Statistical Design and Analysis of Gene Expression Experiments

First Lecture!

An Overview

Central Dogma: DNA→RNA→Protein

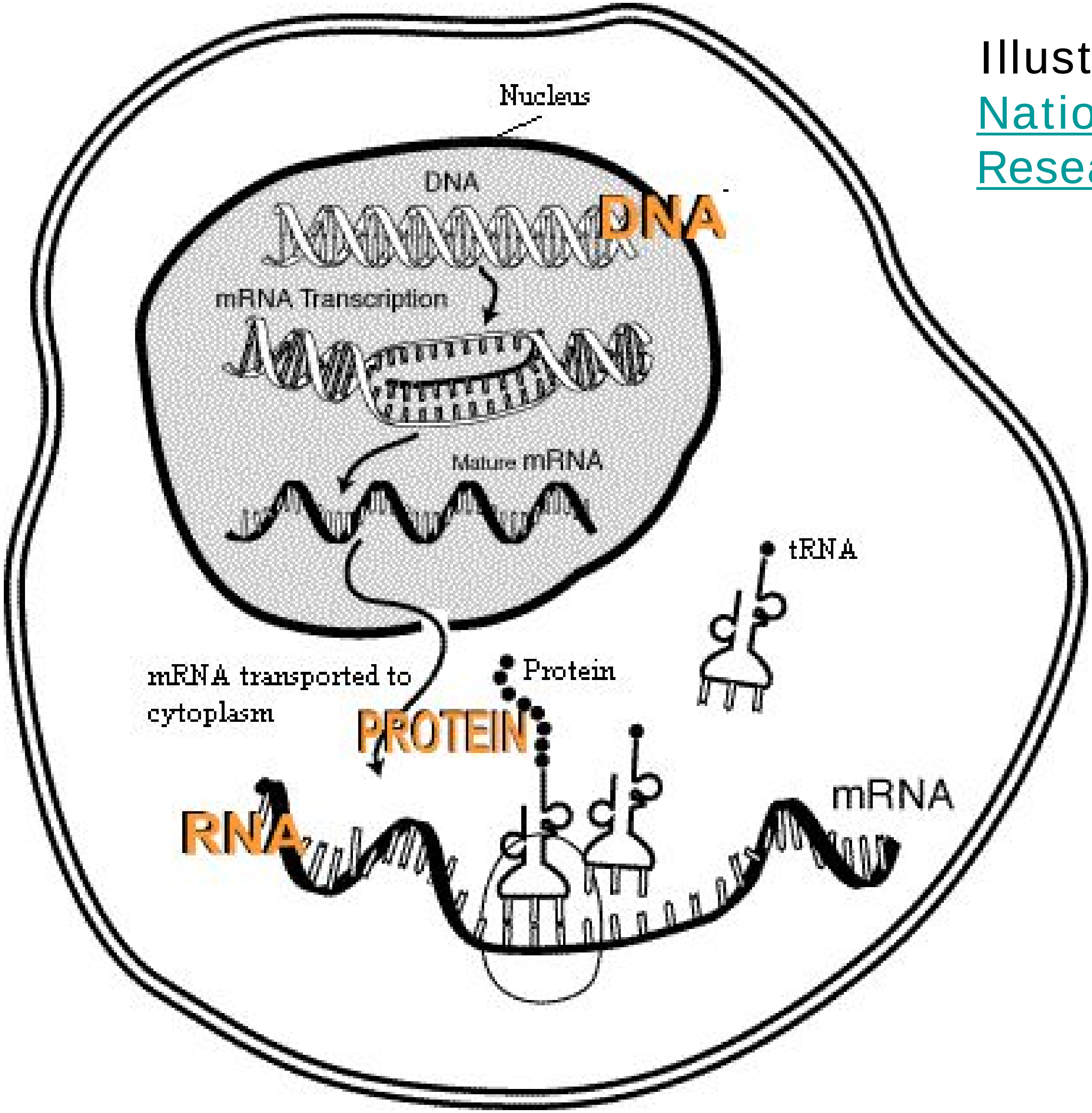


Illustration provided by the [National Human Genome Research Institute]

DNA

↓ (transcription)

RNA

↓ (translation)

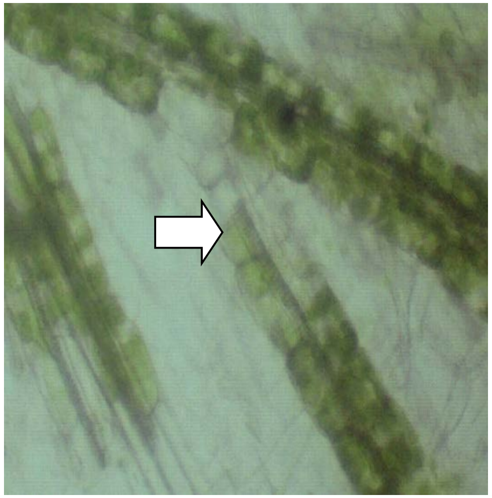
Protein

Gene Expression Data

- Monitoring gene expression helps understand the cellular mechanisms for all biological processes.
  - gene function
  - gene network

- RNA-seq and microarray technologies allow measuring expression levels (abundance of mRNA transcripts) of thousands of genes simultaneously.

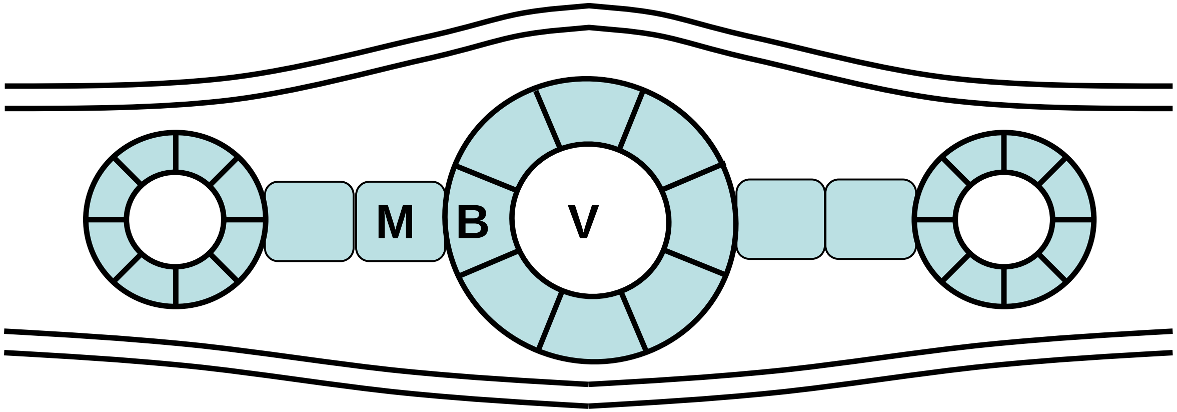
Example 1: Sawers *et al*, 2007, BMC Genomics



bundle sheath strands

mesophyll protoplasts

- Goal: To detect genes that are differentially expressed in Bundle Sheath (B) and Mesophyll (M) cells.



Example 1: Sawers *et al*, 2007, BMC Bioinformatics

- A little more complication:

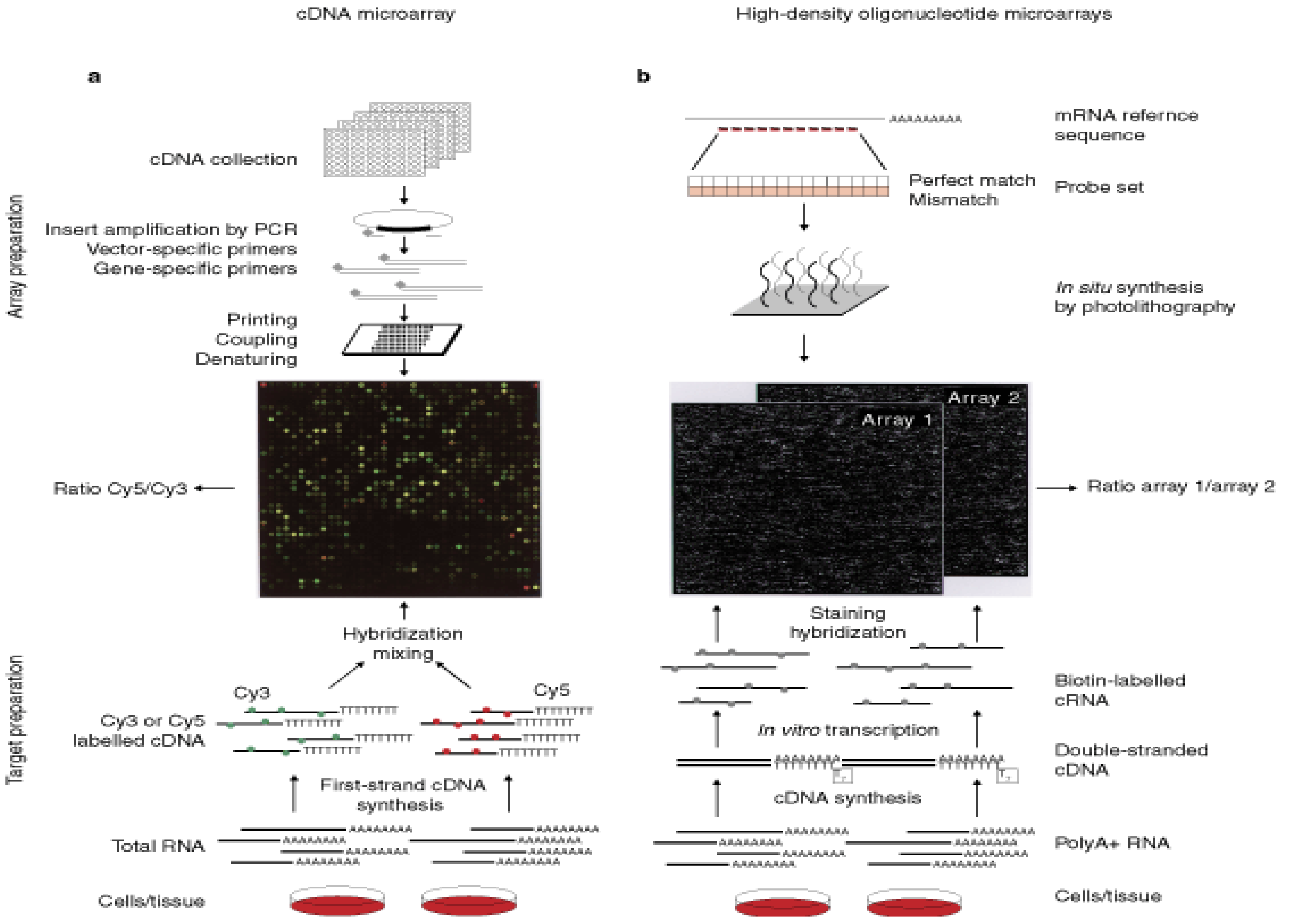
  The procedure for extracting mRNA for the two cells are different. The one to extract mRNA from M cells introduces stress.

- Solution:

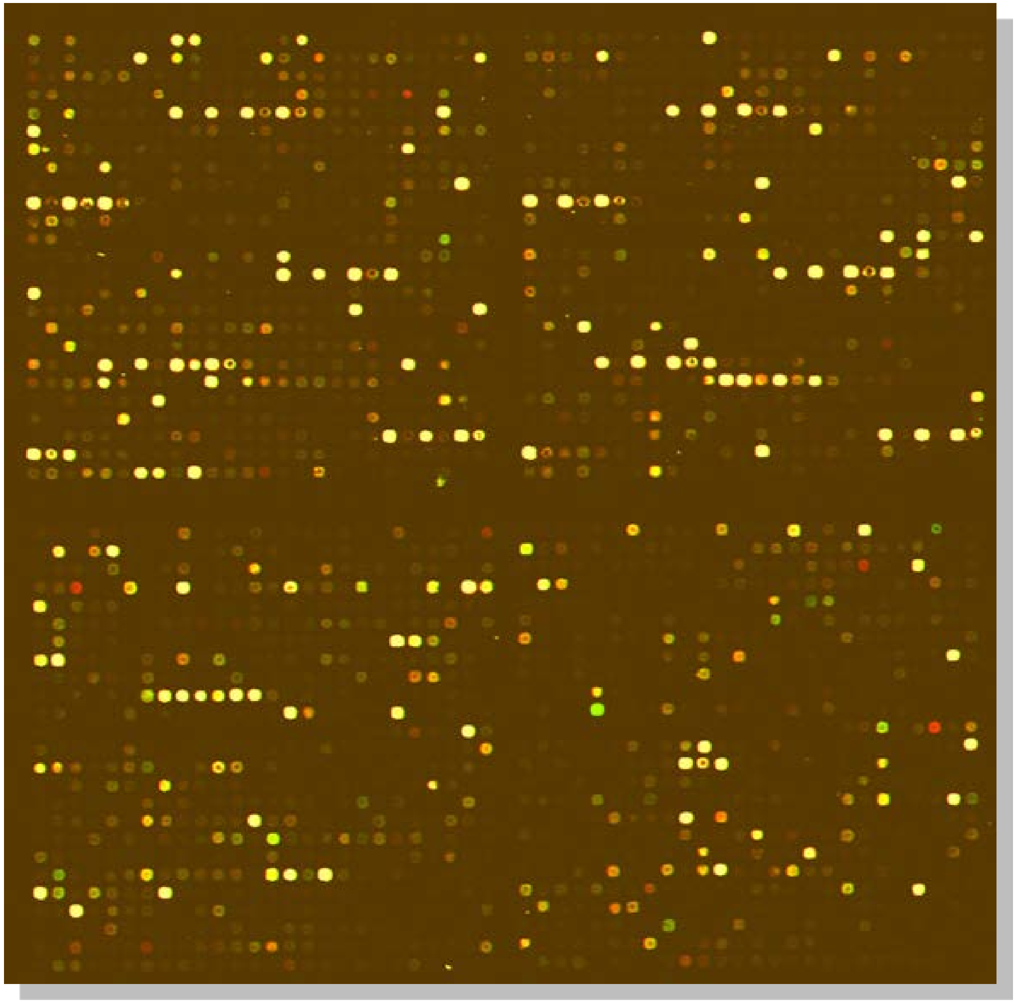
  Add two more treatment groups: samples with both M and B cells going through extraction of mRNA with and without stress.

→B, M, Stress and Total (4 treatment groups)

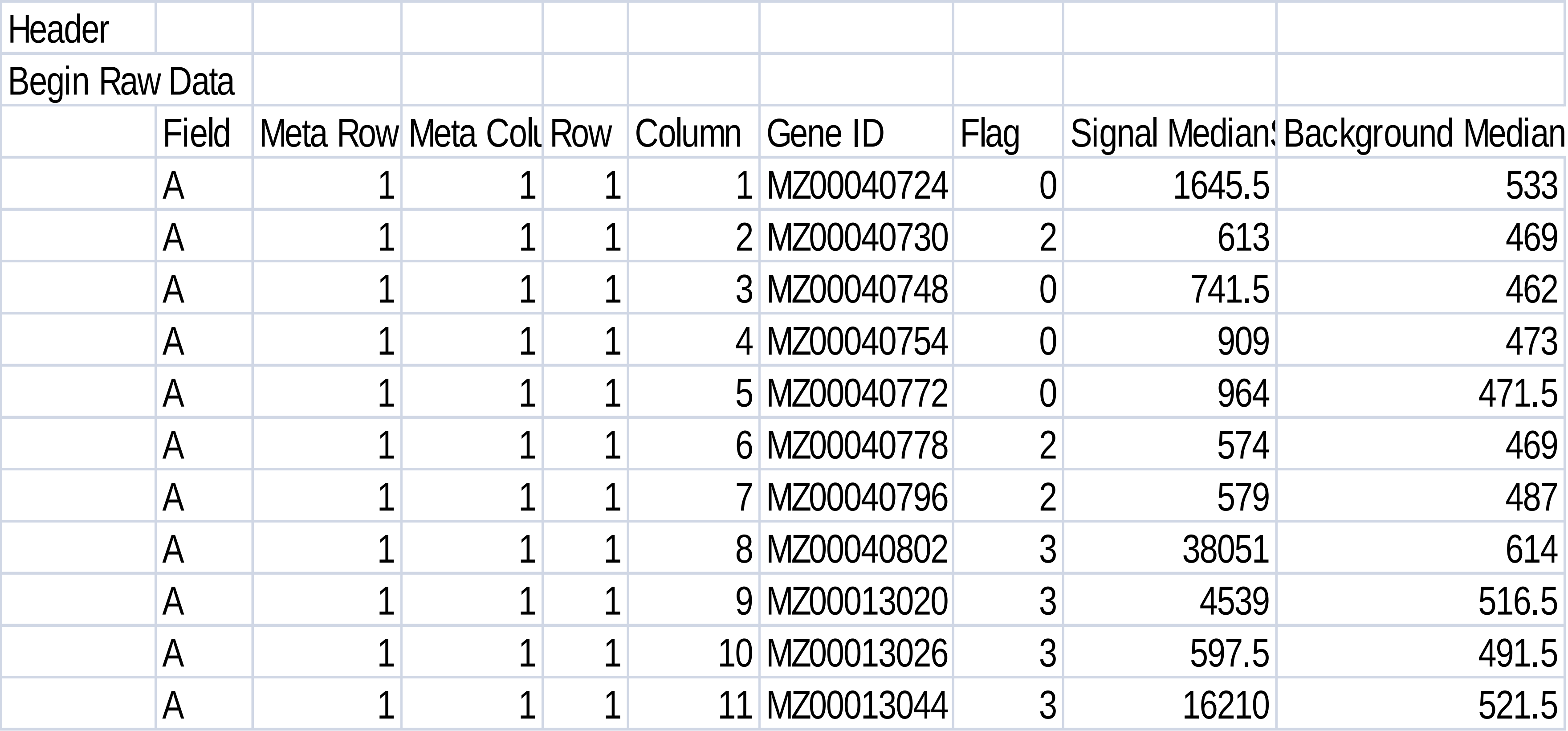
Performing the experiment (*Nature* cell biol. 2001 3:8)



After the bench work…(2-color microarray)



The data table looks like



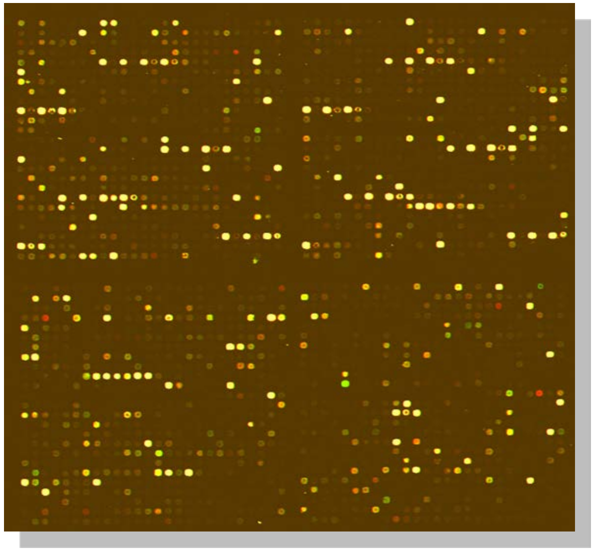
Header									
Begin Raw Data									
	Field	Meta Row	Meta Colu	Row	Column	Gene ID	Flag	Signal Median	Background Median
	A	1	1	1	1	MZ00040724	0	1645.5	533
	A	1	1	1	2	MZ00040730	2	613	469
	A	1	1	1	3	MZ00040748	0	741.5	462
	A	1	1	1	4	MZ00040754	0	909	473
	A	1	1	1	5	MZ00040772	0	964	471.5
	A	1	1	1	6	MZ00040778	2	574	469
	A	1	1	1	7	MZ00040796	2	579	487
	A	1	1	1	8	MZ00040802	3	38051	614
	A	1	1	1	9	MZ00013020	3	4539	516.5
	A	1	1	1	10	MZ00013026	3	597.5	491.5
	A	1	1	1	11	MZ00013044	3	16210	521.5

Microarray analysis

- Image processing
- Background correction
- Transformation
- Normalization
  - remove sources of systematic variation
- Fit linear models
- Multiple testing
- Clustering analysis, Gene set testing, etc.

Testing in Microarray

- With microarray experiments, biologists often want to detect genes differentially expressed between different treatments or conditions



Gene ID	Control			Treatment		
1	0.5	0.6	0.45	1.3	1.4	1.25
2	0.9	1.0	0.7	1.0	0.8	0.9
…		…			…	

Normalized Signal Intensities (NSI)

Detecting differentially expressed genes

- Model the mean for NSI, e.g., $E(Y_{ijk}) = \mu+\tau_i+\delta_j$

  μ represents overall mean of NSI.

  τ_i represent the effects of treatments i on mean NSI.

  δ_j represents the effects of j-th dye (Cy3 and Cy5) on mean NSI

- Construct statistical test for parameters that we are interested in, e.g., what are the difference in gene expression (τ_1 - τ_2)?

  $\tau_1 - \tau_2 \neq 0$ means differential expression.

Detecting differentially expressed genes

- There are some random effects that are unknown:

  slide effects

  other effects introduced in the experiment (such as biological replicate effects)
  residual random effects that include any sources of variation unaccounted for by other terms

Detecting differentially expressed genes

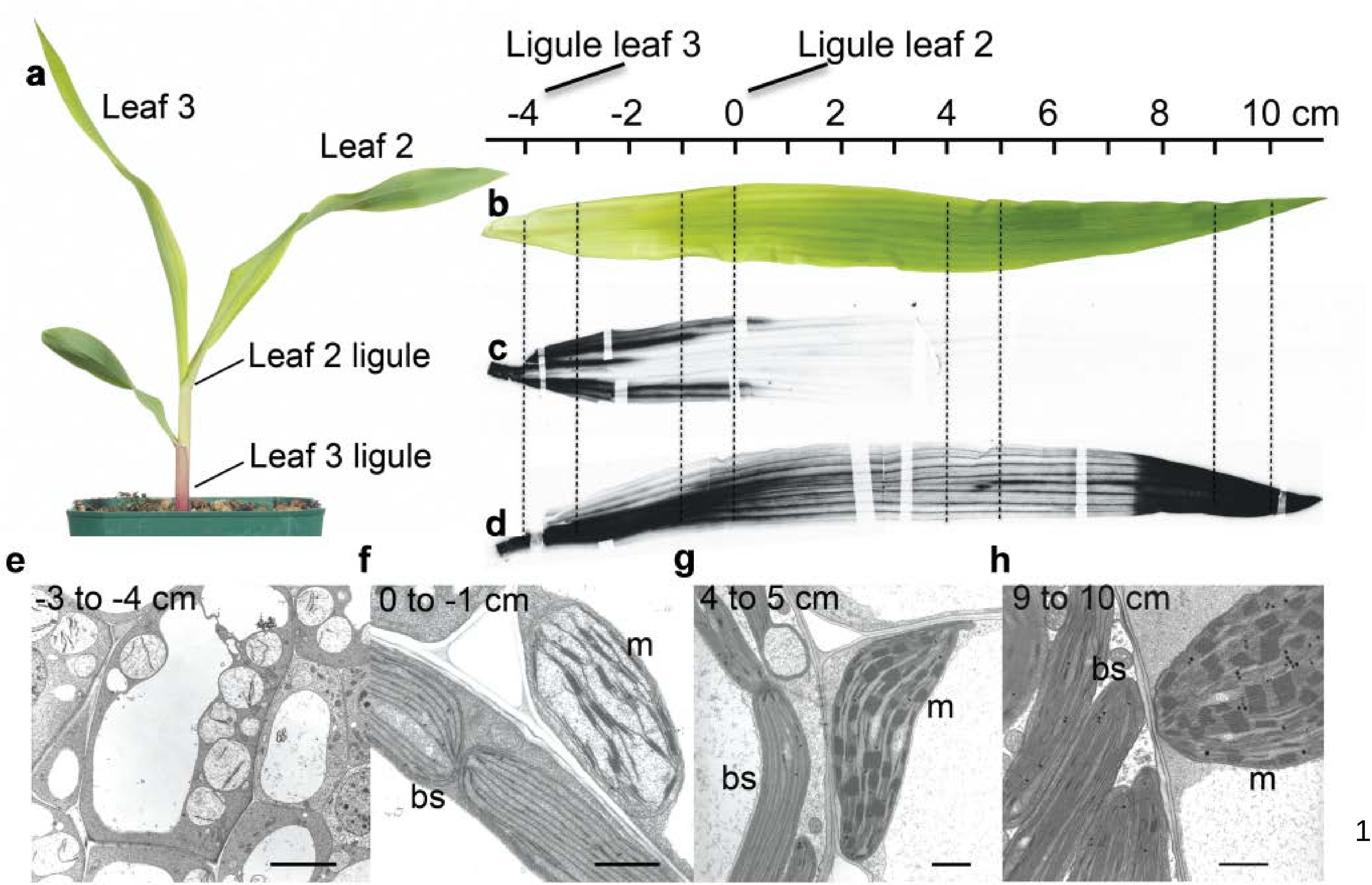
- Model for normalized signal intensities (NSI): $Y_{ijk}=\mu+\tau_i+\delta_j+s_k+e_{ijk}$ for each gene g

  i: treatment index

  j: dye index

  k: slide index
- Model the random effects and perform tests to get a p-value or construct confidence intervals

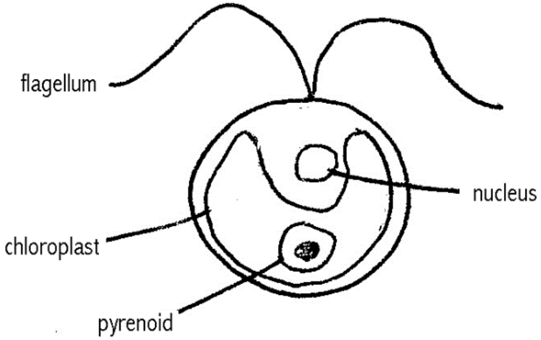
Example 2: Li et al, 2010, Nature Genetics
The developmental dynamics of the maize leaf transcriptome



Example 3: Fang et al. 2012, The Plant Cell (slide from Wei Fang)

- The model organism *Chlamydomonas reinhardtii*
- A unicellular green alga with flagellum and chloroplast.
- Concentrate CO_2 to solve the problem:
  - C4 pathway in plants
  - The CO_2 Concentrating Mechanism (**CCM**) in *Chlamydomonas*

http://web.mst.edu/~microbio/BIO221_2009/C_reinhardtii.html



http://en.academic.ru/dic.nsf/enwiki/314153

Example 3: Fang et al. 2012, The Plant Cell

- Study transcriptome regulation by CO2 and by the transcription regulator CIA5 (CCM1).

- Experiment design:
  - Wild type: 137c (cc125); *cia5*: point mutation in 137c background.
  - 4 hours induction under: High CO_2: ~5% CO_2; Low CO_2: 300 to 400 ppm (air level) CO_2; Very Low CO_2: 100 to 200 ppm CO_2

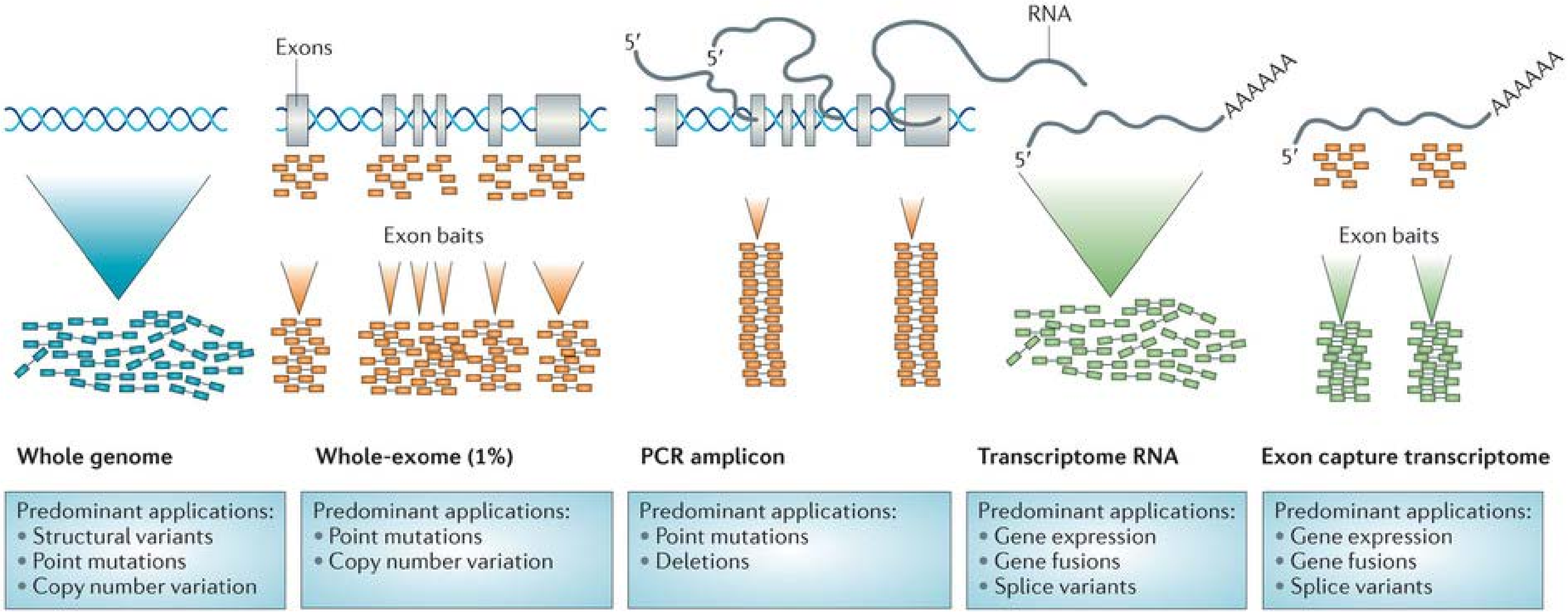
Wild type; High CO_2	Wild type; Low CO_2	Wild type; Very Low CO_2
cia5; High CO_2	*cia5*; Low CO_2	*cia5*; Very Low CO_2

RNA-seq experiments

- Next-generation sequencing (NGS) technology is an ultra-high-throughput technology to measure DNA sequences.

- RNA-seq refers to the method of using NGS technology to measure a set of RNA levels.

Examples of other applications of NGS technologies



Whole genome

Predominant applications:
- Structural variants
- Point mutations
- Copy number variation

Whole-exome (1%)

Predominant applications:
- Point mutations
- Copy number variation

PCR amplicon

Predominant applications:
- Point mutations
- Deletions

Transcriptome RNA

Predominant applications:
- Gene expression
- Gene fusions
- Splice variants

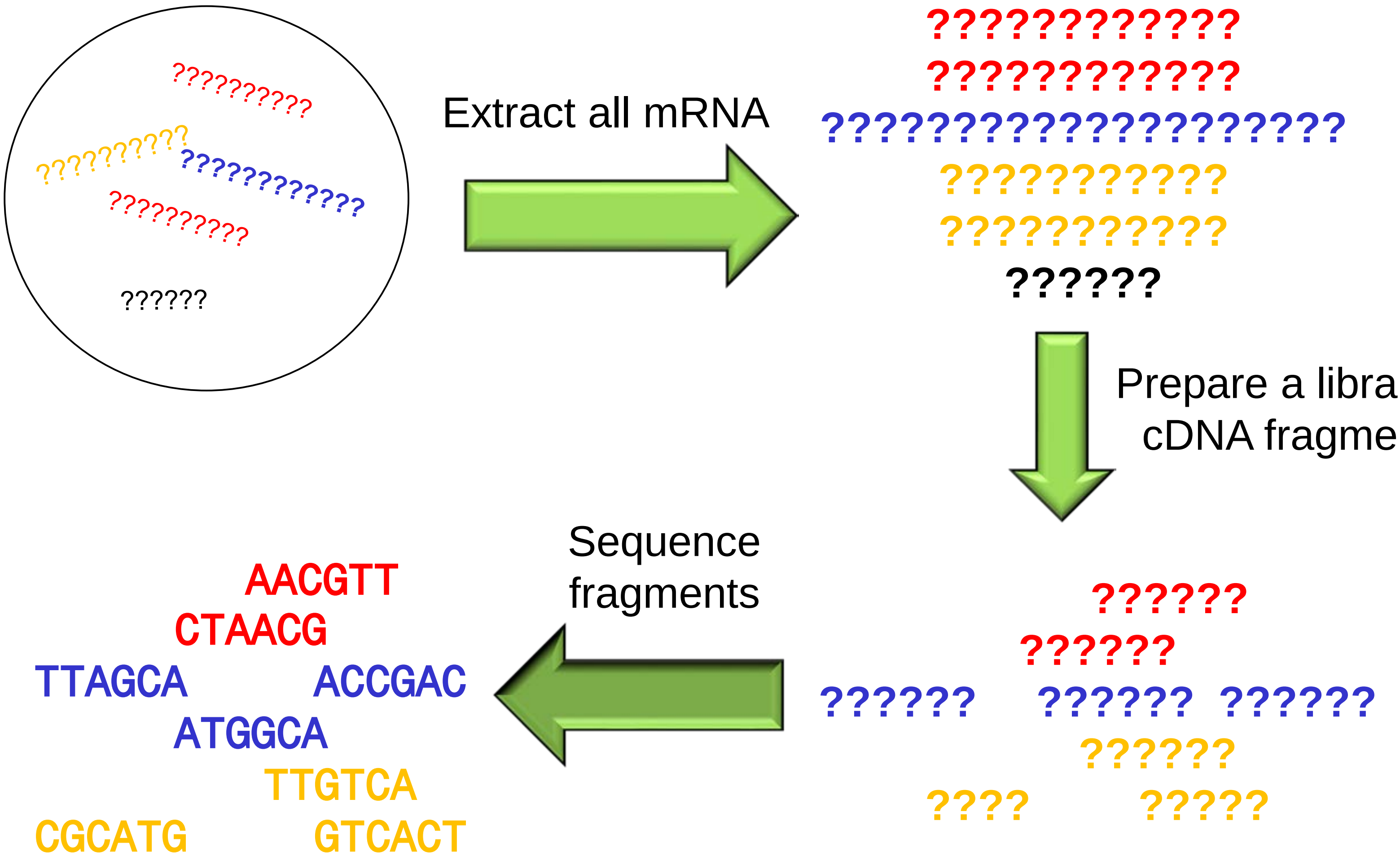
Exon capture transcriptome

Predominant applications:
- Gene expression
- Gene fusions
- Splice variants

http://www.nature.com/nrd/journal/v12/n5/fig_tab/nrd3979_F2.html

Figure 2 of Nature Reviews Drug Discovery 12, 358–369 (2013)

Overview of RNA-seq procedure



??????????
??????????
????????????
??????????
??????

Extract all mRNA

????????????
????????????
????????????????????
???????????
???????????
??????

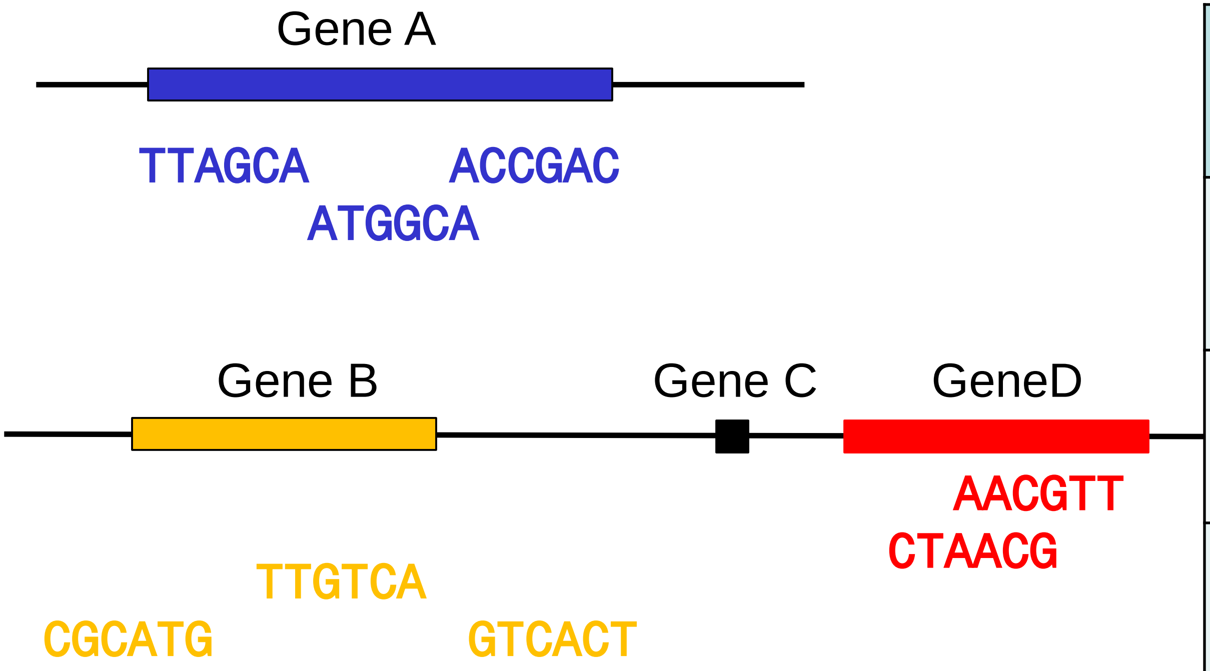
Prepare a library of cDNA fragments

??????
??????
??????   ?????? ??????
??????
????   ?????

Sequence fragments

AACGTT
CTAACG
TTAGCA   ACCGAC
ATGGCA
TTGTCA
CGCATG   GTCACT

Map sequences to genome



Gene ID	T1_rep1
A	3
B	3
C	0
D	2

For a given gene, the number of reads mapped to the gene measures the abundance of its transcripts.

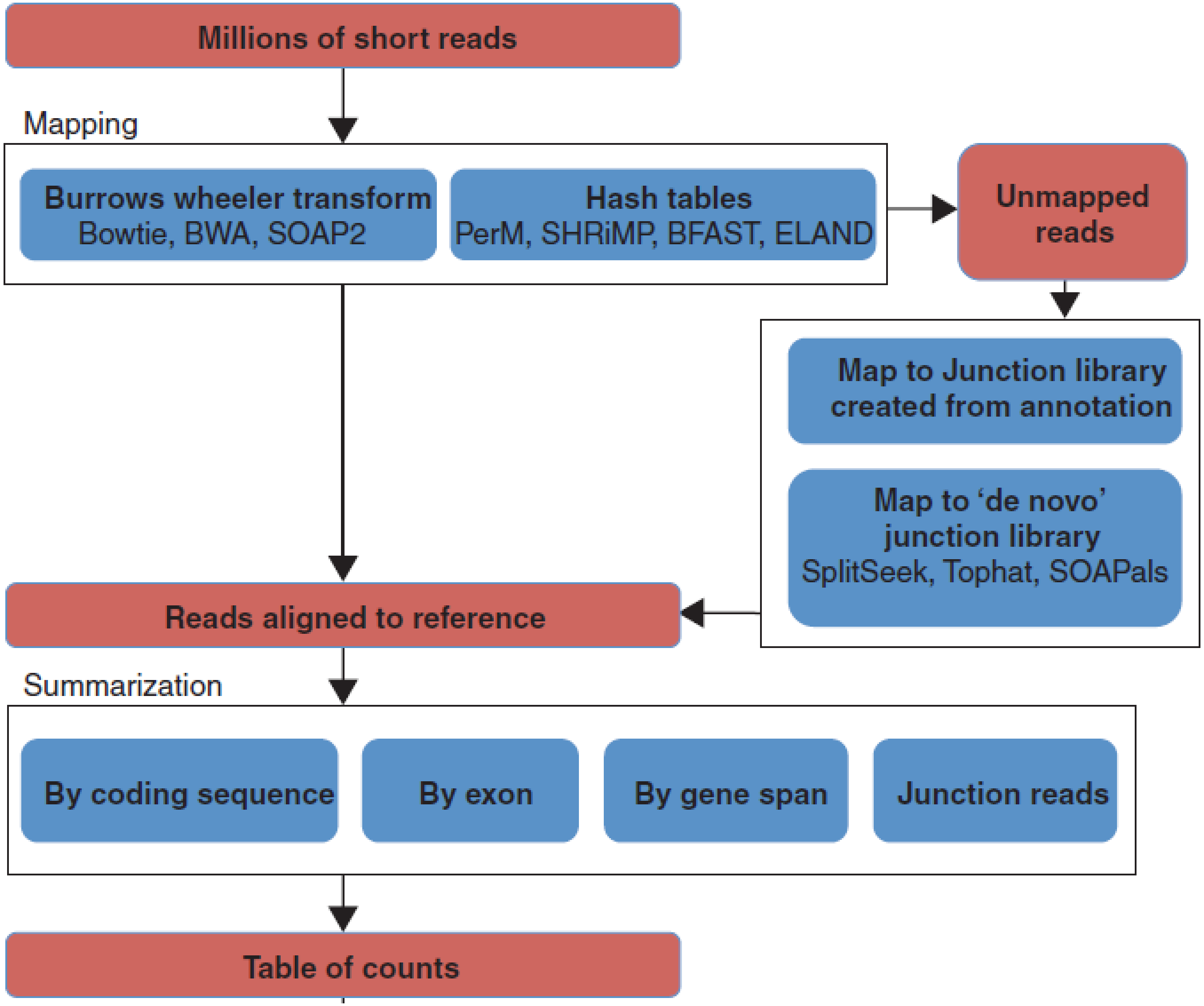
From RNA-seq reads to differential expression results, Oshlack *et al*. *Genome Biology* 2010, **11**:220

Bioinformatics analysis:

Quality monitoring, Base-calling,

Alignment / *de novo* assembly,

Estimating transcript abundance.


Millions of short reads
Mapping
Burrows wheeler transform
Bowtie, BWA, SOAP2
Hash tables
PerM, SHRiMP, BFAST, ELAND
Unmapped reads
Map to Junction library created from annotation
Map to 'de novo' junction library
SplitSeek, Tophat, SOAPals
Reads aligned to reference
Summarization
By coding sequence
By exon
By gene span
Junction reads
Table of counts

Advantages of RNA-seq over microarray

- Not restricted to known genes or genome
- Wider measurable range of expression levels
- Less noisy, low technical variation
- Higher throughput
- Details about transcriptional features
  - Novel transcripts
  - Isoform detection (alternative splicing)
  - Allele-specific expression

Disadvantages of RNA-seq over microarray

- Complex bioinformatics and statistical analysis
- Evolving technologies and analysis methodologies (not as mature)
- Not free of bias
  - Transcript length affects the power of DE detection.
  - Sequence composition etc. may introduce measurement bias.

Bioinformatics analysis pipeline → table of counts

Gene ID	T1_rep1	T1_rep2	…	T2_rep1	…	T2_rep_n
A	3	5	.	23	.	35
B	3	6	.	5	.	2
…	…		.	.	.	…
G	2	0	.	450	.	239

After obtaining the count table

- Statistical analysis:
  - Fit generalized linear models
  - Normalization to remove sources of systematic variation
  - Multiple testing
  - Clustering analysis
  - Gene set testing, etc.

Proposed models for RNA-seq data

- Let Y_{gij} denote the read count mapped to treatment *i*, replicate *j* for a given gene *g* and C_{ij} denote the normalization factor.
- Probability Model 1: Y_{gij} ~ Poisson($C_{ij}\mu_{gij}$)
- Probability Model 2: Y_{gij} ~ *Neg Binomial*($C_{ij}\mu_{gij}$, φ_g)
- Generalized linear model for a gene: $h(\mu_{ij}) = \mu + \tau_i + \delta_j$

  where h(.) is a link function

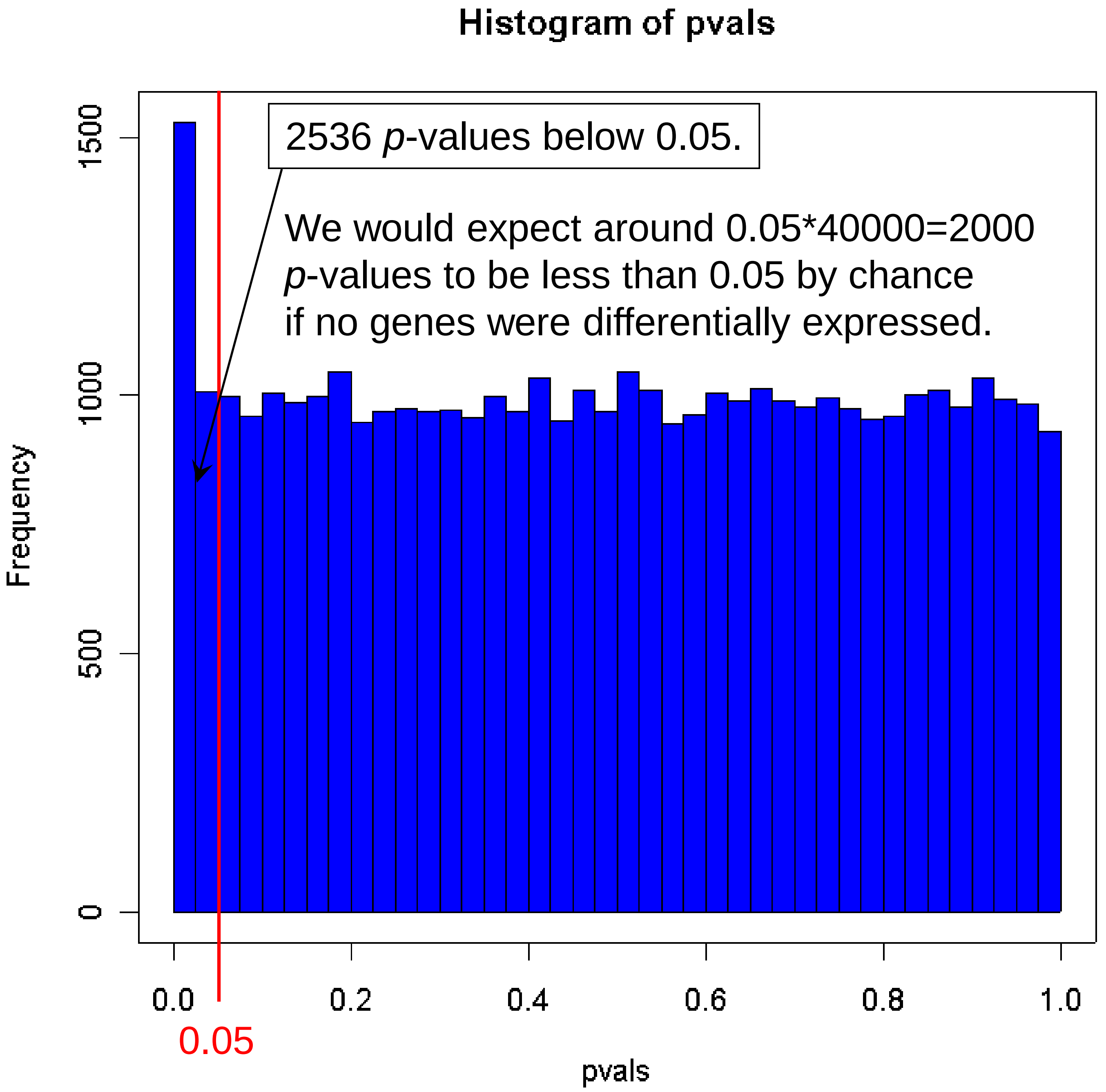
  i: treatment index; j: replicate index
- What are the difference in gene expression ($\tau_1 - \tau_2$)?

  $\tau_1 - \tau_2 \neq 0$ means differential expression.

Detecting differentially expressed genes

- Perform tests for each gene and obtain a p-value.

- For both RNA-seq and microarray data, there is a "small *n*, large *p*" problem due to the relatively few replicates and huge number of genes.
  - Empirical Bayes Tests that borrow information across genes to achieve higher power.

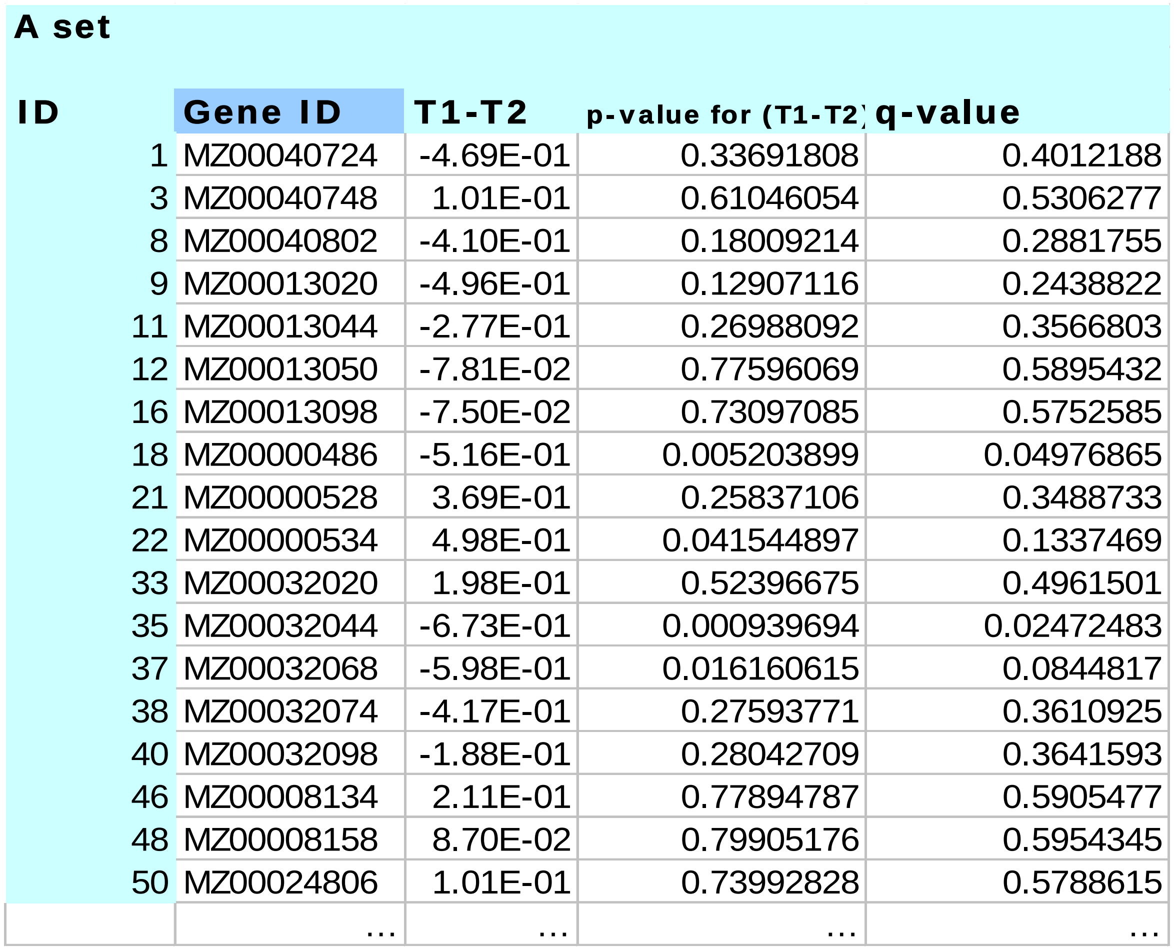


2536 *p*-values below 0.05.

We would expect around 0.05*40000=2000
p-values to be less than 0.05 by chance
if no genes were differentially expressed.

Detecting differentially expressed genes

- Control false discovery rate (FDR) for multiple testing and get a list of differentially expressed genes.



A set

ID	Gene ID	T1-T2	p-value for (T1-T2)	q-value
1	MZ00040724	-4.69E-01	0.33691808	0.4012188
3	MZ00040748	1.01E-01	0.61046054	0.5306277
8	MZ00040802	-4.10E-01	0.18009214	0.2881755
9	MZ00013020	-4.96E-01	0.12907116	0.2438822
11	MZ00013044	-2.77E-01	0.26988092	0.3566803
12	MZ00013050	-7.81E-02	0.77596069	0.5895432
16	MZ00013098	-7.50E-02	0.73097085	0.5752585
18	MZ00000486	-5.16E-01	0.005203899	0.04976865
21	MZ00000528	3.69E-01	0.25837106	0.3488733
22	MZ00000534	4.98E-01	0.041544897	0.1337469
33	MZ00032020	1.98E-01	0.52396675	0.4961501
35	MZ00032044	-6.73E-01	0.000939694	0.02472483
37	MZ00032068	-5.98E-01	0.016160615	0.0844817
38	MZ00032074	-4.17E-01	0.27593771	0.3610925
40	MZ00032098	-1.88E-01	0.28042709	0.3641593
46	MZ00008134	2.11E-01	0.77894787	0.5905477
48	MZ00008158	8.70E-02	0.79905176	0.5954345
50	MZ00024806	1.01E-01	0.73992828	0.5788615
	…	…	…	…

Other analyses

- Cluster analysis

- Relating the gene expressions with biological functional categories → Gene Set Enrichment Test

- Biological validation:
  - Real Time-PCR
  - Other knowledge or experiments?

Reading assignment

- Chapter one of the book: Statistical Analysis of Next Generation Sequencing Data (An earlier version was published on J Proteomics Bioinform. 2010; 3(6): 183–190. doi: 10.4172/jpb.1000138)

- From RNA-seq reads to differential expression results, Oshlack *et al*. *Genome Biology* 2010, **11**:220 http://genomebiology.com/2010/11/12/220

Resources for RNA-seq readings

- Nature.com subject areas on Next-generation sequencing http://www.nature.com/subjects/next-generation-sequencing
- Current Topics in Genome Analysis 2016 https://www.genome.gov/12514286/current-topics-in-genome-analysis-2016/
- Papers in Bioinformatics Journal about NGS data analysis http://www.oxfordjournals.org/our_journals/bioinformatics/nextgenerationsequencing.html