

GENE EXPRESSION OMNIBUS (GEO), NCBI

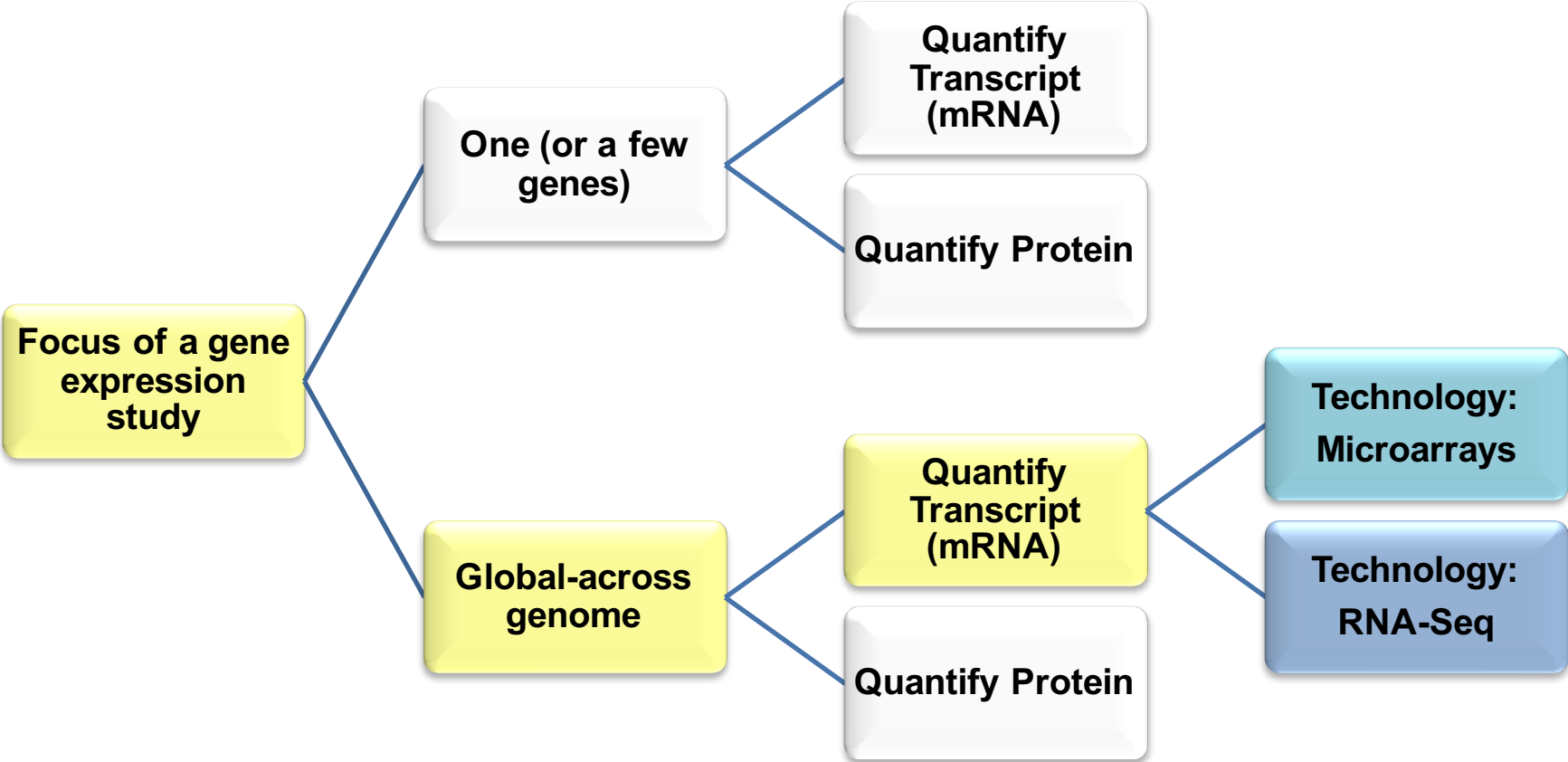
GEO Databases, Tools, and Gene
Expression data at NCBI

Hands-on Practice Outline

(Detailed steps are in the “BTEP2016_GEO_HO” document)

- **Exercise 1:** Explore a Study that was submitted to GEO
 - GEO2R
- **Exercise 2:** Explore a Curated GEO DataSets record:
 - Analysis Tools and Individual Gene Profiles
 - Gene- and Biosystems databases
- **Exercise 3:** The contents of the GEO DataSets database and RNA-Seq studies
- **Exercise 4:** Processed RNA-Seq data in the Gene database (with a quick detour to UniGene)
- **Exercise 5 (Optional):** A tour of BioSample-, BioProject-, and SRA databases

Gene Expression Studies-1

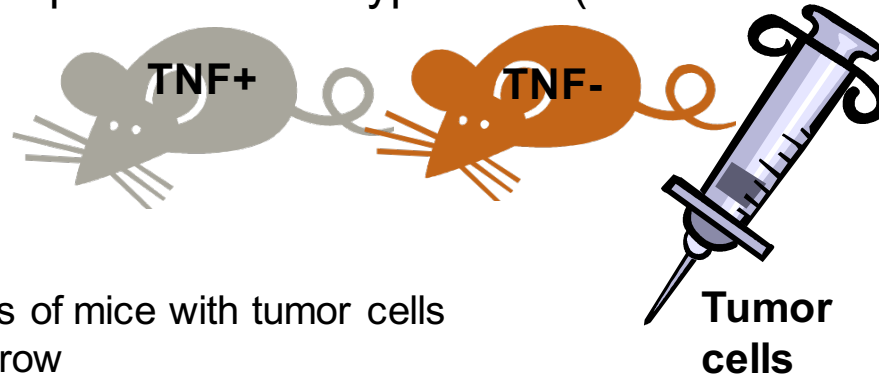


Gene Expression Studies-2

- First **gene expression arrays** studies entered in the NCBI database in 2001.
- First **RNA-Seq (expression profiling by high throughput sequencing)** NCBI entry in 2006, accelerated since 2008.
 - Generates a large amount of data
 - Requires automated (software) processing and statistical analysis
 - Requires large computing power and data-storage capacities
- Historically: **Expressed Sequence Tags (EST)** was the first approach to assess gene expression across a genome.
 - EST is a partial sequence of a transcript
 - First GenBank entries in 1992 (the **EST database**)
 - ESTs can be used as a semi-quantitative assay of gene expression.

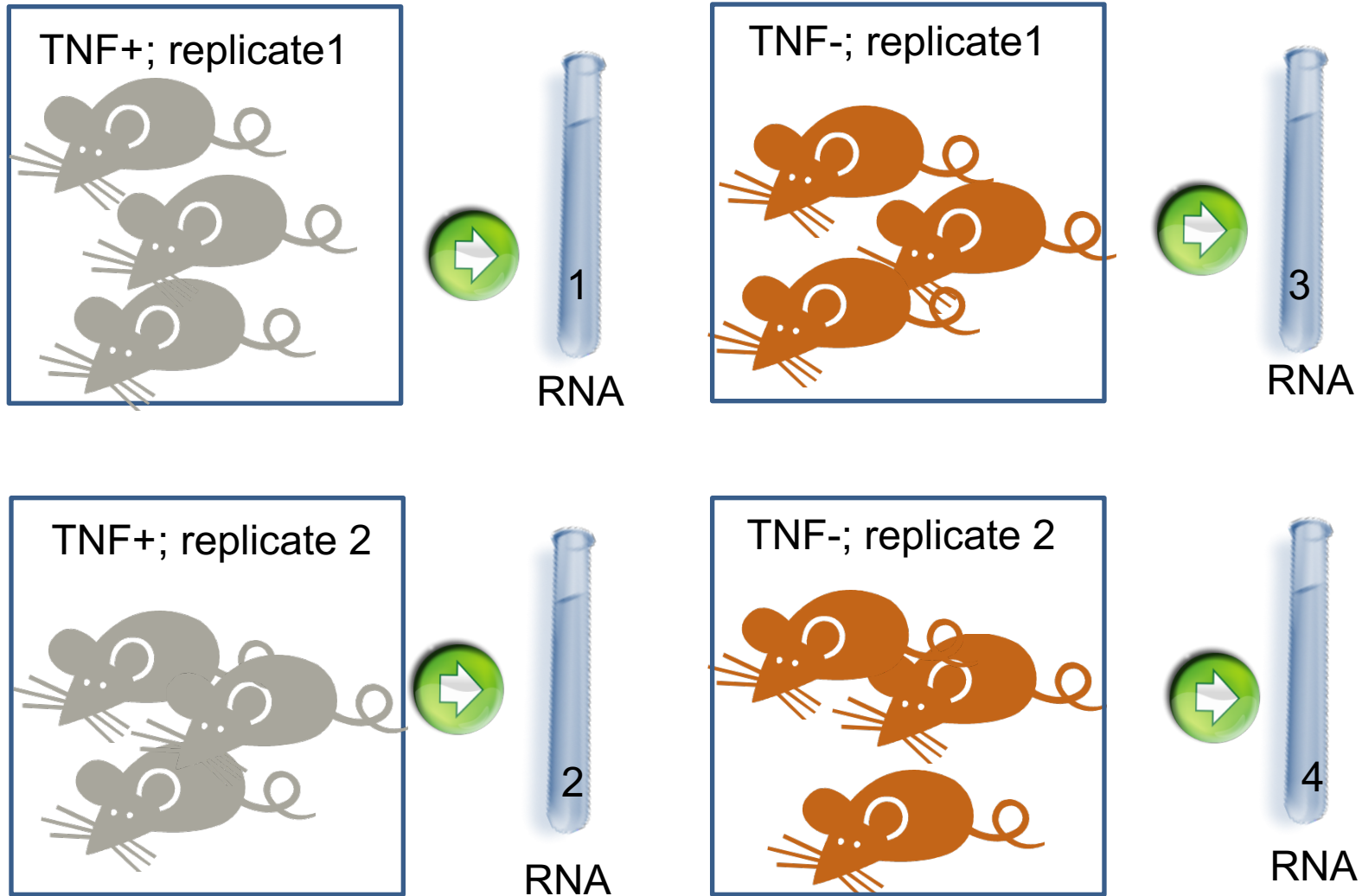
A study example -1

- Dr. Pitroda's team studies the TNF gene.
- Question: How does TNF affect growth of melanoma tumor cells and global gene expression (expression of all genes) in the tumor cells?
- Approach: study tumors in mice that are missing TNF (TNF knockout) and compare with wild type mice (those that have TNF)

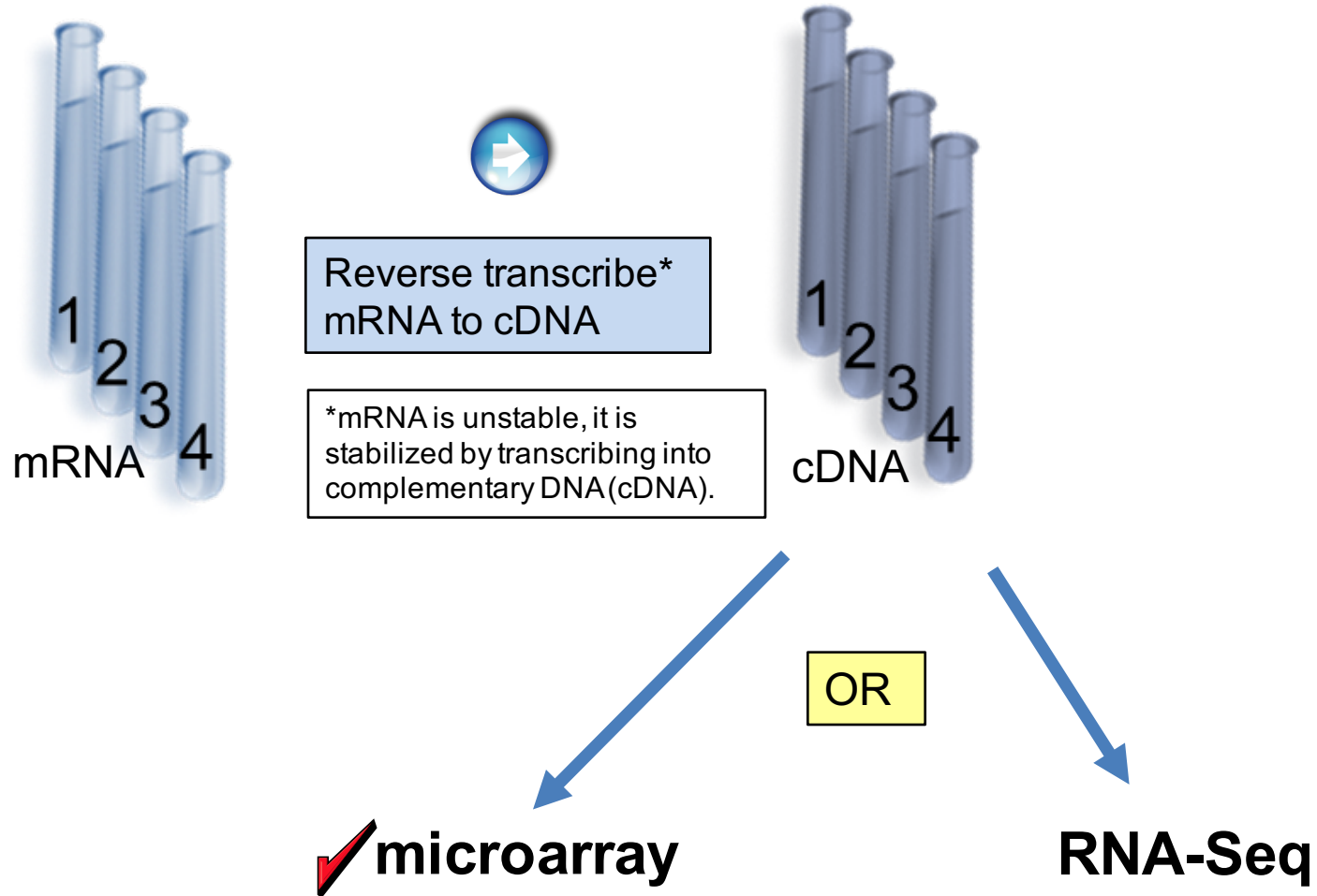


- Experiment:
 - Inject both kinds of mice with tumor cells
 - Let the tumor grow
 - Harvest the tumors from each kind of mice
 - Separate endothelial cells from the rest and isolate their mRNA.

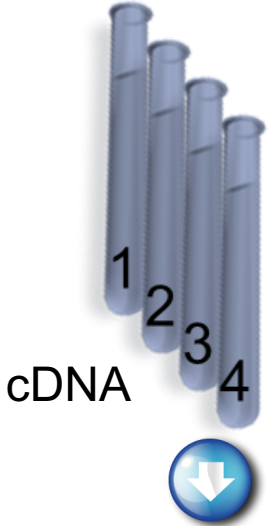
A Study Example -2



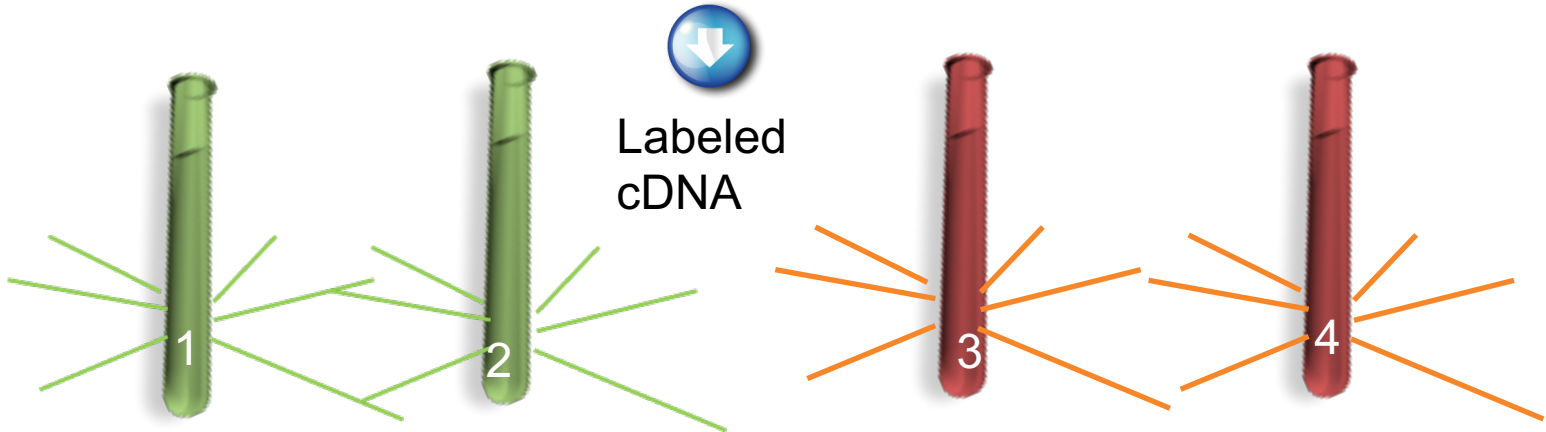
A Study Example - 3



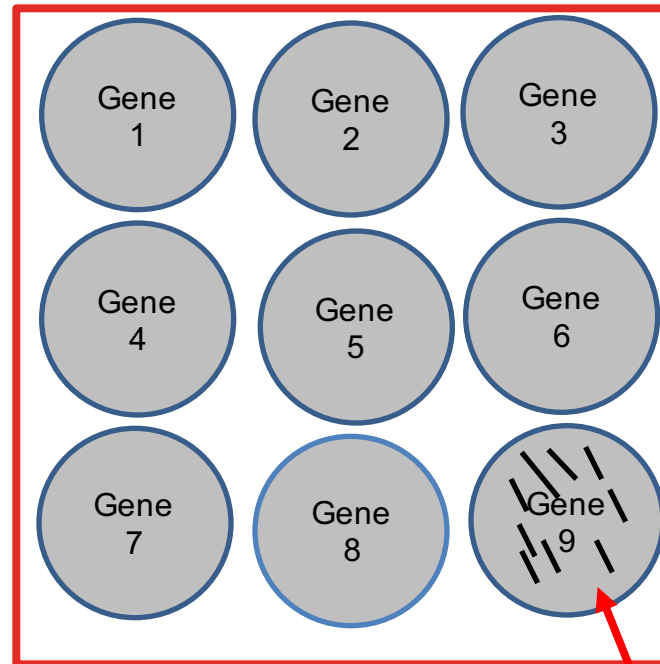
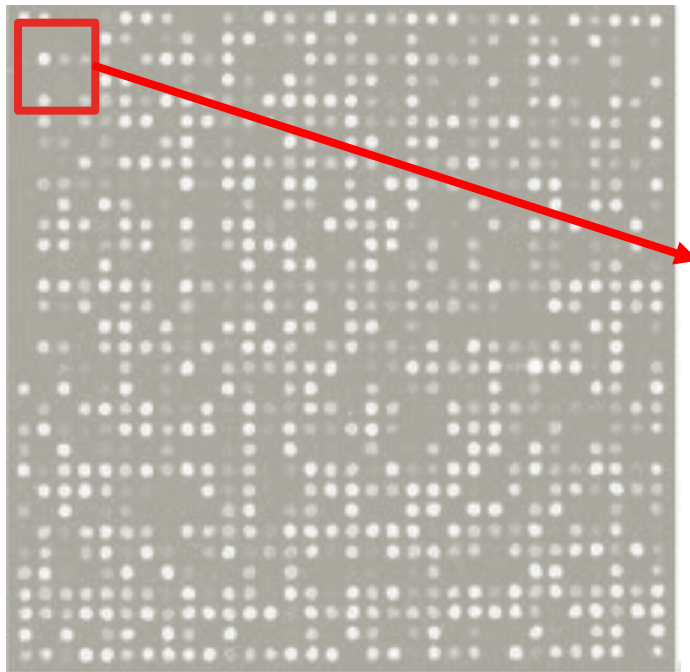
Microarray -1



Attach fluorescent molecules to the cDNA molecules: green for the Wild-type (TNF+) samples; red for the knockout (TNF-) samples.



Microarray -2



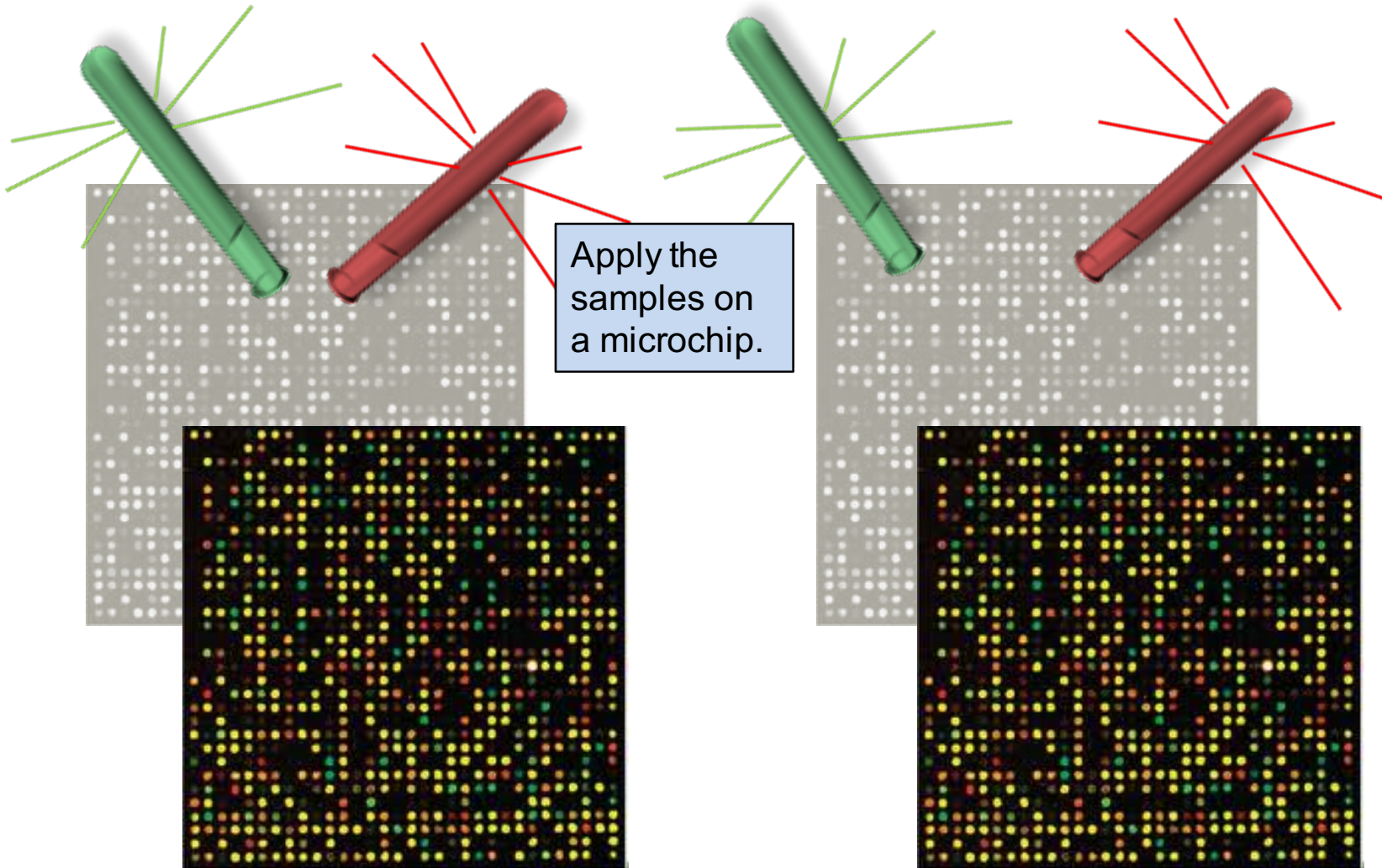
A microarray chip is a tiny plate with spots. Each spot contains a gene probe (many short pieces of DNA, called *oligonucleotides* attached to the plate) that will bind mRNA (cDNA) expressed from that same gene.

Oligonucleotides
(in situ
oligonucleotide)

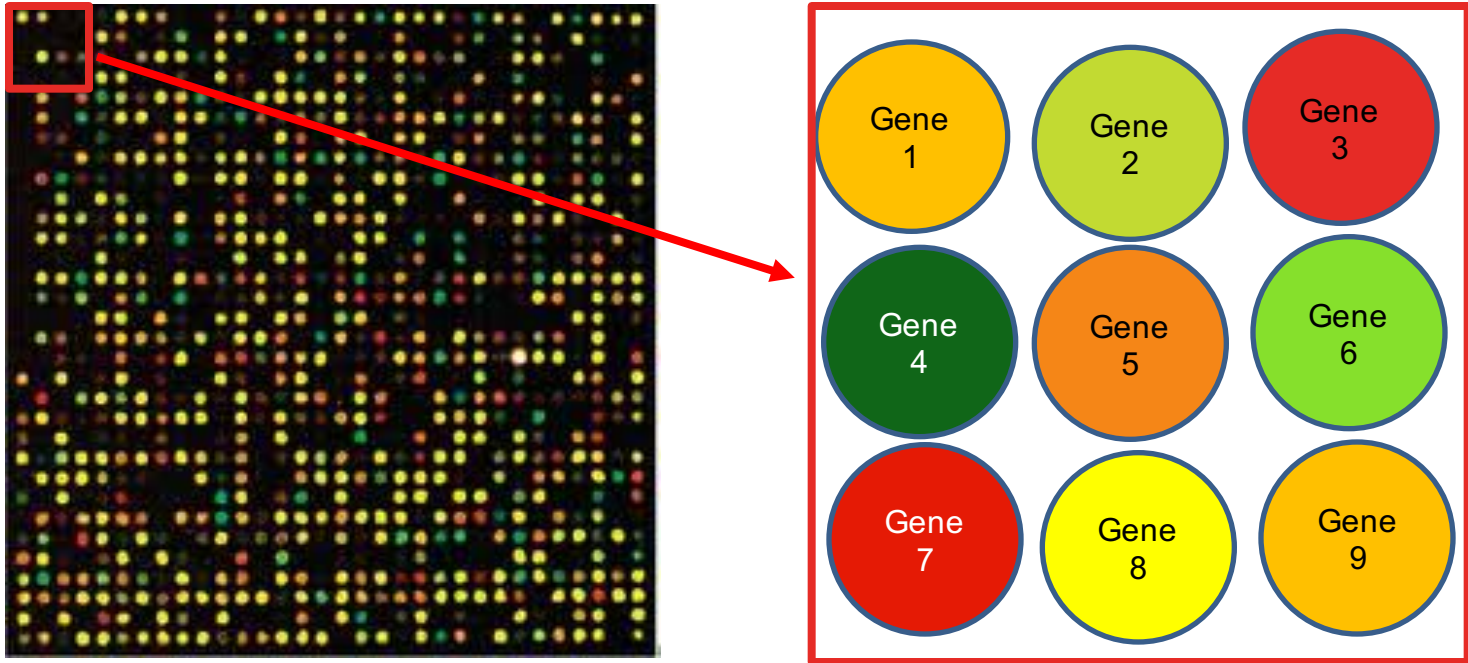
Microarray - 3

1= TNF+ rep. I 3= TNF- rep. I

2= TNF+ rep. II 4= TNF- rep. II

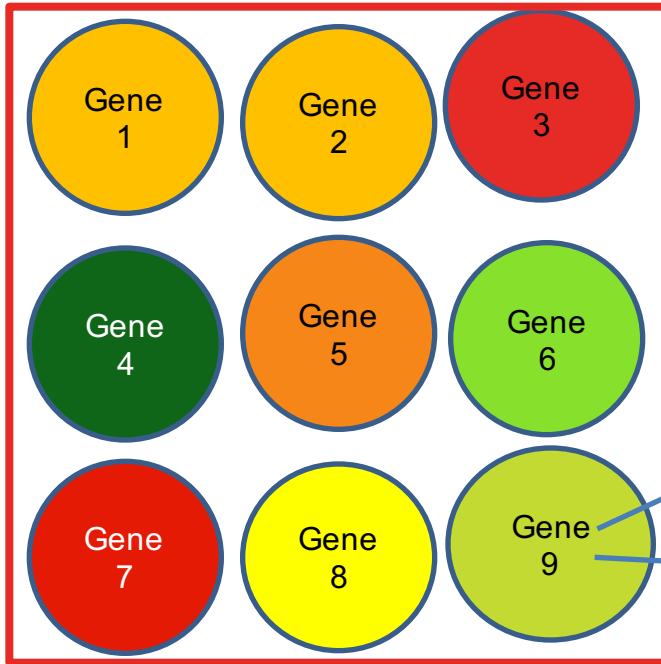


Microarray – 4



Each spot emits two types of fluorescent light. Its intensity can be measured automated equipment...

Microarray – 5

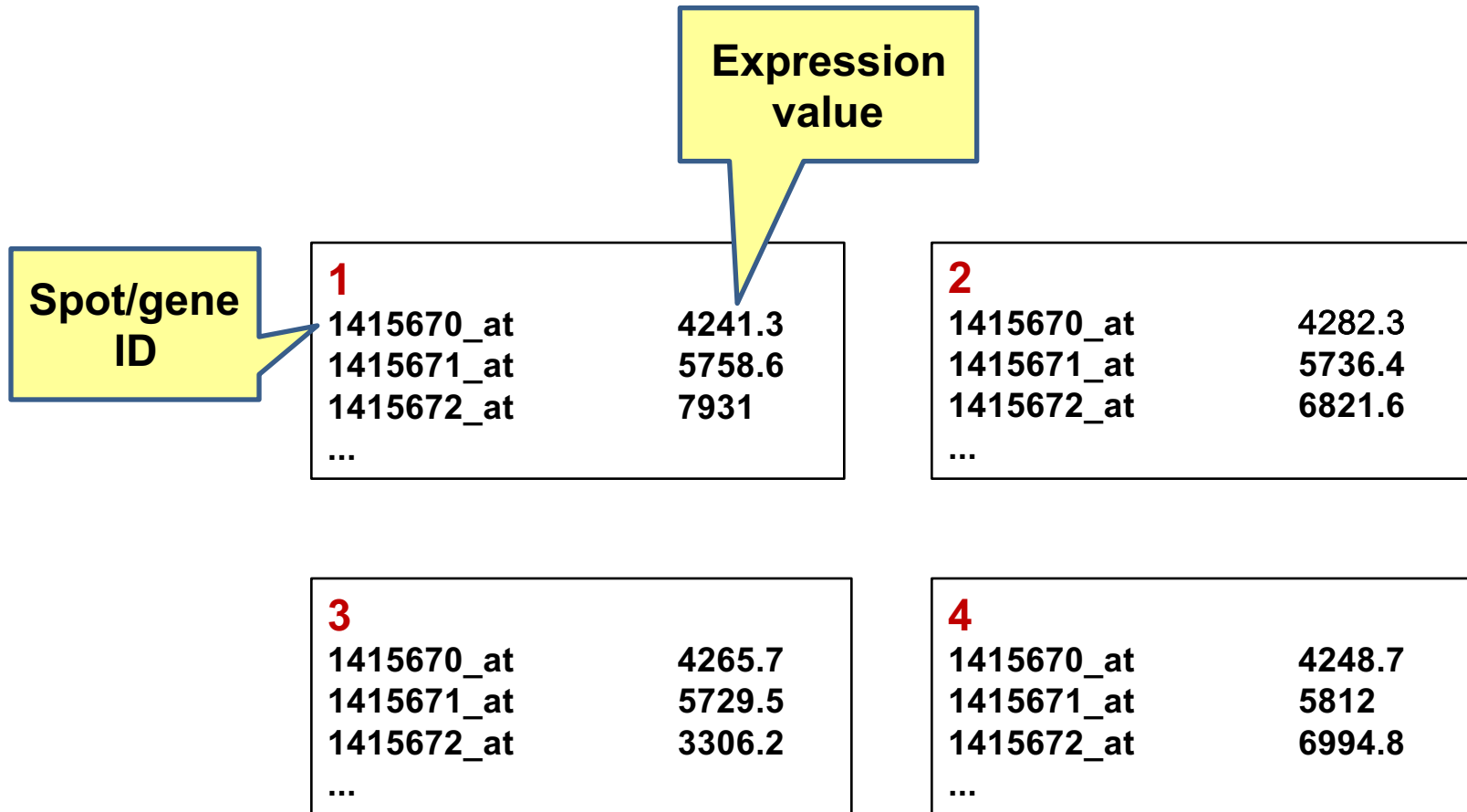


TNF+ 6133.7

TNF- 4282.3

...and given numerical values.

Microarray – 6



Data Submission to GEO

Dr. Pitroda submits the study to GEO

GEO wants to know/have the following from the submitter:

- A general information about the study: what was done, why and how?
→ Dr. Pitroda submits a single **series** record.
- Specifics about the samples, and the obtained data for **each** sample
→ Dr. Pitroda submits **4 sample** records (each record with **values** for all of the genes on the platform).
- Information about the array
→ Dr. Pitroda submits a **platform** record.

A blue scroll graphic with a gradient from light blue to dark blue, featuring a shadow and a rolled-up edge on the right side.

Submitted study

A yellow starburst graphic with a dark blue outline and a gradient from light yellow to dark yellow.

Go to Exercise 1

The Concept of Curation in GEO

- GEO curators select a single study (family) and curate it:
 - family= the series entry + its samples + its platform
 - One Family → one curated DataSet entry (GDS)
- Some of other curation efforts at NCBI:
 - Redundant GenBank sequences → a single reference sequence
 - Redundant submitted SNPs → a single reference SNP
 - Redundant PubChem substances → a single (reference) compound
- GEO does NOT have (normal) gene expression references (studies submitted to GEO are too variable to generate reference).
- Only microarray studies curated.
- Curation efforts currently stalled.



Curated study



Go to Exercise 2

Exercise 1 and 2: Summary

GEO Terminology

GEO DataSets database

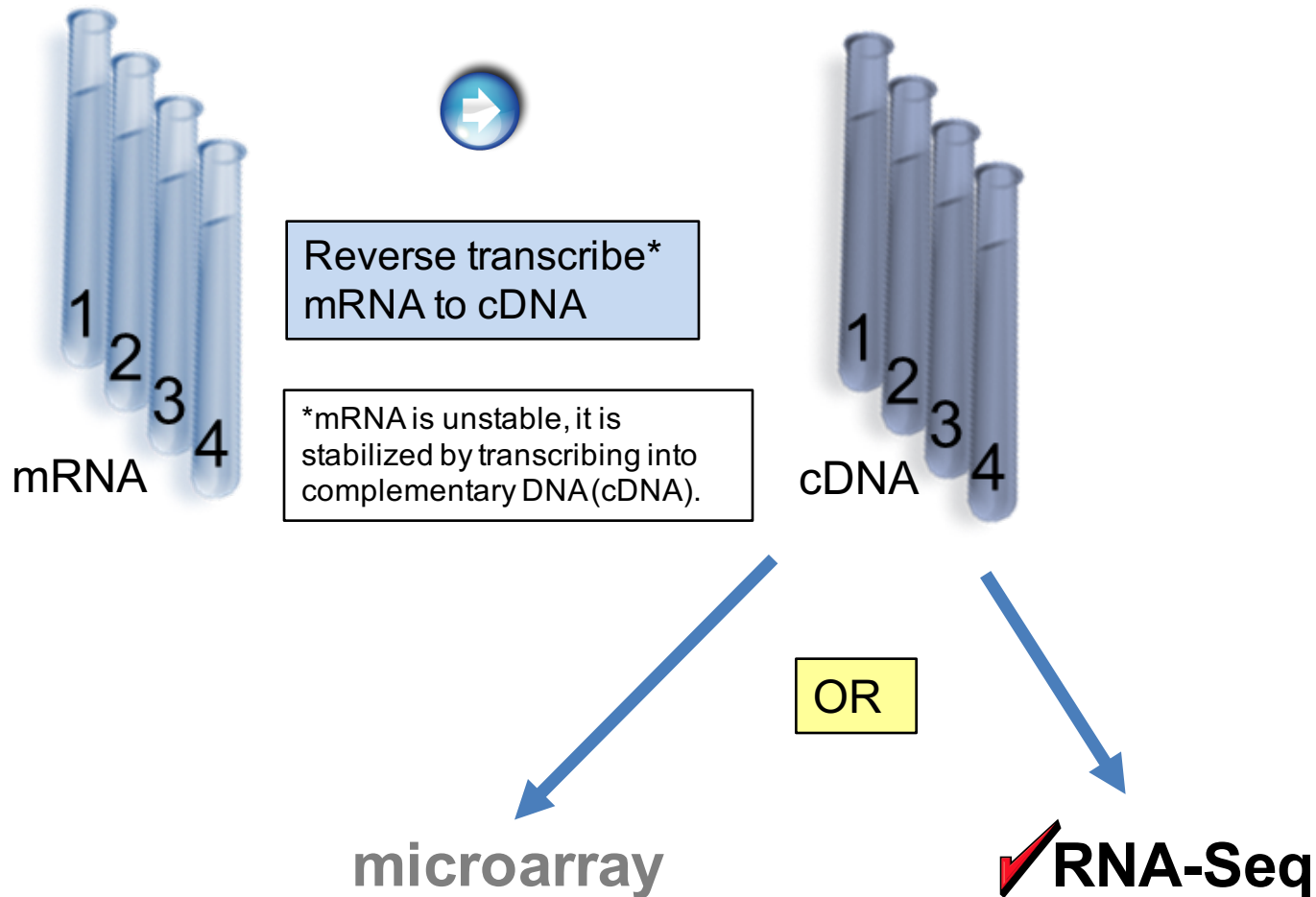
GEO Entry (record) types:

- Series=GSE; describes overall design of the study*
- Sample=GSM; describes individual samples; one record per sample)*
- Platform=GPL; describes technology platform used in a study*
- DataSet=GDS; curated studies*

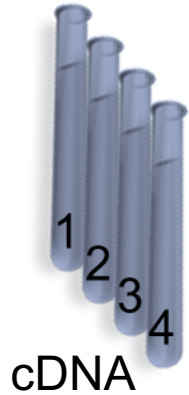
GEO Profiles database

- Individual gene profiles from curated (GDS) studies*

Wet-Bench Example - 4



RNA-seq -1



cDNA



Fragment cDNA and attach adapters (ligate short oligos to cDNA)



cDNA libraries



Proceed to sequencing platform (Illumina HiSeq200).

The adapters allow fragments to attach to a *flow cell* (glass slide with oligos that are complimentary to the adapters).



Clonally amplify the fragments, so to obtain clusters.



The clusters are sequenced in a massively parallel process. Illumina uses “sequencing by synthesis” where each incorporated base excites light that is captured and recorded in a nascent read.



<http://www.illumina.com/technology/next-generation-sequencing.html>

Bottom right: “In-Depth NGS Introduction” PDF

http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Millions of reads are exported to a data file.



RNA-seq -2

```
>gnl|SRA|SRR1644186.21.2 HWI-ST378R:177:D1WAEACXX:1:1101:3231:2194  
TGTCTCTAAGCTGAAGGTTTGGGAAGCGGTTGTTTCTGAAATCTTCCAATAATCTACTGC  
GGCCAGAAGGCATAATATCAGACCTATTATACCGAAGCCG
```

A read is a short piece of sequence.
Image: a single read display on the web at NCBI.

How are the reads processed?

The reads are aligned (mapped) on the reference genome.

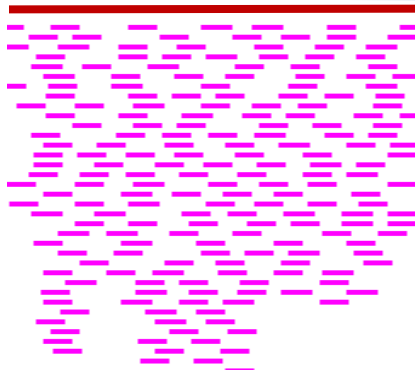
An aligner, for example the HISAT (one of several software tools) is used.



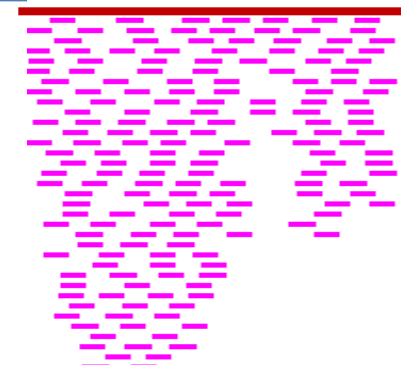
exon 1

intron

exon 2



Genome (DNA); "gene A region"



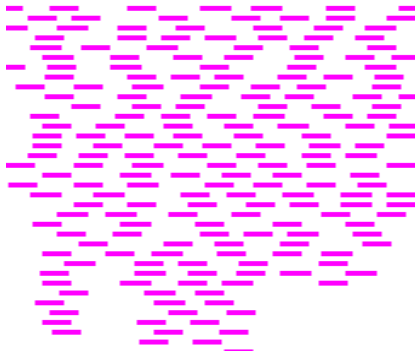
Intron-spanning read

exon 1

exon 2

mRNA

RNA-seq - 3



Genome (DNA); “gene A region”



How are the reads quantified?

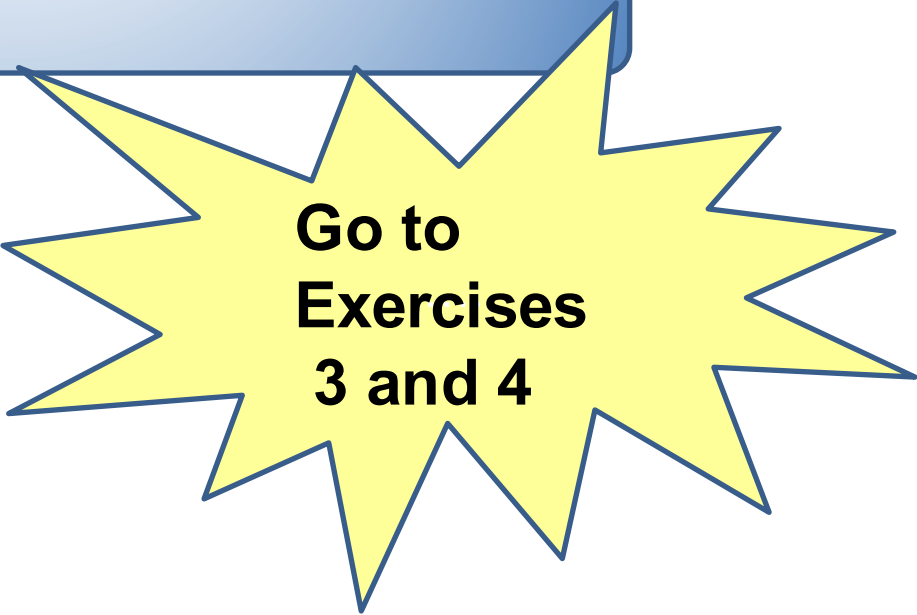
The reads are assembled into a gene model, counted and expressed as **FPKM**: *fragments per kilobase of transcripts per million reads*.*

An example of a software tool for assembling and quantifying reads is Stringtie.

*For example, a software counts 100 reads for the gene A model. The total number of reads in the experiment is two millions, so it has to divide by 2 (50 reads). The assembled transcript of gene A is 5 kb in length, so it also has to divide by 5. The reported FPKM for gene A will be 10.
The software also counts 100 reads for gene B. But it is a shorter gene, 2kb in length. The FPKM for gene B will be 25. (Gene B expression is higher than that of gene A.)



RNA-Seq



**Go to
Exercises
3 and 4**

The Big Picture

NCBI Gene-Expression Databases

