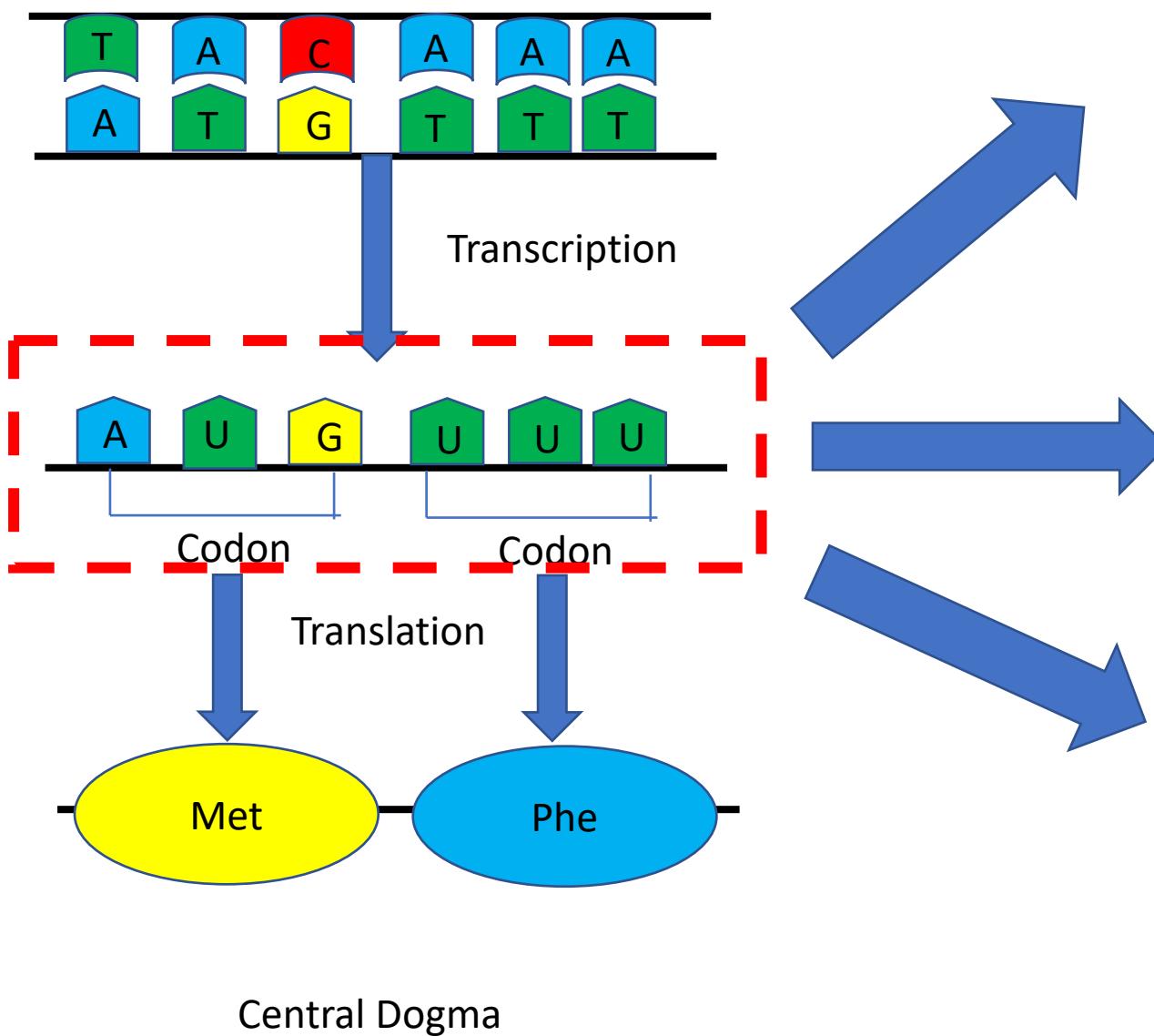
Transcriptomics : Microarrays and RNASeq



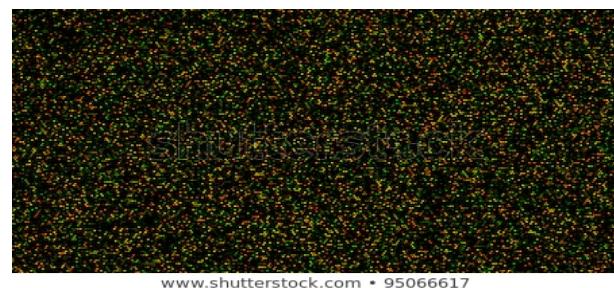
<https://bi-ctsicn.github.io/CBU/>

bioinformatics@childrensnational.org

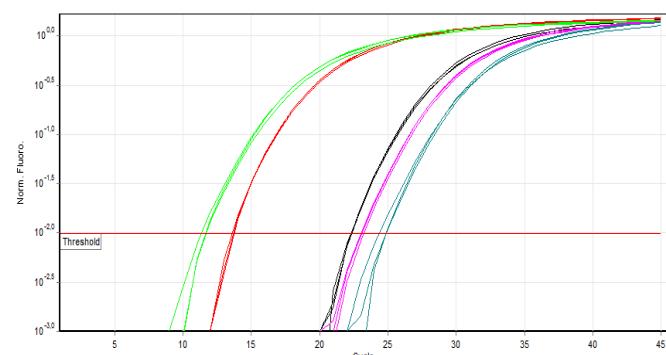
Transcriptomics: What and How?



RNASeq

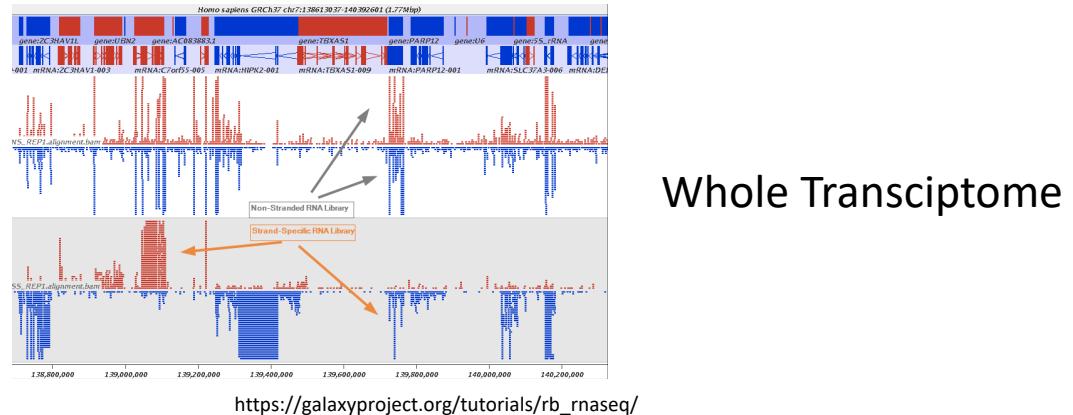
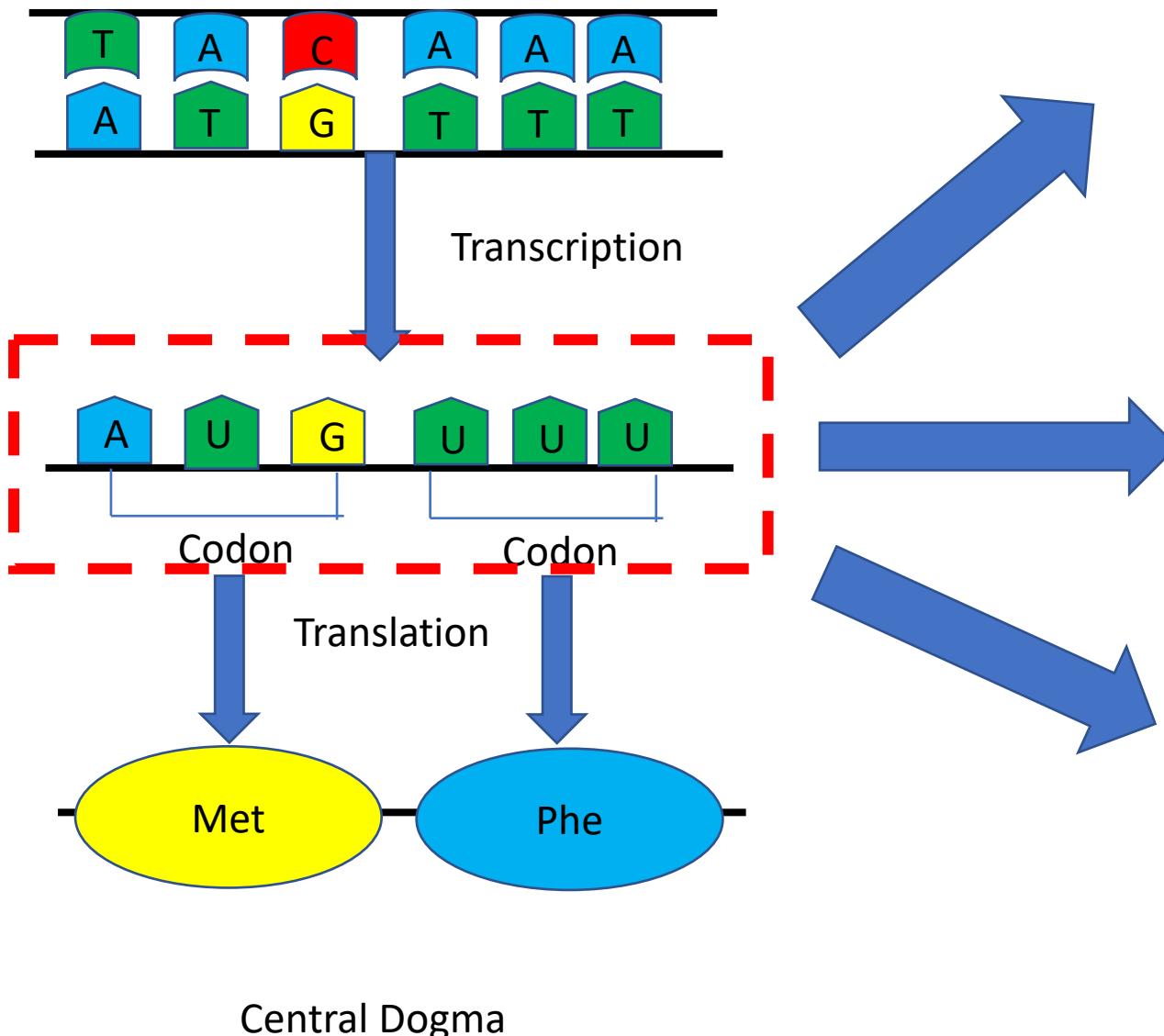


Microarray

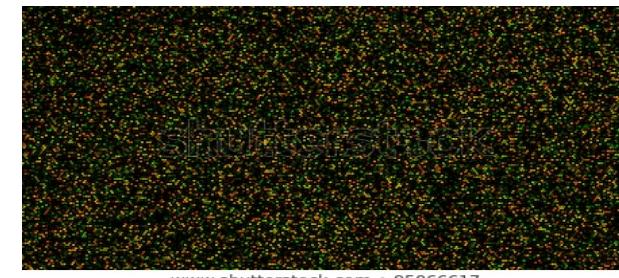


Quantitative PCR

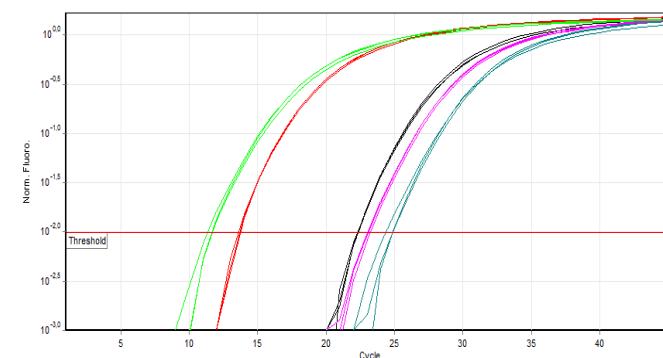
Transcriptomics: What and How?



Whole Transcriptome

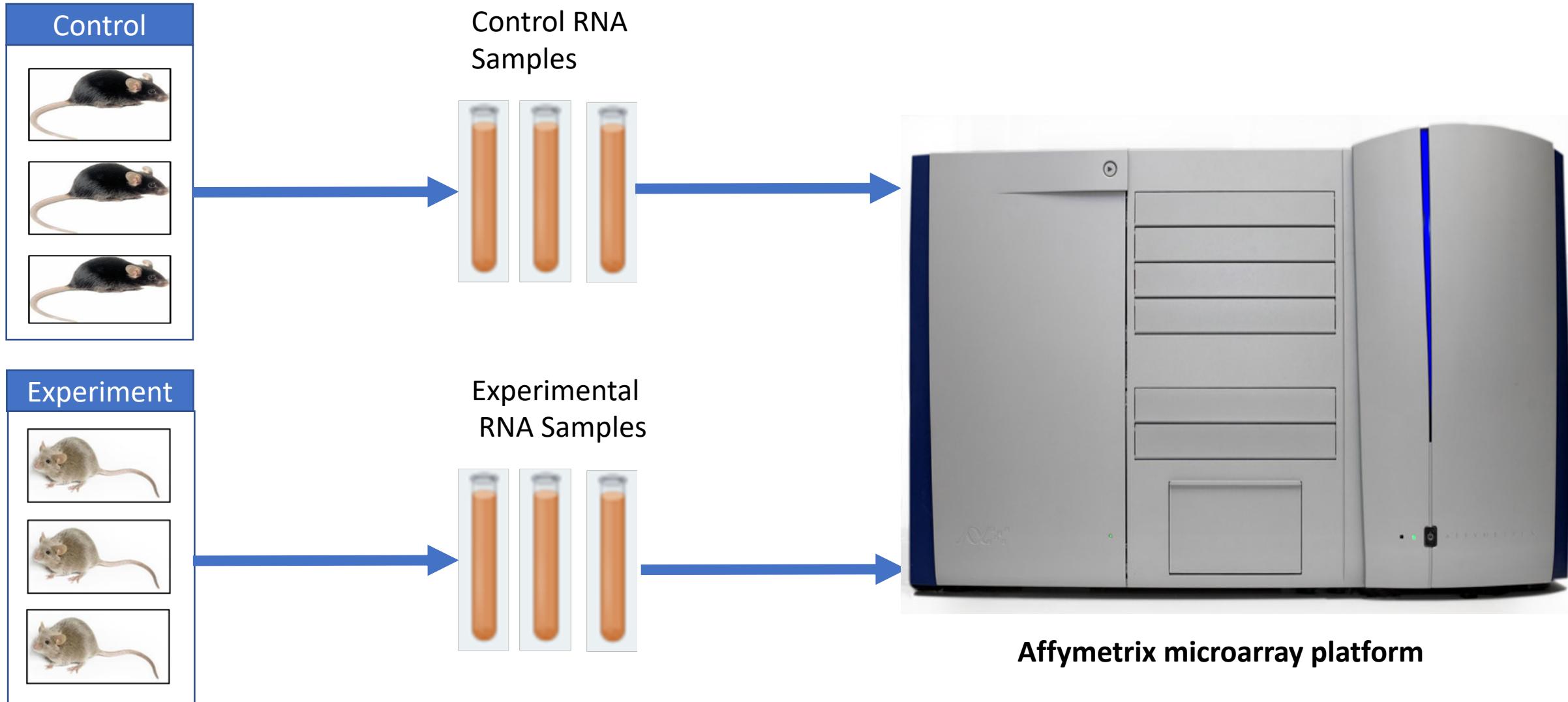


Targeted

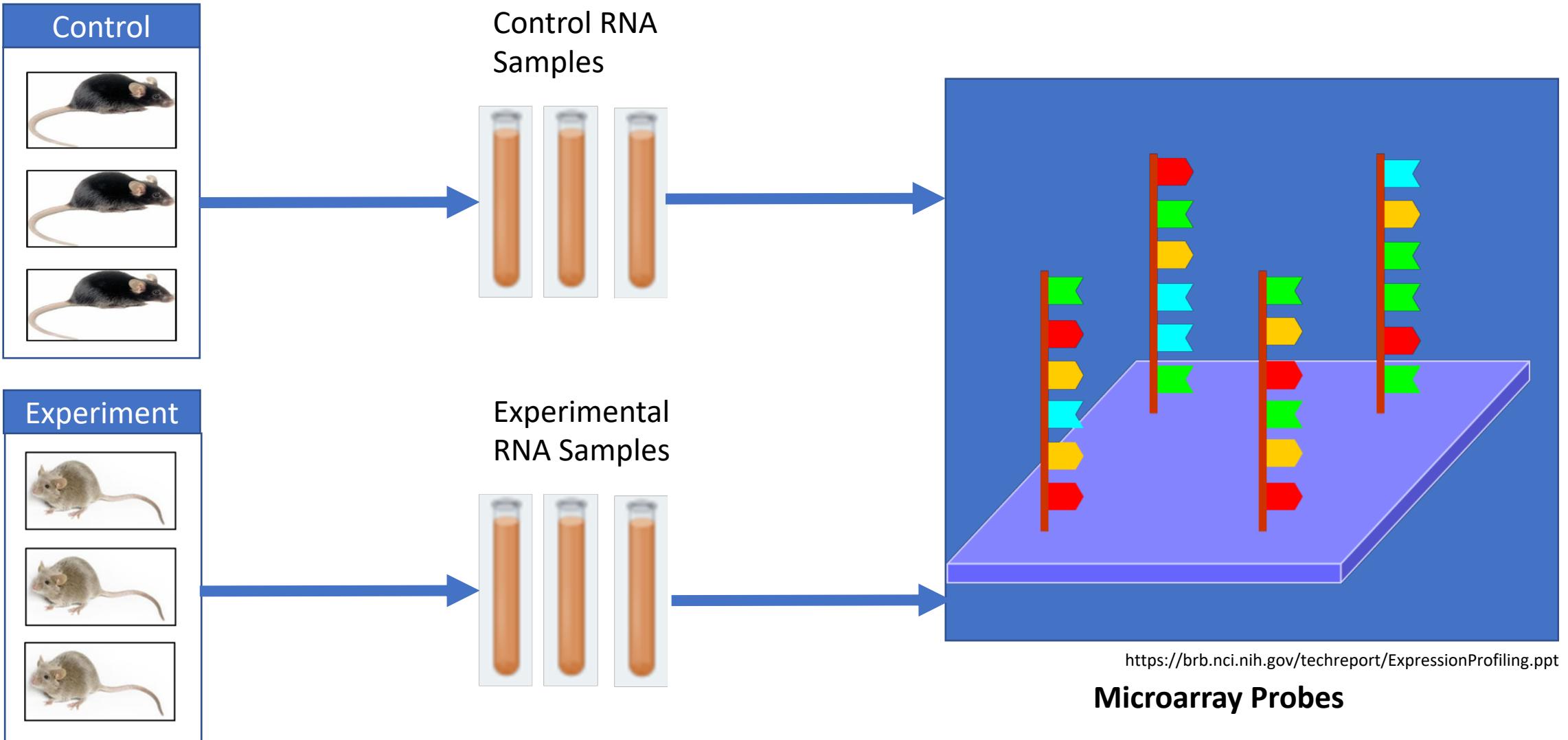


Targeted

Workflow of a microarray experiments



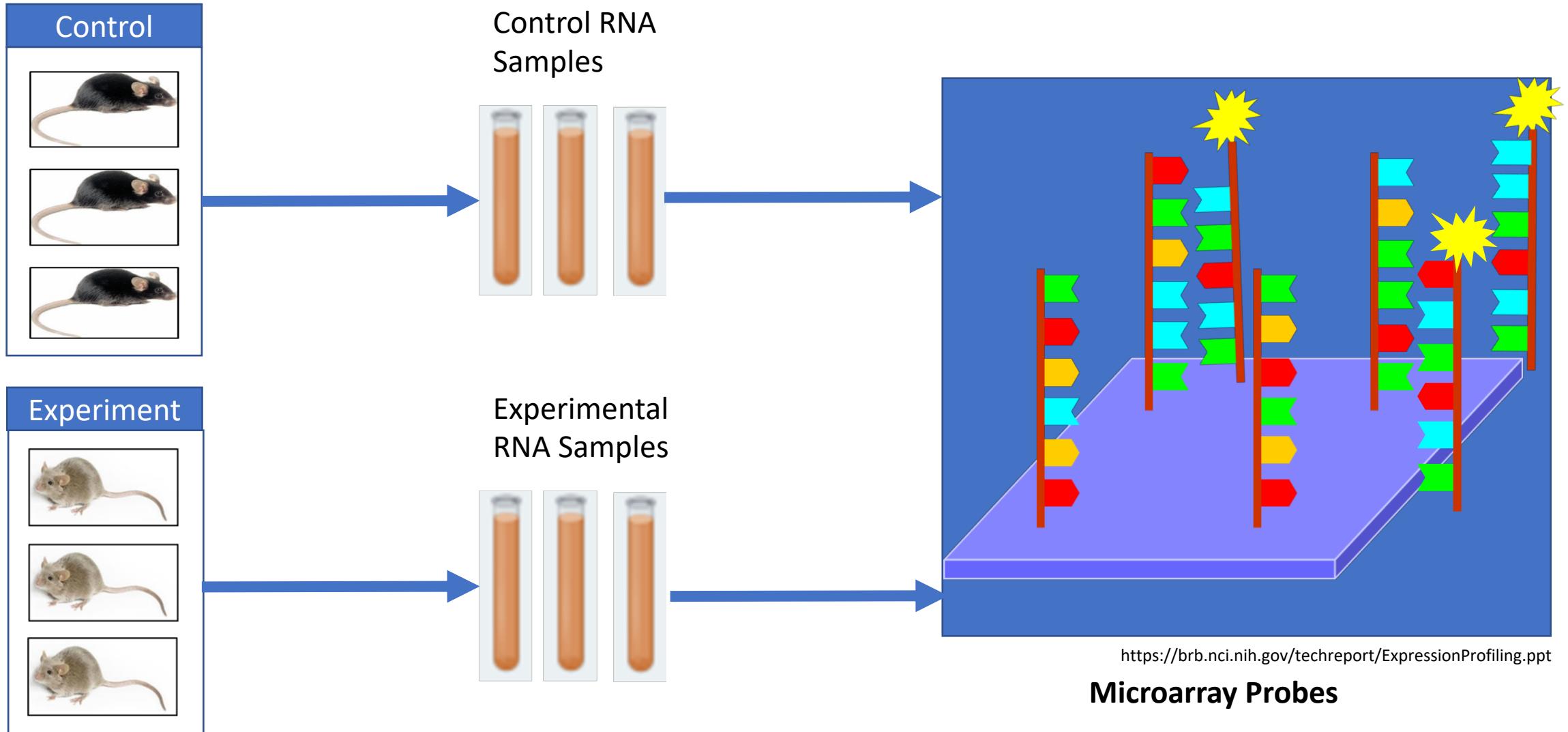
Workflow of a microarray experiments



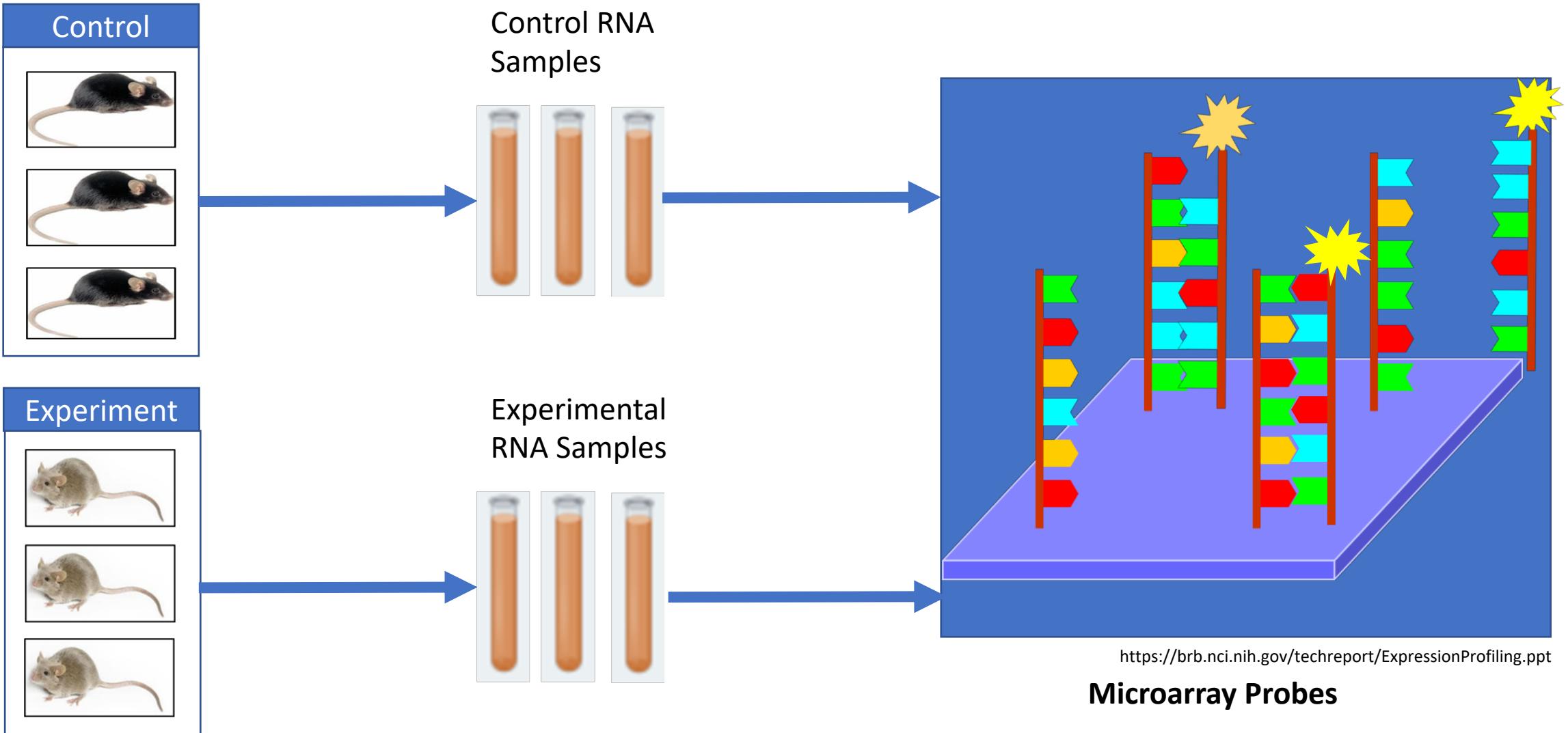
<https://brb.nci.nih.gov/techreport/ExpressionProfiling.ppt>

Microarray Probes

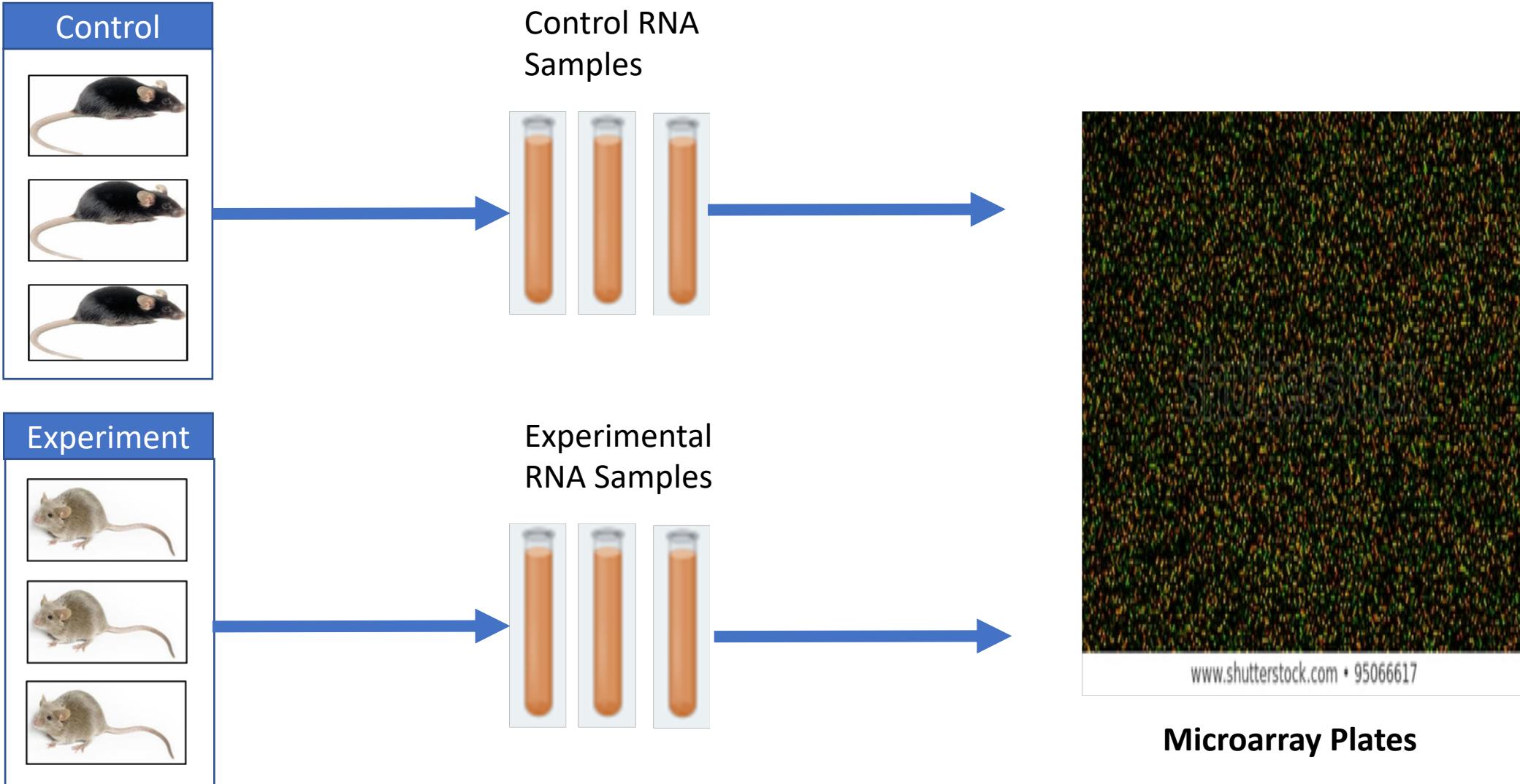
Workflow of a microarray experiments



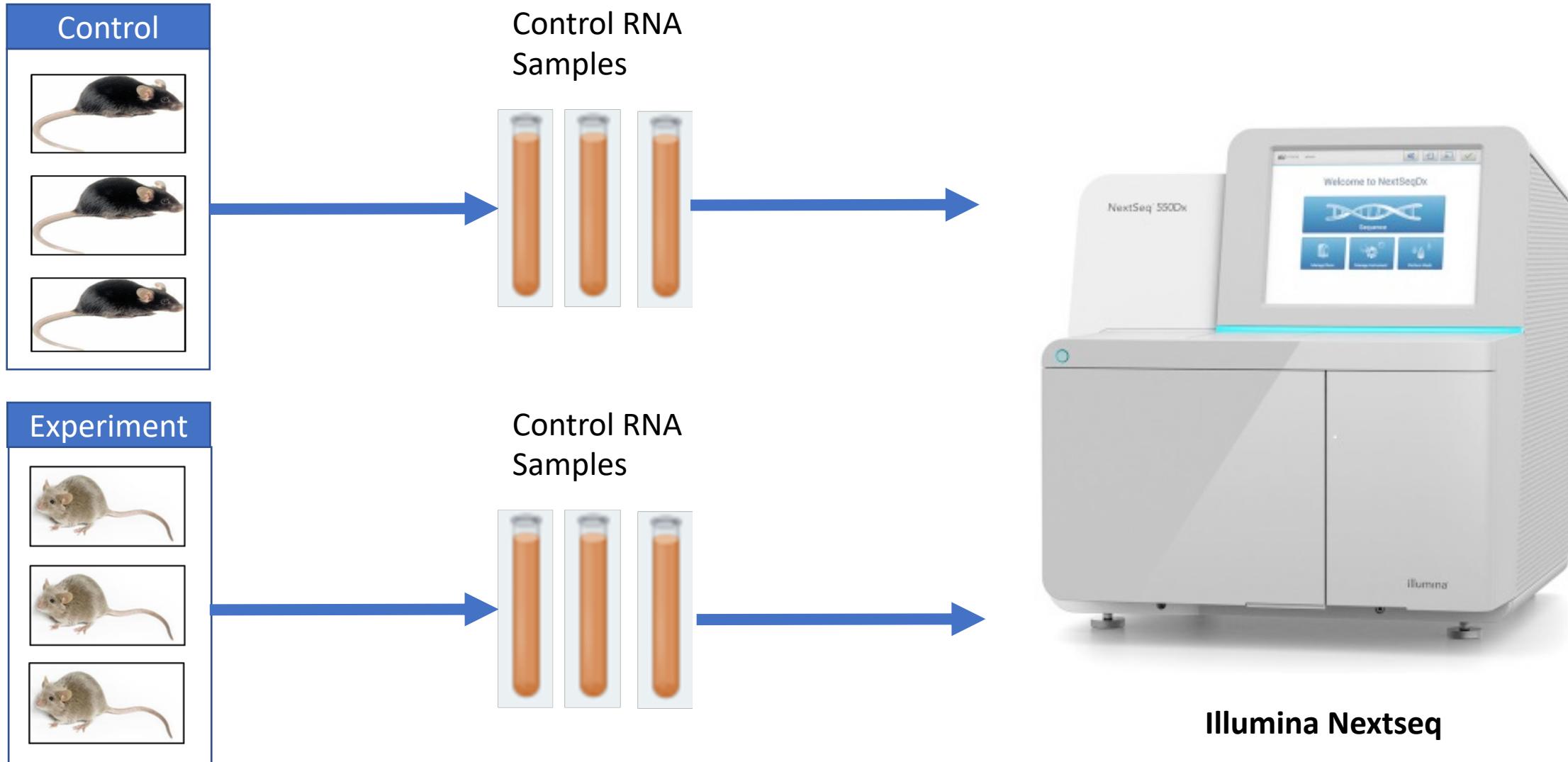
Workflow of a microarray experiments



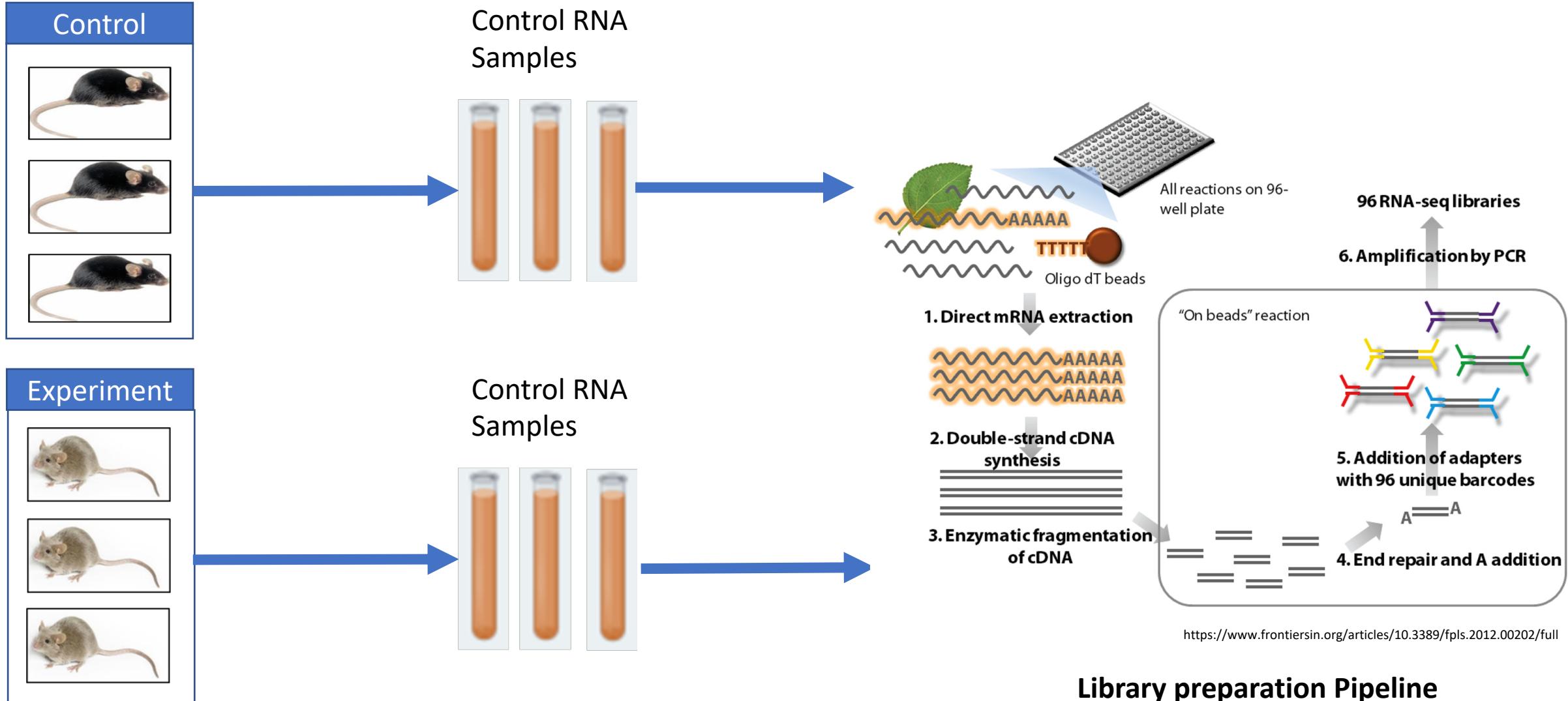
Workflow of a microarray experiments



Workflow of a RNAseq experiments



Workflow of a RNAseq experiments



Is Microarray Dead?

MENU ▾

EMM Experimental &
Molecular Medicine

Article | OPEN | Published: 03 August 2018

Differentially expressed genes related to major depressive disorder and antidepressant response: genome-wide gene expression analysis

Hye In Woo, Shinn-Won Lim, Woojae Myung, Doh Kwan Kim ✉ & Soo-Youn Lee ✉

Experimental & Molecular Medicine **50**, Article number: 92 (2018) | Download Citation ↴

MENU ▾

SCIENTIFIC REPORTS

Article | OPEN | Published: 22 August 2018

Genome-wide mRNA expression analysis of peripheral blood from patients with obsessive-compulsive disorder

Yuqing Song, Yansong Liu, Panpan Wu, Fuquan Zhang ✉ & Guoqiang Wang ✉

Scientific reports **8**, Article number: 12583 (2018) | Download Citation ↴

Not Yet!!

MENU ▾

SCIENTIFIC REPORTS

Article | OPEN | Published: 19 March 2018

Analysis of microRNA and Gene Expression Profiles in Alzheimer's Disease: A Meta-Analysis Approach

Shirin Moradifard, Moslem Hoseinbeyki, Shahla Mohammad Ganji ✉ & Zarrin Minuchehr ✉

Scientific Reports **8**, Article number: 4767 (2018) | Download Citation ↴



this article | search | submit a manuscript | register

BMC Bioinformatics. 2018; 19: 296.

Published online 2018 Aug 8. doi: [10.1186/s12859-018-2308-x](https://doi.org/10.1186/s12859-018-2308-x)

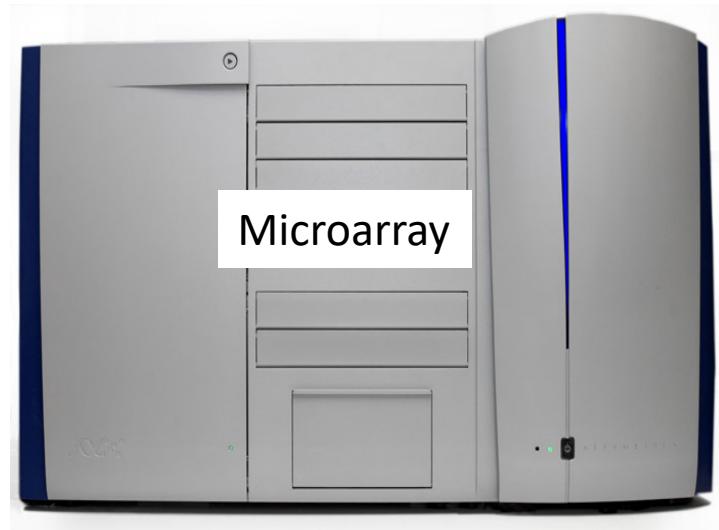
PMCID: PMC6083570

PMID: 30089462

BART: bioinformatics array research tool

[Maria Luisa Amaral](#), [Galina A. Erikson](#), and [Maxim N. Shokhirev](#) ✉

► Author information ► Article notes ► Copyright and License information [Disclaimer](#)



- Organism well annotated.
- Less cost per sample. Cost depends on number of probes used
- Evaluates expression of known genes.
- Relative abundance
- Analysis fast. Finishes in hours
- High error and background noise, due to issues with hybridization.

- Organism does not need to be well annotated
- Higher cost per sample. Cost depends on depth of sequencing.
- Evaluates expression of known as well as unknown gene.
- Absolute abundance
- Analysis slow. Takes days
- Low error and background noise.

Life cycle of a Genomic Experiment



- What is the ultimate goal of the experiment?
- Has this study been done before?
- Is the organism well annotated?
- Number of replicates/samples?

Life cycle of a Genomic Experiment



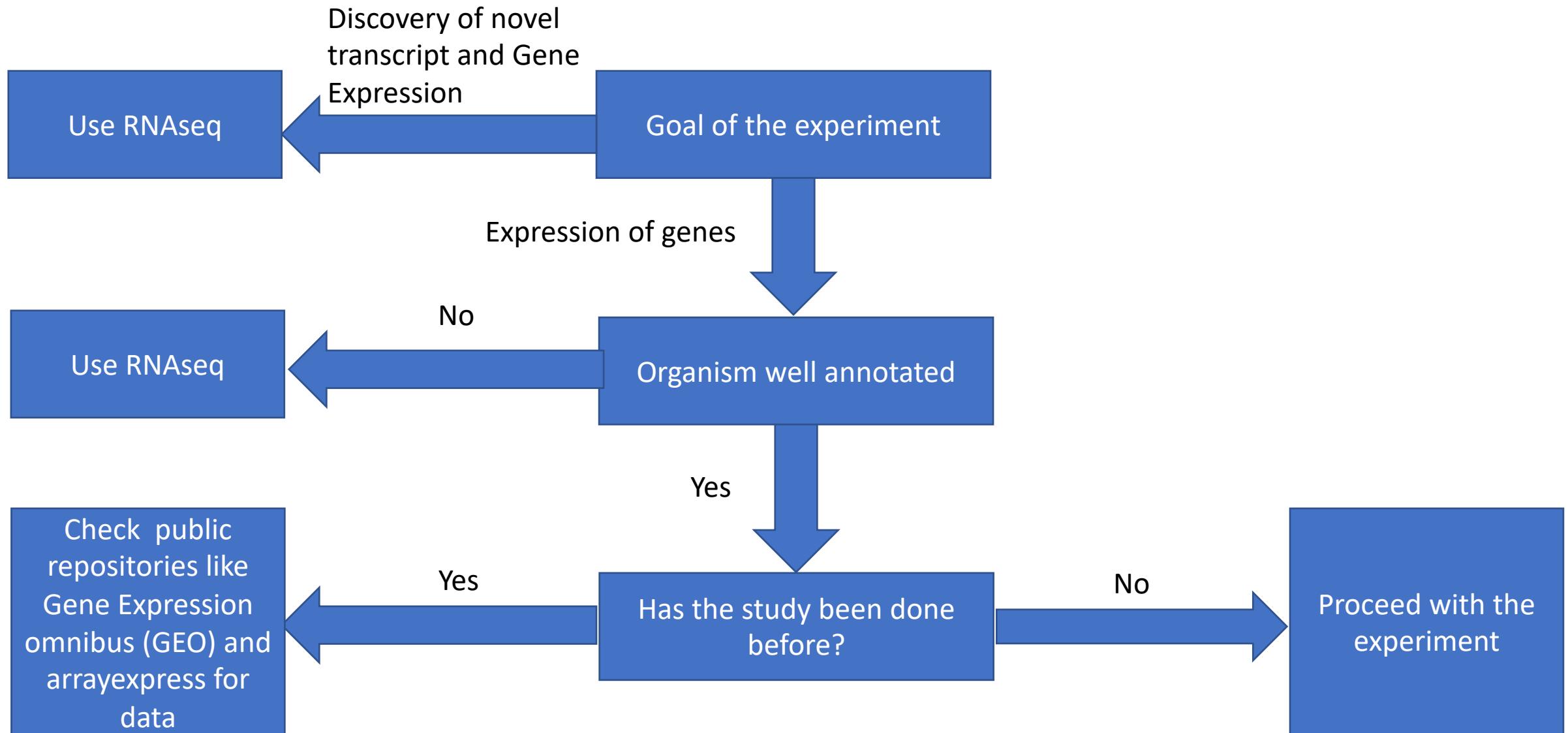
- Extraction of DNA sample depends upon the tissue/cell used. What kit to use?
- How is the quality of the RNA?
- What is the quantity of the RNA?

Life cycle of a Genomic Experiment



- Is this a novel experiment?
- If not is there available literature for the pathway analyzed.
- If novel is it on a known tissue/ cell type.
- What platform used

Design of an Experiment



Power of an Experiment

- More sample=More accuracy of experiment.
- **Power** of a binary hypothesis test is the probability that the test rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_1) is true.
- **Statistical power** ranges from 0 to 1, and as statistical power increases, the probability of making a type II error (wrongly failing to reject the null) decrease
- Parameters to determine Power are:
 - I) mean number of false positive
 - II) The anticipated number of undifferentially expressed genes in the experiment
 - III) The specified power level for an individual gene, which represents the expected proportion of differentially expressed genes that will be declared as such by the tests postulated
 - IV) Mean difference in log-expression between treatment and control conditions as under the alternative hypothesis H_1 .
 - V) The anticipated standard deviation of the difference in log-expression between treatment and control conditions

<https://sph.umd.edu/department/epib/sample-size-calculation-completely-randomized-treatment-control-designs>

Replicates – why?

Ideally, if the experiment were repeated with new, independently obtained samples, the effect would likely be observed again.

Variation in data – are they actual biological changes or are just caused by chance?

Replication for Reproducibility!

Without replicates, what do you miss?

- Identification of random Variation
- Accuracy of measurement

Replicates

- **Biological replicates** measure a quantity from **difference sources** under the same conditions
 - e.g., Tumors from 5 different people with lung cancer may show similar gene expression patterns. These replicates are useful to show what is similar in your replicates and how they are different from a different set of conditions (ie. treated, normal).
- **Technical replicates** measure quantity from **1 source**. This measures the reproducibility of the results. The differences are based only on technical issues in the measurement.
 - e.g., sequence the same sample twice but get different results

Replicates



Technical replicates are: S1 /S2 and S5 / S6

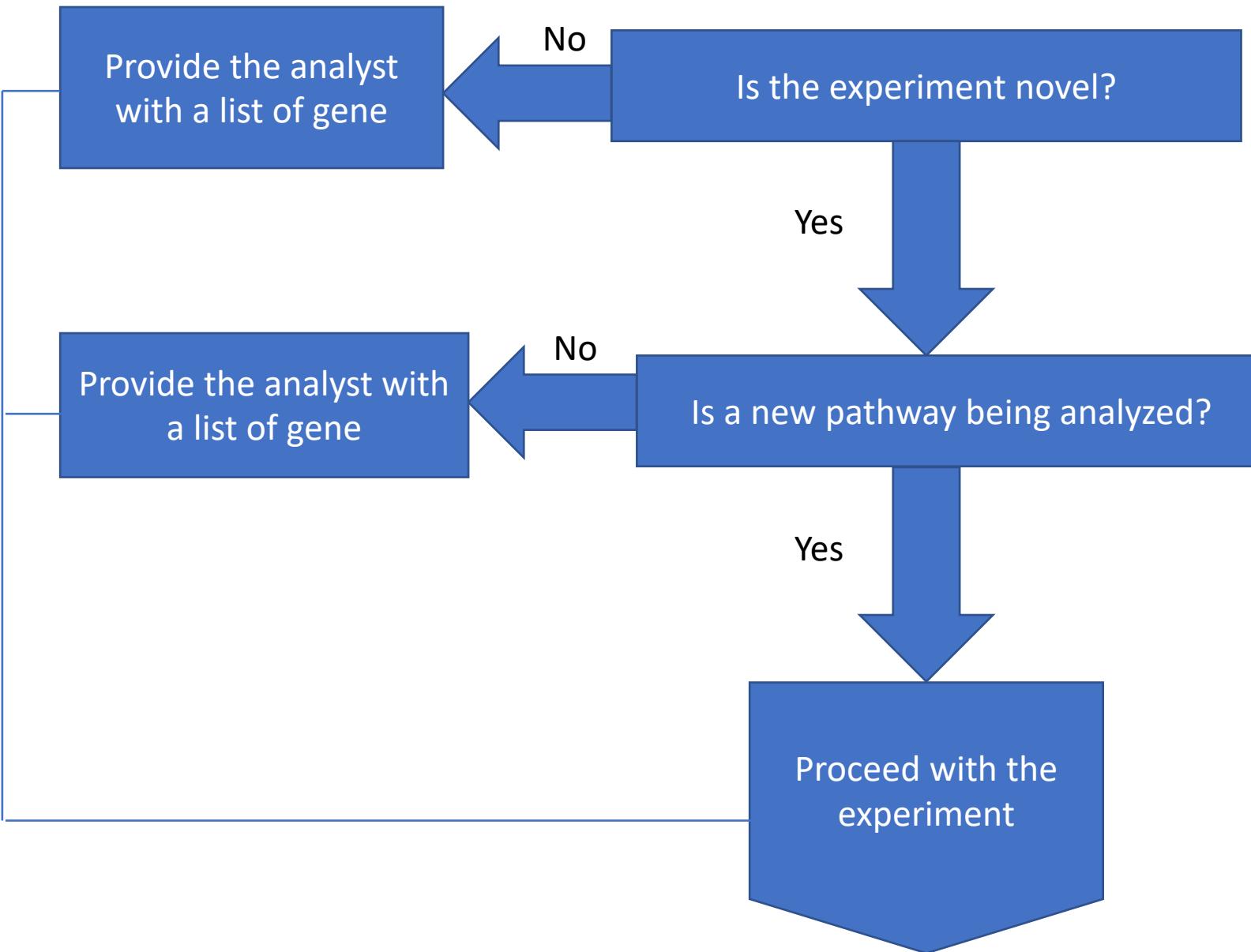
Biological replicates are: S3 – S1/S2 and S4 - S5/S6

To make inferences about the population you need ***biological replicates***

Microarray Experiment (Wet Lab) (Susan/Karuna)

- Sample Amount
- Quality of Samples
- Kits Required
- Microarray platforms available on-site and at Georgetown
- Cost
- Any other things

Genomic Experiment (Data Analysis)



Files and Information Required for analysis microarray

➤ Design of the experiment

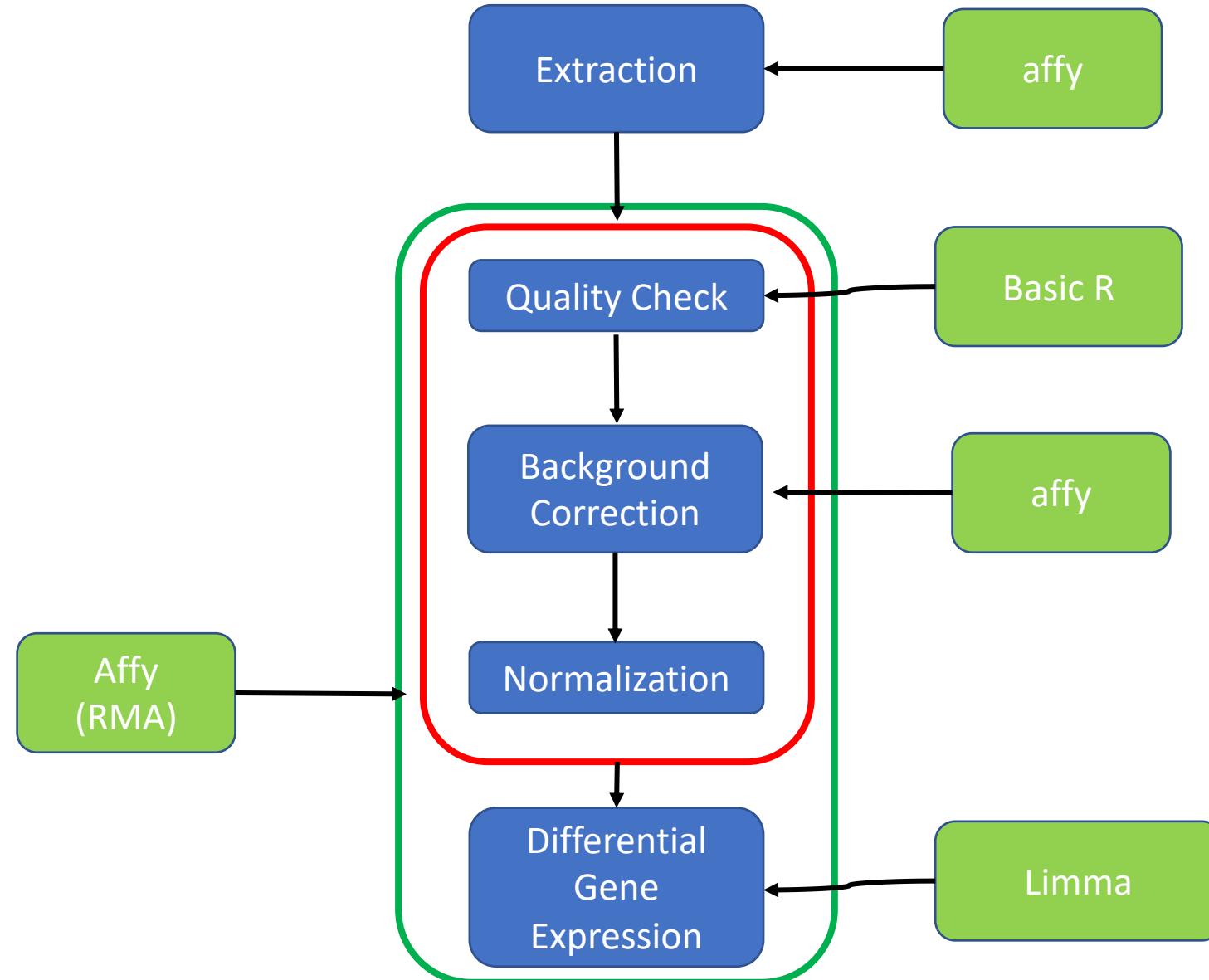
- **Number of samples used?**
- **Does your question require biological or technical replicates?**
- **What kind of probes needed to answer your question?**

➤ Files and Information Required

- **Raw intensity files:** If done in-house, the .CEL files from Affymetrix platform. If done elsewhere and done on other platforms, the platform name and raw data from the same.
- **Sample Names:** Names associated to each sample.
- **Condition:** Condition associated to each sample.

Microarray Analysis Pipeline

1. Extraction
2. Quality Check
3. Background Correction
4. Differential Gene Expression



Introduction to TACC

The screenshot shows the TAC 4.0 software interface. At the top, there is a dark blue header bar with the "appliedbiosystems" logo on the left, the "TAC 4.0" title in the center, and standard window control icons on the right. Below the header is a menu bar with three items: "New Analysis" (highlighted in blue), "Open Existing Result", and "Preferences". Underneath the menu bar is a toolbar with several buttons: "Import Files" (with a dropdown arrow), "Remove Selected File(s)", "Add New Column", "Sample Attributes" (with a dropdown arrow), "Clear Sorting", and "Show Graph" (with a dropdown arrow). The main workspace is a large white area with a thin gray border. In the top-left corner of this workspace, there is a small table with one row and two columns, labeled "File Name". The rest of the workspace is empty. In the center of the workspace, there is a faint gray text message: "Drag-Drop Files or Click \"Import Files\" to select files for analysis." At the bottom of the interface, there is a footer bar with the "Result Name:" label followed by the text "Analysis_1", the "Output Folder:" label followed by the path "C:\Users\Vilain Lab\Documents\TACX\AnalysisResults", a "Browse..." button, and a "Run Analysis" button.

appliedbiosystems

TAC 4.0

New Analysis Open Existing Result Preferences

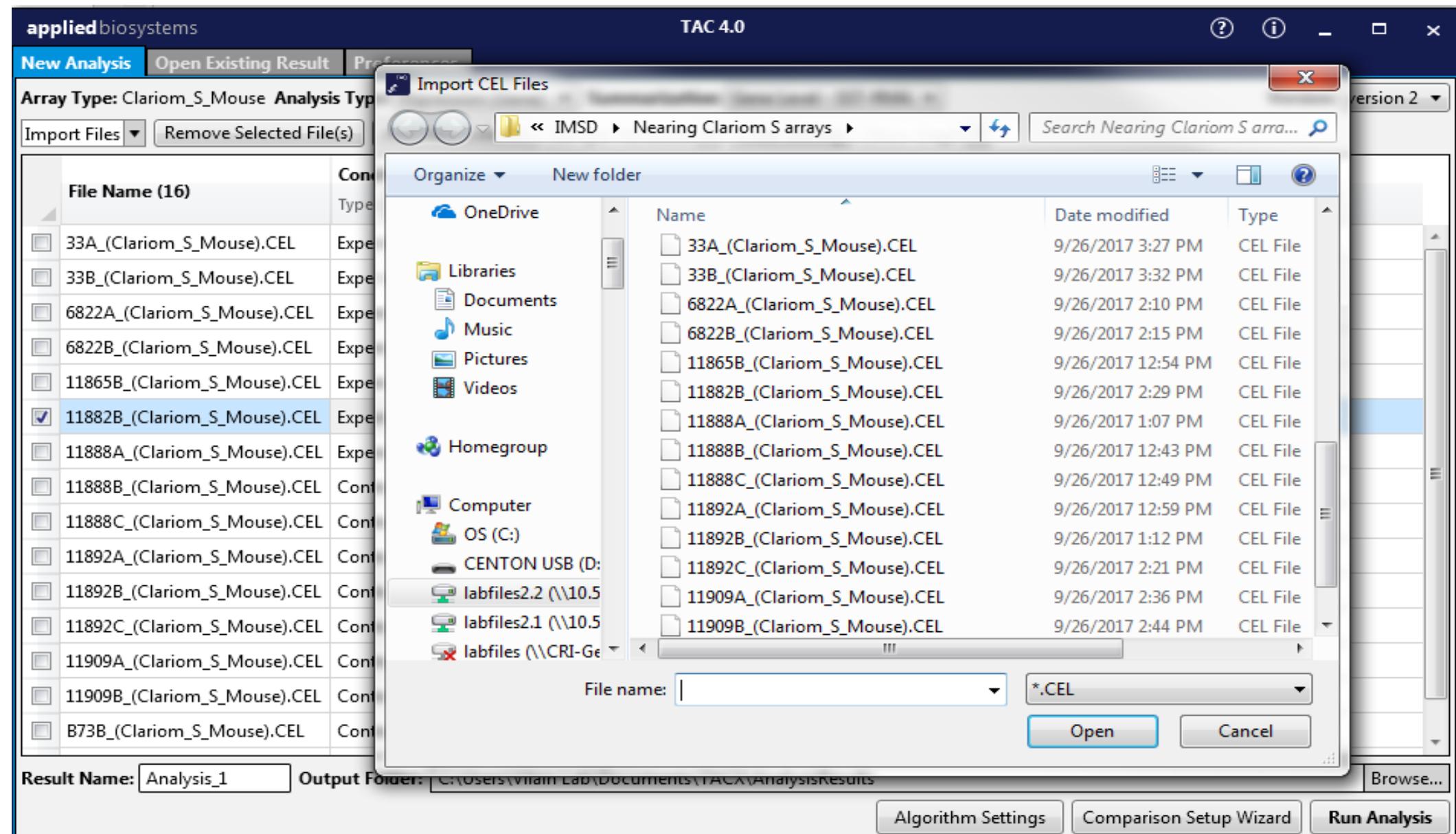
Import Files ▾ Remove Selected File(s) Add New Column Sample Attributes ▾ Clear Sorting Show Graph ▾

File Name

Drag-Drop Files or Click "Import Files" to select files for analysis.

Result Name: Analysis_1 Output Folder: C:\Users\Vilain Lab\Documents\TACX\AnalysisResults Browse... Run Analysis

Importing CEL Files



Apply Condition

applied biosystems TAC 4.0

New Analysis Open Existing Result Preferences

Array Type: Clariom_S_Mouse Analysis Type: Expression (Gene) Summarization: Gene Level - SST-RMA Version: version 2

Import Files Remove Selected File(s) Add New Column Sample Attributes Clear Sorting Show Graph

File Name (16)	Condition
33A_(Clariom_S_Mouse).CEL	Experimental
33B_(Clariom_S_Mouse).CEL	Experimental
6822A_(Clariom_S_Mouse).CEL	Experimental
6822B_(Clariom_S_Mouse).CEL	Experimental
11865B_(Clariom_S_Mouse).CEL	Experimental
<input checked="" type="checkbox"/> 11882B_(Clariom_S_Mouse).CEL	Experimental
11888A_(Clariom_S_Mouse).CEL	Experimental
11888B_(Clariom_S_Mouse).CEL	Control
11888C_(Clariom_S_Mouse).CEL	Control
11892A_(Clariom_S_Mouse).CEL	Control
11892B_(Clariom_S_Mouse).CEL	Control
11892C_(Clariom_S_Mouse).CEL	Control
11909A_(Clariom_S_Mouse).CEL	Control
11909B_(Clariom_S_Mouse).CEL	Control
B73B_(Clariom_S_Mouse).CEL	Control

Result Name: Analysis_1 Output Folder: C:\Users\Vilain Lab\Documents\TACX\AnalysisResults Browse...

Algorithm Settings Comparison Setup Wizard Run Analysis

Selecting Statistical Methods

Clariom_S_Mouse Configuration X

▲ Gene-Level Default Filter Criteria

Gene-Level Fold Change < - or >

Gene-Level P-Value <

Gene-Level Use FDR:

Gene-Level FDR <

▲ Limma

Anova Method:

▲ Is Expressed Criteria

A Probeset (Gene/Exon) is considered expressed if \geq % samples have DABG values below DABG Threshold.

DABG <

▲ Pos/Neg AUC Threshold

Pos/Neg AUC Threshold:

▼ Exploratory Grouping Analysis

Files and Information Required for analysis RNAseq

➤ Design of the experiment

- **Number of samples used?**
- **Does your question require biological or technical replicates?**
- **What is the coverage required for ?**

➤ Files and Information Required

- **Raw Read files:** The output from the sequencers Fastq is needed. If done in other facility, fastq (or bam) has to be provided to the analysts, along with information about the sequencers, spike ins used, library type,etc.
- **Sample Names:** Names associated to each sample.
- **Condition:** Condition associated to each sample.

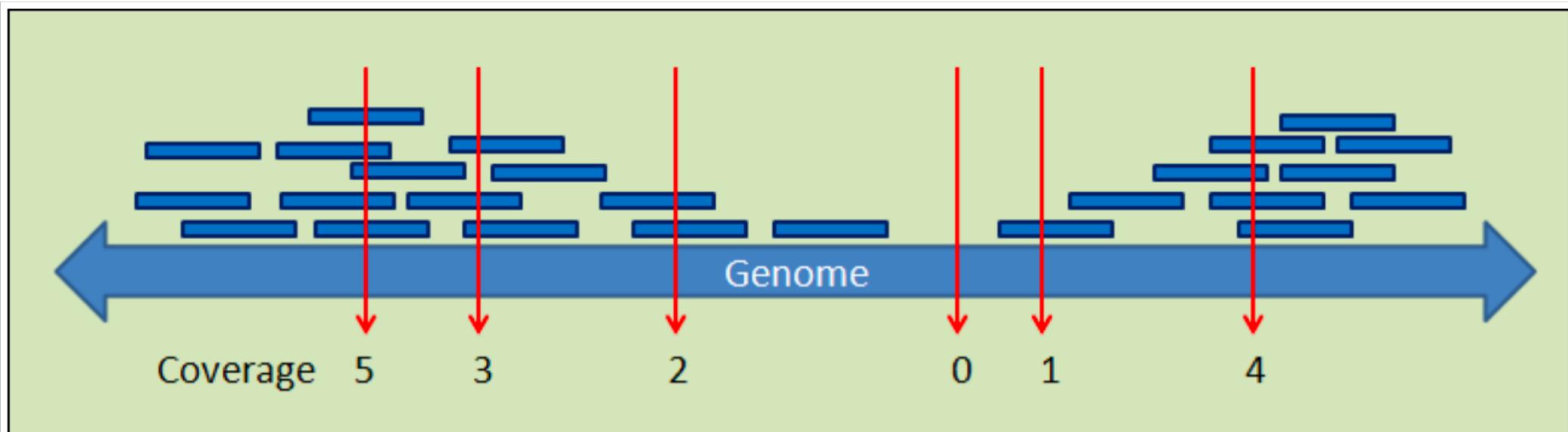
Read Length

- Read length refers to the number of base pairs that are read at a time.
 - For a read length of 50 base pairs, **single end reads** would read 50 base pairs from each fragment,
 - while **paired end reads** would consist of 2 x 50bp reads, covering up to 100 base pairs on the same fragment.

While longer read lengths give you more accurate information on the relative positions of your bases in a genome, they are more expensive than shorter ones.

Coverage

- Coverage refers to the average number of times a single base is read during a sequencing run.
 - Eg: If the coverage is 100 X, this means that on average each base was sequenced 100 times.
- The more frequently a base is sequenced, the more reliable a base is called, resulting in better quality of your data.



Coverage

Lander / Waterman equation

$$\text{Coverage (X)} = \frac{L \text{ (read length)} * N \text{ (number of reads)}}{G \text{ (genome length)}}$$

Coverage Calculators

- <http://core-genomics.blogspot.com/2016/05/how-many-reads-to-sequence-genome.html>
- <http://apps.bioconnector.virginia.edu/covcalc/>

Coverage (RNA)

A more useful metric for RNA-Seq is determining the total number of mapped reads.

- It is important to distinguish between **total reads** and **mapped reads**, as not all reads will map onto a reference genome

So, the number of usable reads will be less than the number of actual reads.

- The number of reads that will map depend on the
 - ❖ library type
 - ❖ quality of sample
 - ❖ how complete the reference genome is

Coverage (RNA)

Coverage needed for a RNA is not always uniform:

- ❖ Different transcripts are expressed at different levels, meaning ***more reads will be captured from highly expressed genes*** while ***fewer reads will be captured by genes expressed at low levels***
- ❖ Alternate expression

Coverage (RNA)

Recommended RNA-Seq Parameters

Optimal sequencing depth for RNA-Seq will vary based on the scientific objective of study but here are some general recommendations based on sample type and application:

Sample Type	Reads Needed for Differential Expression (millions)	Reads Needed for Rare Transcript or De Novo Assembly (millions)	Read Length
Small Genomes (i.e. Bacteria / Fungi)	5	30 - 65	50 SR or PE for positional info
Intermediate Genomes (i.e. Drosophila / C. Elegans)	10	70 - 130	50 – 100 SR or PE for positional info
Large Genomes (i.e. Human / Mouse)	15 - 25	100 - 200	>100 SR or PE for positional info

Sample and Library Preparation

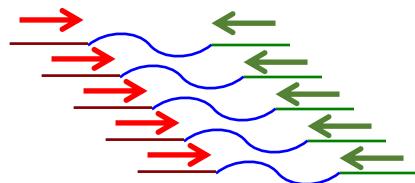
- ❑ Single end vs Paired end
- ❑ PolyA vs Ribodepletion

Single end and Paired end

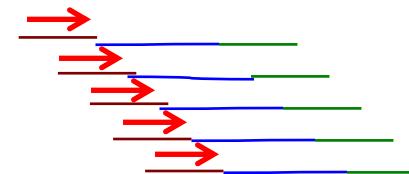
The sequencer instrument read from one end to the other end, and then start another round of reading from the opposite end.

The sequencer reads from one end of a fragment to the other end.

**PE = paired end
(mate pairs)**



SE = single end



PE sequencing provides additional positioning information in the genome

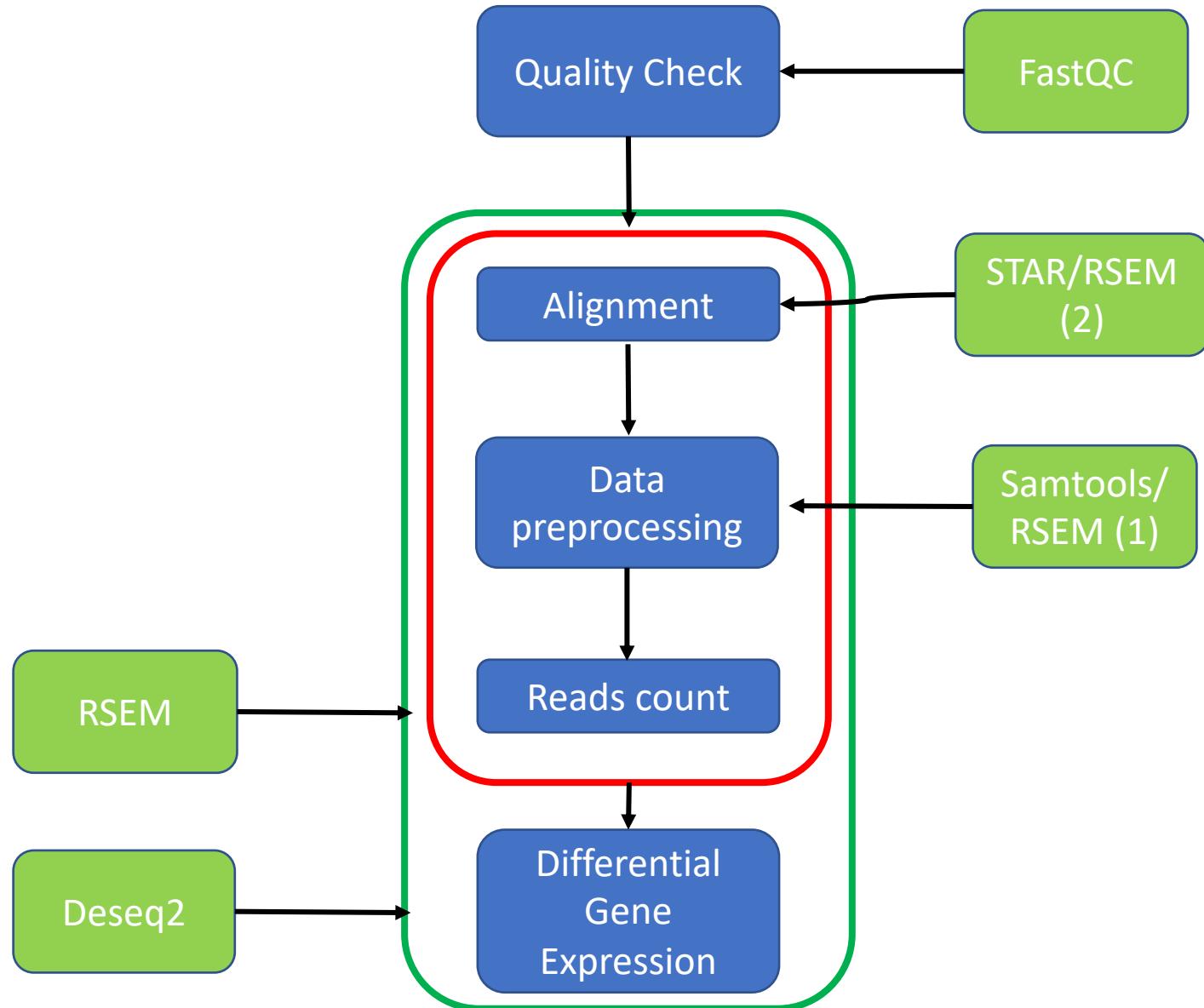
PolyA or RiboDepletion? (RNA)

- Ribosomal RNA (rRNA) constitutes >70% of the purified total cell RNA.
- RiboDepletion removes specifically ribosomal RNA, leaving all other RNA transcripts, however it is not 100% efficient.
- PolyA selection is very efficient, but it will only select polyadenylated RNA, therefore many long, non coding RNAs will be lost.

Poly A	Ribo-depletion
Eukaryotes mostly	Prokaryotes
mRNA	mRNA along with non-coding RNA like lncRNA etc
3 prime bias	-

RNA-Seq Analysis Pipeline

1. Quality check
2. Alignments
3. Data Pre-processing
4. Reads Counts
5. Differential Gene Expression



Fastq Format

FASTQ format is a

- Text-based format
- Stores :
 - **Biological sequence**
 - **Corresponding quality scores**

Size of fastq files depend on:

- **Type of experiment** – WGS > Exome > RNA
- **Type of Genome** – Human > Mouse > Bacteria
- **Coverage** – more the coverage greater the size of fastq file

FASTA Format

```
>unique_sequence_ID My sequence is pretty cool  
ATTCATTAAAGCAGTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTATGATAAAAGAATAC
```

FASTQ Format

sequence identifier and an *optional* description

@unique_sequence_ID

ATTCATTAAAGCAGTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTATGATAAAAGAATAC

+

-- (DD--DDD/DD5 : *1B3&) -B6+8@+1 (DDB:DD07/DB&3 ((+:?=8*D+DDD+B)*) B . 8CDBDD4DDD@@D

raw sequence letters

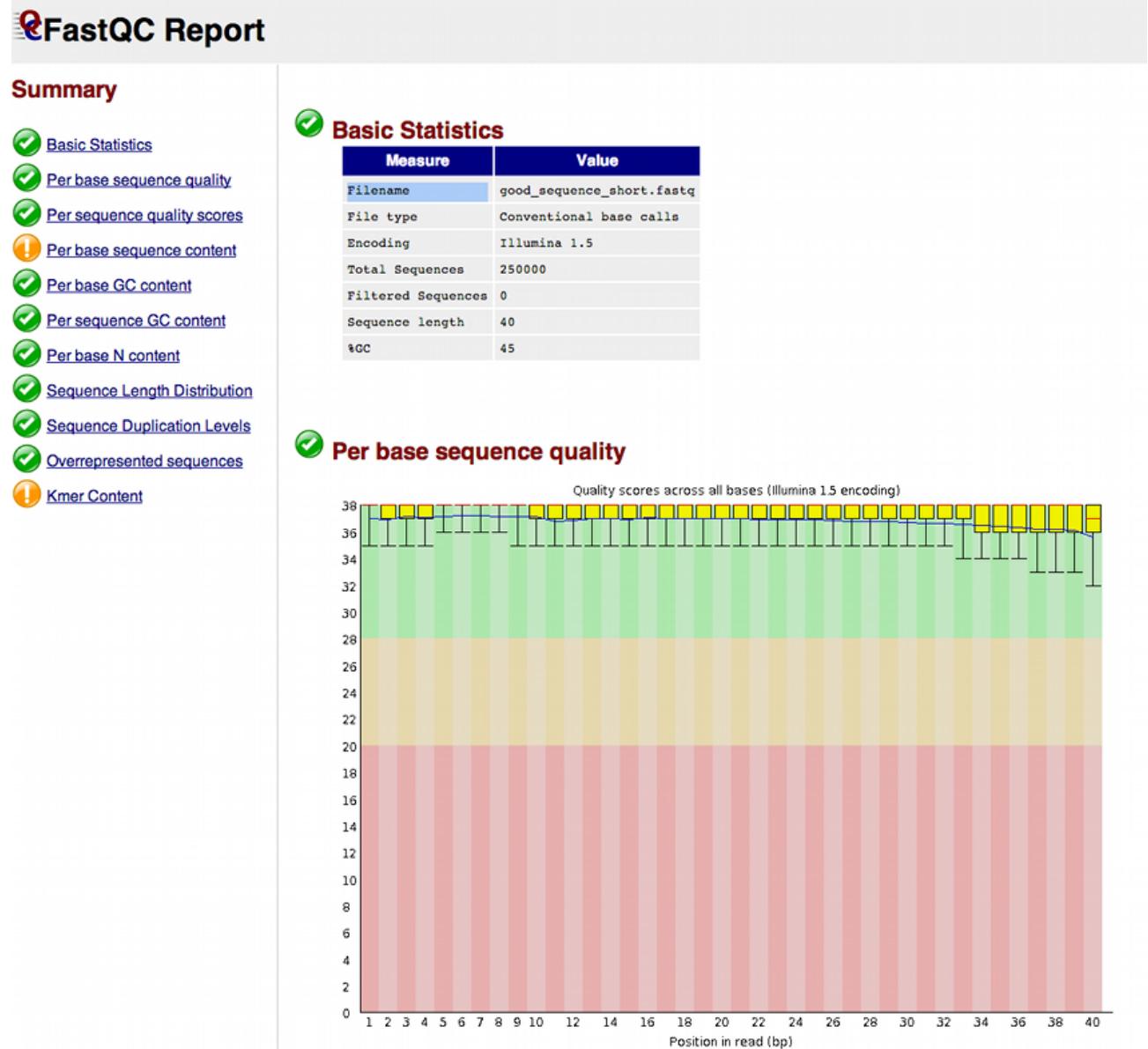
quality values for the sequence

Quality Check

Tool: FastQC

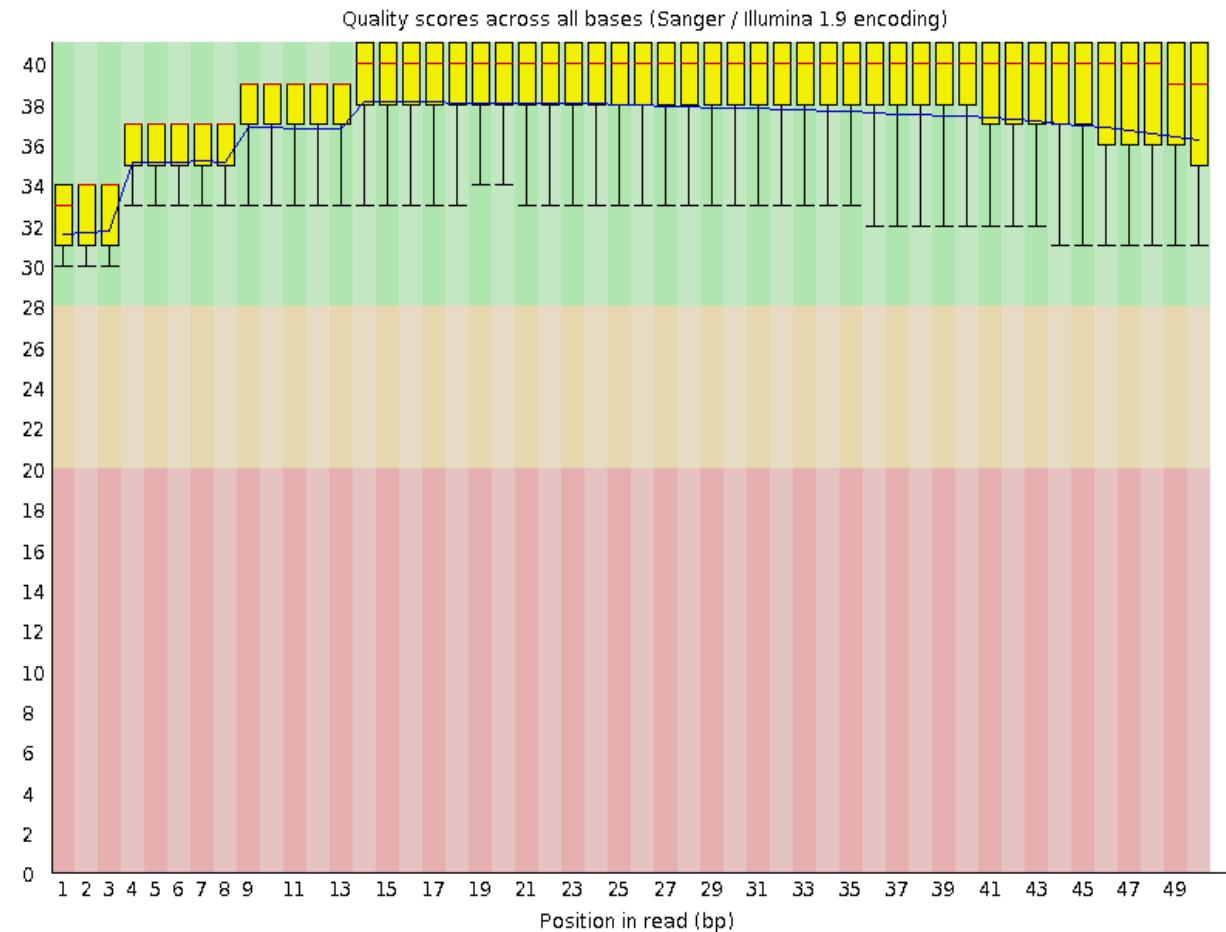
Input: FastQ files

Output: HTML file

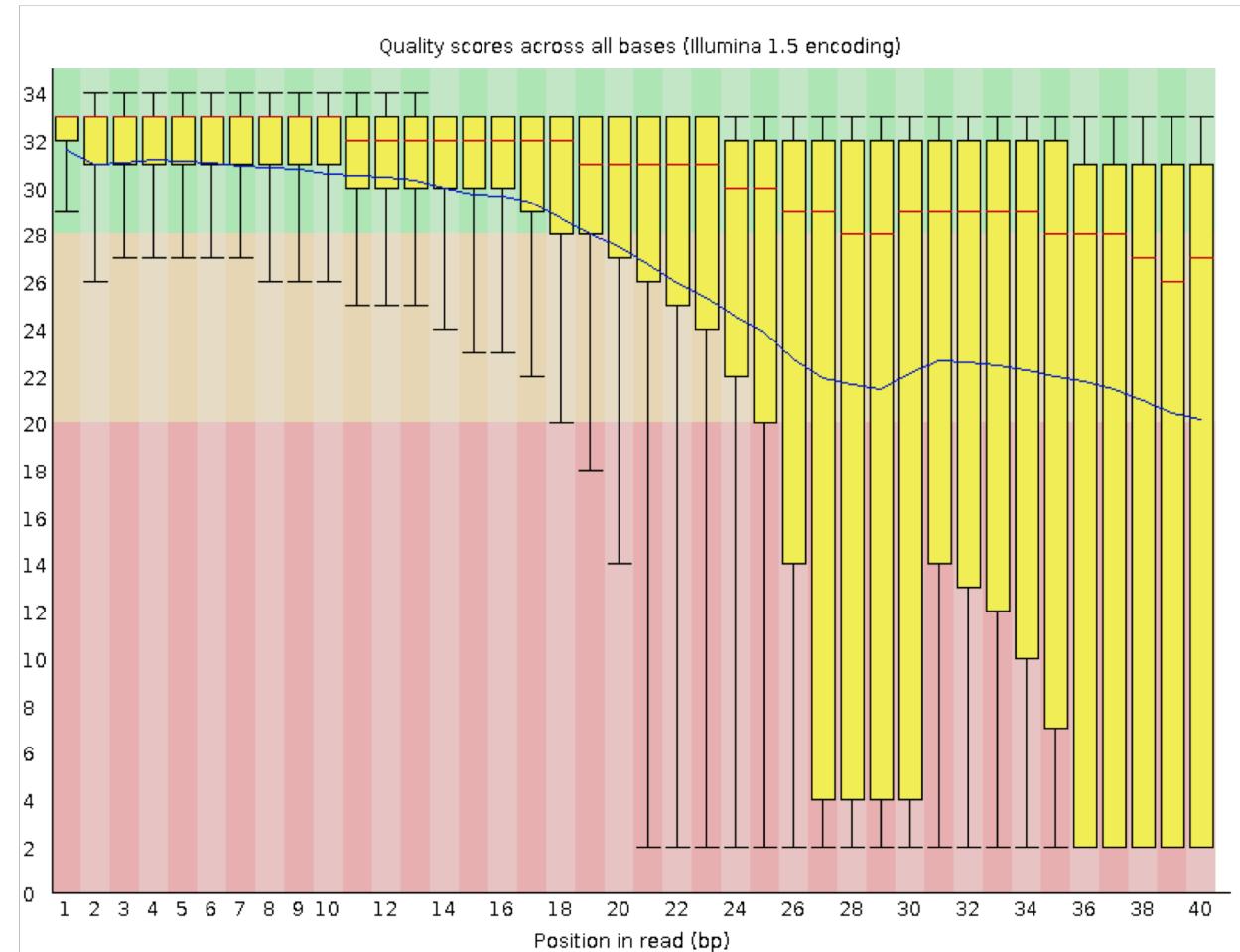


FastQC : Good vs Bad

Good Quality



Bad Quality

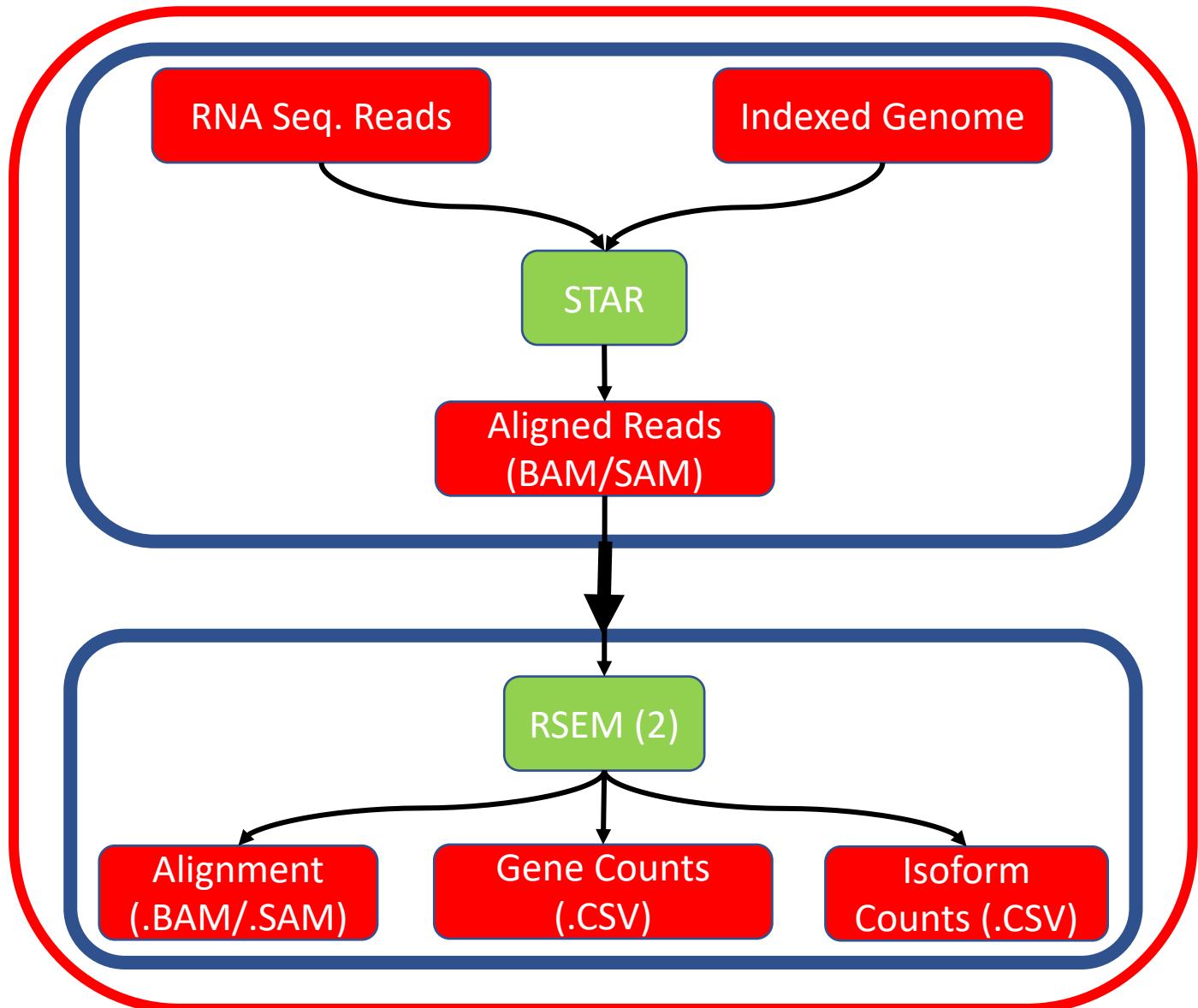


Alignments and Reads count

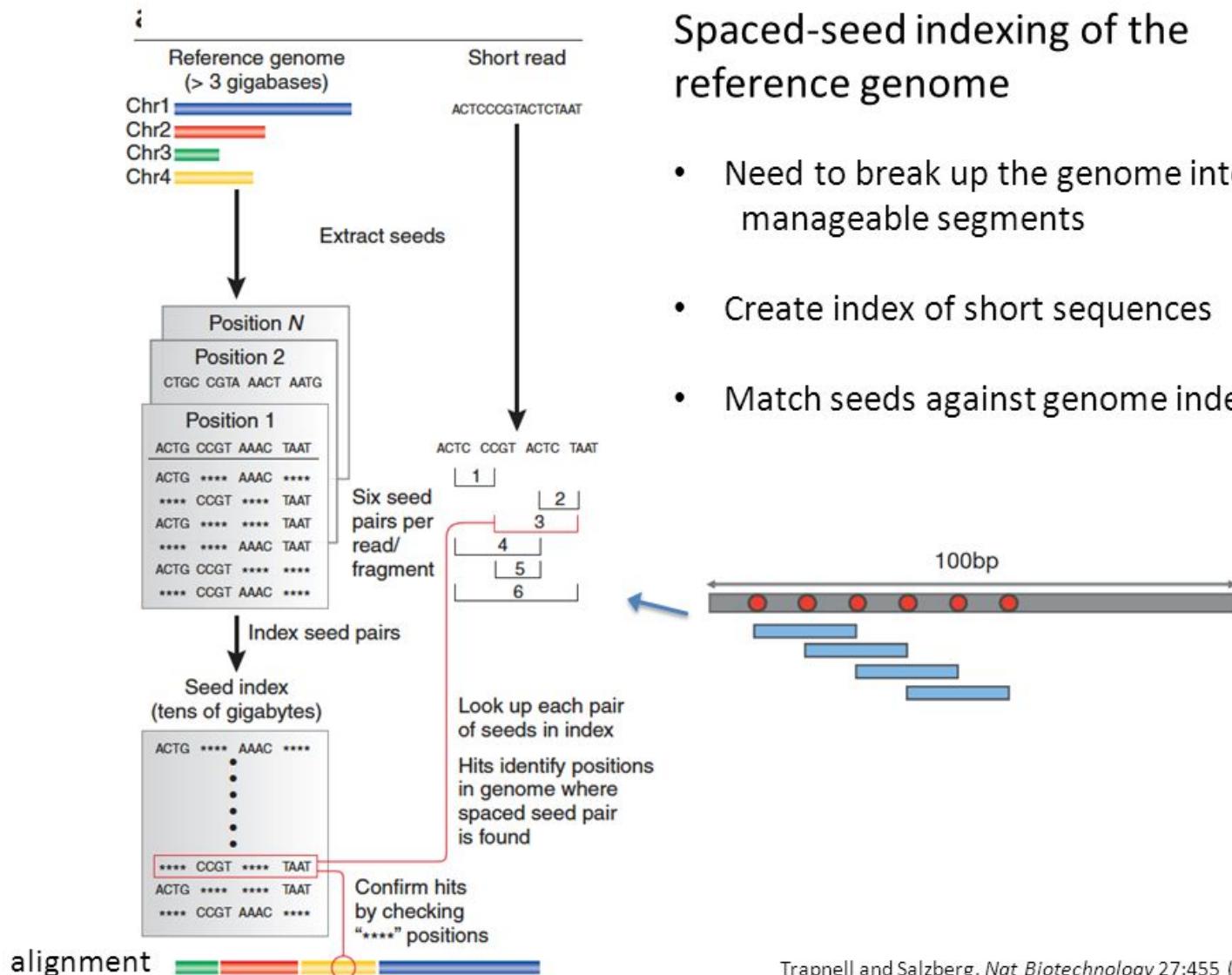
Tools : STAR and RSEM(2)

Input: RNA Seq. Reads (.fastQ)
& Indexes (previous step
output)

Output: .BAM/.SAM files



Indexing genome



Trapnell and Salzberg, *Nat Biotechnology* 27:455 (2009)

Aligned reads

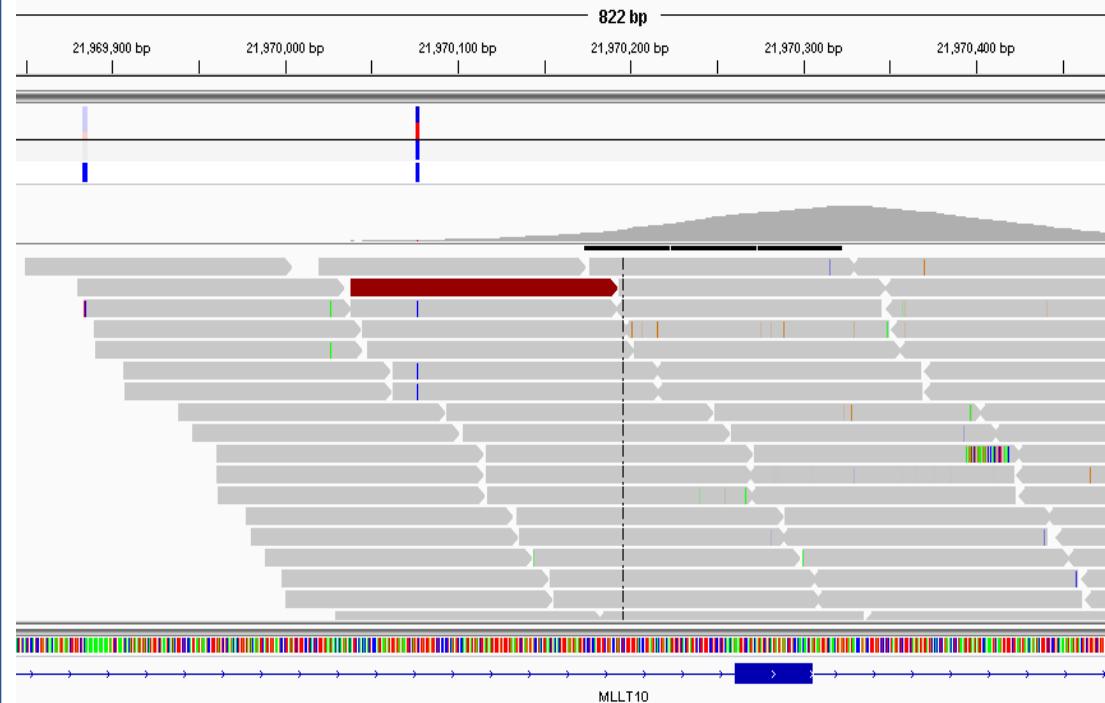
View by Samtools

```

@HD VN:1.0 SO:coordinate
@SQ SN:chr20 LN:64444167
@PG ID:TopHat VN:2.0.14 CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-realign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr20 /data/user446/mapping_tophat/L6_18_GTAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714 16 chr20 190930 3 100M * 0 0
CCGTGTTAACGGTGGATCGGGTCACCTCCAGCTAGGCTTAGGGATTCTTAGTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTAGCCCCCTCT
C BBDCDDCCDDDCDDDDCDCCDBC?DDDDDDDDDDDDDDCDCCDCC?DDDDDDDDDDDDDDCCCCEDDC?DDDDDDDDDDDDDDDDDBDHFFFFDC@AS:i:-15 XM:i:3 X0:i:0 XG:i:0 MD:Z:55C20C13A9 NM:i:3 NH:i:2 CC:Z:= CP:i:55352714 HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961 16 chr20 193953 50 100M * 0 0
TGCTGGATCATCTGGTTAGTGGCTTGTACTCAGAGGACCTCGTCCCCTGGGGCAGTGGACCTTCAGTGTATTCCCCTGACATAAGGGGCATGGACAGA
G DCDDDDDEDDDDDDCDDDDDDCCDDDCDDDEEC>DFFEJJJJJIGJJJJHGBHHGJIIJJJJJJGJJJJJJHJJJJJJHHHHHHFFFFCCCA
AS:i:-16 XM:i:3 X0:i:0 XG:i:0 MD:Z:60G16T18T3 NM:i:3 NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030 16 chr20 270877 50 100M * 0 0
GGCTTATTGGAAAAAGGAATAGCAGATTAATCAGAAATTCCACCTGGCCAGCACCAACCAGAAAAGAAGGGAAAGAACAGGAAAAAACCA
C DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHFGDJJIHJJIIJJJJIIIGFJJIIHIIJJJJJIGHHFAGHFJHFGGHFFFDD@BB
AS:i:-11 XM:i:2 X0:i:0 XG:i:0 MD:Z:0A85G13 NM:i:2 NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699 0 chr20 271218 50 50M4700N50M * 0
0 GTGGCTTCCACAGGAATGTTGAGGATGACATCCATGTCGGGTGACTTGGGCTCCGAAGCAGAACATCTCAAATGACCTCTG
accepted_hits.sam

```

View by IGV

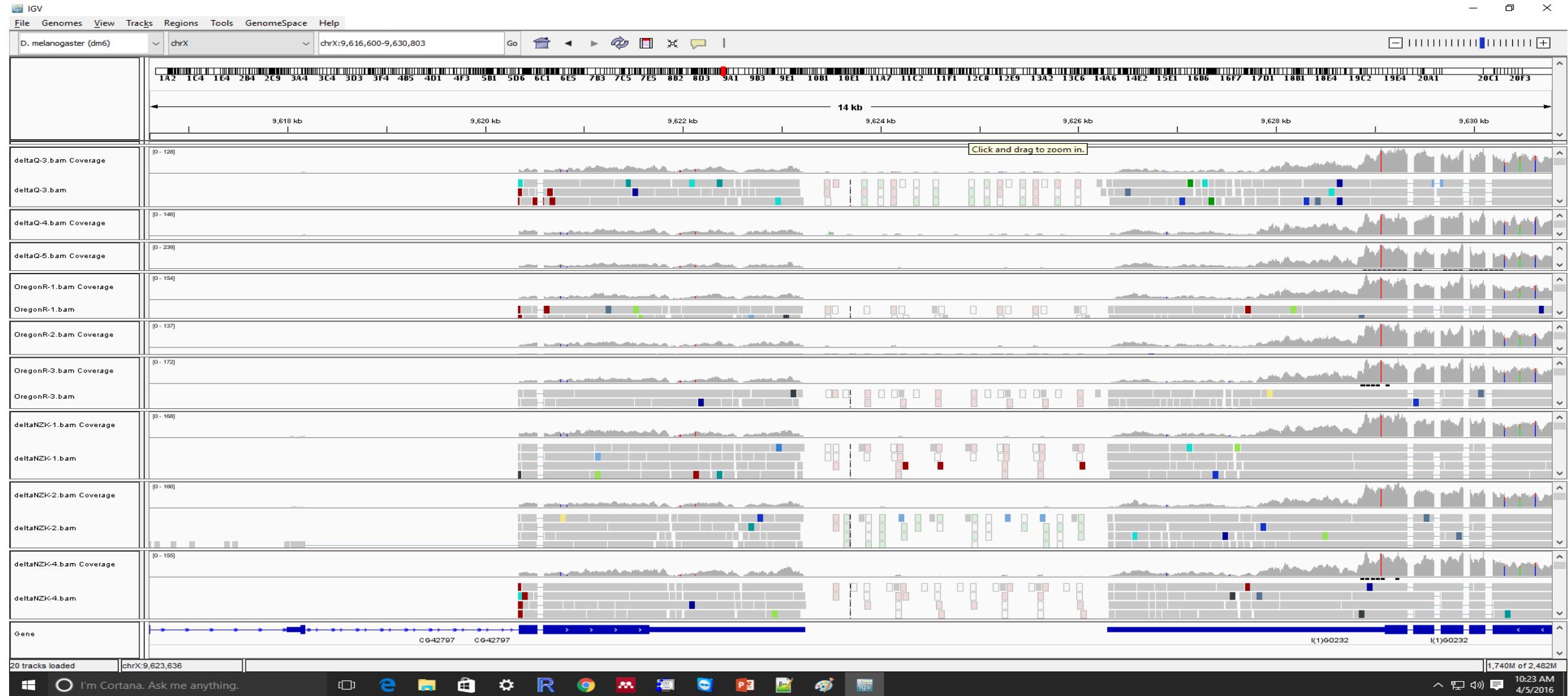


Samtool Headers

samtools view -H Konzo82_aligned.sam

```
[sbhattachary3@login3 sam_01212019]$ samtools view -H Konzo82_aligned.sam
@SQ SN:1 LN:248956422
@SQ SN:10 LN:133797422
@SQ SN:11 LN:135086622
@SQ SN:12 LN:133275309
@SQ SN:13 LN:114364328
@SQ SN:14 LN:107043718
@SQ SN:15 LN:101991189
@SQ SN:16 LN:90338345
@SQ SN:17 LN:83257441
@SQ SN:18 LN:80373285
@SQ SN:19 LN:58617616
@SQ SN:2 LN:242193529
@SQ SN:20 LN:64444167
@SQ SN:21 LN:46709983
@SQ SN:22 LN:50818468
@SQ SN:3 LN:198295559
@SQ SN:4 LN:190214555
@SQ SN:5 LN:181538259
@PG ID:bwa PN:bwa VN:0.7.17-r1194-dirty CL:bwa mem -M -t 16 /lustre/groups/vilaingrp/fastq/Homo_sapiens.GRCh38.dna.primary_assembly.fa /lustre/groups/vilaingrp/KonzoData_1/EricVilainKonzo-80421341/Konzo82/Konzo82_R1.fastq.gz /lustre/groups/vilaingrp/KonzoData_1/EricVilainKonzo-80421341/Konzo82/Konzo82_R2.fastq.gz
```

Integrated Genome Viewer (IGV)



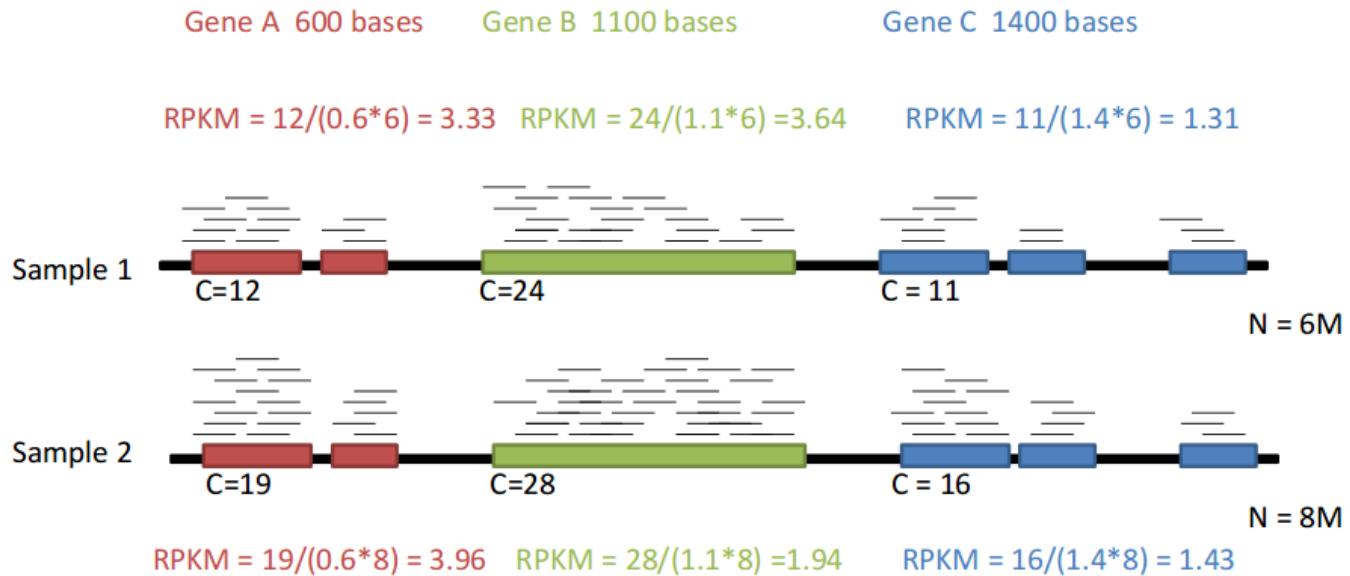
Read Counts

- Raw reads: Number of reads that align to a reference sequence in the genome. It depends on amount of fragments sequenced and length of the reference sequence.
- Counts Per Million (CPM): It is the raw counts (X_i) scaled by number of fragments sequenced (N) times one million.

$$\text{CPM}_i = \frac{X_i}{N} = \frac{X_i}{\overline{N}} \cdot 10^6$$

Read Counts (contd...)

RPKM Example



- Reads per kilobase of exons per million (RPKM) is a normalized read count.

$$RPKM = \frac{\text{number of reads of the region}}{\frac{\text{total reads}}{1,000,000} \times \frac{\text{region length}}{1,000}}$$

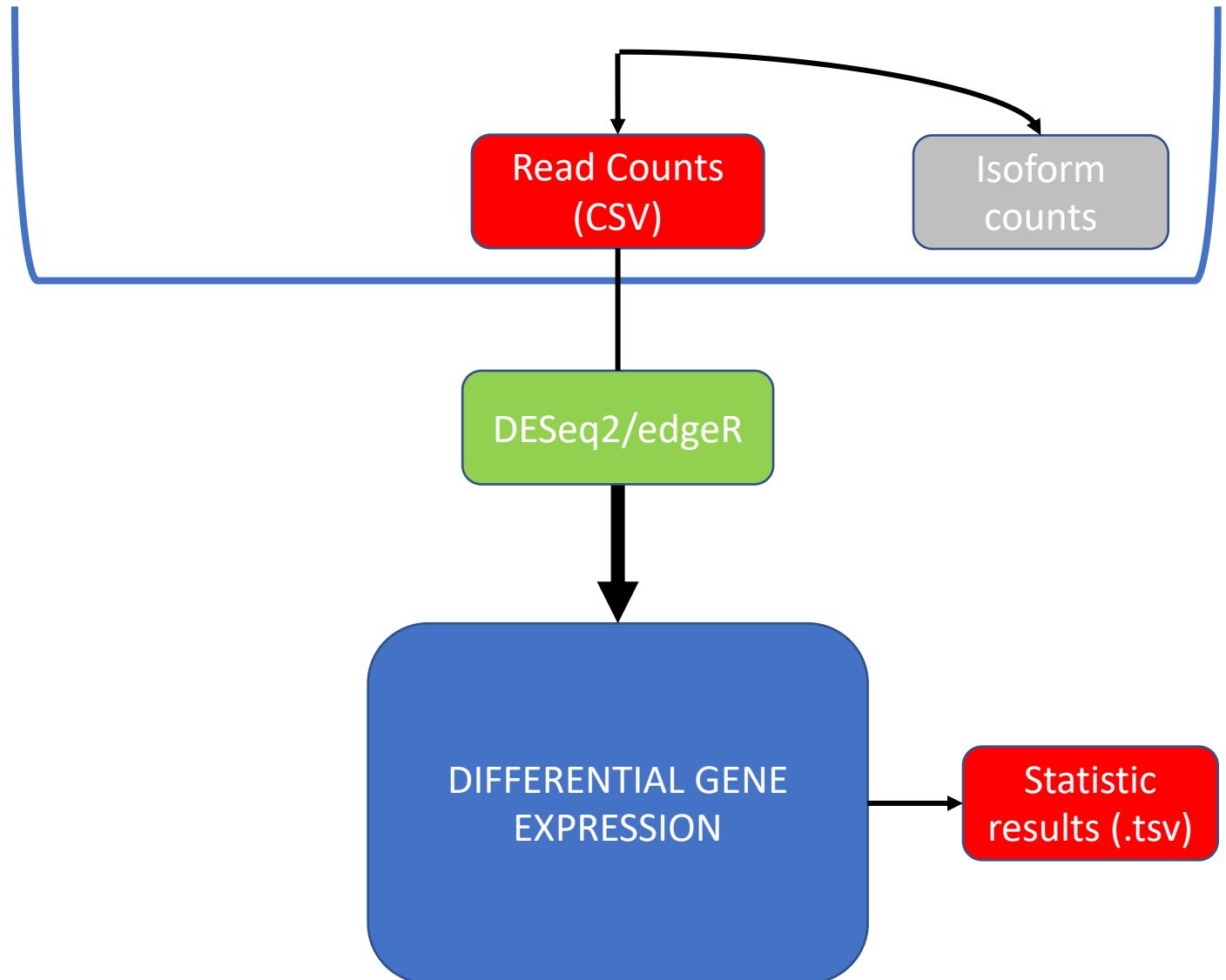
- Fragments per kilobase of exons per million (FPKM) is similar to RPKM, only it works for paired end reads.

Differential Gene Expression

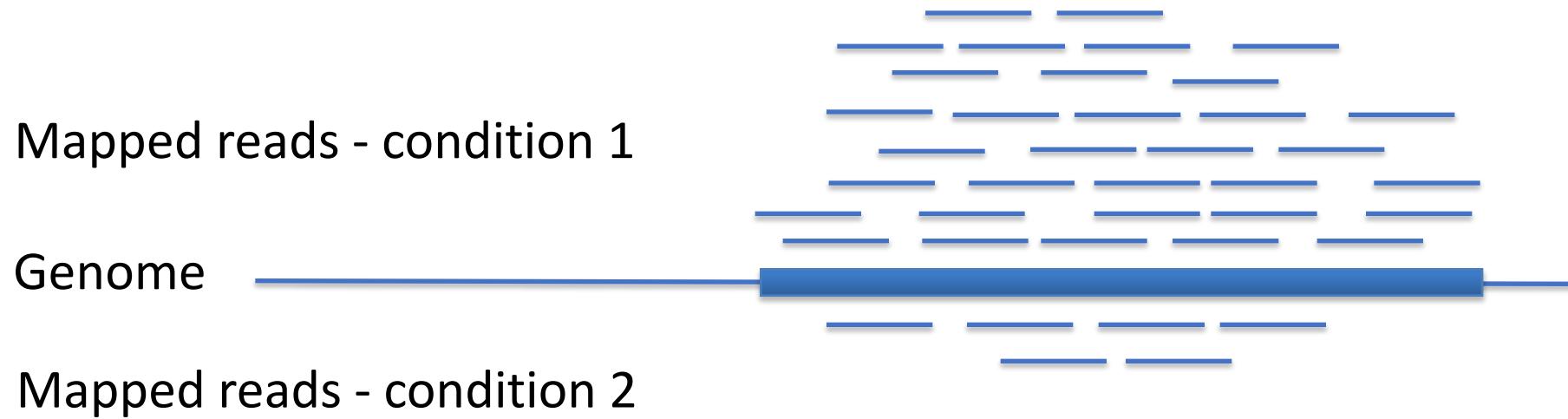
Tool: DESeq2, edgeR

Input: Reads count (.CSV)

Output: TSV files and



Differential Expression



Differential Expression (contd..)

- Counting reads
- Statistical significance testing

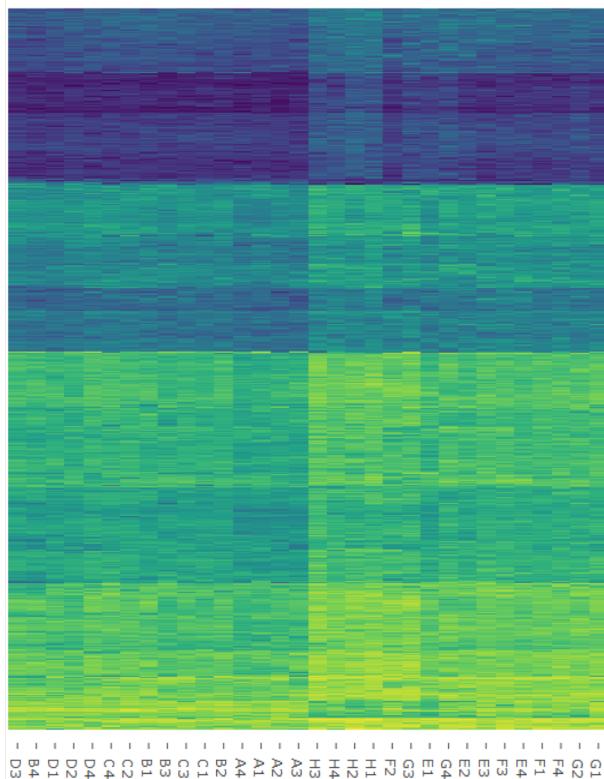
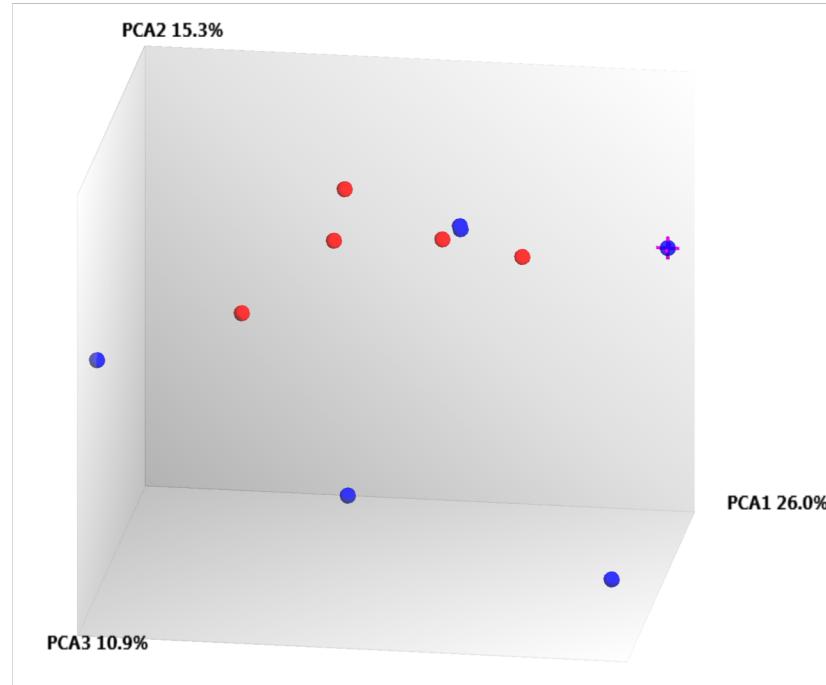
	Sample_A	Sample_B	Fold_Change	Significant?
Gene A	1	2	2-fold	No
Gene B	100	200	2-fold	Yes

More Counts = more Statistical significance

Example: 5000 total reads per sample.
Observed 2-fold differences in read counts.

	SampleA	Sample B	Hypothesis Test (P-value)
geneA	1	2	1.00
geneB	10	20	0.098
geneC	100	200	< 0.001

Outputs from Microarray and RNAseq

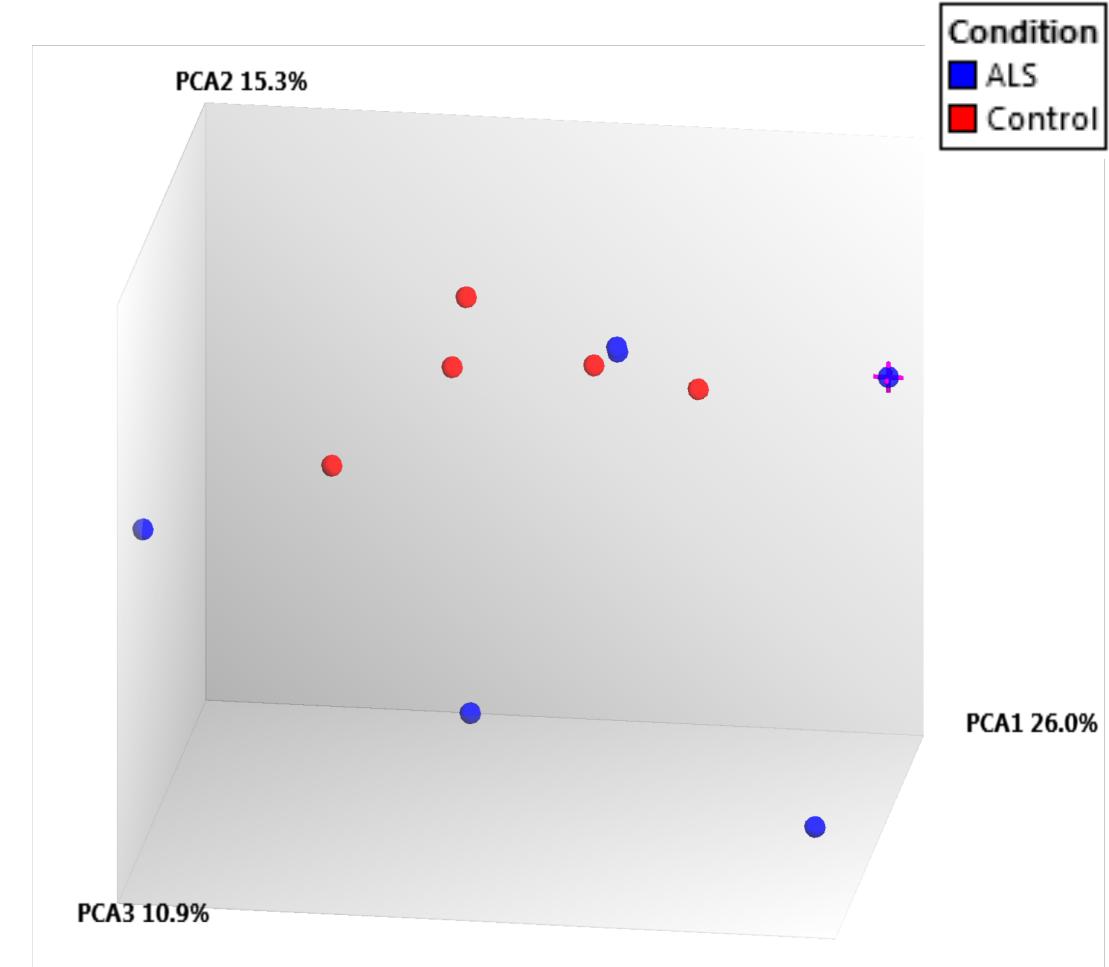


test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2fold	test_stat	p_value	q_value	significance
1	XLOC_000	XLOC_000 CG33281	2L:332423	deltaQ	OregonR	OK	7.78404	28.3487	1.86469	3.89904	5.00E-05	0.00347601	yes
2	XLOC_000	XLOC_000 CG33282	2L:332790	deltaQ	OregonR	OK	1.46714	5.18562	1.82151	3.08678	5.00E-05	0.00347601	yes
3	XLOC_000	XLOC_000 CG2772	2L:344599	deltaQ	OregonR	OK	15.7892	33.2996	1.07657	2.46754	0.0001	0.00623576	yes
4	XLOC_000	XLOC_000 pgant4	2L:347260	deltaQ	OregonR	OK	3.81965	1.71272	-1.15715	-2.10786	0.00045	0.022476	yes
5	XLOC_000	XLOC_000 CG3355	2L:465140	deltaQ	OregonR	OK	15.2984	2.46559	-2.63338	-3.80939	5.00E-05	0.00347601	yes
6	XLOC_000	XLOC_000 Sgs1	2L:493718	deltaQ	OregonR	OK	48.9408	25.6161	-0.93395	-1.85136	0.00095	0.037339	yes
7	XLOC_000	XLOC_000 Gpdh	2L:594364	deltaQ	OregonR	OK	89.1781	179.681	1.01074	2.93463	5.00E-05	0.00347601	yes
8	XLOC_000	XLOC_000 Muc26	2L:615196	deltaQ	OregonR	OK	49.2003	20.3772	-1.27171	-2.76563	5.00E-05	0.00347601	yes
9	XLOC_000	XLOC_000 CG15818	2L:741090	deltaQ	OregonR	OK	31.8493	17.9784	-0.825	-1.95182	0.00105	0.0427014	yes
10	XLOC_000	XLOC_000 CG7164,Uro	2L:778008	deltaQ	OregonR	OK	13.686	36.0323	1.39659	2.98212	5.00E-05	0.00347601	yes
11	XLOC_000	XLOC_000 Aldh	2L:937030	deltaQ	OregonR	OK	199.624	328.681	0.719399	2.16988	0.00015	0.00871949	yes
12	XLOC_000	XLOC_000 CG3841	2L:958947	deltaQ	OregonR	OK	10.0911	24.0034	1.25016	2.94501	5.00E-05	0.00347601	yes
13	XLOC_000	XLOC_000 ApoltP	2L:963417	deltaQ	OregonR	OK	10.3169	6.35646	-0.69872	-2.08466	5.00E-05	0.00347601	yes
14	XLOC_000	XLOC_000 CG33301	2L:100495	deltaQ	OregonR	OK	24.7792	99.4039	2.00417	5.07914	5.00E-05	0.00347601	yes
15	XLOC_000	XLOC_000 CR44874	2L:100513	deltaQ	OregonR	OK	0.43958	1.70197	1.953	2.27683	0.00065	0.0297238	yes
16	XLOC_000	XLOC_000 Cpr31A	2L:100542	deltaQ	OregonR	OK	6.31093	17.4553	1.46774	2.94322	5.00E-05	0.00347601	yes
17	XLOC_000	XLOC_000 CG17134	2L:108042	deltaQ	OregonR	OK	36.6715	75.0104	1.03243	2.59761	5.00E-05	0.00347601	yes
18	XLOC_000	XLOC_000 NimC2	2L:139801	deltaQ	OregonR	OK	42.4803	17.2581	-1.29953	-2.91512	0.00025	0.0133971	yes
19	XLOC_000	XLOC_000 Adh,Adh	2L:145991	deltaQ	OregonR	OK	56.877	965.215	0.772912	2.16993	0.0005	0.0244976	yes
20	XLOC_001	XLOC_001 CG15155,CG17681,CG5783	2L:181529	deltaQ	OregonR	OK	50.2681	124.125	1.30408	2.63595	5.00E-05	0.00347601	yes
21	XLOC_001	XLOC_001 CG17325	2L:188125	deltaQ	OregonR	OK	134.215	227.034	0.758362	2.13804	0.0004	0.0200761	yes
22	XLOC_001	XLOC_001 CG9259	2L:210893	deltaQ	OregonR	OK	25.9753	5.20173	-2.32008	-4.57379	5.00E-05	0.00347601	yes
23	XLOC_001	XLOC_001 CG42460,Sfp23F	2L:338735	deltaQ	OregonR	OK	97.3734	40.2101	-1.27597	-2.57545	5.00E-05	0.00347601	yes
24	XLOC_001	XLOC_001 tim	2L:349397	deltaQ	OregonR	OK	2.23157	4.79765	1.10427	2.39393	0.0001	0.00623576	yes
25	XLOC_001	XLOC_001 CG16712	2L:369621	deltaQ	OregonR	OK	108.423	200.422	0.886362	2.01938	0.0006	0.0284488	yes
26	XLOC_001	XLOC_001 Sgs1	2L:493718	deltaQ	OregonR	OK	125.233	58.323	-1.10242	-2.06432	0.0008	0.035327	yes
27	XLOC_001	XLOC_001 CG9498	2L:634939	deltaQ	OregonR	OK	11.4523	26.0854	1.18761	2.6796	5.00E-05	0.00347601	yes
28	XLOC_001	XLOC_001 CG67025	2L:769110	deltaQ	OregonR	OK	27.8459	48.1527	0.79015	2.01797	0.0006	0.0284488	yes
29	XLOC_001	XLOC_001 CG7214	2L:774368	deltaQ	OregonR	OK	13.5935	26.837	0.981304	2.09078	0.0005	0.0244976	yes
30	XLOC_001	XLOC_001 CG7203	2L:775216	deltaQ	OregonR	OK	21.7555	52.889	1.28159	2.76209	5.00E-05	0.00347601	yes
31	XLOC_002	XLOC_002 CG9463,CG9465	2L:876526	deltaQ	OregonR	OK	31.923	18.3205	-0.80114	-1.91992	0.00125	0.0488921	yes
32	XLOC_002	XLOC_002 CG4382	2L:959887	deltaQ	OregonR	OK	4.6955	1.29471	-1.85865	-2.67568	5.00E-05	0.00347601	yes
33	XLOC_002	XLOC_002 CG7299	2L:106834	deltaQ	OregonR	OK	83.4625	211.766	1.34237	3.95222	5.00E-05	0.00347601	yes
34	XLOC_002	XLOC_002 CG17108	2L:106925	deltaQ	OregonR	OK	24.7576	43.8733	0.825473	2.04522	0.00055	0.0266932	yes
35	XLOC_002	XLOC_002 CG31869	2L:107720	deltaQ	OregonR	OK	11.1676	5.46428	-1.03121	-2.32458	0.0001	0.00623576	yes
36	XLOC_002	XLOC_002 AstC	2L:110761	deltaQ	OregonR	OK	3.61998	12.8274	1.82517	3.39604	5.00E-05	0.00347601	yes
37	XLOC_002	XLOC_002 CG16957	2L:133953	deltaQ	OregonR	OK	8.18801	2.37806	-1.78373	-2.85575	5.00E-05	0.00347601	yes
38	XLOC_002	XLOC_002 CG3306,CG8997	2L:138279	deltaQ	OregonR	OK	1632.48	474.126	-1.78372	-2.60345	5.00E-05	0.00347601	yes

Differential Expression Table

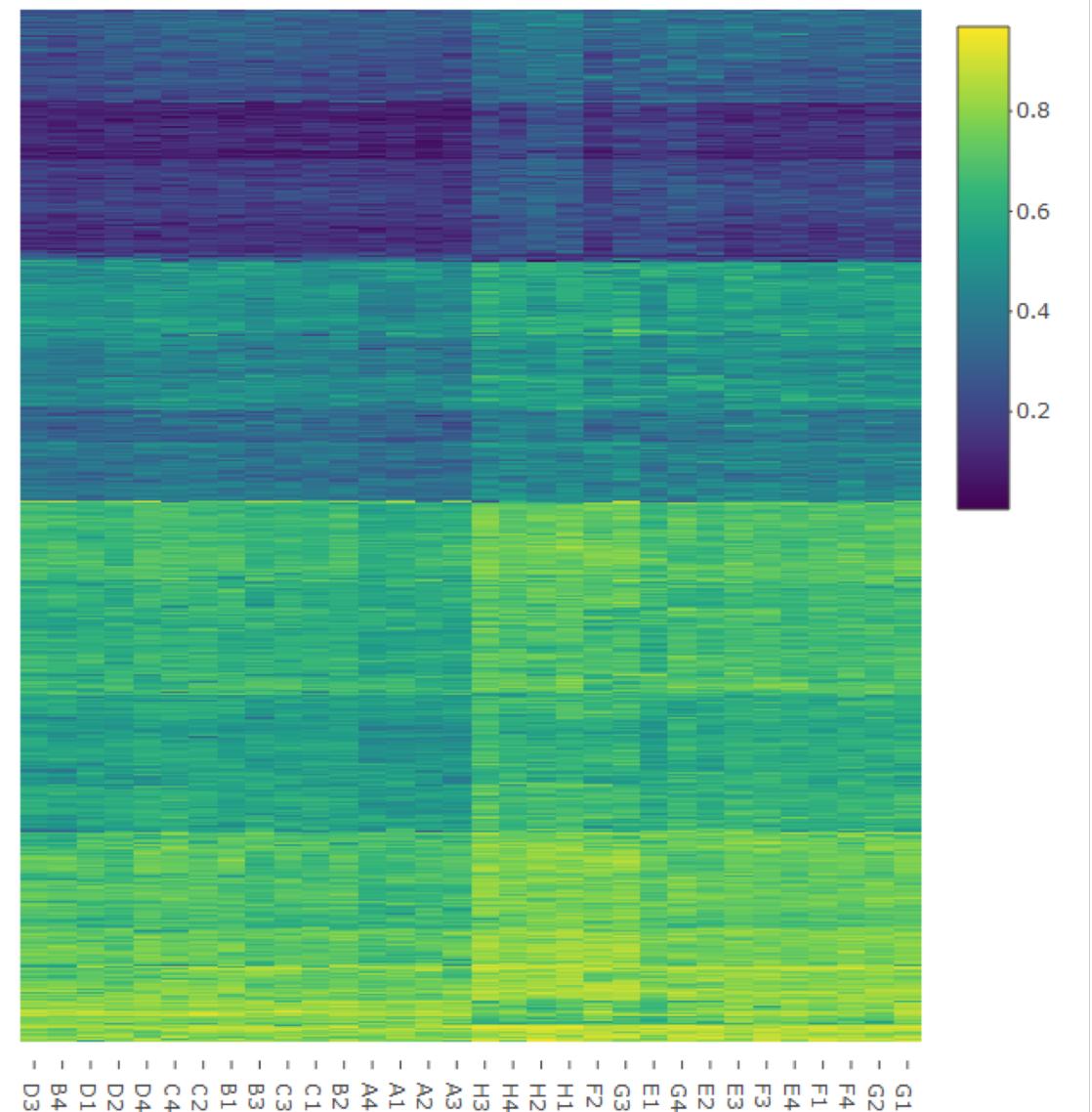
Principle component Analysis (PCA) Plot

- Reduces the number of components for a condition.
- Helps in visualizing variabilities within samples of the same condition.
- In the figure on the right, there are 6 experimental (**Blue**) and 5 control conditions (**Red**).
- We can see, 5 control samples are quite variable, whereas 2 out of 6 experimental samples have least variability.
- Also there is variability between the experimental and the control samples.

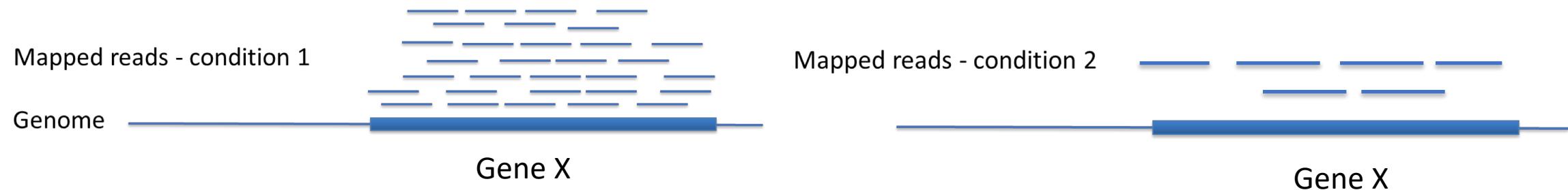


Heatmaps

- Heatmaps are visual representation of the expression of genes.
- In the image on the right, color ranges from dark blue (low expression) to light yellow.
- Expression can be represented either raw (raw intensity in case of microarray; raw read counts in case of RNA-seq) or scaled.
- Scaling of expression data can be divided into 2 parts:
 - ✓ Logarithmic scaled
 - ✓ Z-score values



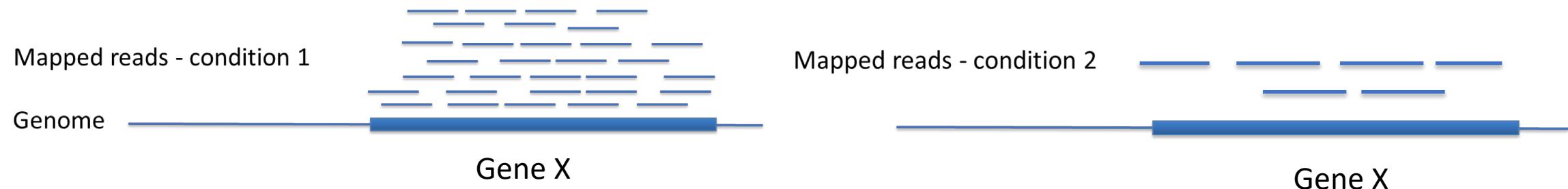
Differential Expression tables



Expression of Gene in
condition 1 and
condition 2

Gene	Condition I	Condition II	Log fold change	T-test	P-value	Q-value
Gene X	100	10	2.3	282.59	2.2e-16	2.2e-14

Differential Expression tables



Change in expression
in log based 2 format

Gene	Condition I	Condition II	Log fold change	T-test	P-value	Q-value
Gene X	100	10	2.3	282.59	2.2e-16	2.2e-14

T-test

- T-test is a statistical hypothesis testing to determine, if two conditions are significantly different or not.
- For a Transcriptomic expression:
 - Null Hypothesis (H_0)= Means of expression of the gene X in 2 conditions are **equal**
 - Alternative Hypothesis (H_A)= Means of expression of the gene X in 2 conditions are **not equal**.

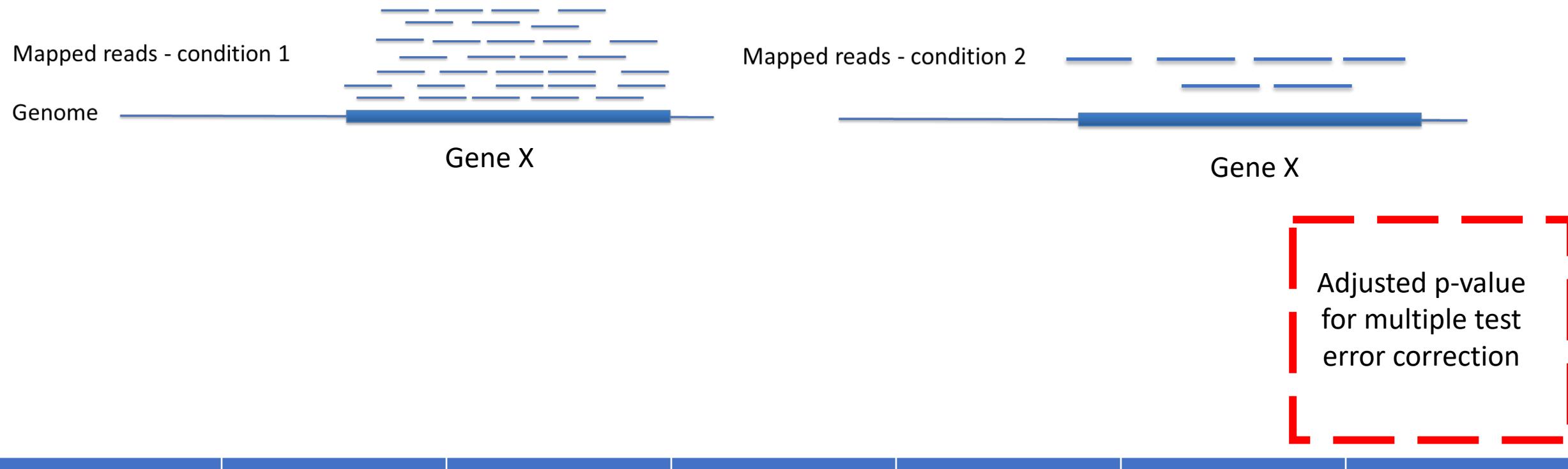
$$t\text{-value} = \frac{\text{Mean of expression Gene X condition1} - \text{Mean of Expression Gene X Condition 2}}{\text{Standard error (SE)}}$$

- Positive t-values means condition 1 has a higher value than condition 2; whereas negative value means expression of gene in condition 2 more than in 1.

P-values

- P-value is the probability of occurrence of a given event.
- In case of a t-test, higher p-value would signify the higher chance of Null Hypothesis being true. So a p-value of 0.99 means there is 99% chance that the 2 genes have same mean expression.
- Lower the p-value, lesser the chance of the null hypothesis being true, and higher chance of alternative hypothesis being true. A p-value of 0.05, would mean 5% chance of the 2 genes having, same mean expression across the two conditions.
- For biological experiments threshold for p-value is 0.05.
- P-value depends on t-value and degrees of freedom, which is generally number of samples -1.

Differential Expression tables



Gene	Condition I	Condition II	Log fold change	T-test	P-value	Q-value
Gene X	100	10	2.3	282.59	2.2e-16	2.2e-14

Adjusted P-values

- Multiple hypothesis testing leads to the rejection of True Positives. So, the p-value is not the measure of the significance of the test.
- To reduce the error rate p values need to be corrected.
- Most of the differential expression tools calculate adjusted p-value by 2 methods
 - ❖ **Bonferroni Correction:**
 - ✓ Adjusted p-value = $\frac{p\text{-value for the experiment}}{\text{Number of samples}}$
 - ✓ Compare it with adjusted p-value threshold (0.05). If less; than alternative hypothesis significant.
 - ❖ **Benjamini Hochberg False Discovery Rate (FDR):**
 - ✓ Sort the frequency in ascending order and rank them, i.e. lowest p-value is rank 1, next rank 2 and so on.
 - ✓ $\text{FDR}(q\text{-value}) = \left(\frac{\text{Rank of the } p\text{-value}}{\text{Number of samples}}\right) * \text{False Discovery rate}$
 - ✓ Compare it with q-value threshold (0.05). If less; than alternative hypothesis significant.
- For Multiple testing q-value is the best way to measure significance.

<https://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503>

What to do with significant genes?

Significant Gene List

Functional Annotation



Interaction Database

