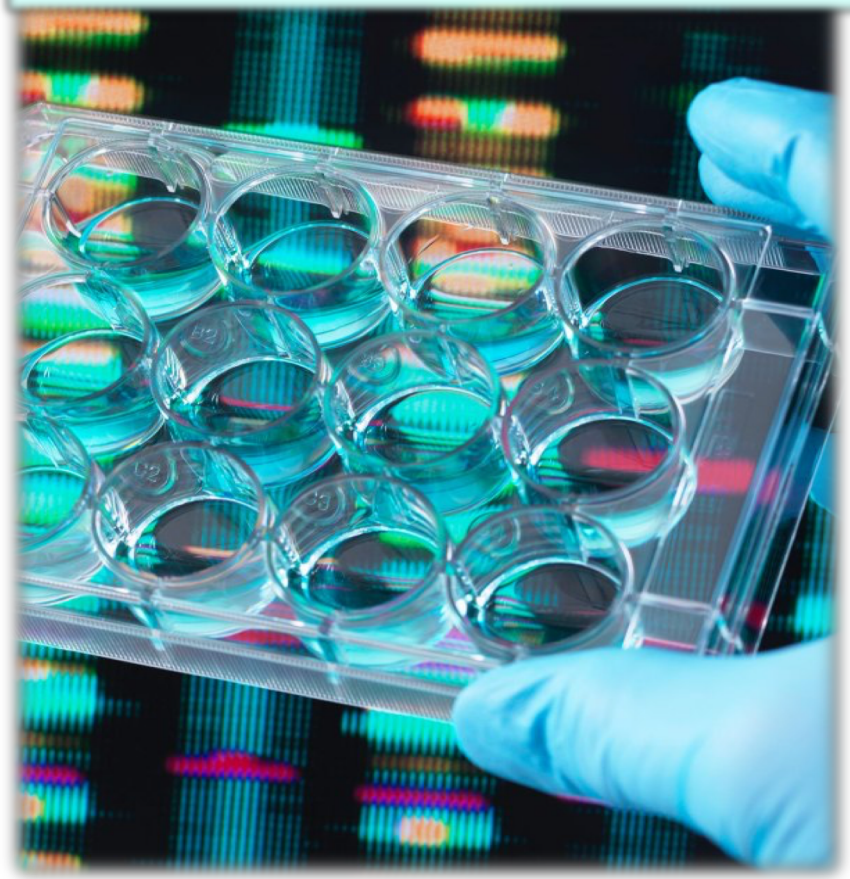
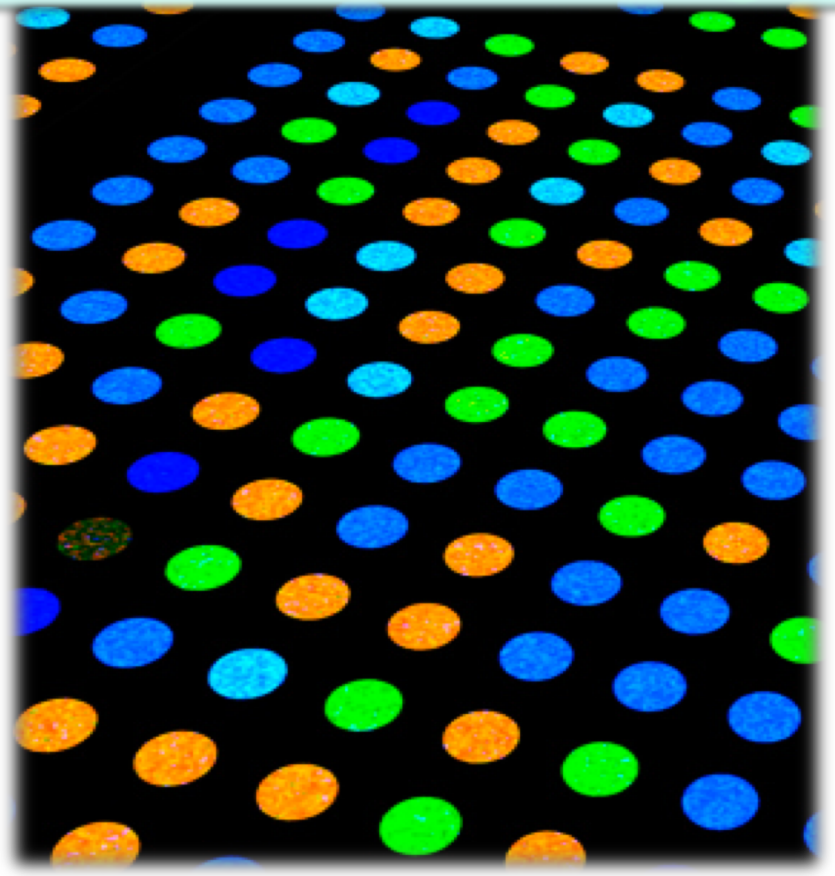
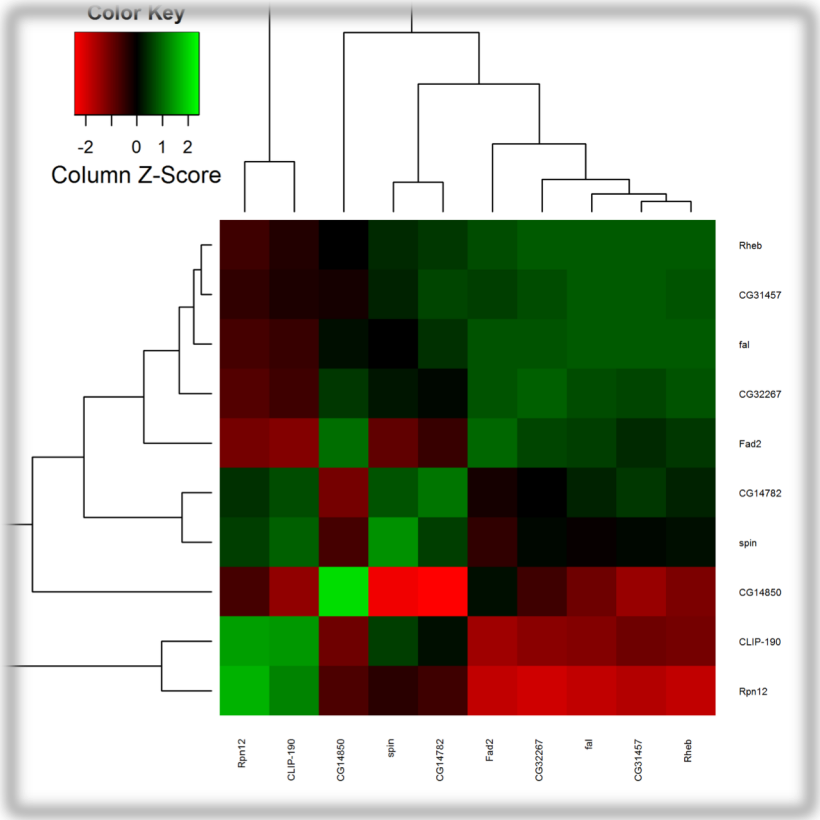
# Transcriptomics : Microarrays and RNASeq

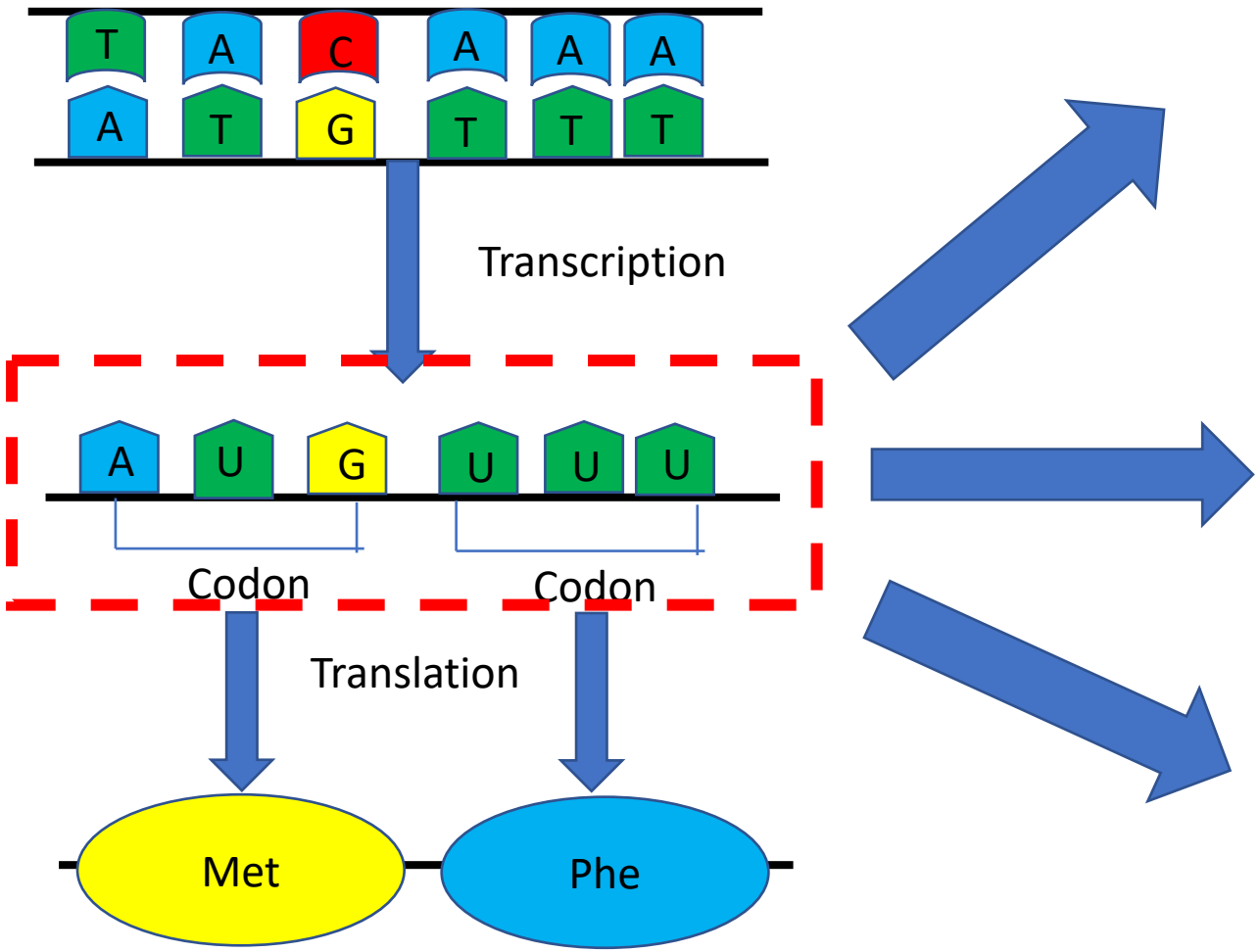




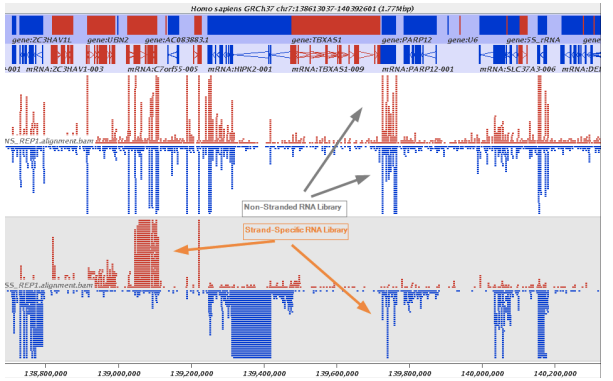



https://bi-ctsicn.github.io/CBU/
bioinformatics@childrensnational.org

# Transcriptomics: What and How?

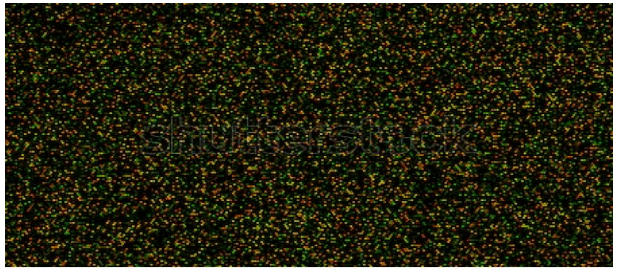



Central Dogma



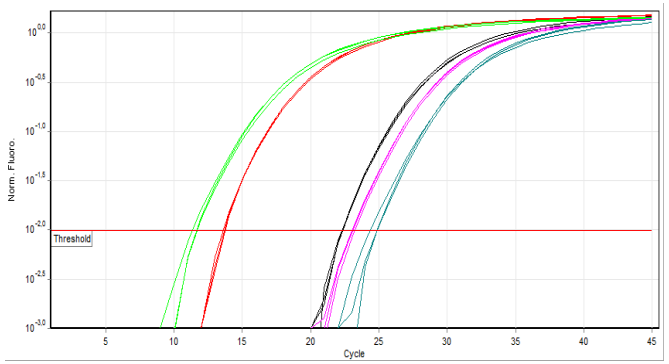

https://galaxyproject.org/tutorials/rb_rnaseq/

RNASeq



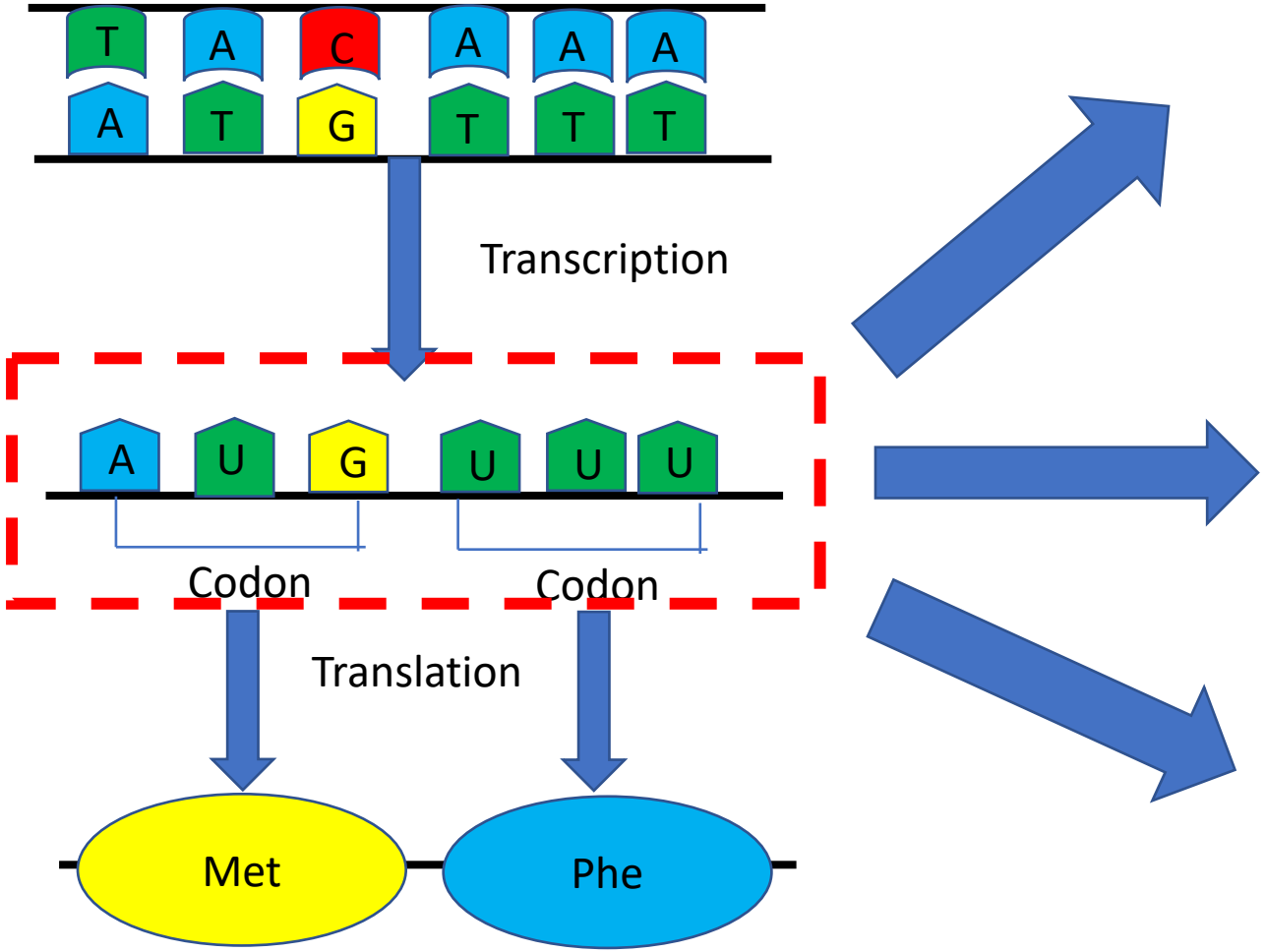
www.shutterstock.com • 95066617

Microarray

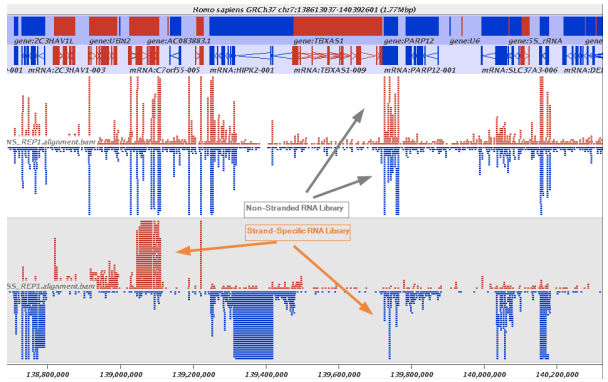



Quantitative PCR

# Transcriptomics: What and How?



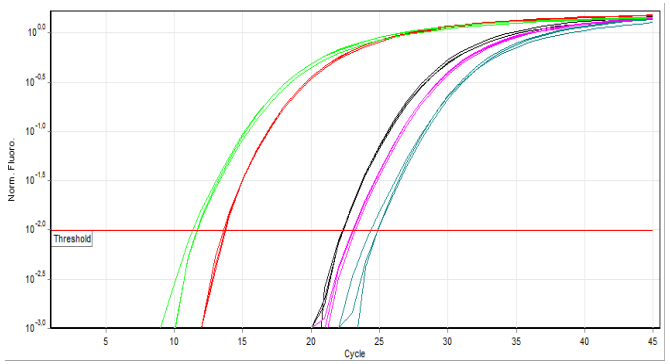

Central Dogma



https://galaxyproject.org/tutorials/rb_rnaseq/

Whole Transcriptome

Whole Transcriptome or Targeted

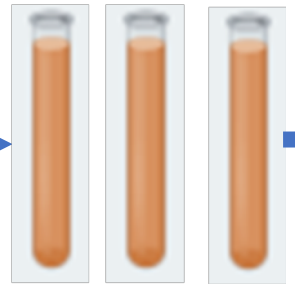



Targeted

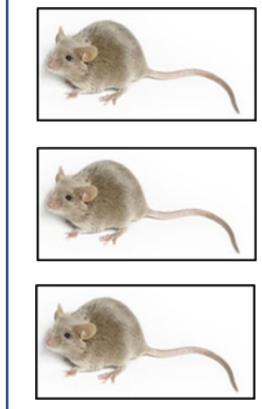
# Workflow of a microarray experiments

Control

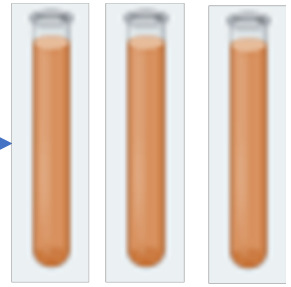
Control RNA Samples



Experiment

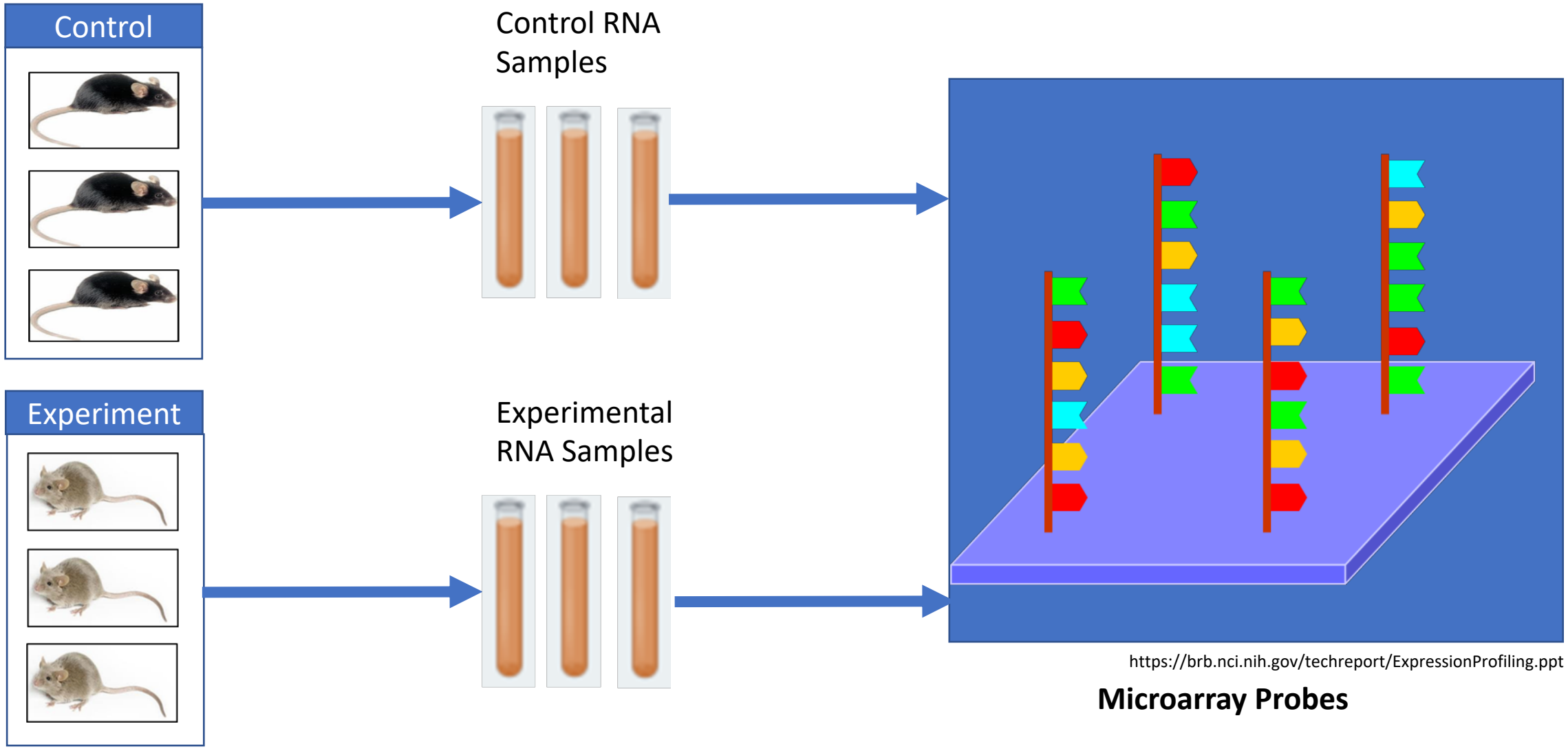


Experimental RNA Samples

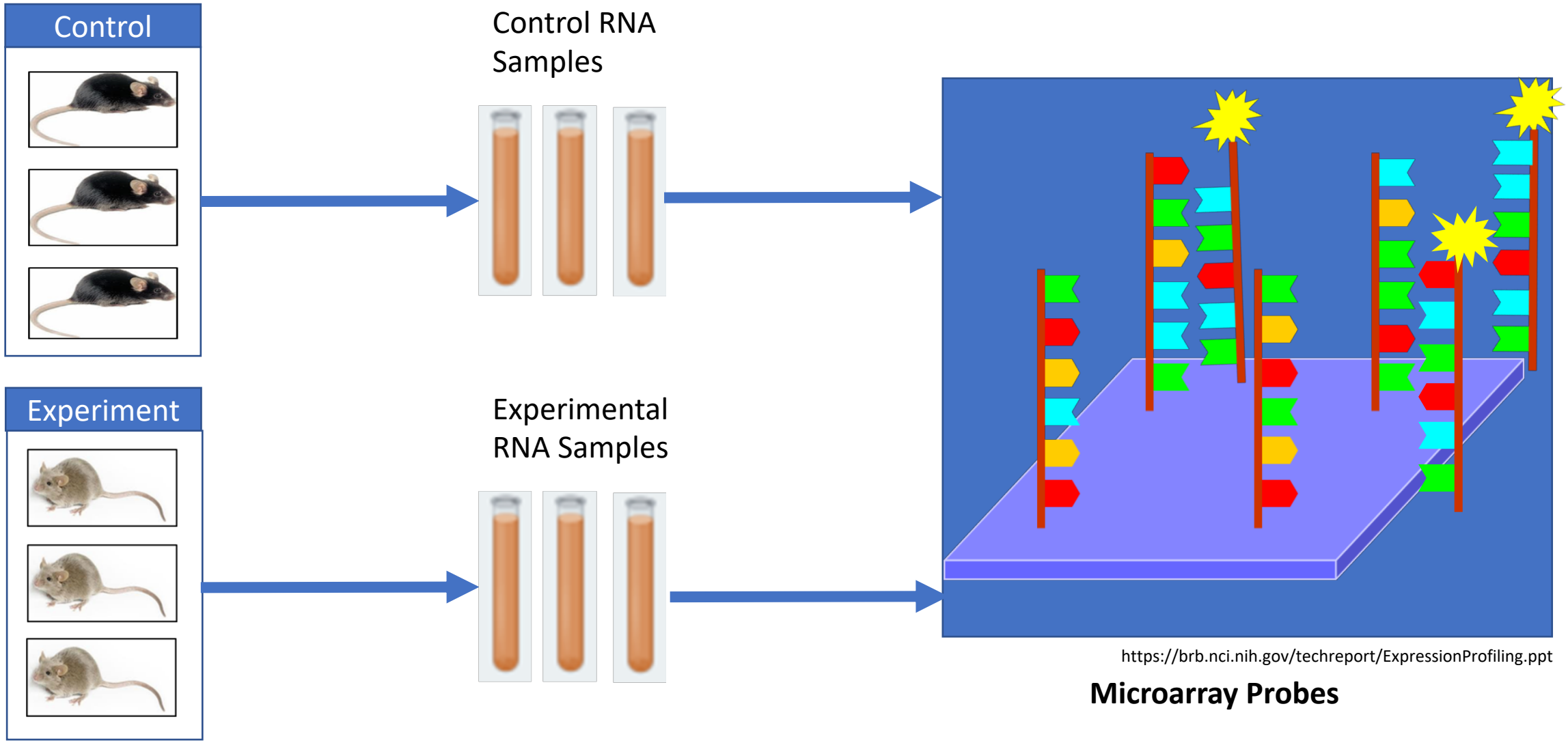


**Affymetrix microarray platform**

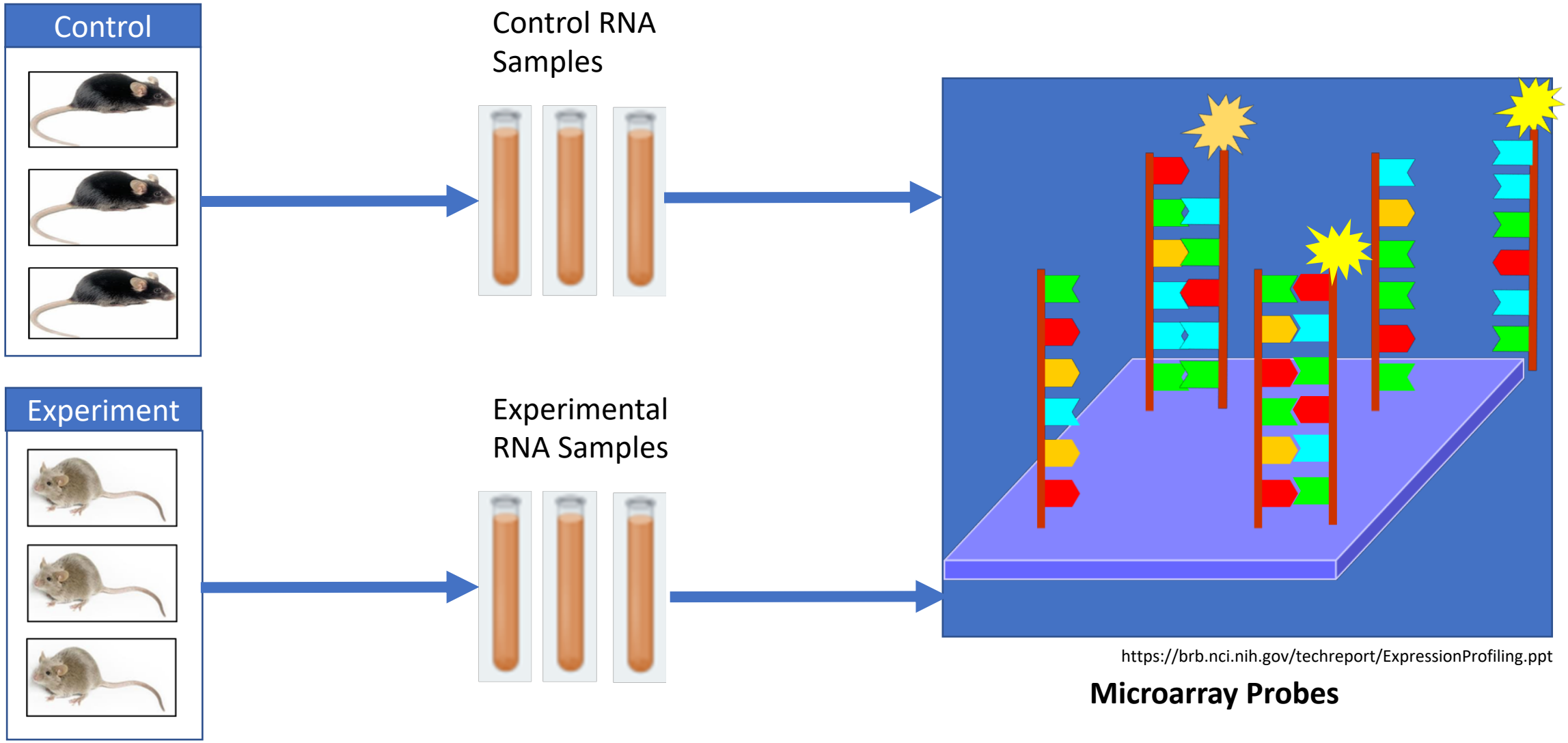
# Workflow of a microarray experiments



# Workflow of a microarray experiments



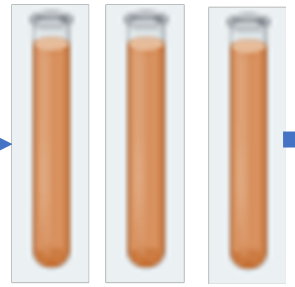
# Workflow of a microarray experiments



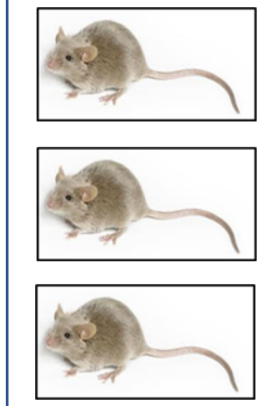
# Workflow of a microarray experiments

Control

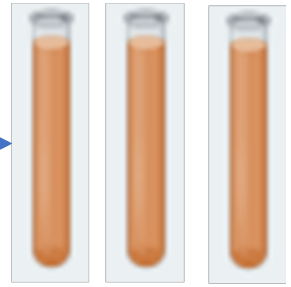
Control RNA Samples

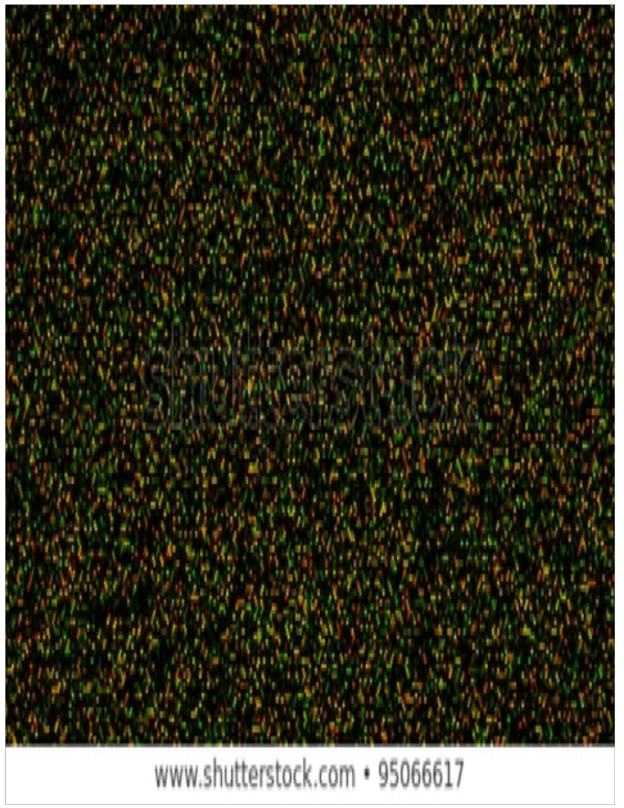


Experiment



Experimental RNA Samples

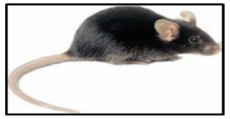




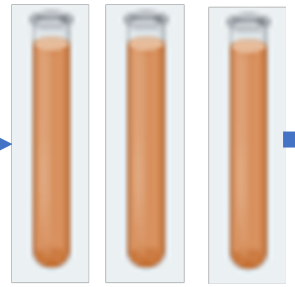

**Microarray Plates**

# Workflow of a RNAseq experiments

Control

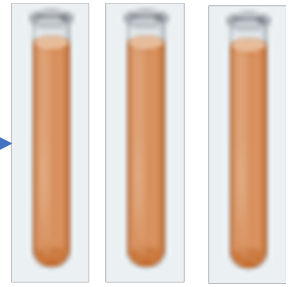


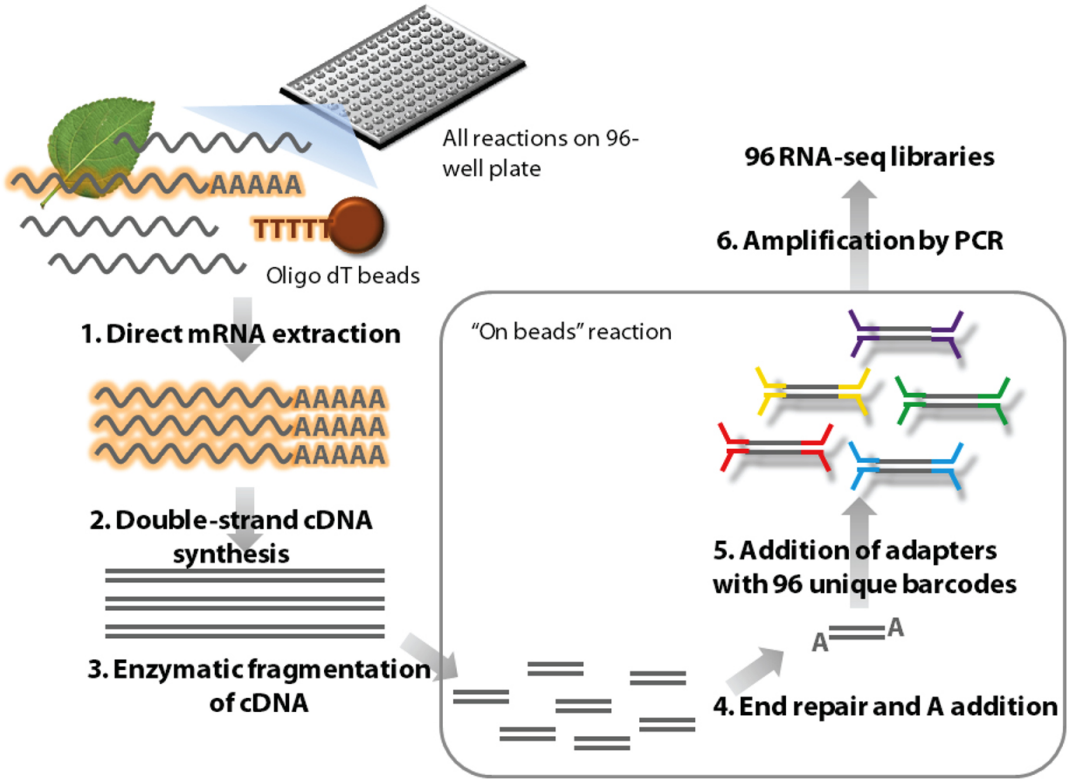
Control RNA Samples



Experiment

Control RNA Samples





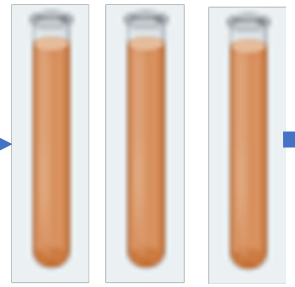

https://www.frontiersin.org/articles/10.3389/fpls.2012.00202/full

**Library preparation Pipeline**

# Workflow of a RNAseq experiments

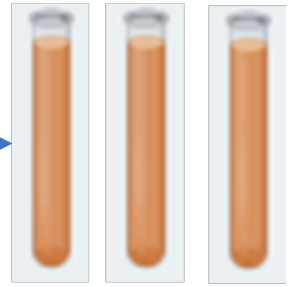
Control

Control RNA Samples



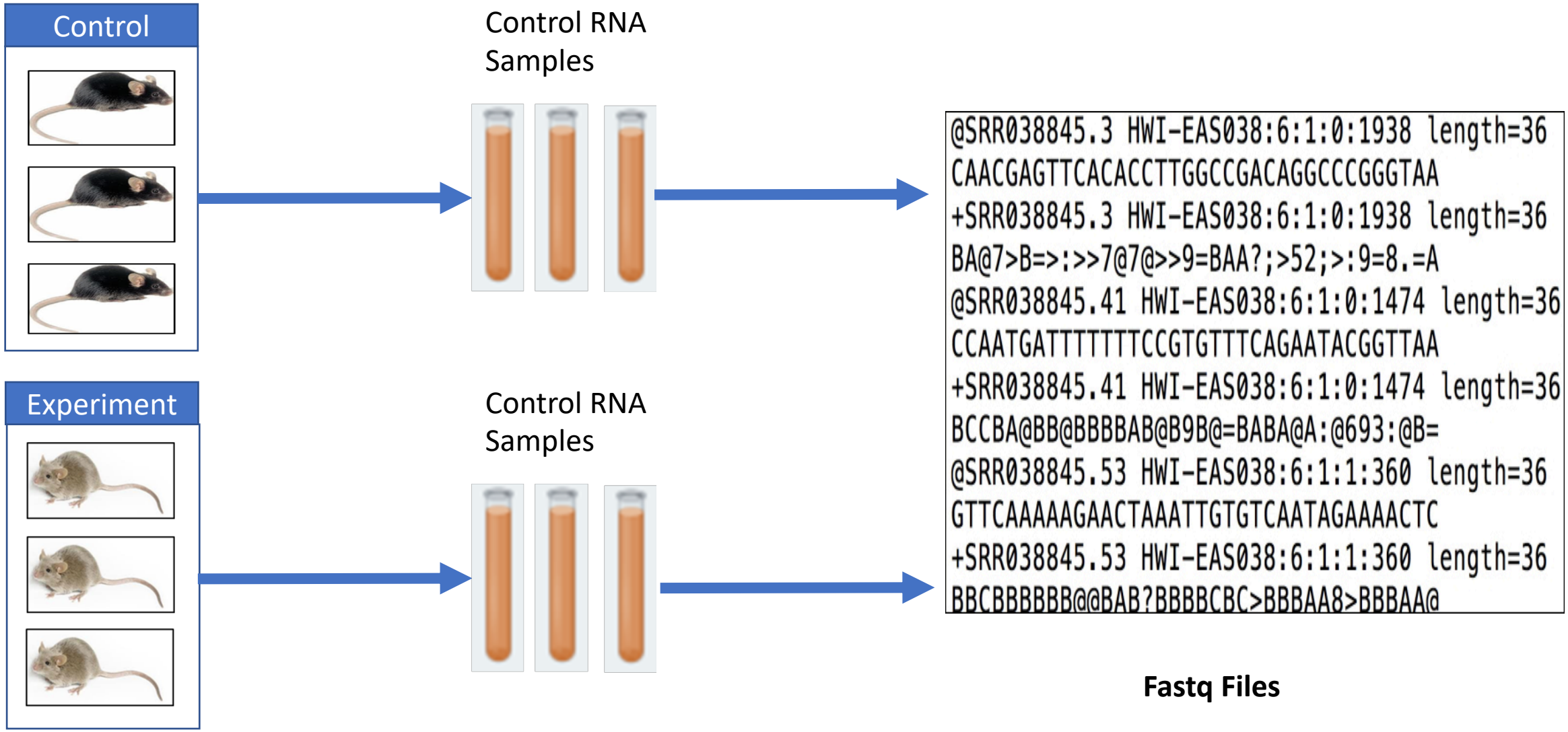
Experiment

Control RNA Samples




**Illumina Nextseq**

# 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 Citatio

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

BioMed Central The Open Access Publisher | BMC Bioinformatics | this article | search | submit a manuscript | register

BMC Bioinformatics. 2018; 19: 296.
Published online 2018 Aug 8. doi: 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 annotatated.
- Less cost per sample. Cost depends on number of probes used
- Evaluates expression of known genes.
- Relative abundance. Intensity.
- 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/non-coding transcripts .
- Absolute abundance. Read counts.
- 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?

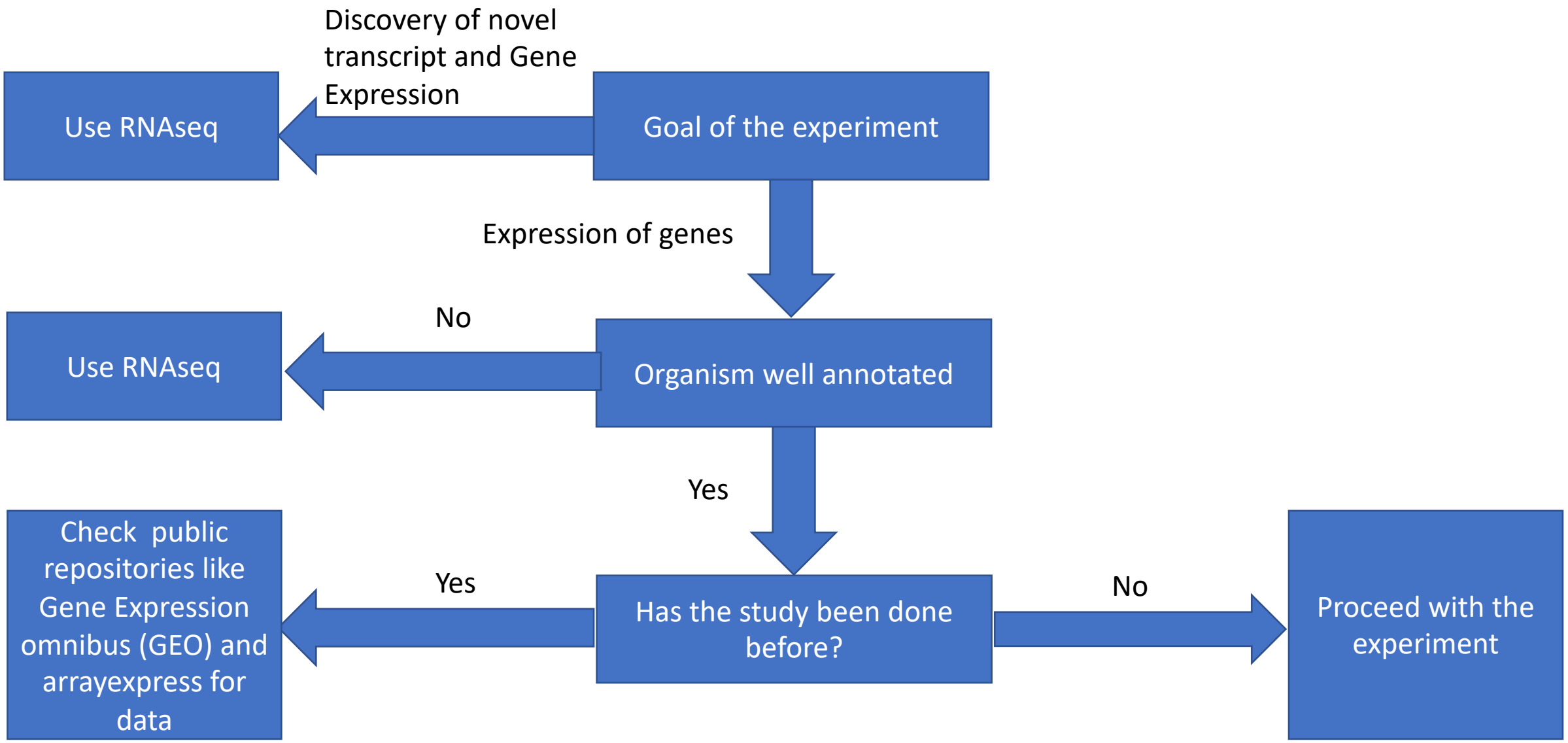
# Life cycle of a Genomic Experiment


- 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 (H0) when a specific alternative hypothesis (H1) 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

IV) Mean difference in log-expression between treatment and control conditions as postulated under the alternative hypothesis H1.

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?**

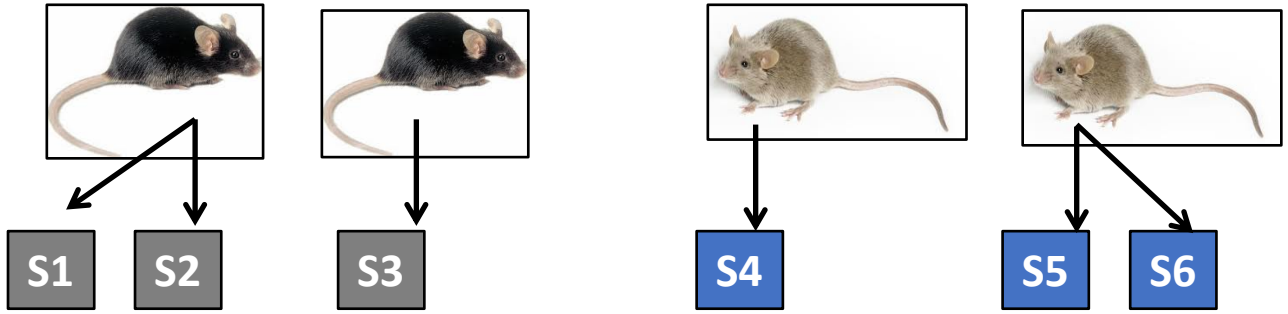
- Identification of random Variation
- Accuracy of measurement

# 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**

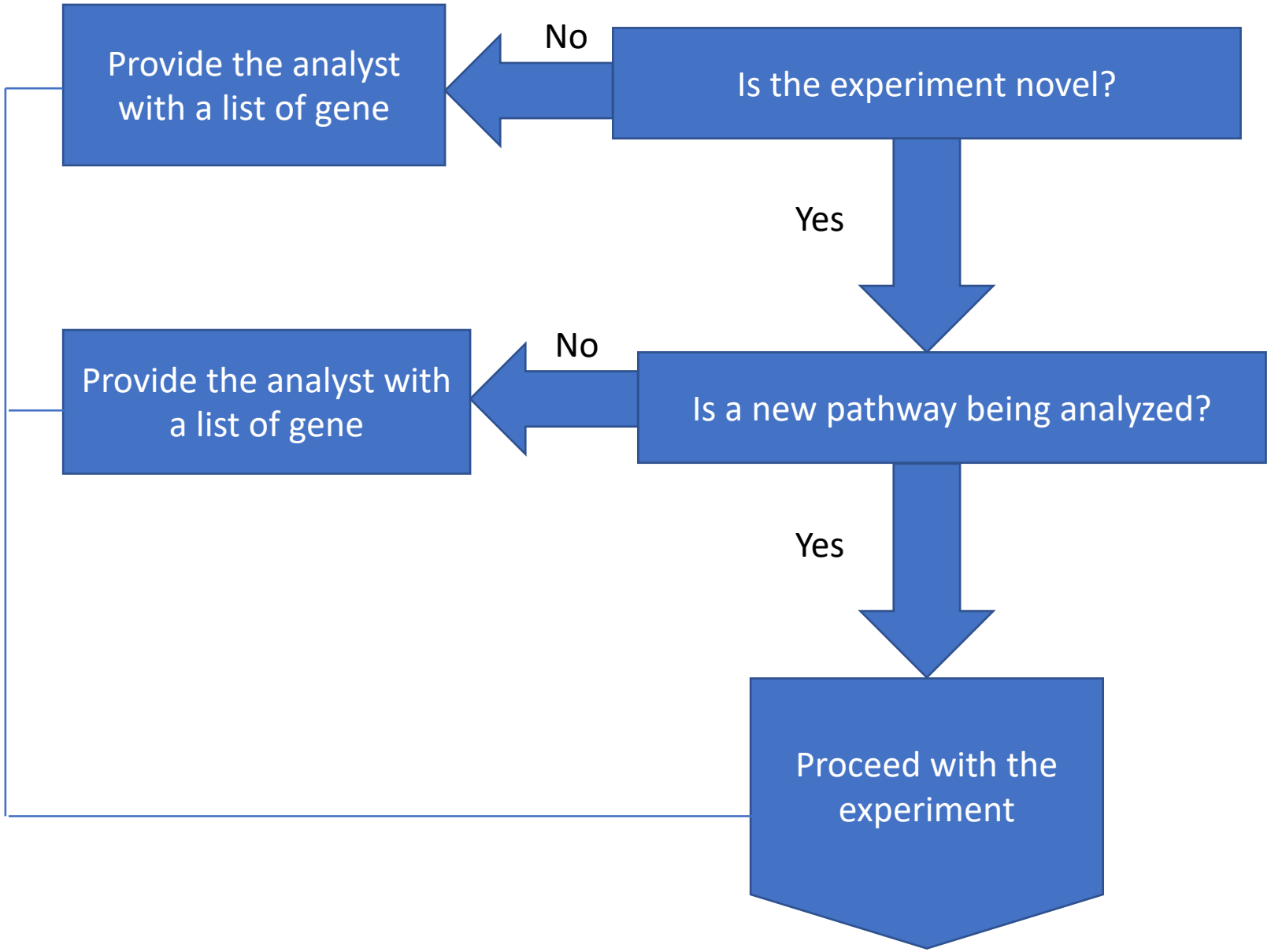
To make inferences about  the population you need ***biological replicates***

# Transcriptomics in the Genomics Core

- **Array-based Assays**
  - $200-800/sample, depending on type of array
  - Affymetrix mRNA and miRNA Expression Profiling Microarrays, 3’ ivt and exome
  - RIN ≥ 6, 260/280/230 ratios, quantity as low as 100 ng, but more is better (~300 ng)

- **RNA Sequencing**
  - Priced per run ~$200-1,800/sample, depth and coverage increase cost
  - Illumina miRNA Sequencing
  - Illumina RNA Sequencing (PolyA, Custom, Whole Transcriptome, Depleted (Mito/Ribo) Whole Transcriptome)
  - Oxford Nanopore GridION Sequencing (Long Read)
  - RIN ≥8, 260/280/230 ratios, ≥500 ng preferred; >1 ug for depletion methods (lose 90%)

# Genomic Experiment (Data Analysis)



# Files and Information Required for analysis: Microarray

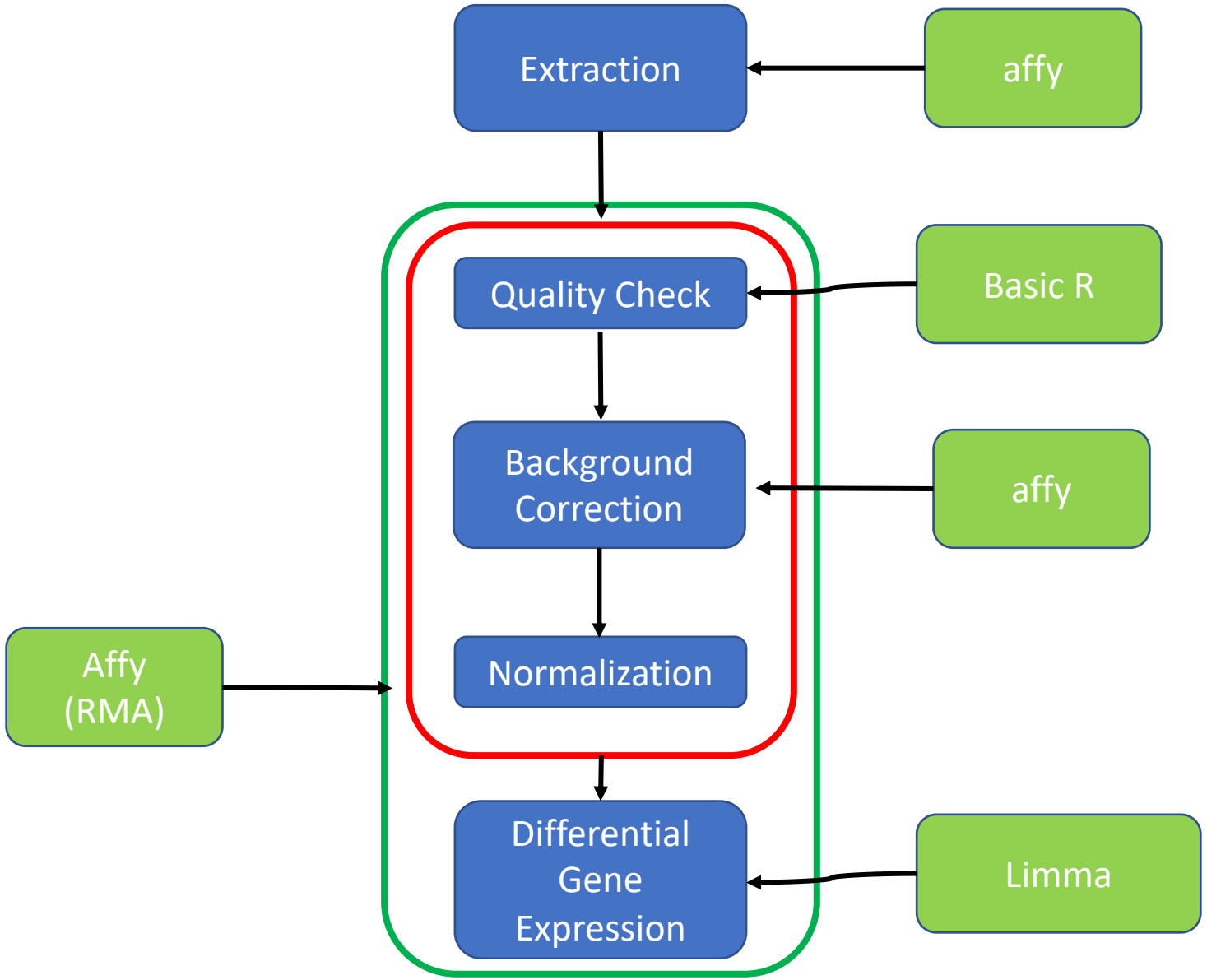
- **Design of the experiment**
  - **Number of samples used?**
  - **Does your question require biological or technical replicates?**
  - **What kind of probes needed to answer your question?**
- **Files and Information Required**
  - **Raw intensity files:** If done in-house, the .CEL files from Affymetrix 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/Group:** Condition associated to each sample. Example: Sample 1-4 are control, Sample 5-8 are experimental.

# Clariom S and D arrays

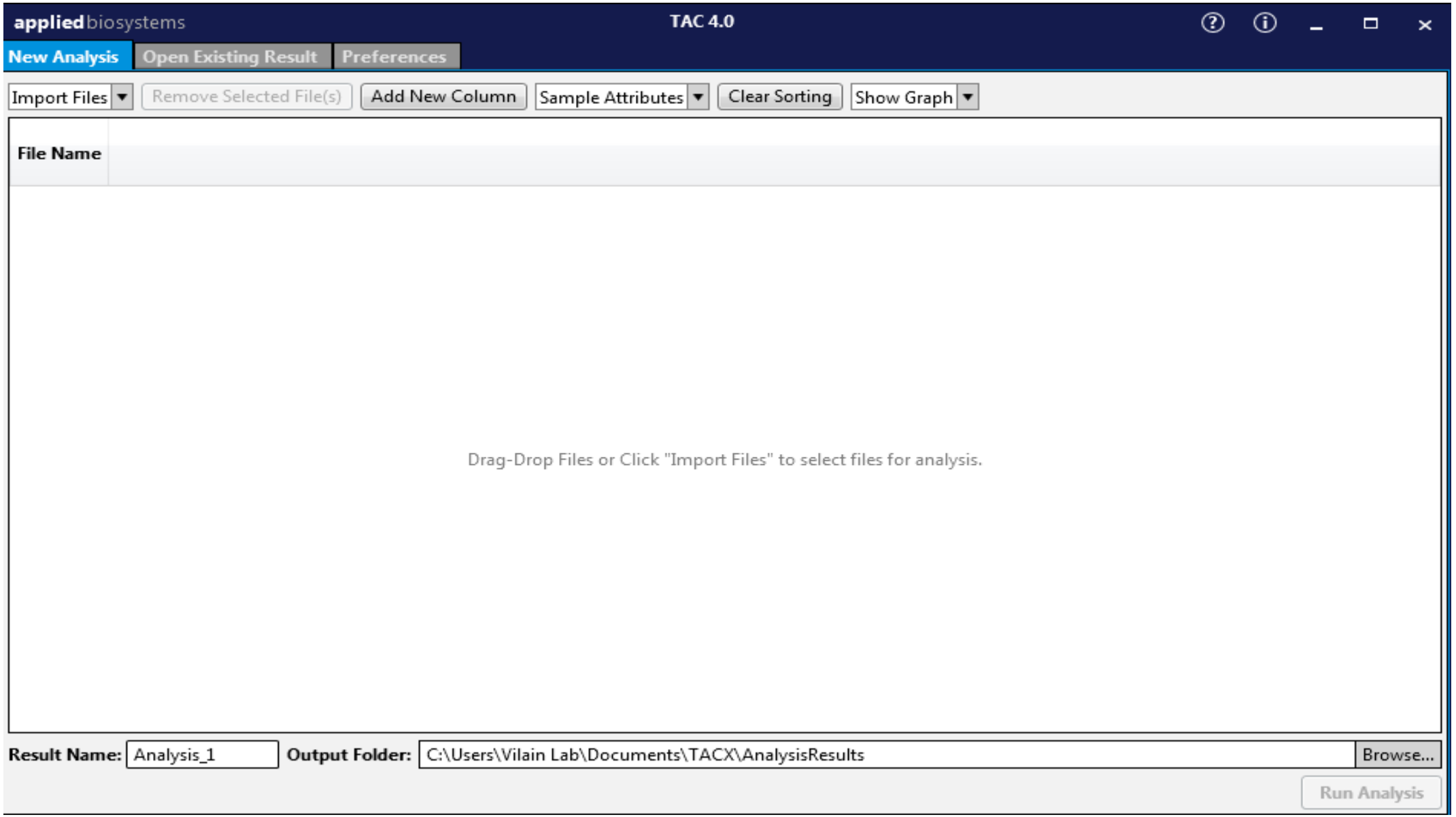
| | Clariom D Assay | Clariom D Pico Assay | Clariom S Assay | Clariom S Pico Assay |
|---|---|---|---|---|
| **Application(s)** | Deep and broad transcriptome analysis and biomarker discovery | | Gene-level expression profiling of well-annotated genes | |
| **Level of analysis** | Coding and noncoding genes, exons, and alternative splicing, including both well-annotated and speculative transcripts | | Well-annotated genes | |
| **FFPE tissue–compatible** | No | Yes | No | Yes |
| **RNA input minimum** | 50 ng | 0.1 ng (0.5 ng for FFPE) | 50 ng | 0.1 ng (0.5 ng for FFPE) |
| **Part of gene measured** | Whole transcript | | | |
| **Available format(s)** | Cartridge (single sample) | | Cartridge (single sample) Array plates (24 or 96 samples) | |
| **Available species** | Human, mouse, rat | | | |
| **Assay kit includes** | • Clariom D Array<br>• GeneChip WT PLUS Reagent Kit | • Clariom D Array<br>• GeneChip WT Pico Kit | • Clariom S Array<br>• GeneChip WT PLUS Reagent Kit | • Clariom S Array<br>• GeneChip Pico Kit |
| **Instrument (array format)** | GeneChip Scanner 3000 7G System (cartridge) | | GeneChip Scanner 3000 7G System (cartridge) GeneTitan Multi-Channel (MC) Instrument (plates) | |

# Microarray Analysis Pipeline

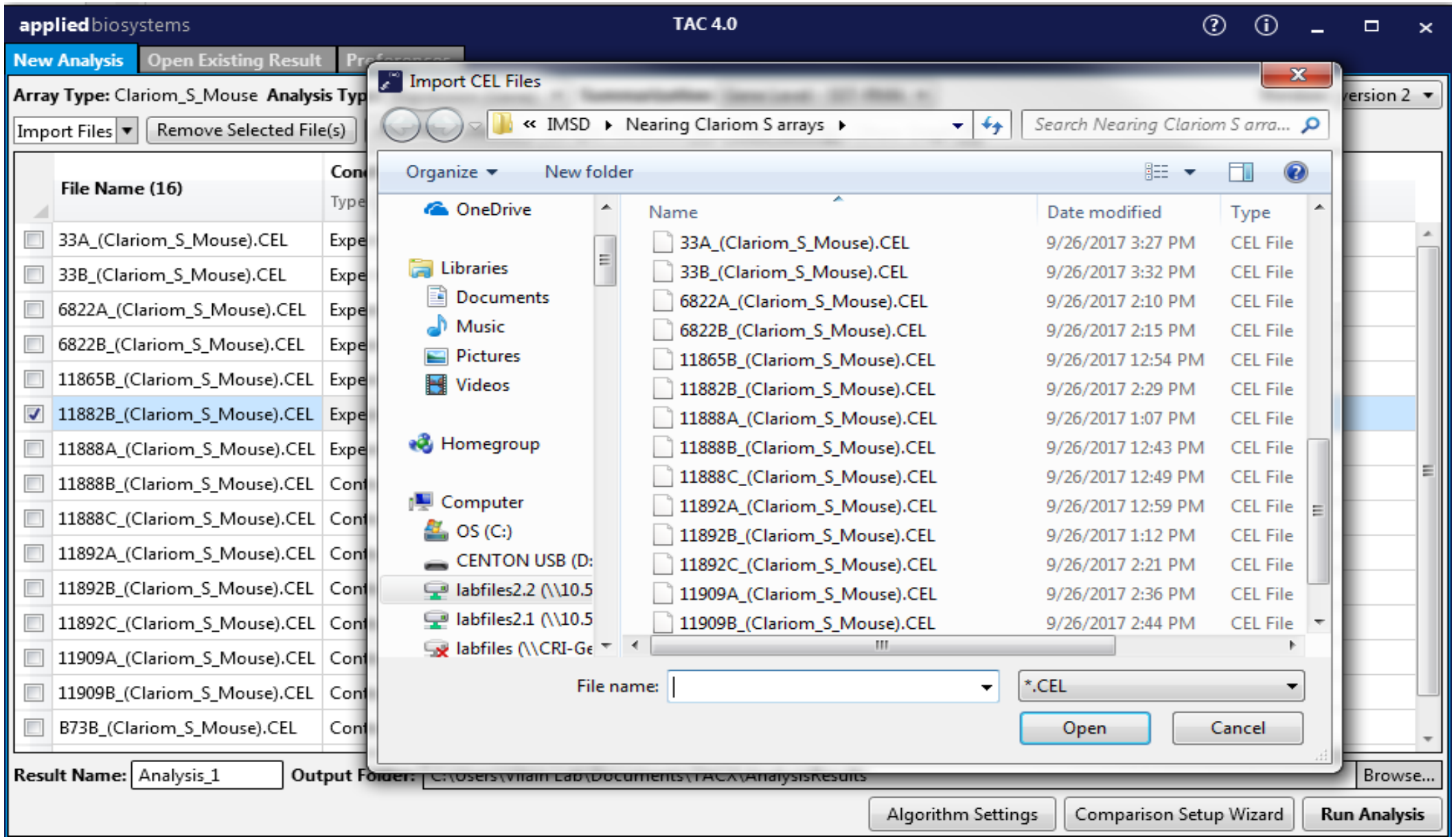
1. Extraction
2. Quality Check
3. Background Correction
4. Differential Gene Expression



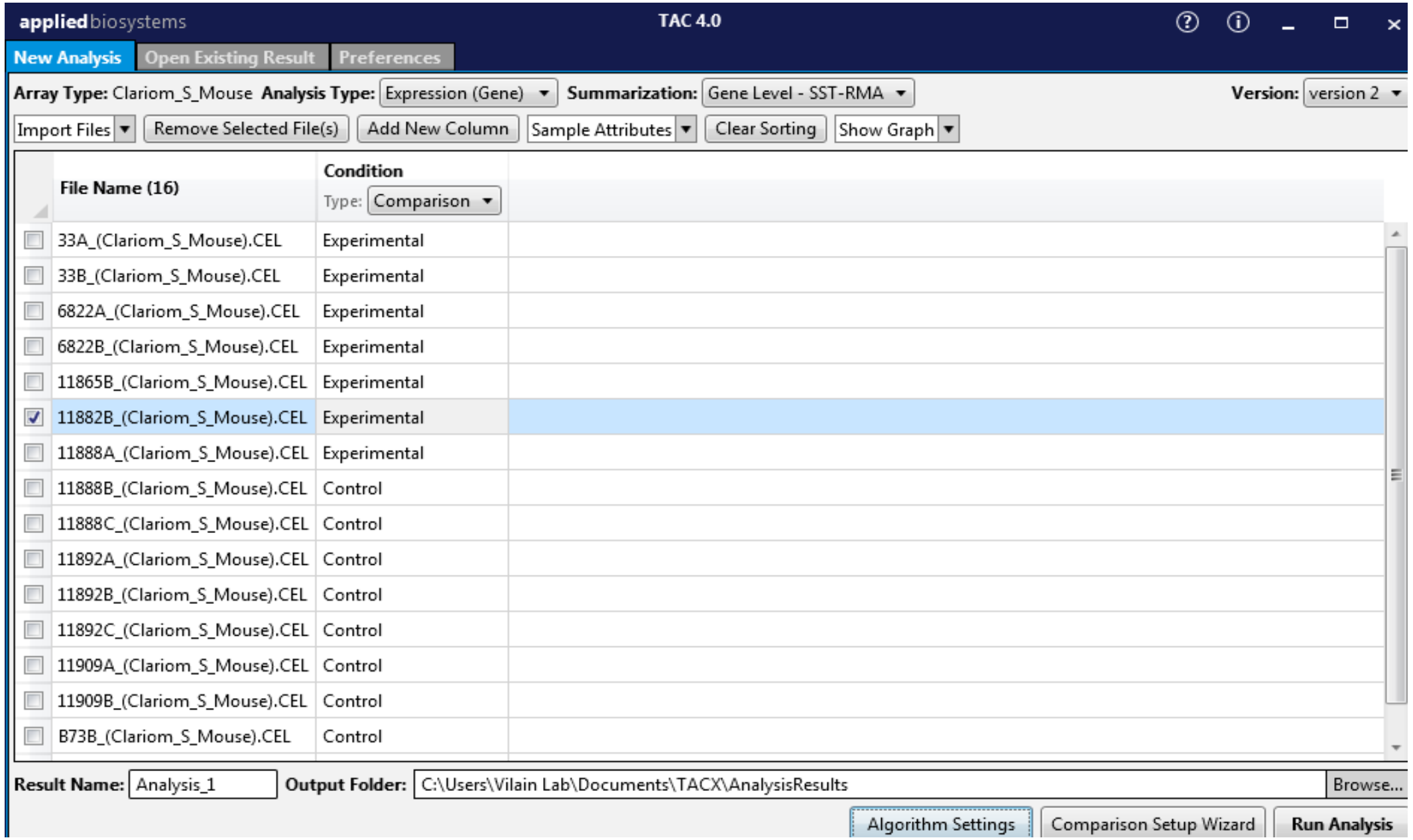
# Introduction to TACC



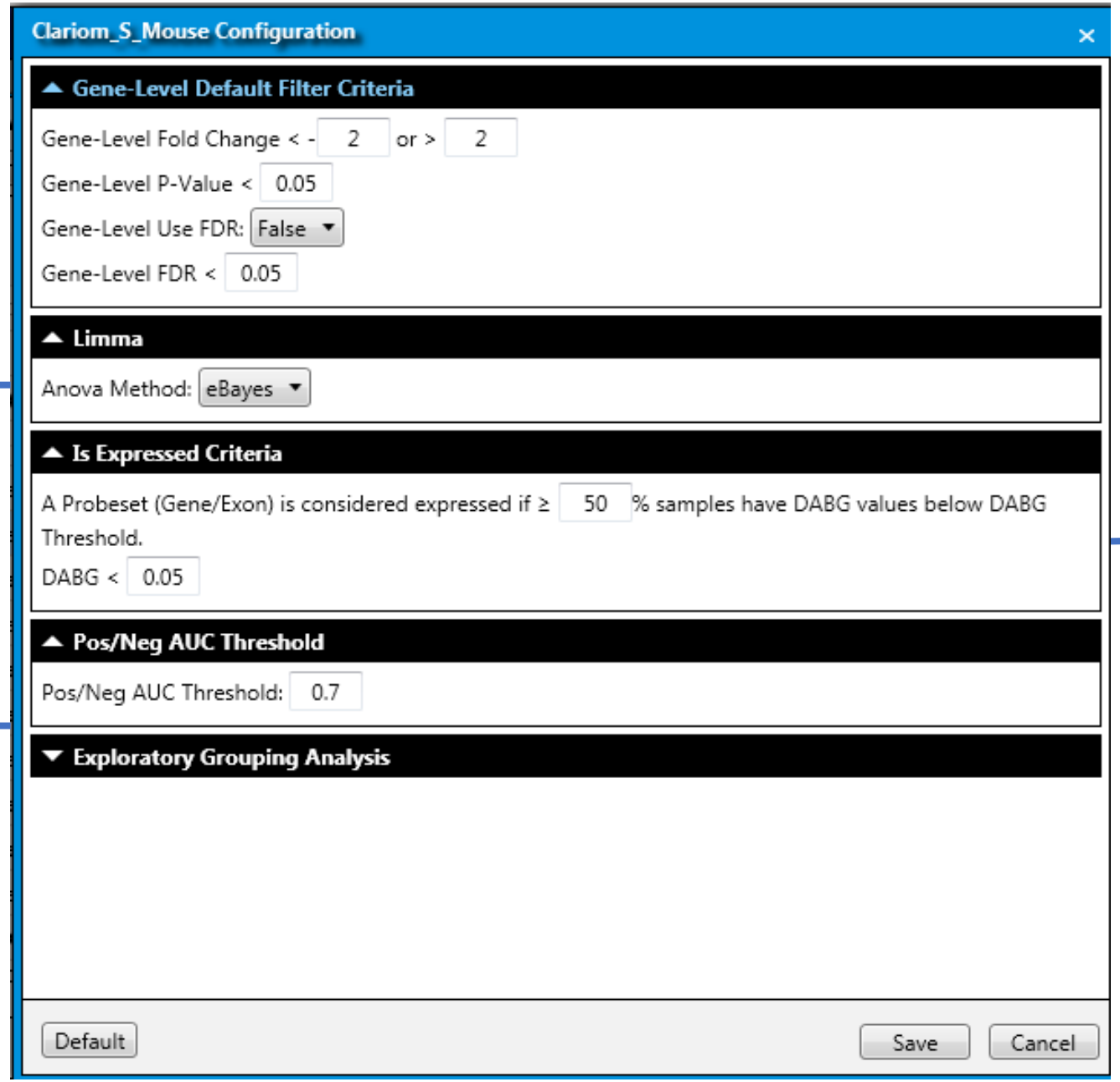
# Importing CEL Files



# Apply Condition



# Selecting Statistical Methods




Method chosen: eBayes or ANOVA

Area under curve: Determines separation ofexons from introns

Threshold to detect genes above background

# Files and Information Required for analysis: RNA-seq

- **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.

# Sample and Library Preparation

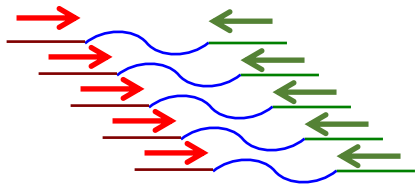
- ❑Single end vs Paired end
- ❑PolyA vs Ribodepletion

# Single end and Paired end

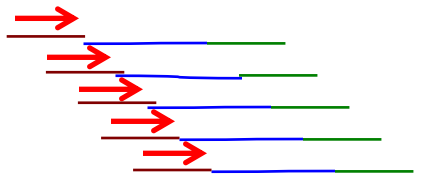
The sequencer instrument read from one end to the other end, and then start another round of reading from the opposite end.

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/eukaryotes |
| mRNA | mRNA along with non-coding RNA like lncRNA etc |
| 3 prime bias | - |

# 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 (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
  - ❖Type of sequencers (Long/Short)

# 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)

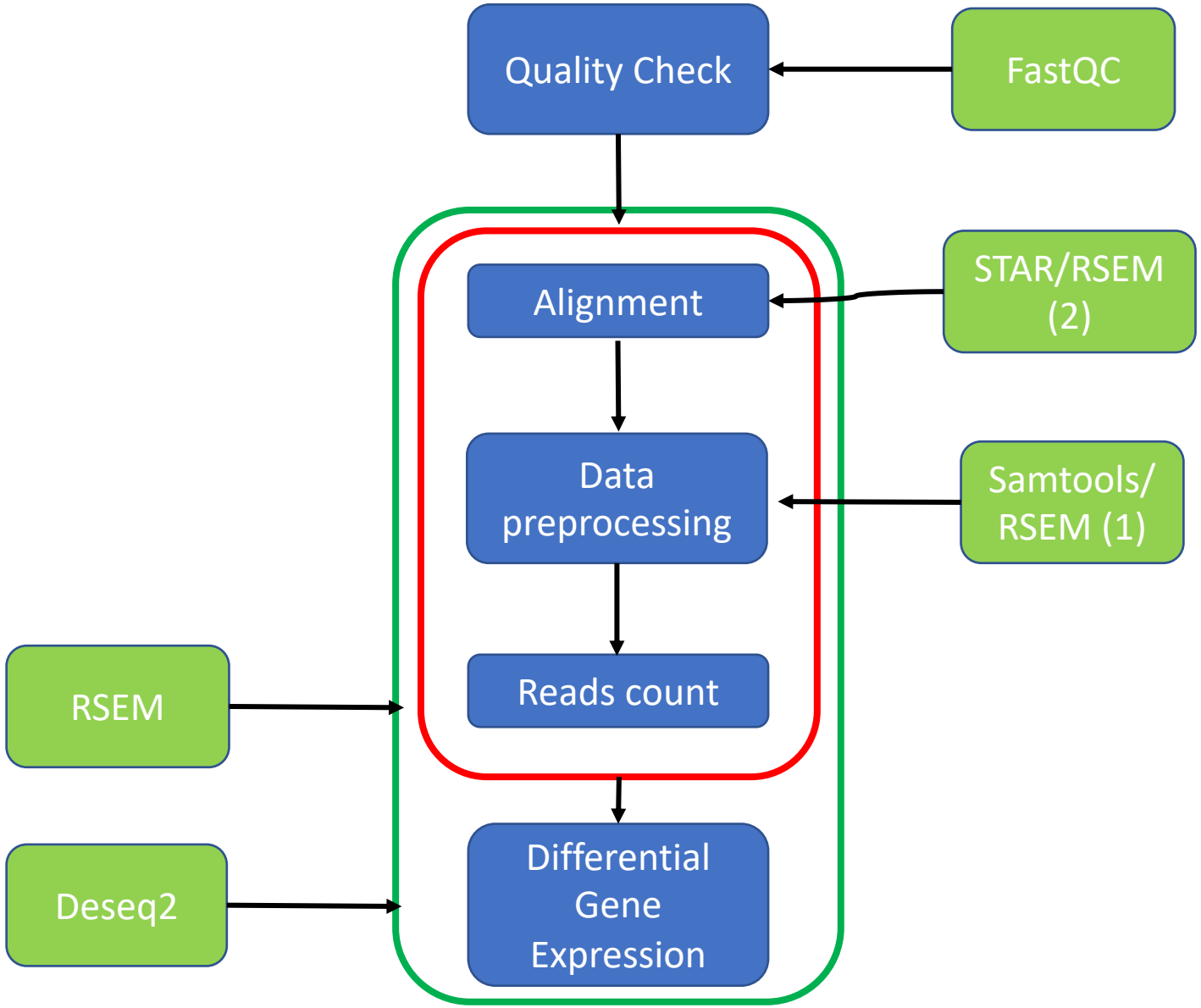
**Recommended RNA-Seq Parameters**

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 |

# 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
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAAGAATAC
```

# FASTQ Format

sequence identifier and an *optional* description

raw sequence letters

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

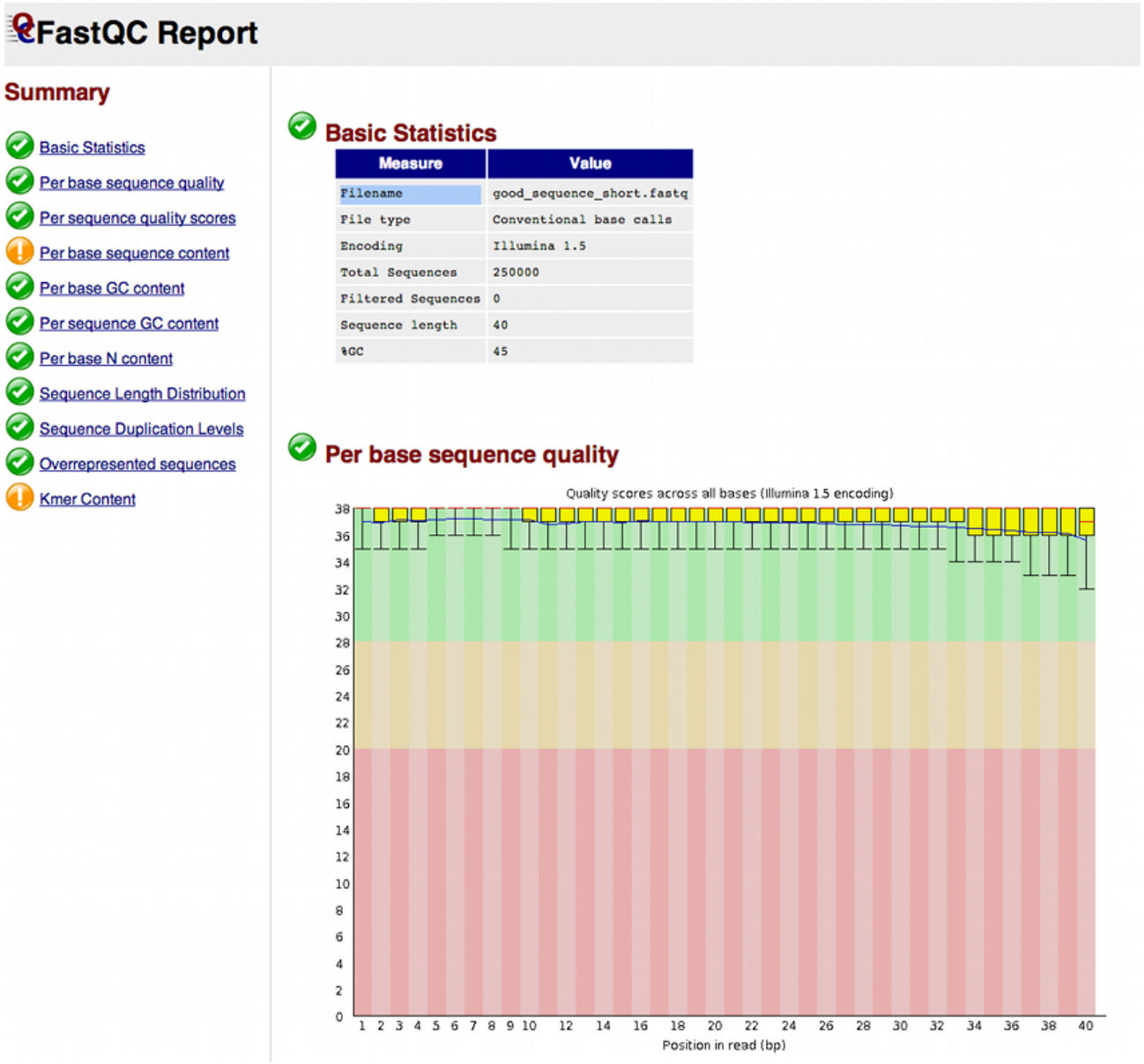
quality values for the sequence

# Quality Check

Tool: FastQC

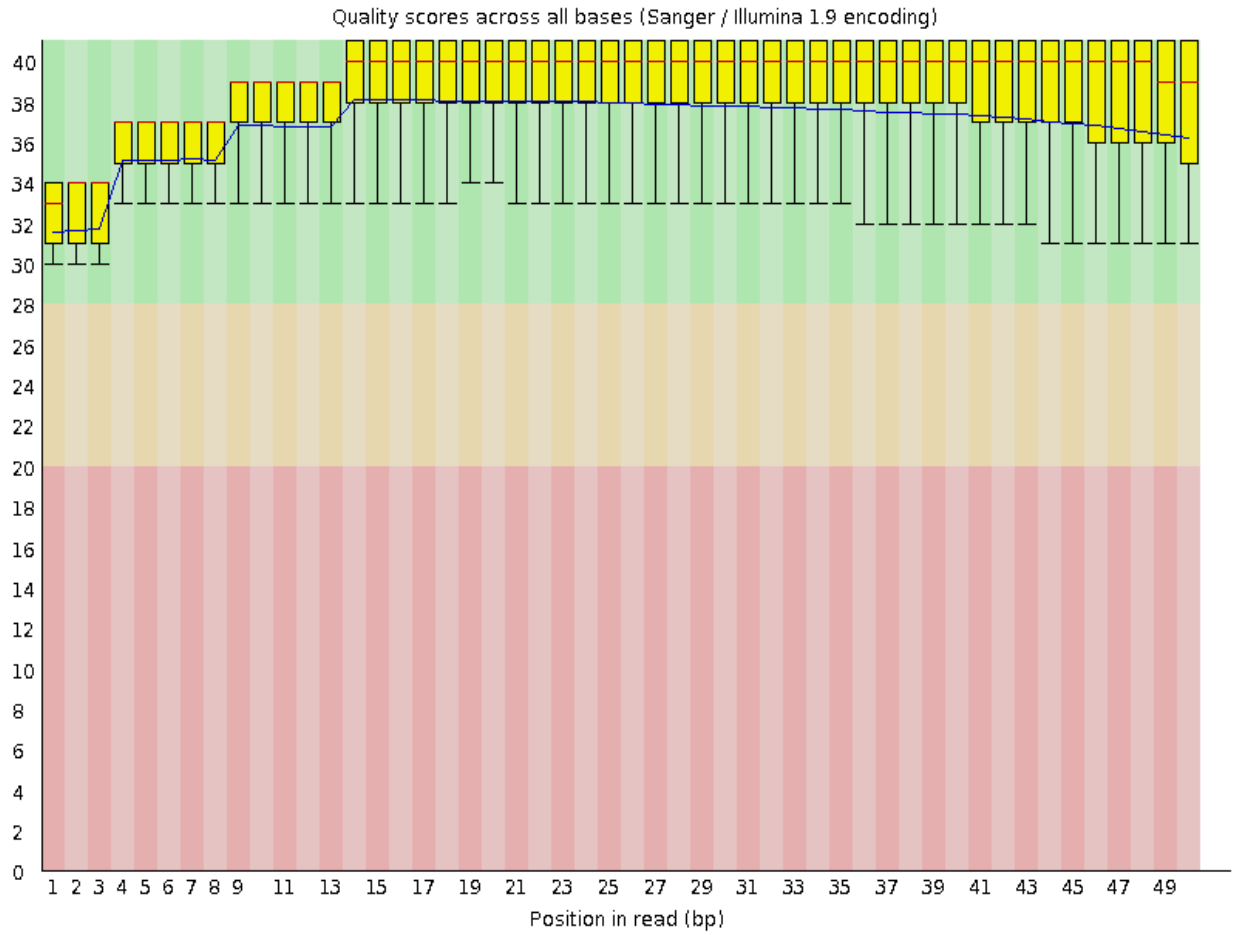
Input: FastQ files

Output: HTML file

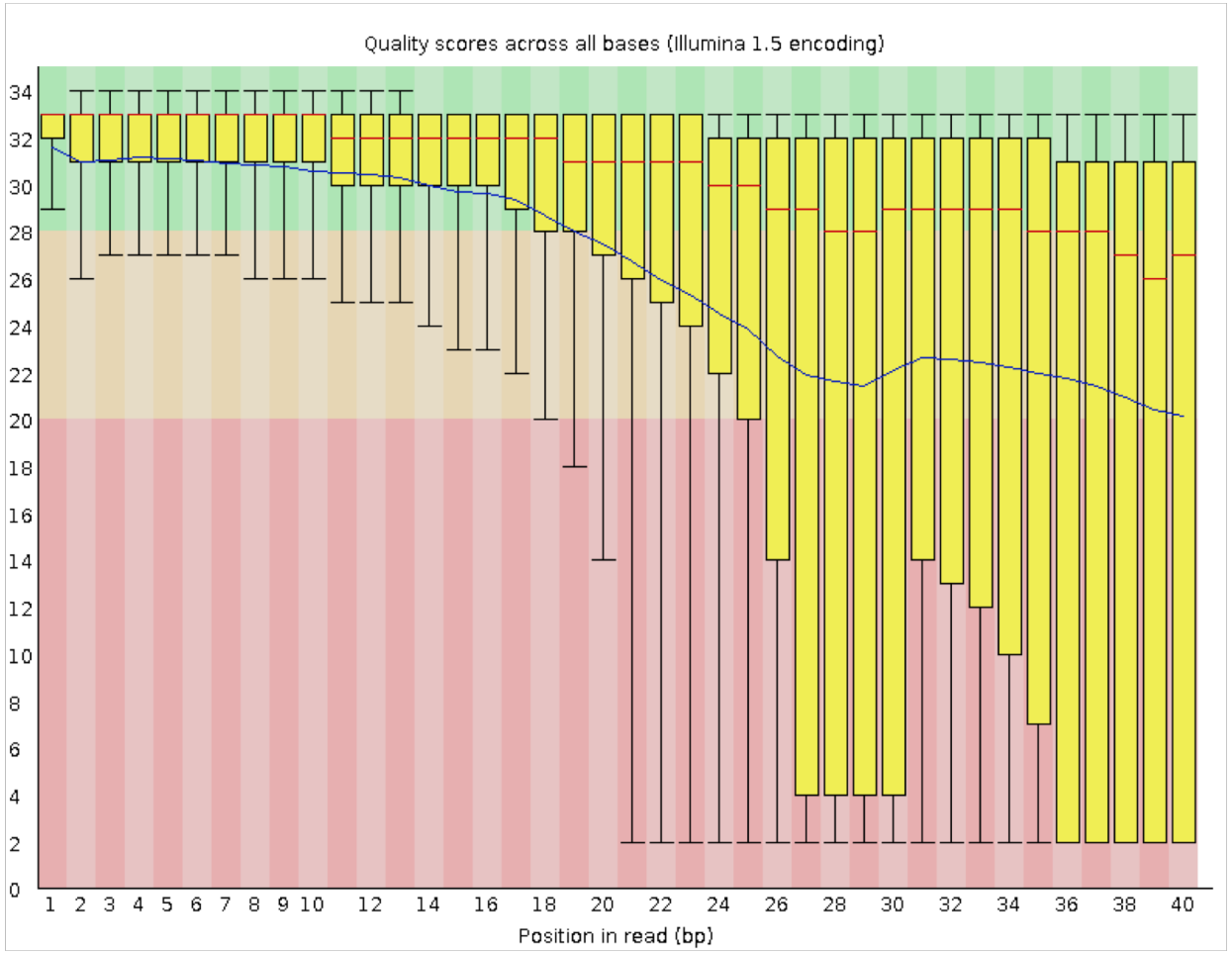


# FastQC : Good vs Bad

## Good Quality




## Bad Quality

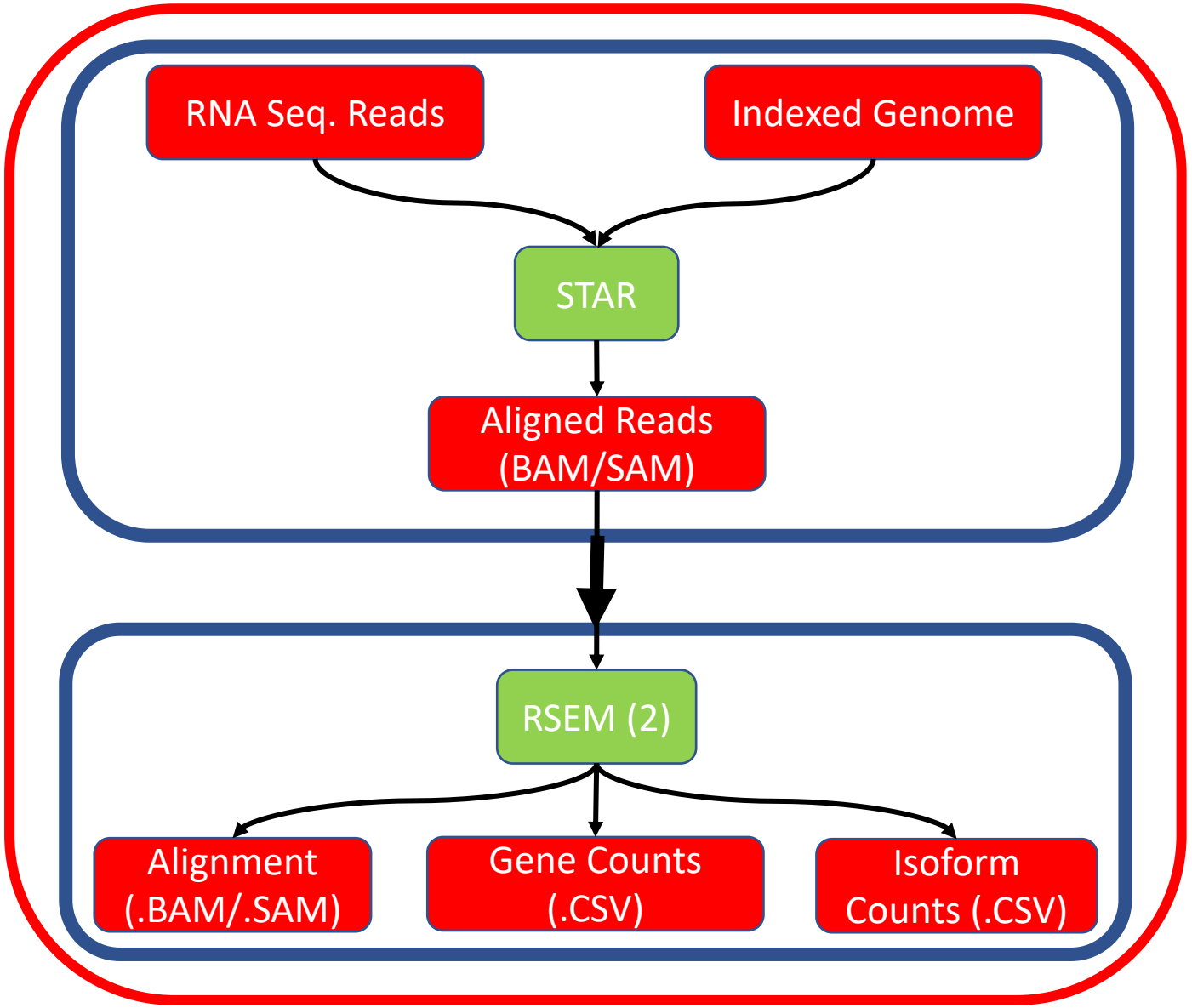


# Alignments and Reads count

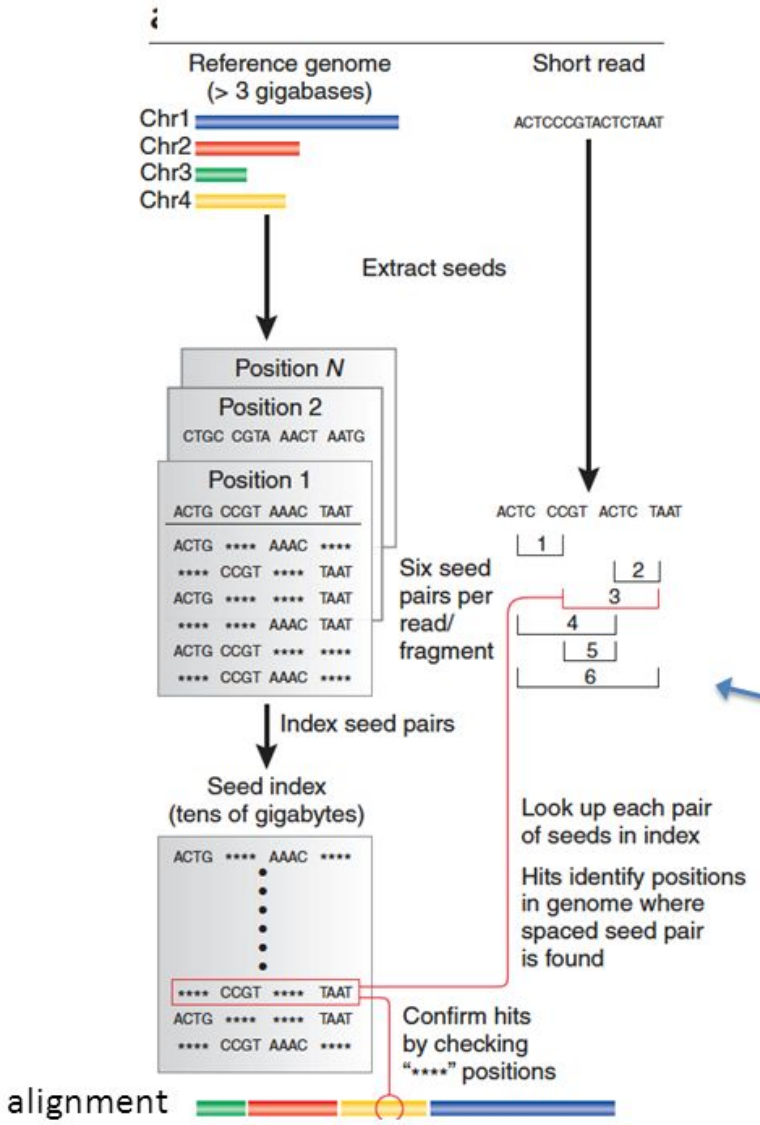
Tools : STAR and RSEM(2)

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

Output: .BAM/.SAM files

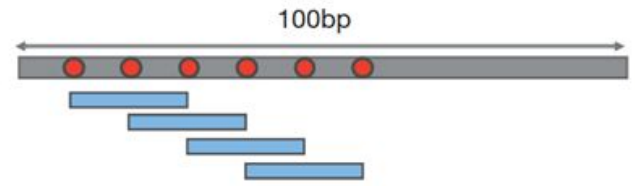


# Indexing genome




## Spaced-seed indexing of the reference genome

- Need to break up the genome into manageable segments
- Create index of short sequences
- Match seeds against genome index



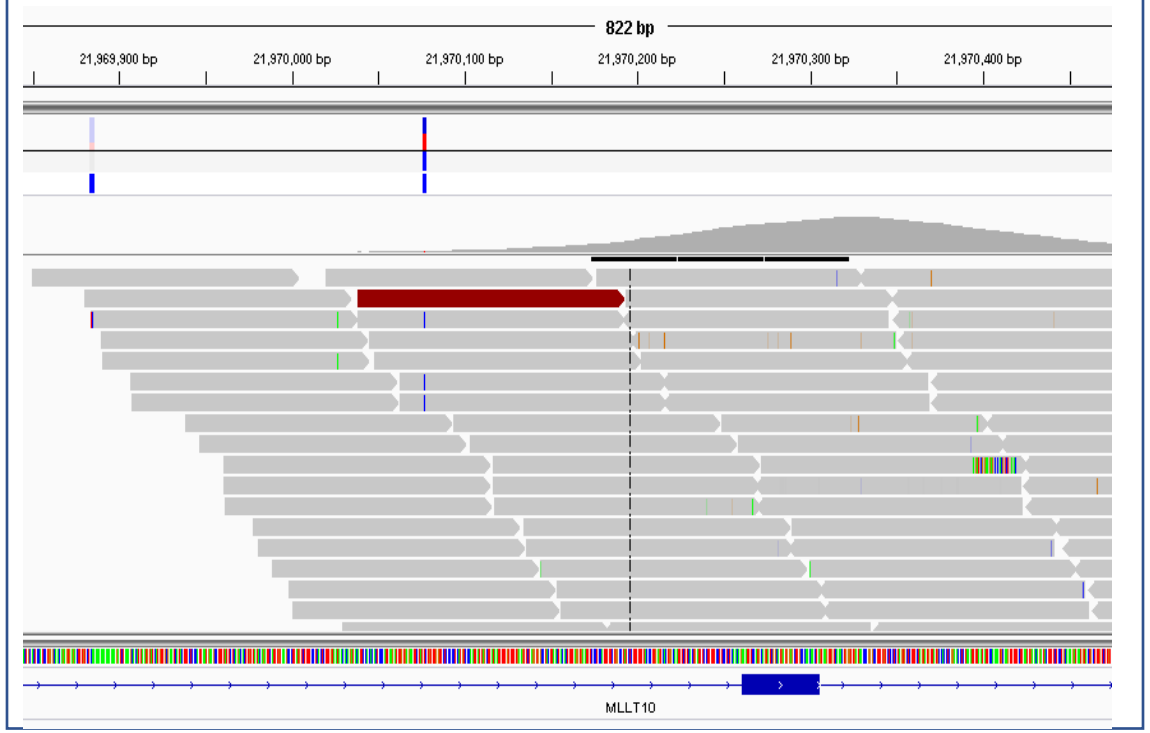

Trapnell and Salzberg, *Nat Biotechnology* 27:455 (2009)

# 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-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO: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
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO: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
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO: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       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

## View by IGV

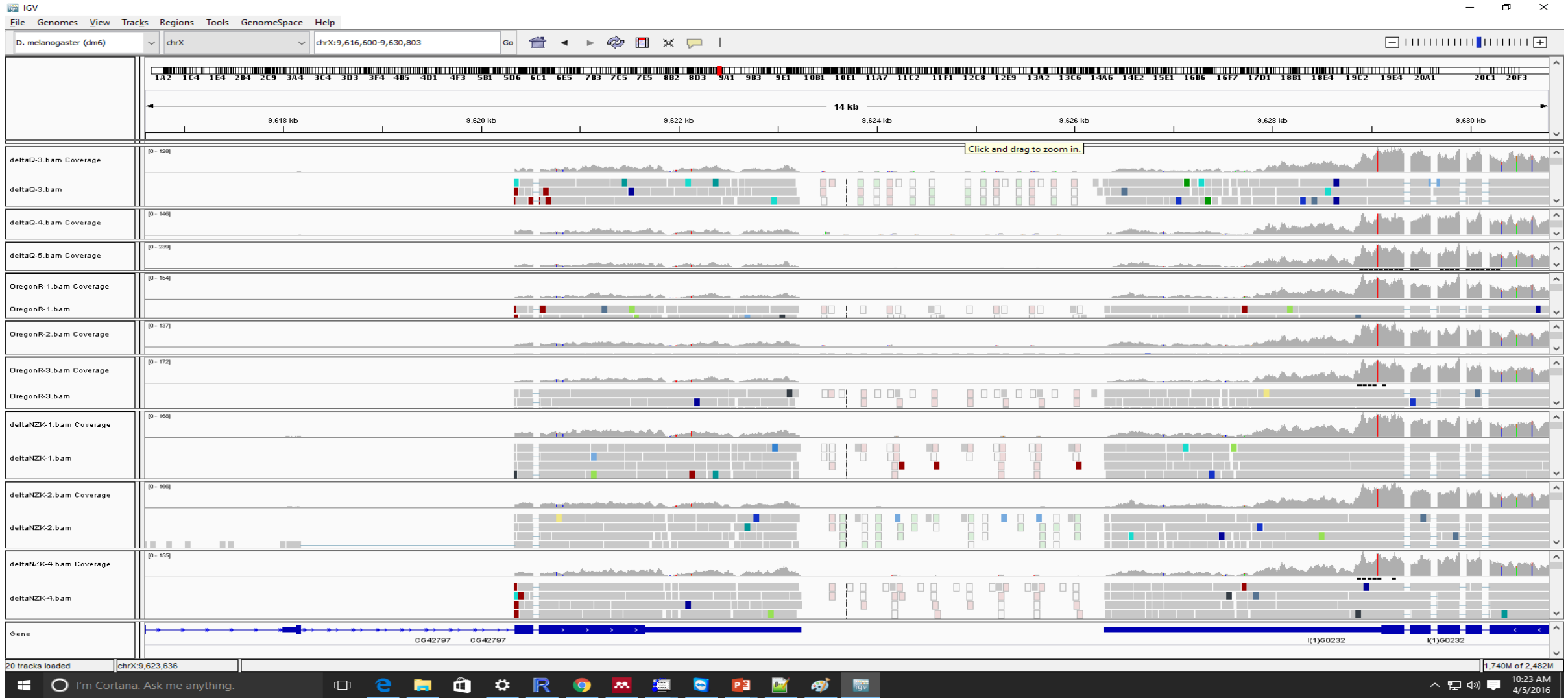


# Samtool Headers

## 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)

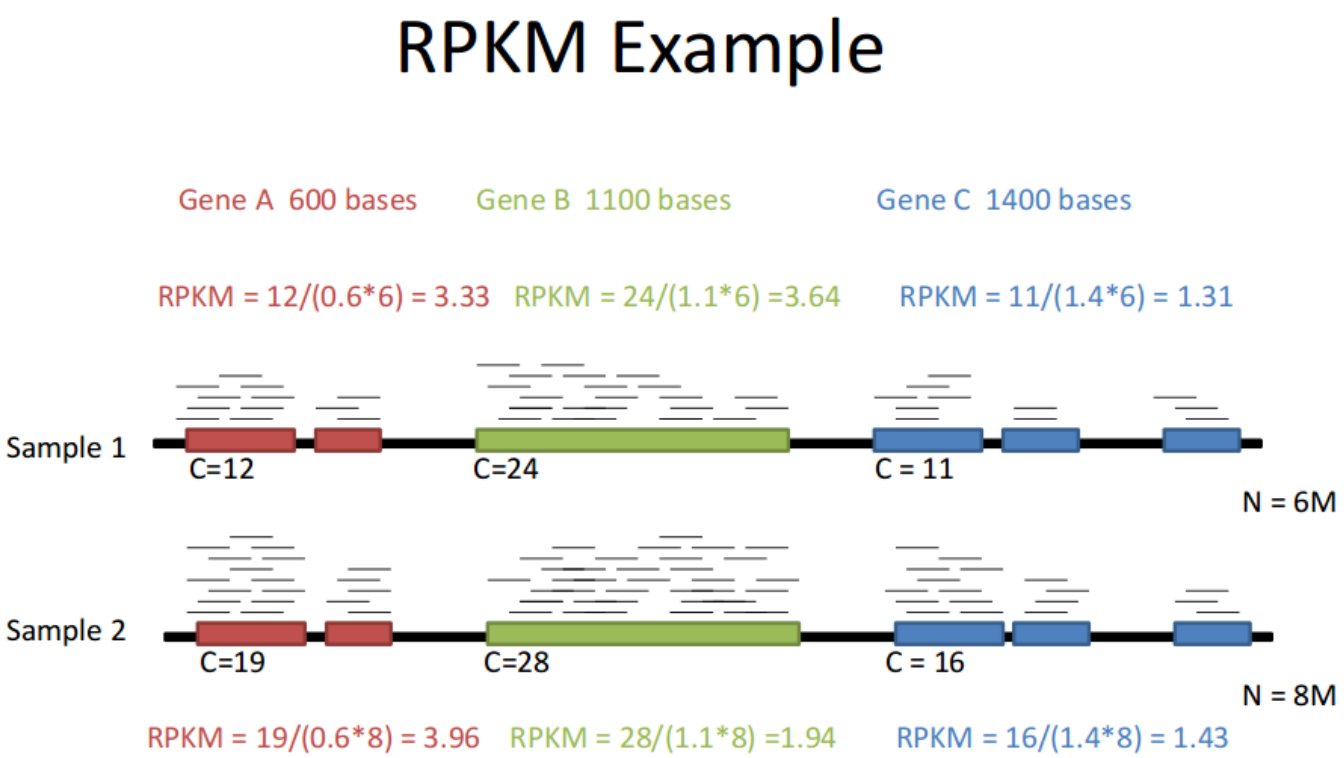


# Read Counts Normalization

- 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.

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

# Read Counts Normalization (contd..)




https://izabelcavassim.wordpress.com/2015/03/09/rpkm-and-fpkm-normalization-units-of-expression/

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

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

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

# Read Counts Normalization (contd..)

**Transcripts Per million**

- Divide the Number of reads of a transcript by the length of that gene in kilo bases (gene length divided by $10^3$). This is the Read per Kilobases (**RPK**).
- Summation of all RPK values in a sample divided by $10^6$, is the "per-million" scaling factor.
- Divide RPK of each gene by the per million scaling factor to get Transcripts per million (**TPM**).

$$\mathrm{TPM}_i = \frac{X_i}{\widetilde{l}_i} \cdot \left( \frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j}} \right) \cdot 10^6$$

$X_i$ : Number of reads of transcript for Gene i

$l_i$ : Length of Gene i

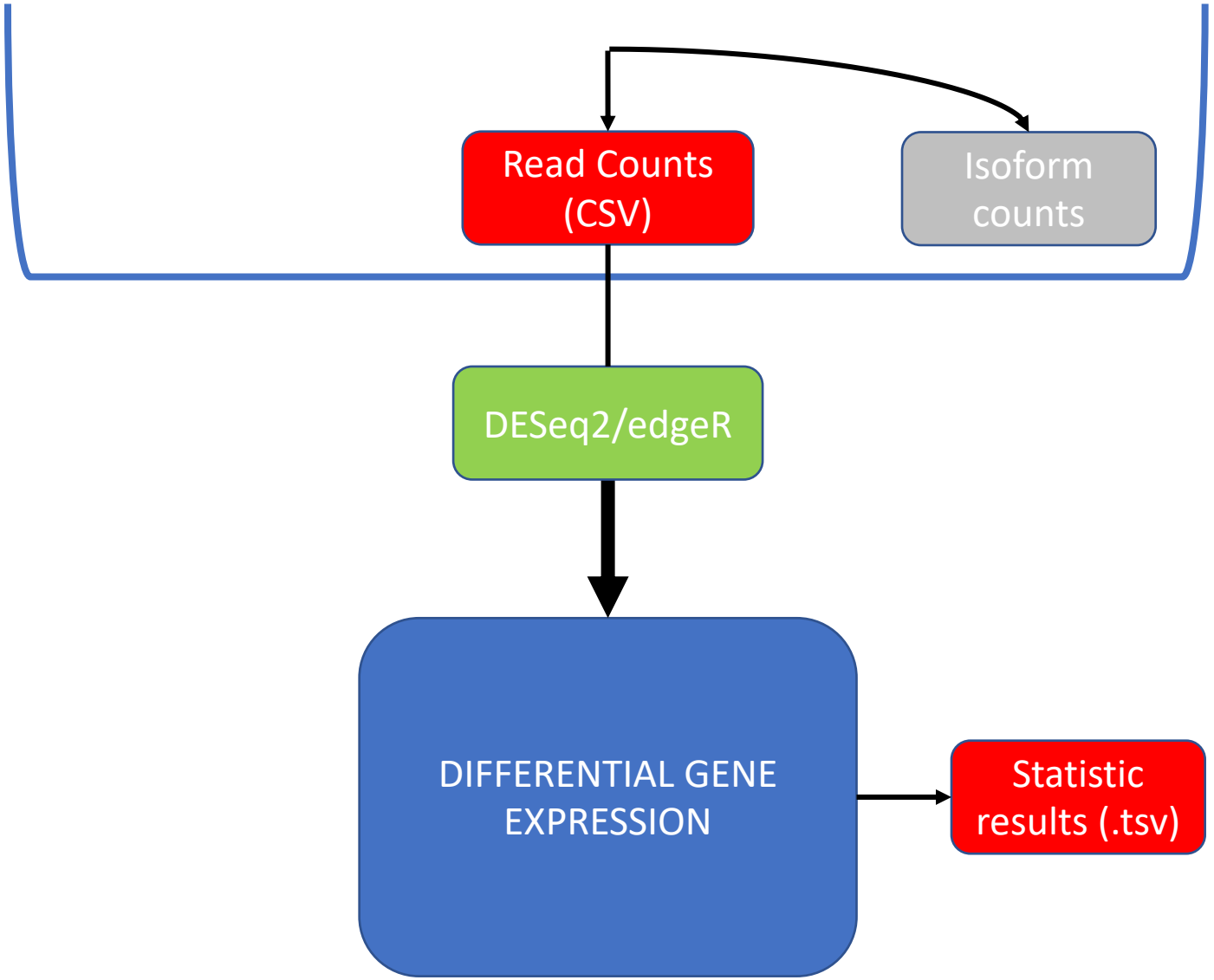
- Due to the normalization technique used TPM values are less variables between samples of the same condition.

# Differential Gene Expression

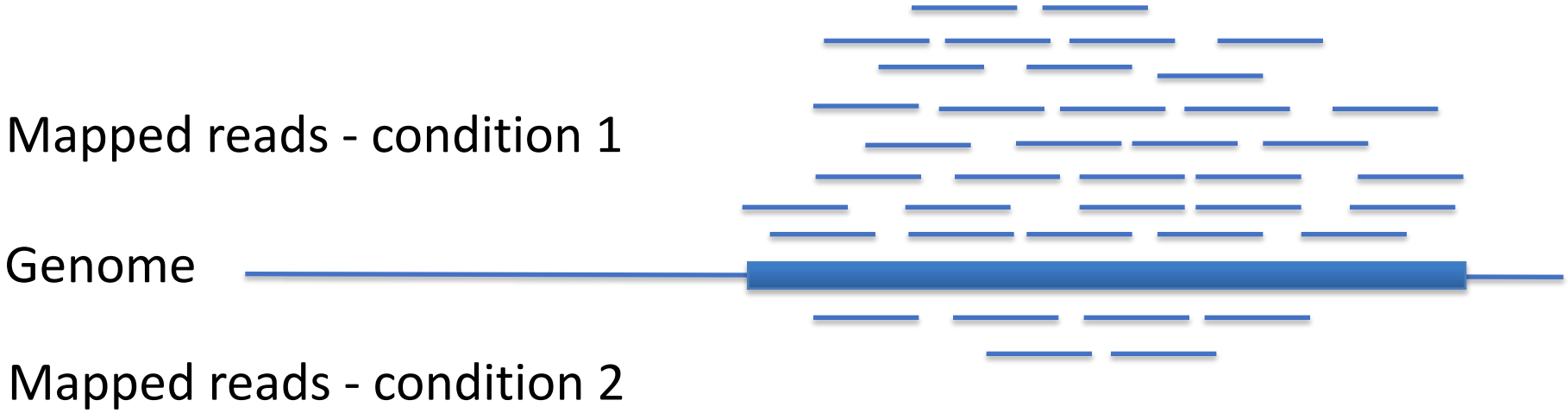
Tool: DESeq2, edgeR

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 |

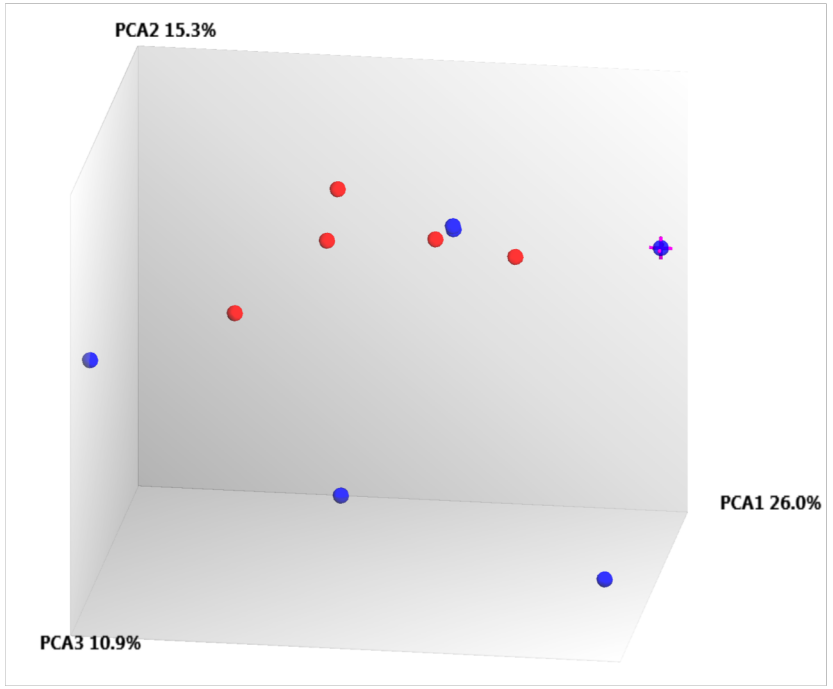
# More Counts = more Statistical significance

Example: 5000 total reads per sample.

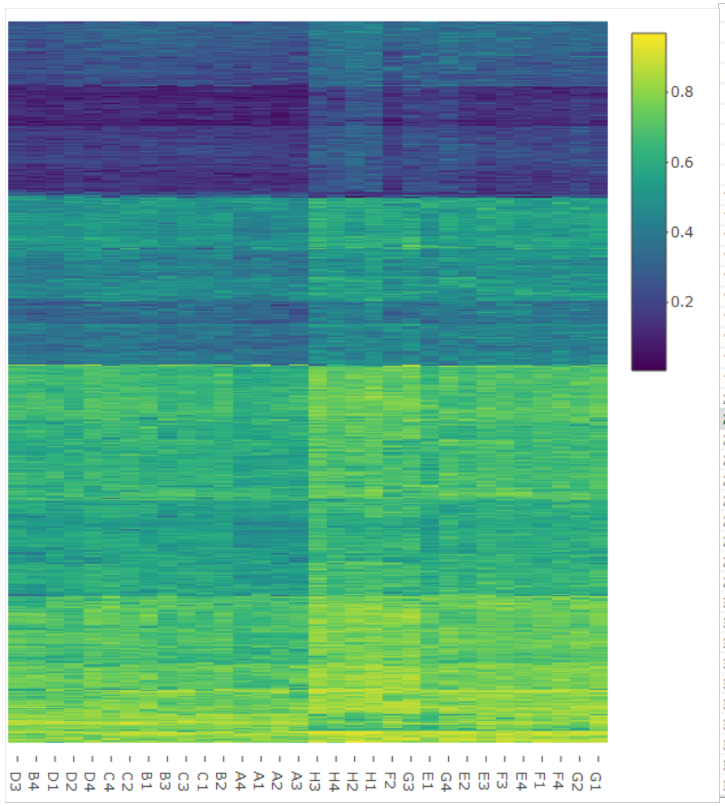
Observed 2-fold differences in read counts.

| | 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




PCA Plot



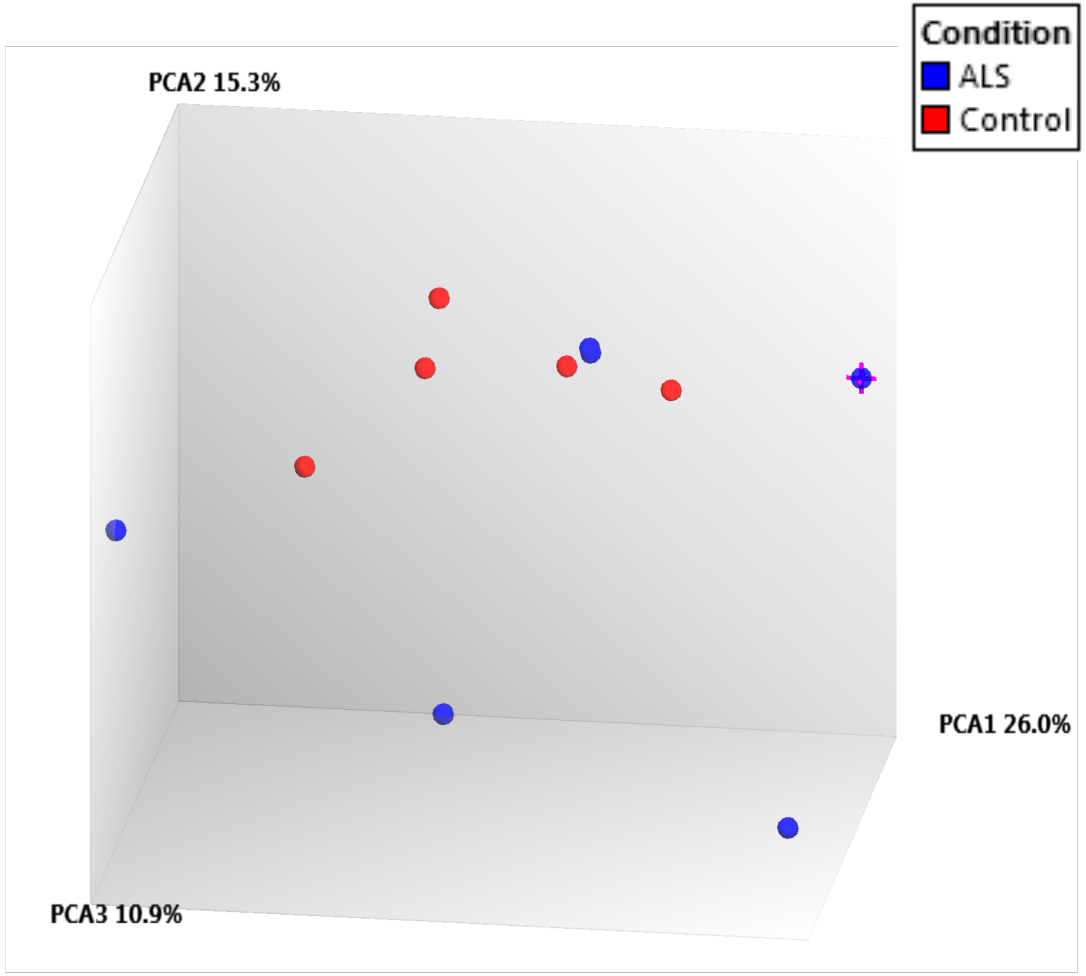

Heatmap

| | test_id | gene_id | gene | locus | sample_1 | sample_2 | status | value_1 | value_2 | log2(fold | test_stat | p_value | q_value | significa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | XLOC_000 | XLOC_000 | CG33281 | 2L:332423 | deltaQ | OregonR | OK | 7.78404 | 28.3487 | 1.86469 | 3.89904 | 5.00E-05 | 0.00347601 | yes |
| 3 | XLOC_000 | XLOC_000 | CG33282 | 2L:332790 | deltaQ | OregonR | OK | 1.46714 | 5.18562 | 1.82151 | 3.08678 | 5.00E-05 | 0.00347601 | yes |
| 4 | XLOC_000 | XLOC_000 | CG2772 | 2L:344599 | deltaQ | OregonR | OK | 15.7892 | 33.2996 | 1.07657 | 2.46754 | 0.0001 | 0.00623576 | yes |
| 5 | XLOC_000 | XLOC_000 | pgant4 | 2L:347260 | deltaQ | OregonR | OK | 3.81965 | 1.71272 | -1.15715 | -2.10786 | 0.00045 | 0.022476 | yes |
| 6 | XLOC_000 | XLOC_000 | CG3355 | 2L:465140 | deltaQ | OregonR | OK | 15.2984 | 2.46559 | -2.63338 | -3.80939 | 5.00E-05 | 0.00347601 | yes |
| 7 | XLOC_000 | XLOC_000 | Sgs1 | 2L:493718 | deltaQ | OregonR | OK | 48.9408 | 25.6168 | -0.93395 | -1.85136 | 0.00095 | 0.0397339 | yes |
| 8 | XLOC_000 | XLOC_000 | Gpdh | 2L:594364 | deltaQ | OregonR | OK | 89.1781 | 179.689 | 1.01074 | 2.93463 | 5.00E-05 | 0.00347601 | yes |
| 9 | XLOC_000 | XLOC_000 | Muc26B | 2L:615196 | deltaQ | OregonR | OK | 49.2003 | 20.3772 | -1.27171 | -2.76563 | 5.00E-05 | 0.00347601 | yes |
| 10 | XLOC_000 | XLOC_000 | CG15818 | 2L:741090 | deltaQ | OregonR | OK | 31.8493 | 17.9784 | -0.825 | -1.95182 | 0.00105 | 0.0427014 | yes |
| 11 | 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 |
| 12 | XLOC_000 | XLOC_000 | Aldh | 2L:937030 | deltaQ | OregonR | OK | 199.624 | 328.681 | 0.719399 | 2.16988 | 0.00015 | 0.00871949 | yes |
| 13 | XLOC_000 | XLOC_000 | CG3841 | 2L:958947 | deltaQ | OregonR | OK | 10.0911 | 24.0033 | 1.25016 | 2.94501 | 5.00E-05 | 0.00347601 | yes |
| 14 | XLOC_000 | XLOC_000 | Apoltp | 2L:963417 | deltaQ | OregonR | OK | 10.3169 | 6.35646 | -0.69872 | -2.08466 | 5.00E-05 | 0.00347601 | yes |
| 15 | XLOC_000 | XLOC_000 | CG33301 | 2L:100495 | deltaQ | OregonR | OK | 24.7792 | 99.4039 | 2.00417 | 5.07912 | 5.00E-05 | 0.00347601 | yes |
| 16 | XLOC_000 | XLOC_000 | CR44874 | 2L:100513 | deltaQ | OregonR | OK | 0.43958 | 1.70197 | 1.953 | 2.27683 | 0.00065 | 0.0297238 | yes |
| 17 | XLOC_000 | XLOC_000 | Cpr31A | 2L:100542 | deltaQ | OregonR | OK | 6.31093 | 17.4553 | 1.46774 | 2.94322 | 5.00E-05 | 0.00347601 | yes |
| 18 | XLOC_000 | XLOC_000 | CG17134 | 2L:108042 | deltaQ | OregonR | OK | 36.6715 | 75.0104 | 1.03243 | 2.59768 | 5.00E-05 | 0.00347601 | yes |
| 19 | XLOC_000 | XLOC_000 | NimC2 | 2L:139801 | deltaQ | OregonR | OK | 42.4803 | 17.2581 | -1.29953 | -2.91512 | 0.00025 | 0.0133971 | yes |
| 20 | XLOC_000 | XLOC_000 | Adh,Adhr | 2L:145991 | deltaQ | OregonR | OK | 564.877 | 965.215 | 0.772912 | 2.16993 | 0.0005 | 0.0244976 | yes |
| 21 | XLOC_001 | XLOC_001 | CG15155,CG17681,CG5783 | 2L:181529 | deltaQ | OregonR | OK | 50.2681 | 124.125 | 1.30408 | 2.65395 | 5.00E-05 | 0.00347601 | yes |
| 22 | XLOC_001 | XLOC_001 | CG17325 | 2L:188125 | deltaQ | OregonR | OK | 134.215 | 227.034 | 0.758362 | 2.13804 | 0.0004 | 0.0200761 | yes |
| 23 | XLOC_001 | XLOC_001 | CG9259 | 2L:210893 | deltaQ | OregonR | OK | 25.9753 | 5.20173 | -2.32008 | -4.57379 | 5.00E-05 | 0.00347601 | yes |
| 24 | 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 |
| 25 | XLOC_001 | XLOC_001 | tim | 2L:349397 | deltaQ | OregonR | OK | 2.23157 | 4.79765 | 1.10427 | 2.39393 | 0.0001 | 0.00623576 | yes |
| 26 | XLOC_001 | XLOC_001 | CG16712 | 2L:369621 | deltaQ | OregonR | OK | 108.423 | 200.422 | 0.886362 | 2.01938 | 0.0006 | 0.0284488 | yes |
| 27 | XLOC_001 | XLOC_001 | Sgs1 | 2L:493718 | deltaQ | OregonR | OK | 125.233 | 58.325 | -1.10242 | -2.06432 | 0.0008 | 0.035327 | yes |
| 28 | XLOC_001 | XLOC_001 | CG9498 | 2L:634939 | deltaQ | OregonR | OK | 11.4523 | 26.0854 | 1.18761 | 2.6796 | 5.00E-05 | 0.00347601 | yes |
| 29 | XLOC_001 | XLOC_001 | CG7025 | 2L:769110 | deltaQ | OregonR | OK | 27.8459 | 48.1527 | 0.79015 | 2.01797 | 0.0006 | 0.0284488 | yes |
| 30 | XLOC_001 | XLOC_001 | CG7214 | 2L:774368 | deltaQ | OregonR | OK | 13.5935 | 26.837 | 0.981304 | 2.09078 | 0.0005 | 0.0244976 | yes |
| 31 | XLOC_001 | XLOC_001 | CG7203 | 2L:775216 | deltaQ | OregonR | OK | 21.7555 | 52.889 | 1.28159 | 2.76209 | 5.00E-05 | 0.00347601 | yes |
| 32 | XLOC_002 | XLOC_002 | CG9463,CG9465 | 2L:876526 | deltaQ | OregonR | OK | 31.923 | 18.3205 | -0.80114 | -1.91992 | 0.00125 | 0.0489021 | yes |
| 33 | XLOC_002 | XLOC_002 | CG4382 | 2L:959887 | deltaQ | OregonR | OK | 4.6955 | 1.29471 | -1.85865 | -2.67568 | 5.00E-05 | 0.00347601 | yes |
| 34 | XLOC_002 | XLOC_002 | CG7299 | 2L:106834 | deltaQ | OregonR | OK | 83.4625 | 211.766 | 1.34327 | 3.95222 | 5.00E-05 | 0.00347601 | yes |
| 35 | XLOC_002 | XLOC_002 | CG17108 | 2L:106925 | deltaQ | OregonR | OK | 24.7576 | 43.8733 | 0.825473 | 2.04522 | 0.00055 | 0.0266932 | yes |
| 36 | XLOC_002 | XLOC_002 | CG31869 | 2L:107720 | deltaQ | OregonR | OK | 11.1676 | 5.46428 | -1.03121 | -2.32458 | 0.0001 | 0.00623576 | yes |
| 37 | XLOC_002 | XLOC_002 | AstC | 2L:110761 | deltaQ | OregonR | OK | 3.61998 | 12.8274 | 1.82517 | 3.39604 | 5.00E-05 | 0.00347601 | yes |
| 38 | XLOC_002 | XLOC_002 | CG16957 | 2L:133953 | deltaQ | OregonR | OK | 8.18801 | 2.37806 | -1.78373 | -2.85575 | 5.00E-05 | 0.00347601 | yes |
| 39 | XLOC_002 | XLOC_002 | CG33306,CG8997 | 2L:138279 | deltaQ | OregonR | OK | 1632.48 | 474.126 | -1.78372 | -2.60345 | 5.00E-05 | 0.00347601 | yes |

Differential Expression Table

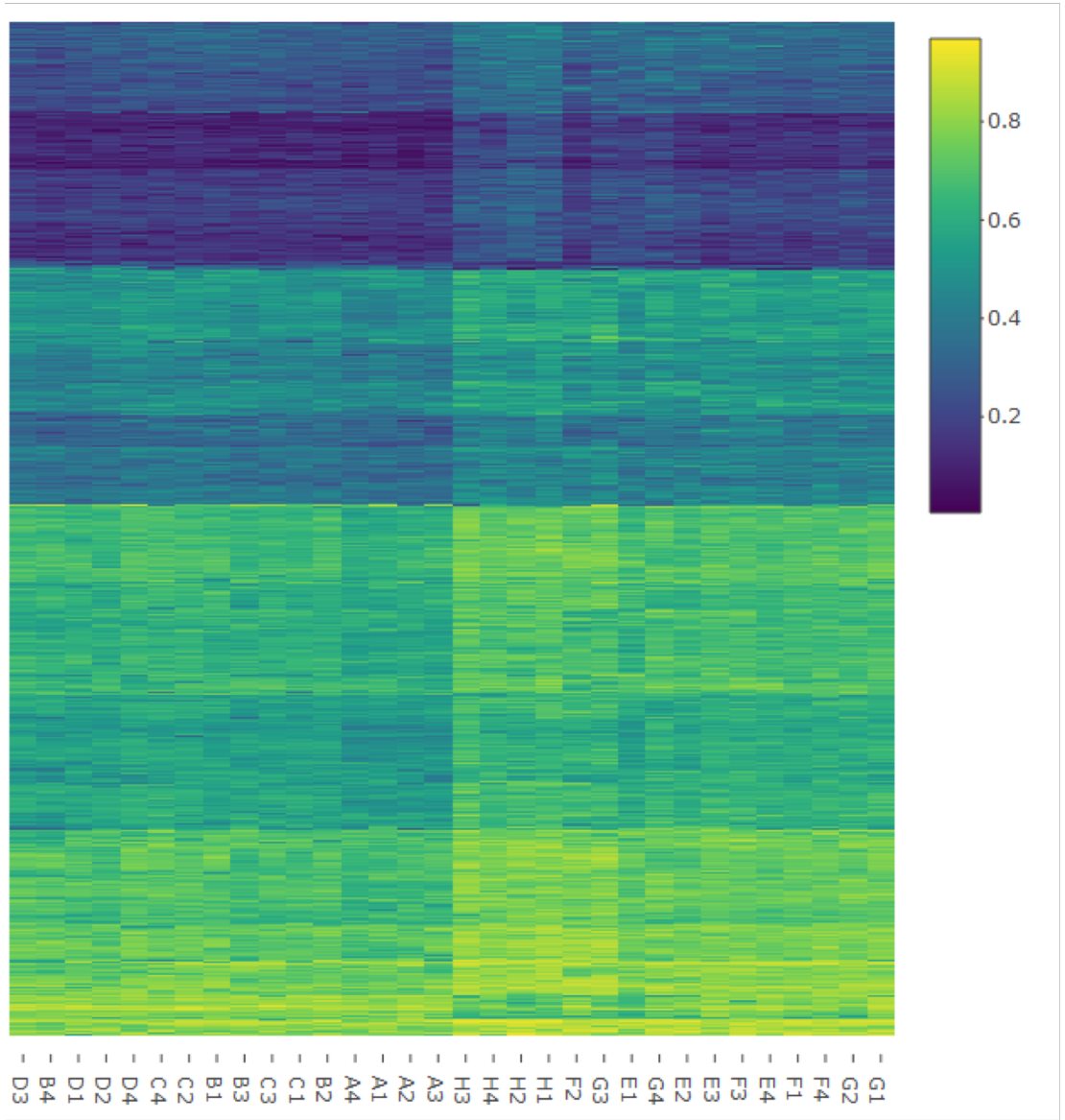
# Principle component Analysis (PCA) Plot

- 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.

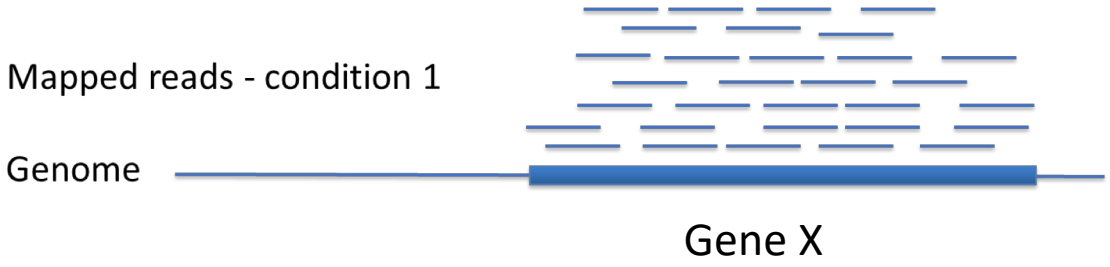


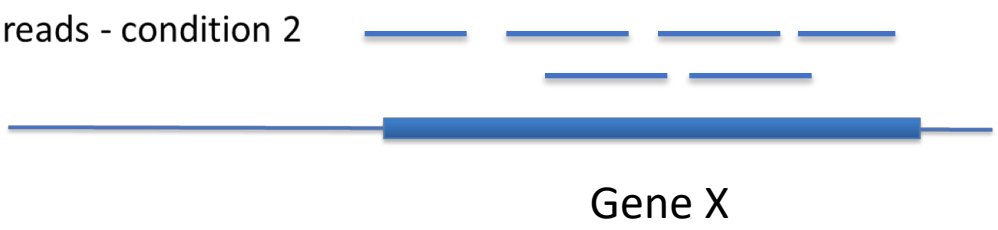
# Heatmaps

- 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 (log or z-score values.
- Hierarchical clustering done, to group genes with similar expression.



# Differential Expression tables

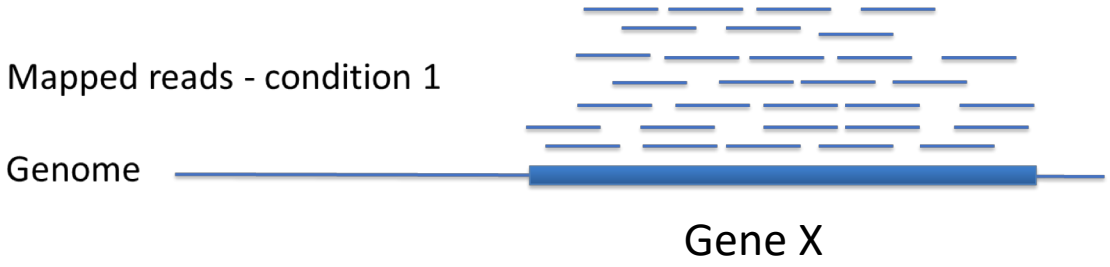


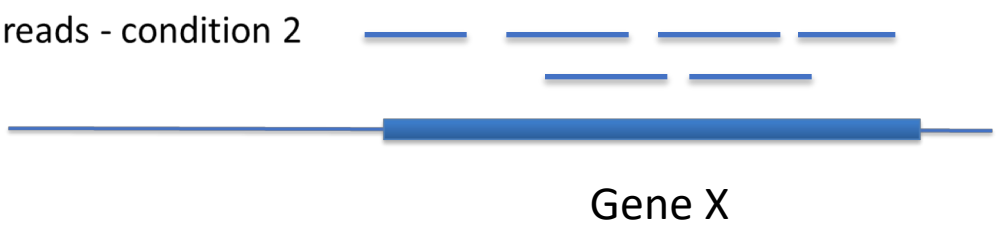




Expression of Gene in condition 1 and condition2

| Gene | Condition I | Condition II | Log fold change | T-value | P-value | Q-value |
|---|---|---|---|---|---|---|
| Gene X | 100 | 10 | 3.32 | 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-value | P-value | Q-value |
|---|---|---|---|---|---|---|
| Gene X | 100 | 10 | 3.32 | 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.**

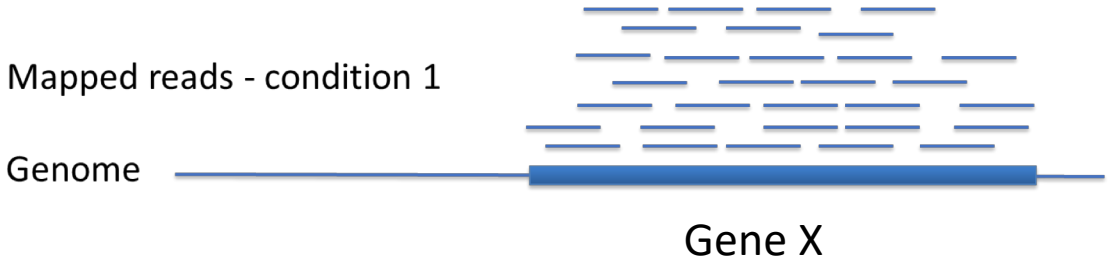
$$t\text{ –value} = \frac{\boldsymbol{Mean\ of\ expression\ Gene\ X\ condition1\ -Mean\ of\ Expression\ Gene\ X\ Condition\ 2}}{\boldsymbol{Standard\ error\ (SE)}}$$

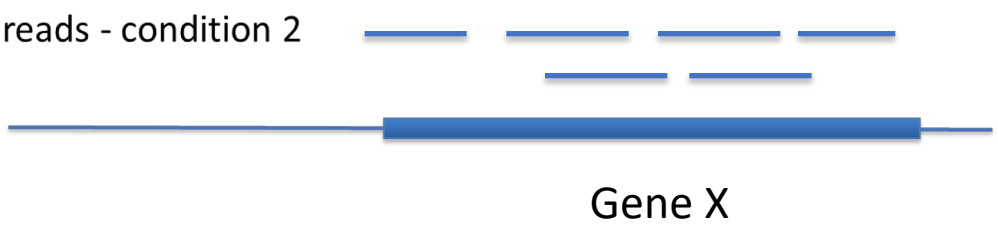
- Positive t-values means condition 1 has a higher value then condition 2; whereas negative value means expression of gene in condition 2 more than in 1.

# 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. This assumes that 95% of the
- P-value depends on t-value and degrees of freedom, which is generally number of samples -1.

# Differential Expression tables




Adjusted p-value for multiple test error correction

| Gene | Condition I | Condition II | Log fold change | T-value | P-value | Q-value |
|---|---|---|---|---|---|---|
| Gene X | 100 | 10 | 3.32 | 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** $= \frac{p-value\ for\ the\ experiment}{Number\ of\ samples}$
    - ✓ **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.**
    - ✓ $\mathbf{FDR(q-value)} = (\frac{Rank\ of\ the\ p-value}{Number\ of\ samples})$*False Discovery rate
    - ✓ **Compare it with q-value threshold (0.05). If less; than alternative hypothesis significant.**

- **For Multiple testing q-value is better than p-value to measure significance.**

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

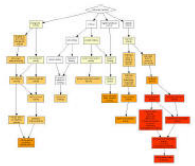
# What to do with significant genes?

Significant Gene List

## Functional Annotation



GORILLA



*Gene Ontology enRIchment anaLysis and visuaLizAtion tool*

## Interaction Database

BioGRID 3.5

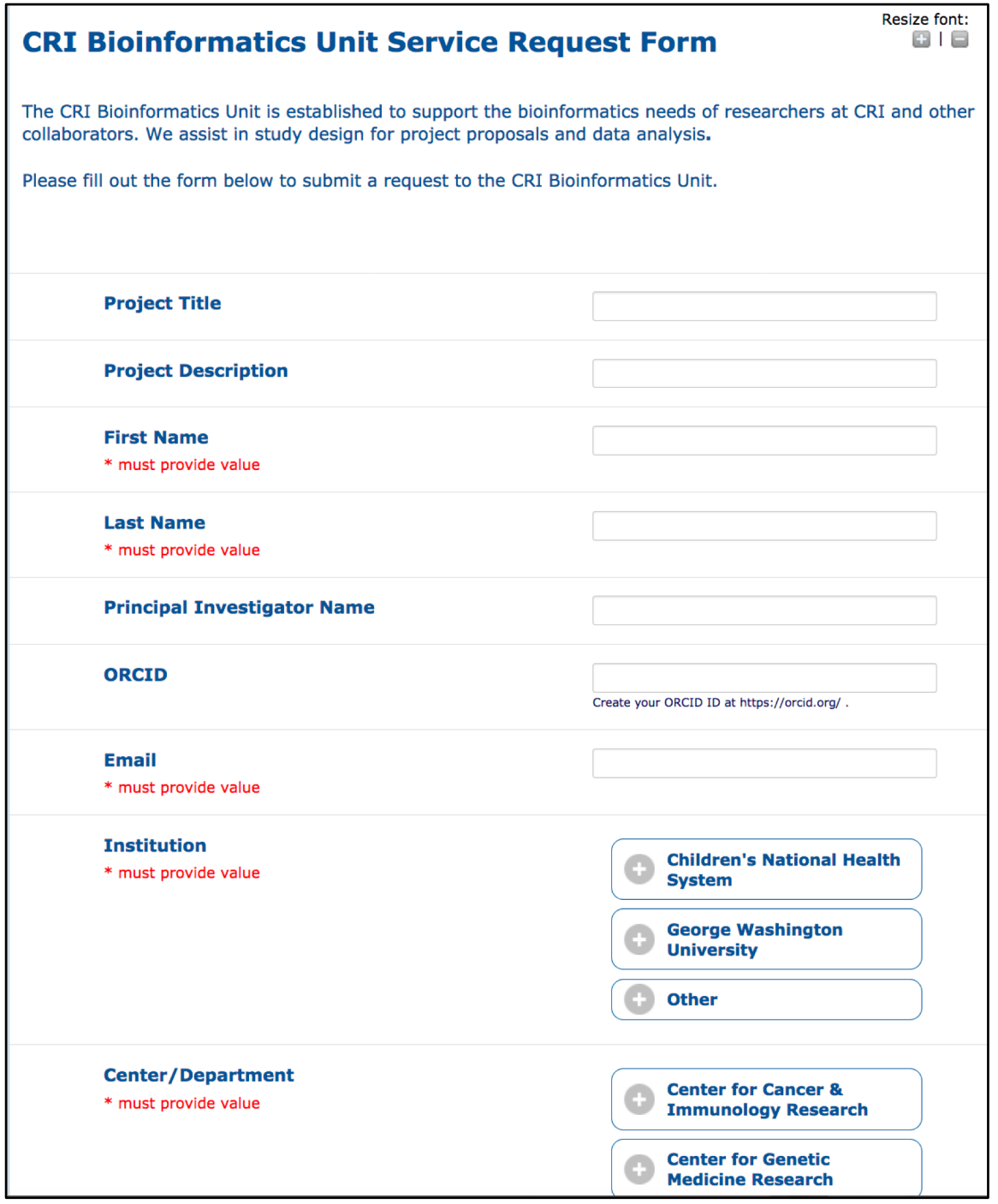
GeneMANIA

STRING

# CBU Service Request Form

https://cri-datacap.org/surveys/?s=3EJP7L8PLK

**Email :** bioinformatics@childrensnational.org



## CRI Bioinformatics Unit Service Request Form

Resize font:

The CRI Bioinformatics Unit is established to support the bioinformatics needs of researchers at CRI and other collaborators. We assist in study design for project proposals and data analysis.

Please fill out the form below to submit a request to the CRI Bioinformatics Unit.

**Project Title**

**Project Description**

**First Name**
* must provide value

**Last Name**
* must provide value

**Principal Investigator Name**

**ORCID**
Create your ORCID ID at https://orcid.org/ .

**Email**
* must provide value

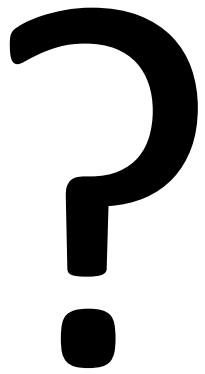
**Institution**
* must provide value

- Children's National Health System
- George Washington University
- Other

**Center/Department**
* must provide value

- Center for Cancer & Immunology Research
- Center for Genetic Medicine Research

# Questions



# References Microarrays


- Types of Microarrays :https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2435252/
- Microarray Analysis: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2762517/
- Limma: https://academic.oup.com/nar/article/43/7/e47/2414268
- Normalization I: http://web.cs.mun.ca/~harold/Courses/Old/CS6754.W06/Diary/ng1032.pdf
- Normalization II: http://www.cs.cmu.edu/~epxing/Class/10810/lecture/recitation7.pdf
- Afyymetrix Clariom S and D : https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/transcriptome-profiling-microarrays/clariom-assays.html
- Empirical Bayes: http://varianceexplained.org/r/empirical_bayes_baseball/
- ANOVA (with Excel): https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/

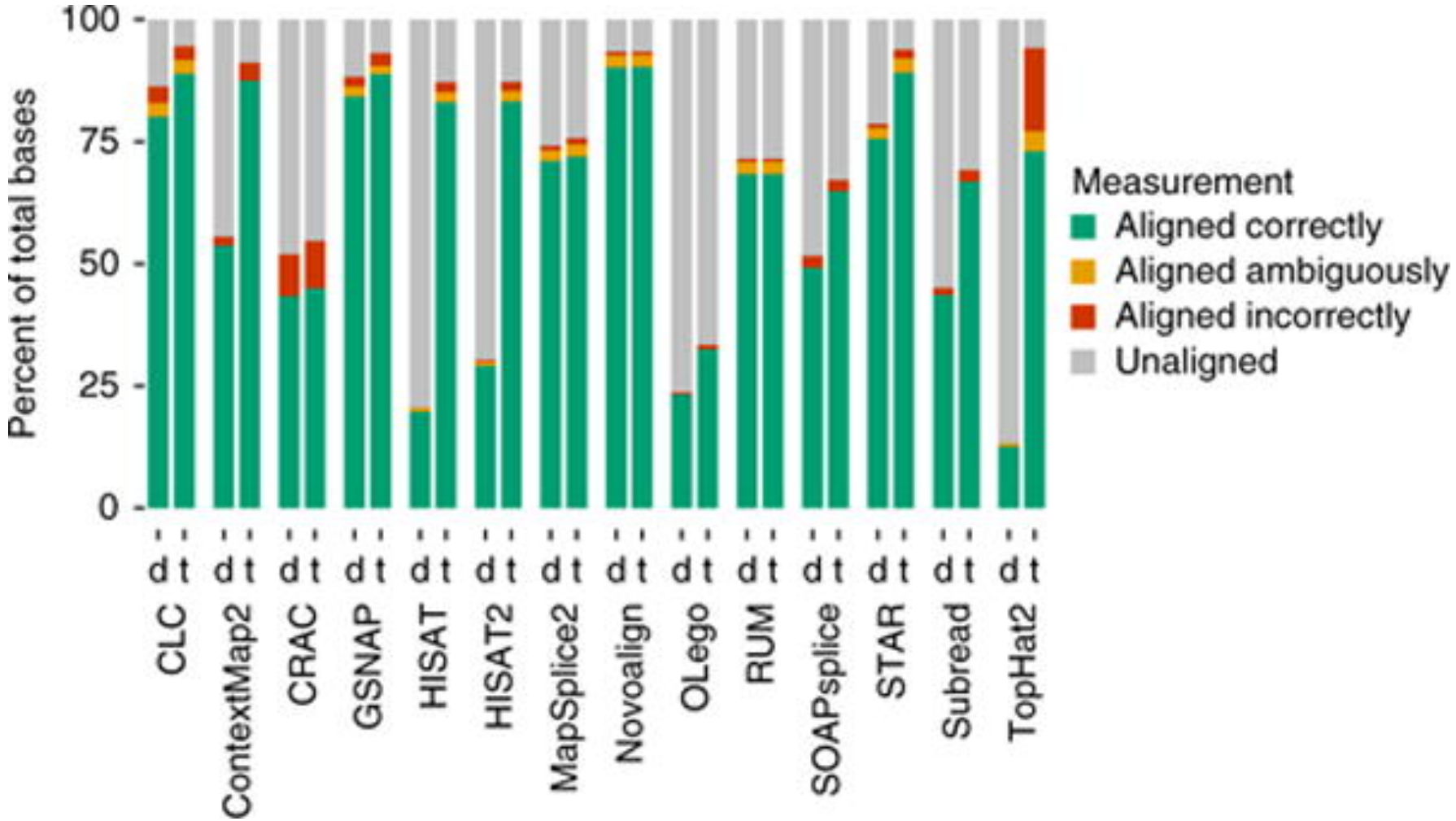
# References RNA-Seq I


- **RNA-seq and Transcriptomics I:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/
- **RNA-seq Best Practices analysis:** https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8
- **RNAseq Transcriptomics pipeline:** http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf
- **RNA-seq Aligners comparison:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5792058/
- **STAR:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/
- **RNA-seq Read count tools Comparisons I:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4673975/
- **RNA-seq Read count tool Comparison:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5003039/

# References RNA-Seq II


- TPM vs RPKM vs FPKM I: https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/
- TPM vs RPKM vs FPKM I: https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/
- RSEM: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-323
- Differential Expression I: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4293378/
- Differential Expression II: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-91
- Deseq2: https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

# RNAseq Aligner Comparisons




- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5792058/