Gene expression and Differentiation

DAVID L. TABB, PH.D.

Overview

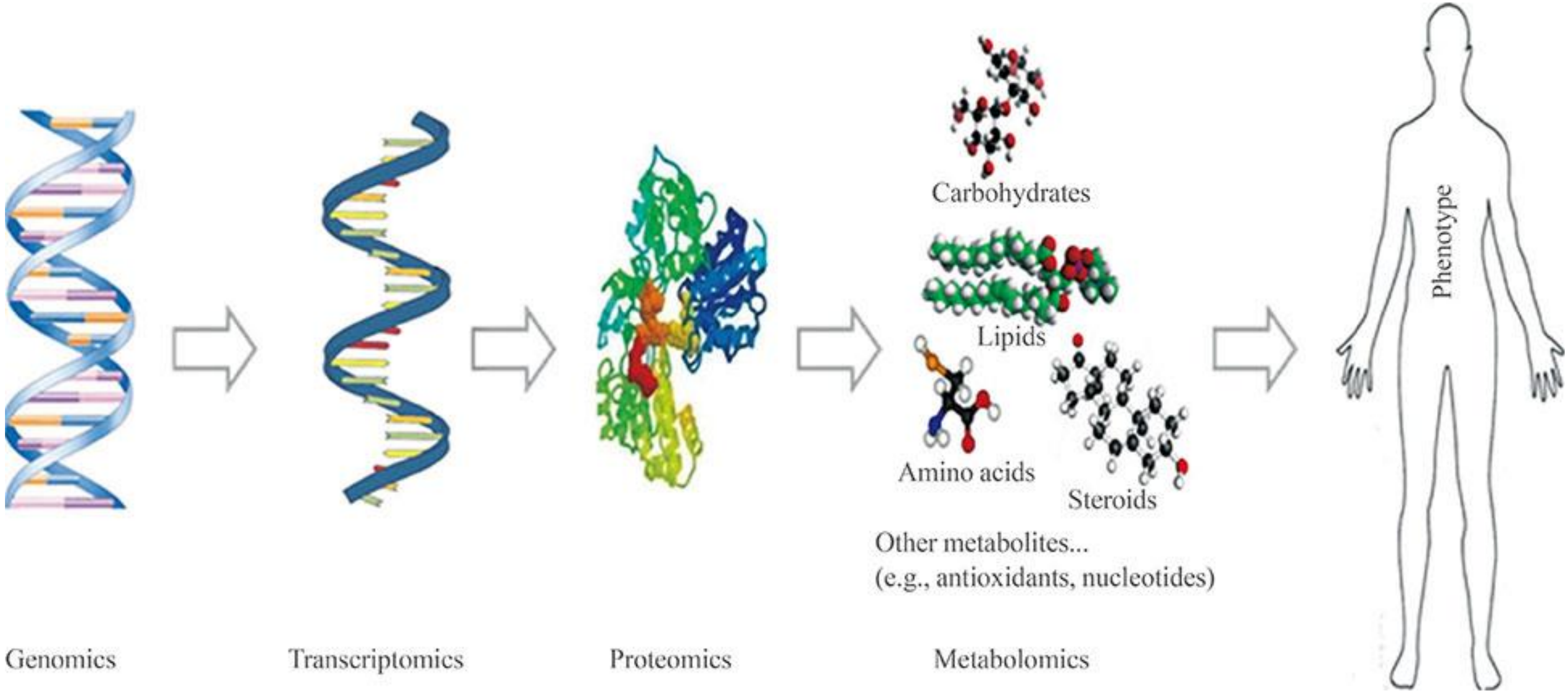
- Why measure transcription?
- Technologies for gene expression:
  - microarrays and beads
  - RNA-Seq via massively-parallel sequencing
- Clustering and difference testing
- Statistical concerns

Gene expression regulation

- Gene expression is highly regulated
  - In mutant cells *vs* wild-type cells
  - In response to stimuli such as drugs, light, or sleep
  - At different developmental stages
  - In different cell types (e.g. muscle cells, fibroblasts)
  - In disease states *vs* healthy
- The number of mRNA copies in a cell for a gene is an indicator of corresponding protein expression level.

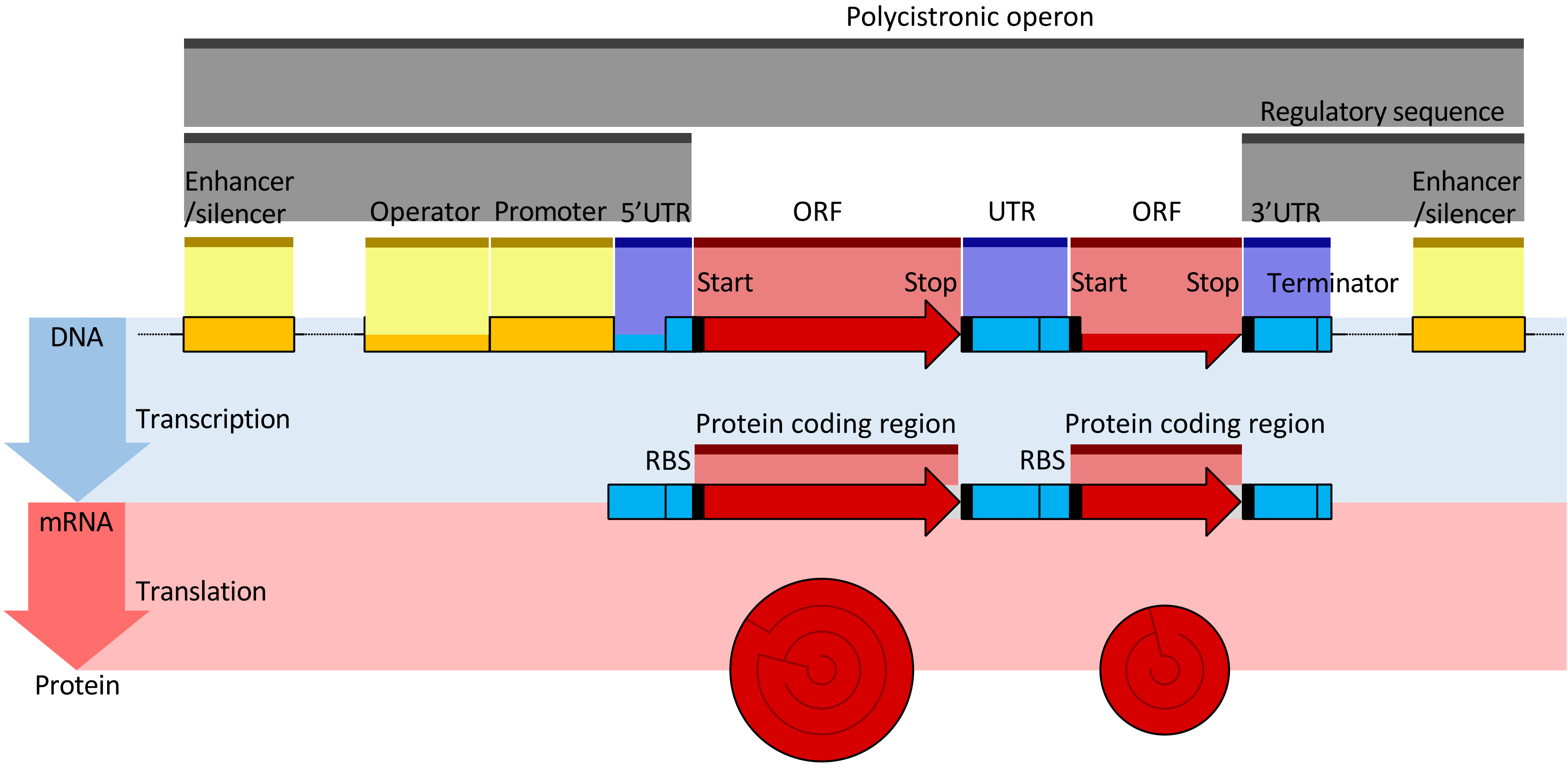
T Lee and RA Young. *Cell* (2013) 152: 1237-1251.

Expression gets us closer to phenotype than genomics.



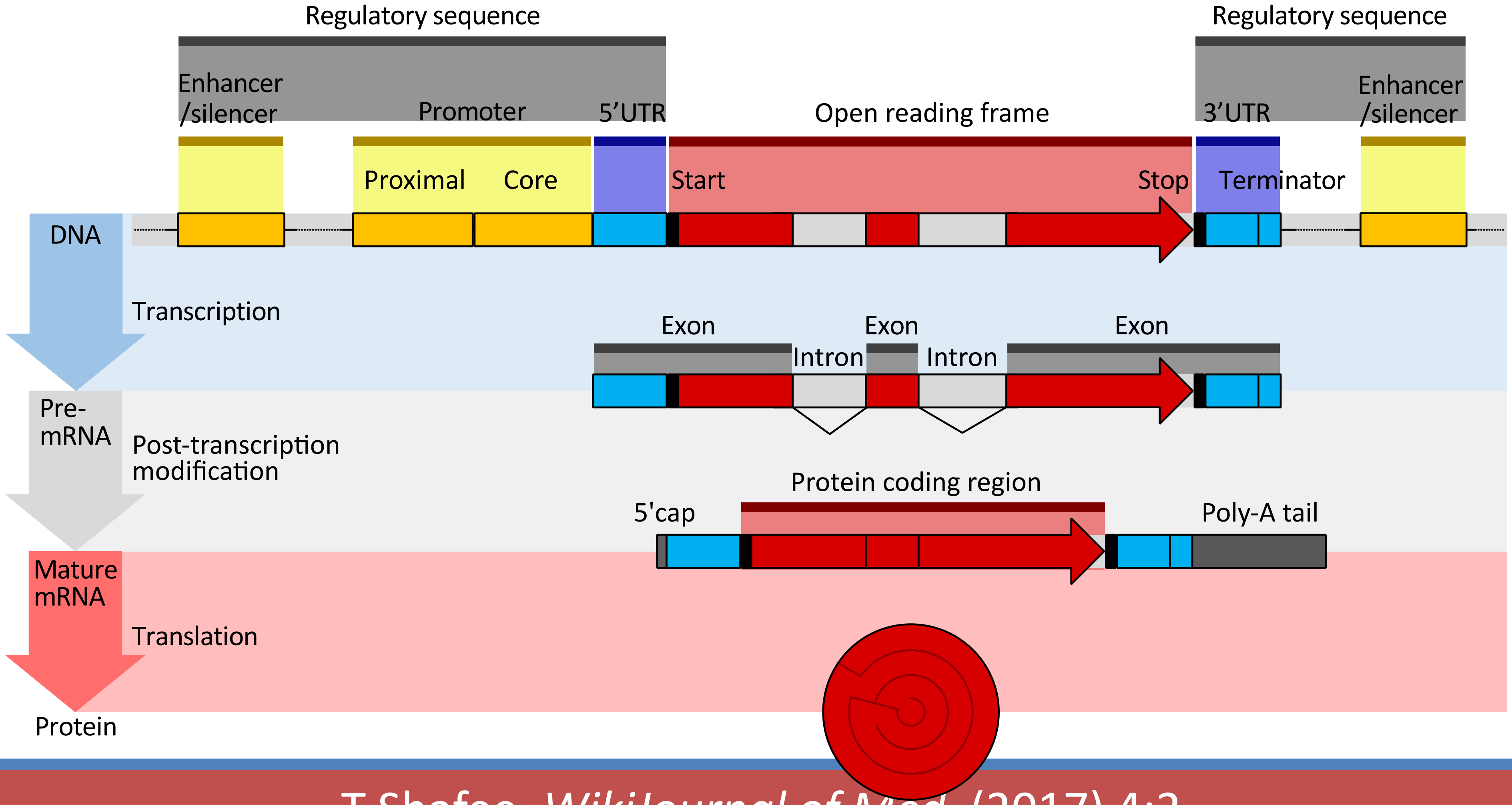
SJ Kim et al. *Endocrinol. Metab*. (2016) 31: 7-16

Operon expression from prokaryotic organisms



T Shafee. *WikiJournal of Med*. (2017) 4:2

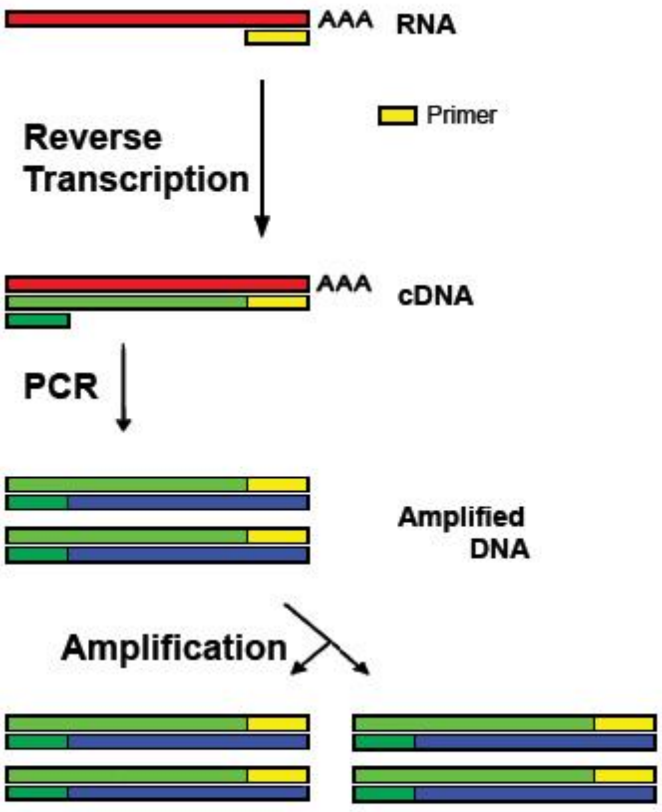
Gene expression from eukaryotic organisms



T Shafee. *WikiJournal of Med*. (2017) 4:2

cDNA produced to serve as more stable analyte

- RNAse is everywhere so producing DNA complement early preserves info.
- Sparse samples may require PCR amplification.



Wikipedia: Jpark623

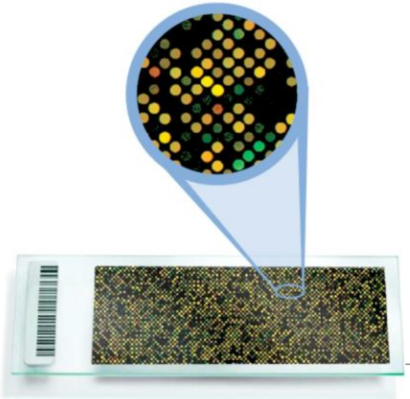
Technologies

High-throughput transcriptome profiling

- **Transcriptome:** set of messenger RNA molecules (“transcripts”) produced in cells
- **Hybridization based approaches:** incubate fluorescently labeled cDNA with microarrays. Intensity reflects abundance.
- **Sequencing based approaches:** directly determine the cDNA sequences. Read count reflects abundance.

Hybridization produces hydrogen bonds between complementary DNA sequences

DNA microarrays



- **DNA microarray:** a solid support (glass slide, silicon chip, etc) on which DNA of known sequence is deposited in a regular grid-like array.

- **Spotted or printed arrays:** DNA feature physically transferred from a plate or reservoir and transferred to a solid support, typically a chemically modified glass microscope slide.

- **Synthesized arrays:** DNA features chemically synthesized *in-situ* on the substrate.

Invented by Stephen Fodor and by Edward Southern

Photolithography manufactures high-density microarrays

- Start with chip bearing primers on which DNA probes can be assembled.
- Shine laser through a mask to deprotect specific spots.
- Add free nucleotides
- Repeat to desired length (e.g. 25 bases)

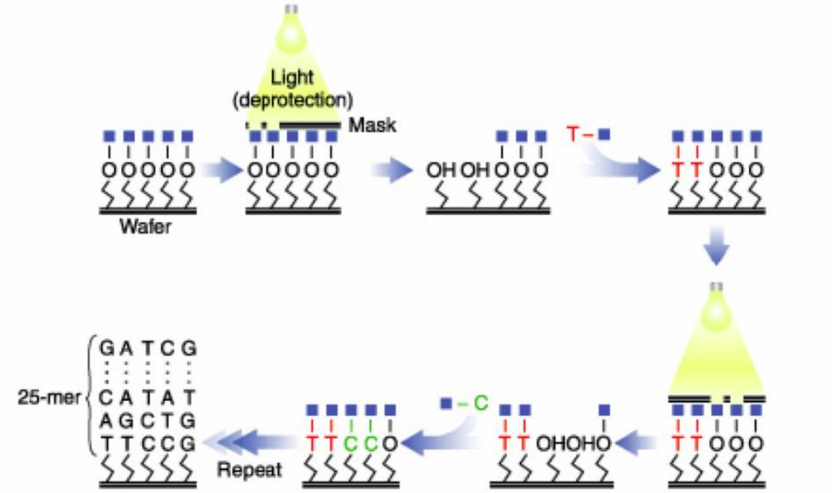
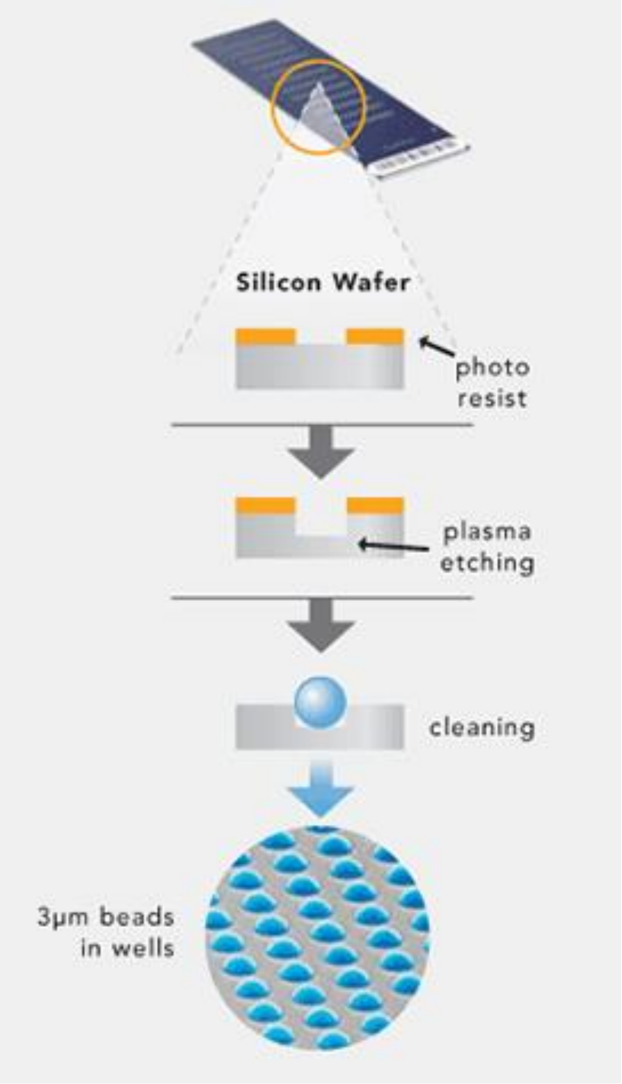


Image from Affymetrix

Bead arrays allow flexible probe palettes

- Each bead is a lawn of the same probe sequences.
- Each bead reports which probes it is wearing.
- Beads are trapped into a grid of wells for reading.
- Multiple beads are measured for each transcript.

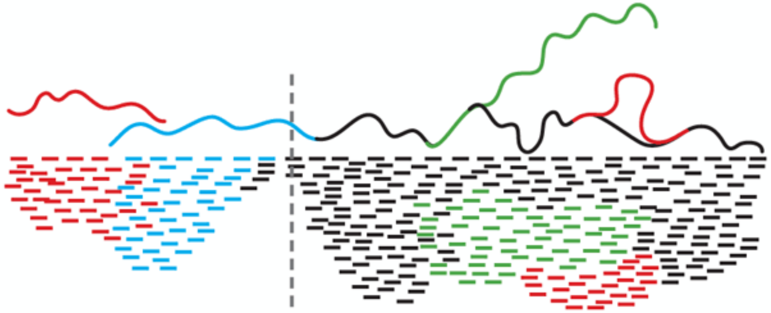


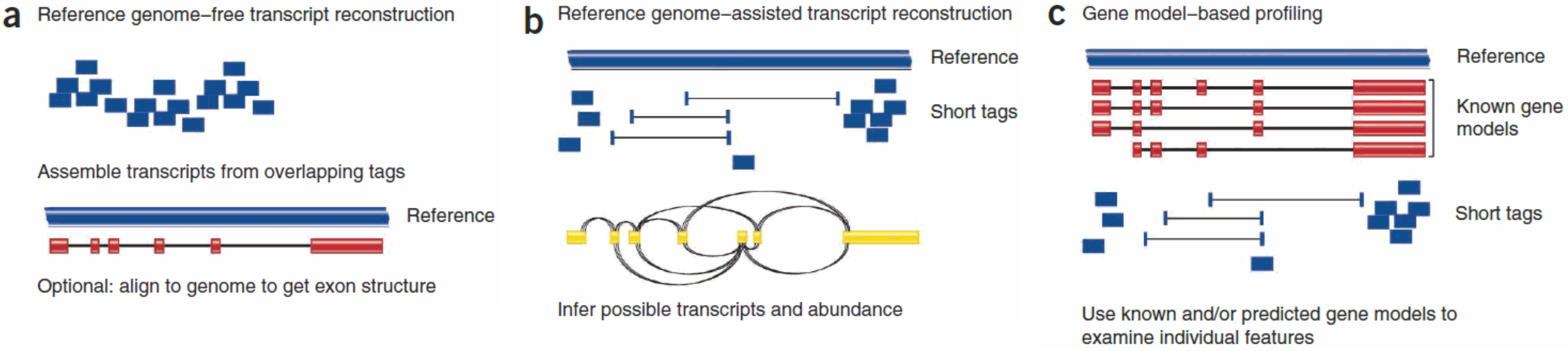
For an example, see GSE60438.

https://emea.illumina.com/science/technology/beadarray-technology.html

RNA-Seq as alternative

Reads throughout transcripts may be assembled to improve gene models or simply aligned to known annotation.

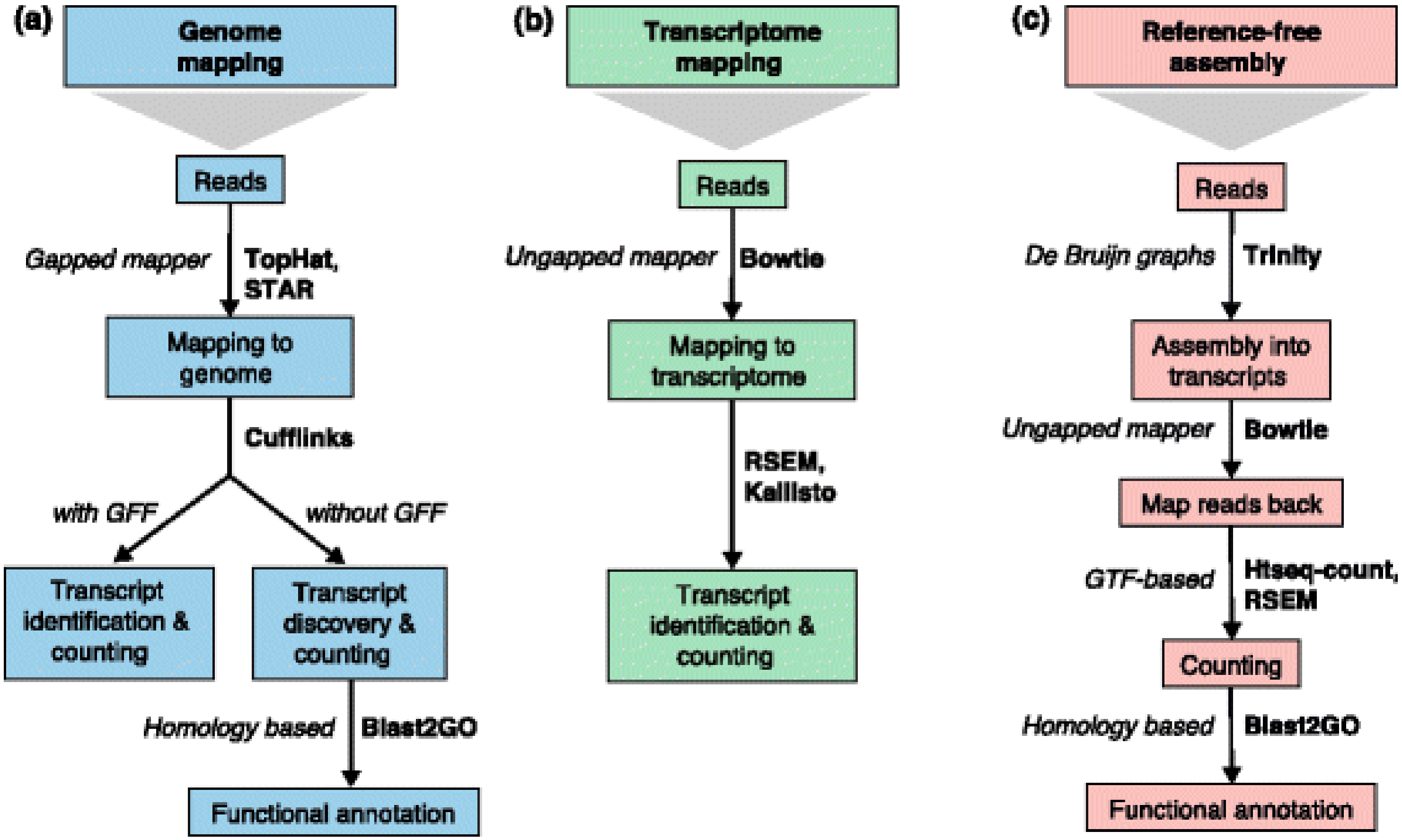



a
Reference genome–free transcript reconstruction
Assemble transcripts from overlapping tags
Reference
Optional: align to genome to get exon structure
b
Reference genome–assisted transcript reconstruction
Reference
Short tags
Infer possible transcripts and abundance
c
Gene model–based profiling
Reference
Known gene models
Short tags
Use known and/or predicted gene models to examine individual features

Top image from Iyer and Chinnaiyan, *Nat. Biotech*. (2011) 29: 599-600

Lower set from Cloonan and Grimmond. *Nat. Methods* (2010) 793-795.

Goals determine your path

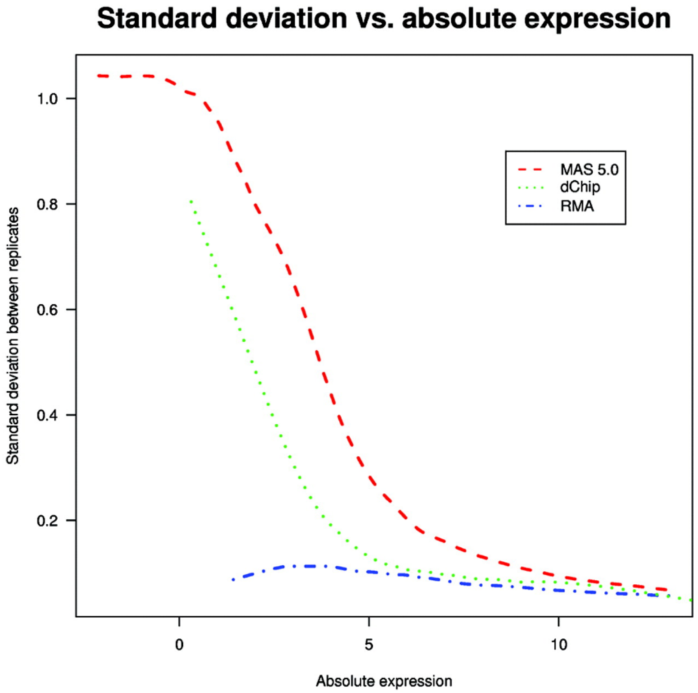


A. Conesa et al. *Genome Biol*. (2016) 17: 13

Data analysis

From fluorescent probe intensity to mRNA quantity

- The more probes hybridized to sample cDNA, the brighter the fluorescence.
- Dim spots have more error.
- Robust Multi-Array Analysis (RMA) adjusts log intensities among arrays as a combo of expression, probe affinity, and measurement error.



RA Irizarry et al. *Nucl. Acids Res*. (2003) 31: e15

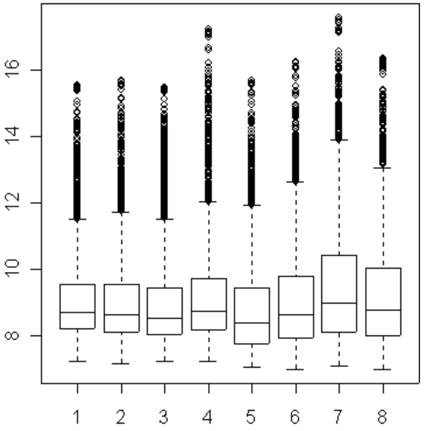
RNA-Seq: reads per kilobase of exon model per million reads

- numReads = number of sequencing reads that map to a particular transcript
- geneLength = length of the transcript sequence (~2000 is common human value)
- totalNumReads= the number of sequencing reads that mapped to any transcript
- $RPKM = \frac{numReads}{\frac{geneLength}{1000} * \frac{totalNumReads}{1000000}}$

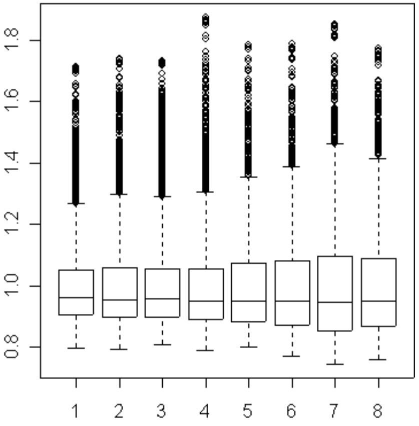
FPKM counts distinct DNA pieces for paired-end experiments.

Normalization (among arrays)

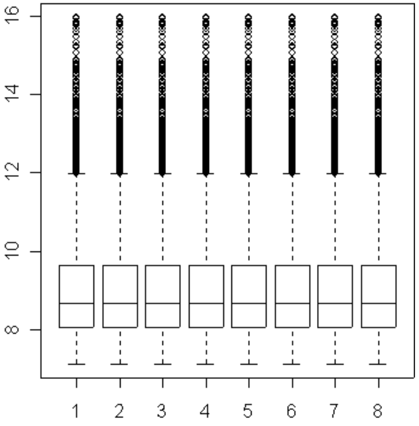
- Adjust the arrays using “housekeeping genes” (not recommended)
- Multiply each array by a constant to make the median intensity the same for each individual array (Global normalization)
- Match the percentiles of each array (Quantile normalization)



Without normalization



Global normalization

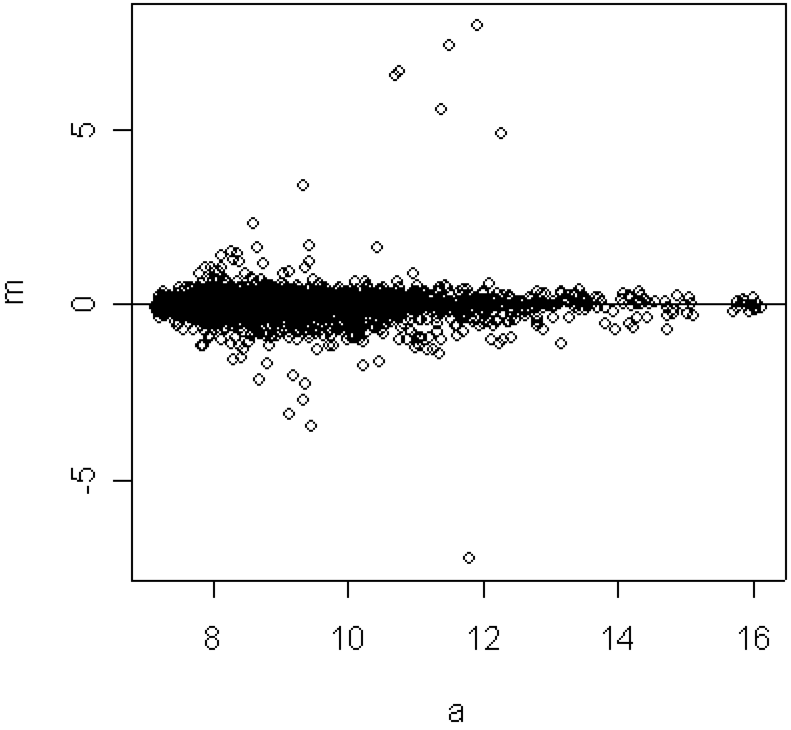


Quantile normalization

J.H. Bullard et al. *BMC Bioinformatics* (2010) 11:94

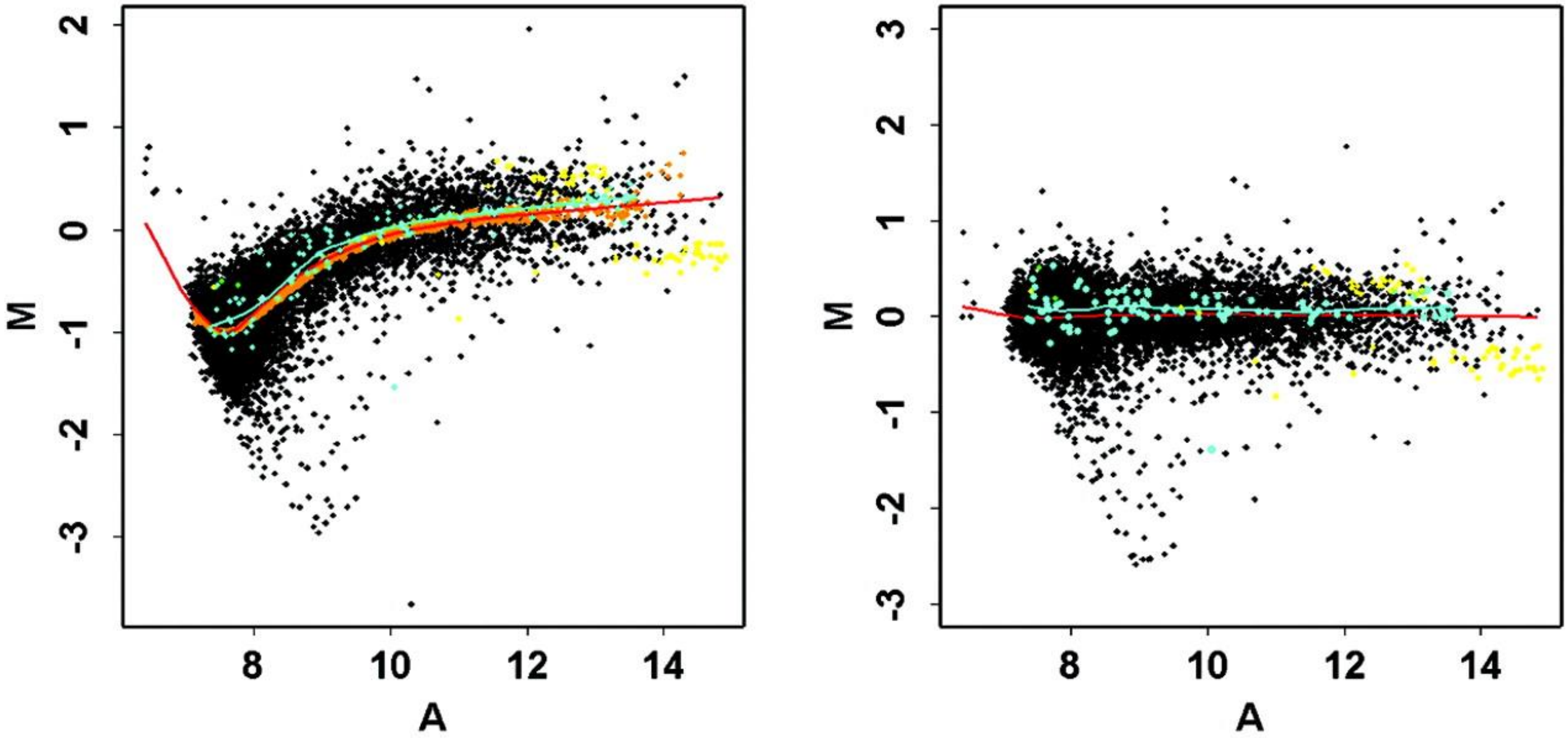
The MA or Bland-Altman plot

M shows the log difference in signal between two sets for a given gene. *A* shows the log average intensity for a given gene. Distortion from the center represents a bias.



http://www.pmean.com/99/arrayNormalization.htm

MA: example of negative bias at low magnitude



intensity log ratios

Average log abundance

after normalization

Y.H. Yang et al. *Nucl. Acids Res*. (2002) 30: e15. Figure 3

Batch Effects

- Instrument performance varies with time.
- Data acquired on samples in week 1 generally differs from data acquired on the same samples in week 2.
- When data must be acquired in different batches from an instrument:
  - Randomly distribute samples to batches.
  - Include the batch in the statistical model.
  - Run some samples in every batch.

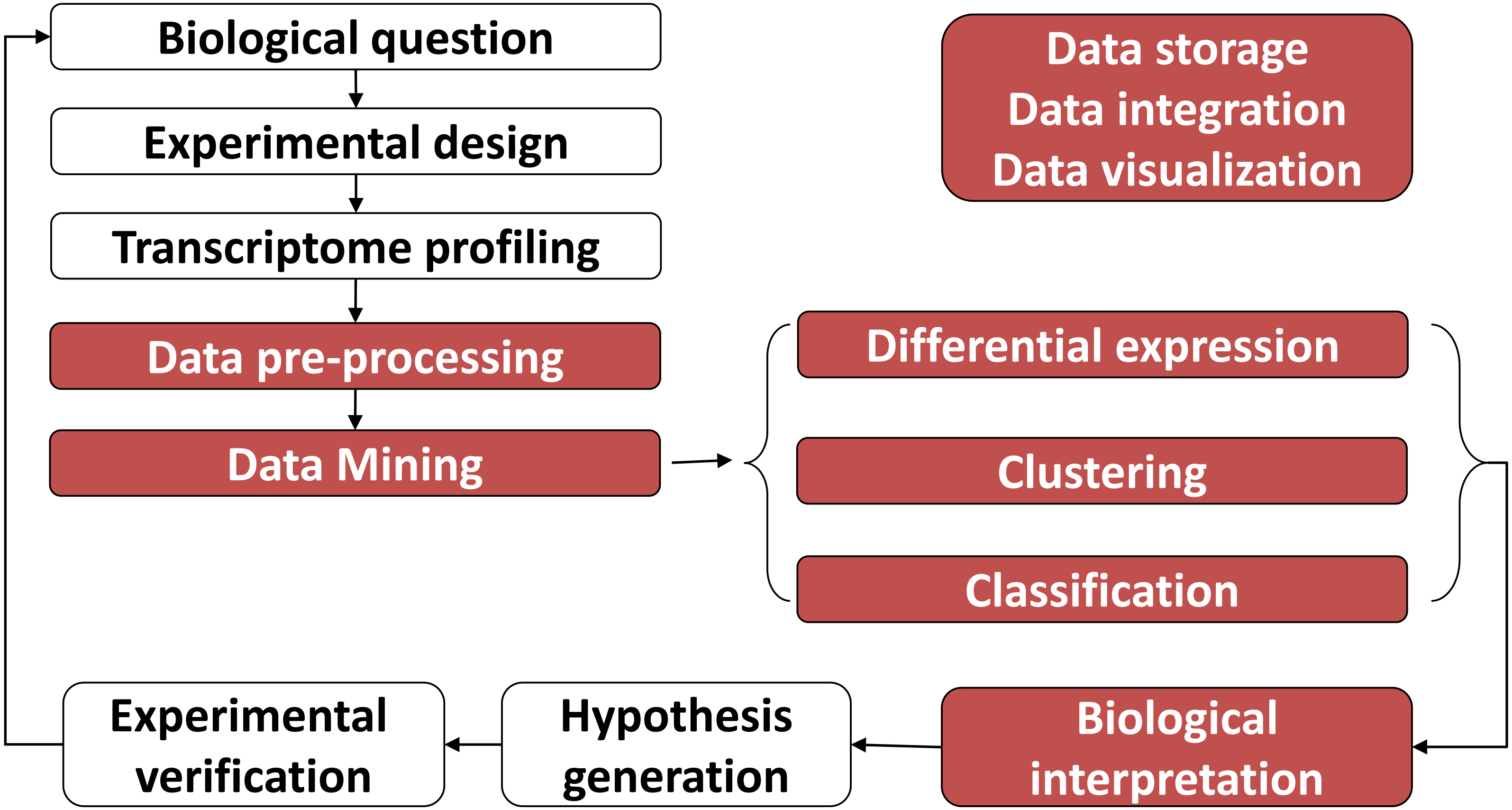
W.E. Johnson et al, *Biostatistics* (2007) 8: 118-127.

Batch effects will ruin your day.

- "They ran all the controls on one day and all the cancers on the next day," Dr. Baggerly said. "This is the worst kind of design when you are using a machine that can be subject to external factors," such as changes in calibration or mechanical breakdown.

https://www.mdedge.com/obgynnews/article/51470/gynecology/controversy-erupts-over-proteomics-studies

Bioinformatics tasks



Three major goals of gene expression studies

Class comparison (what genes differentiate?)

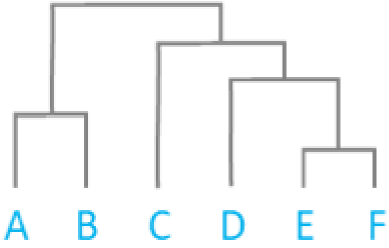
1. Differential expression analysis
2. Input: gene expression data, class label of the samples
3. Output: differentially expressed genes

Class detection (which samples are similar?)

1. Biclustering analysis
2. Input: gene expression data
3. Output: groups of similar samples or genes

Class prediction (which genes predict outcome?)

1. Machine learning techniques
2. Input: training set (expression plus class labels)
3. Output: prediction model with test set evaluation



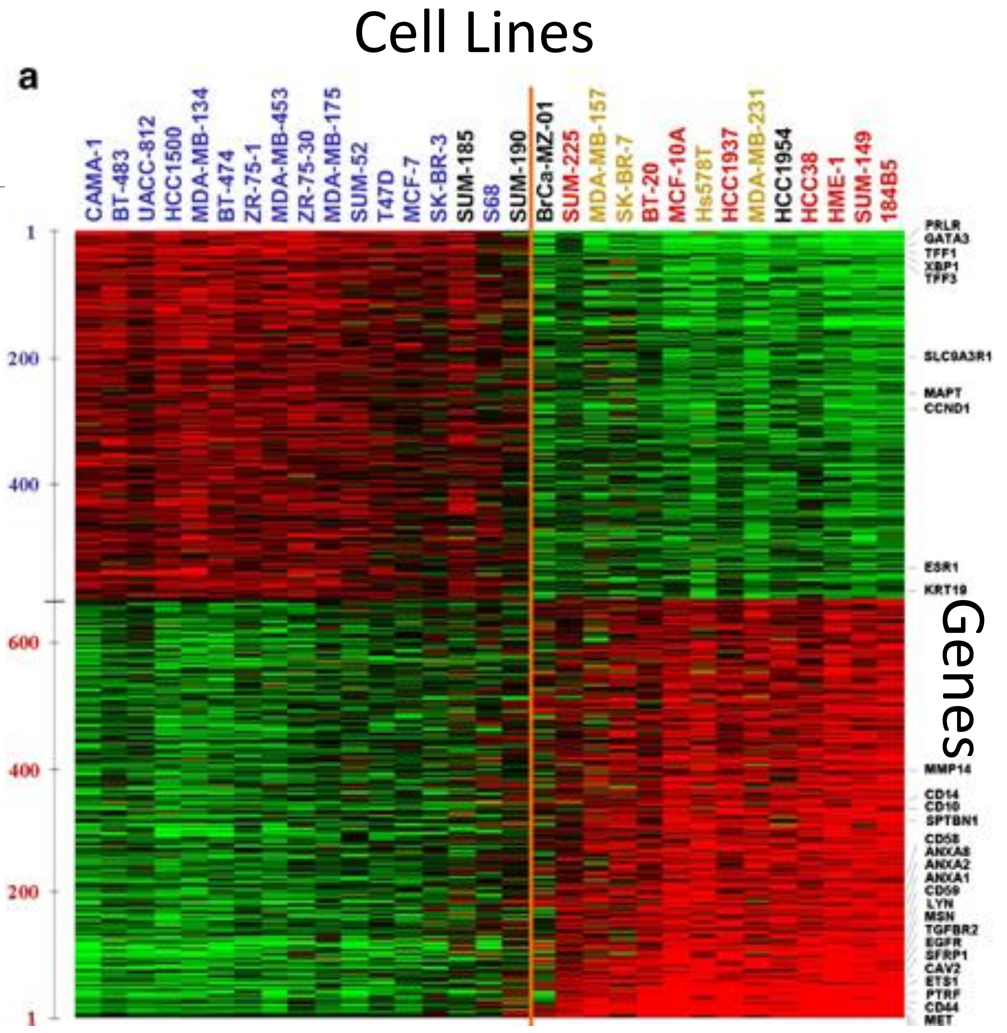
What is clustering?

- Represents a data mining technique to infer a *hierarchy* joining a group of data points.
- Requires a *distance metric*: a strategy such as Euclidean distance for computing the distance between two data points.
- Employs either an *agglomerative* (grouping together) or *divisive* (splitting apart) plan.
- Most *dendrograms* can be cut at various levels to produce desired number of groups.

http://www.informit.com/articles/article.aspx?p=357695&seqNum=4

Biclustering

"a method that simultaneously clusters genes and conditions, finding distinctive ***checkerboard*** patterns in matrices of gene expression data, if they exist."



Y Cheng and GM Church. *ICISMB* (2000)

10.1038/sj.onc.1209254

Y Kluger et al. *Genome Research* (2003) 13:703-716

High performance algorithms

- BBC (Bayesian BiClustering)
  - J.Gu & J.S. Liu, *BMC Genomics* (2008) 9:S4
- Plaid model (overlapping layers)
  - L. Lazzeroni & A. Owen, *Statistica Sinica* (2002) 12: 61-86.
- CPB (Coherent Pattern Bicluster)
  - D. Bozdağ et al. *Bioinfo. & Computat. Bio.* (2009) 151-163.
- QUBIC (QUalitative BIClustering)
  - G. Li et al. *Nucl. Acids Res*. (2009) 37: e101

K. Eren et al. *Briefings Bioinfo*. (2012) 14: 279-292.

Expression is measured for each replicate in both cohorts

Samples

Genes

probe_set_id	HNE0_1	HNE0_2	HNE0_3	HNE60_1	HNE60_2	HNE60_3
1007_s_at	8.6888	8.5025	8.5471	8.5412	8.5624	8.3073
1053_at	9.1558	9.1835	9.4294	9.2111	9.1204	9.2494
117_at	7.0700	7.0034	6.9047	9.0414	8.6382	9.2663
121_at	9.7174	9.7440	9.6120	9.7581	9.7422	9.7345
1255_g_at	4.2801	4.4669	4.2360	4.3700	4.4573	4.2979
1294_at	6.3556	6.2381	6.2053	6.4290	6.5074	6.2771
1316_at	6.5759	6.5330	6.4709	6.6636	6.6438	6.4688
1320_at	6.5497	6.5388	6.5410	6.6605	6.5987	6.7236
1405_i_at	4.3260	4.4640	4.1438	4.3462	4.3876	4.6849
1431_at	5.2191	5.2070	5.2657	5.2823	5.2522	5.1808
1438_at	7.0155	6.9359	6.9241	7.0248	7.0142	7.0971
1487_at	8.6361	8.4879	8.4498	8.4470	8.5311	8.4225
1494_f_at	7.3296	7.3901	7.0886	7.2648	7.6058	7.2949
1552256_a_at	10.6245	10.5235	10.6522	10.4205	10.2344	10.3144
1552257_a_at	10.3224	10.1749	10.1992	10.2464	10.2191	10.2405

Case

Control

Each row comprises a separate test of differences between means.

Do all six expression values come from the same distribution?

Determine probability of a more extreme test statistic

- Under *null hypothesis* (H_0), we assume two distributions have the same mean.
- Under this assumption, we ask how probable a higher or lower test statistic would be *by random chance*.
- Gosset described the t-distribution in 1908; we can compute these probabilities! We call them *p-values*.

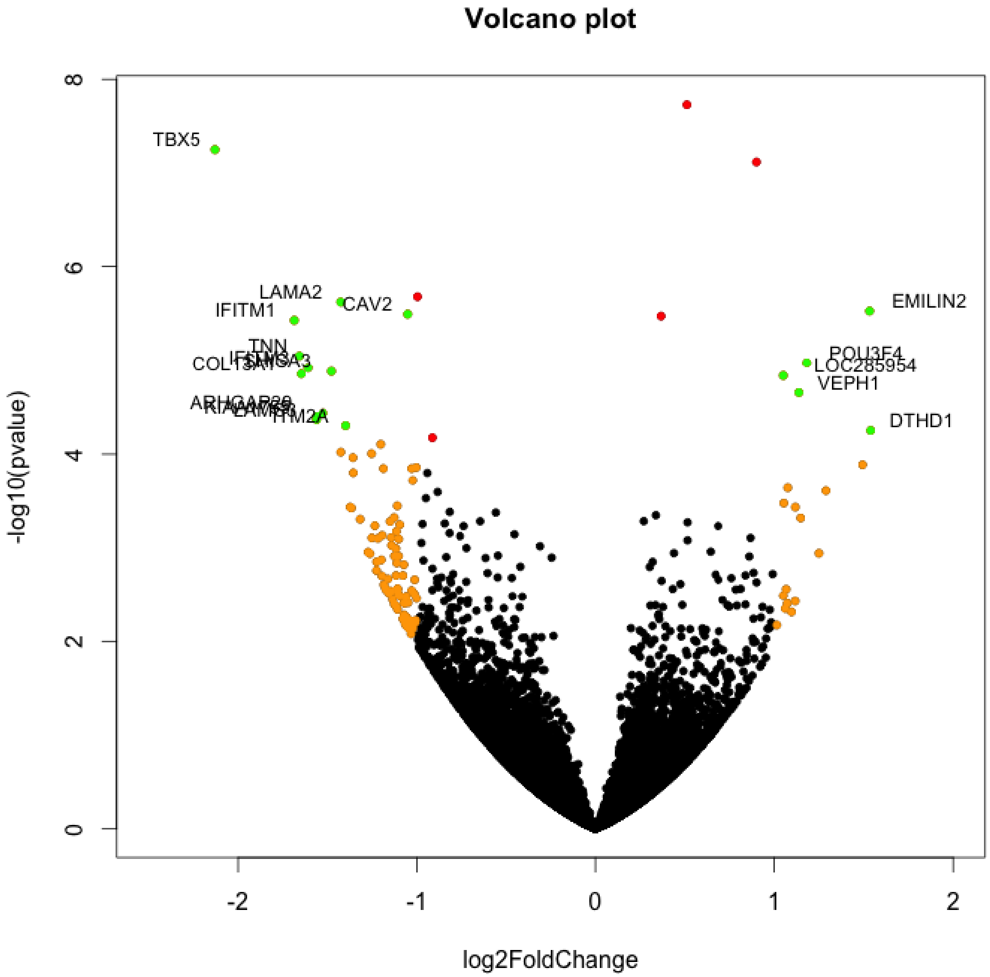
http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-t-tests-t-values-and-t-distributions

Key concepts in difference testing

- Paired: If you have two snapshots of each sample (say, before and after), each value in one cohort pairs with one value in the other.
- One-Sided or Two-Sided: If you hypothesized that values will rise in B than in A rather than fall, use a one-sided test. A two-sided test thinks both increases and decreases are important changes.

The ubiquitous volcano plot

Genes near the center are relatively small fold changes. Genes near the bottom do not yield significant p-values. The green genes are those with both significant p-values and meaningful fold changes.



http://www.gettinggeneticsdone.com/2014/05/r-volcano-plots-to-visualize-rnaseq-microarray.html

Why do we need Multiple Testing Correction?

Each row is a test

probe_set_id	HNE0_1	HNE0_2	HNE0_3	HNE60_1	HNE60_2	HNE60_3
1007_s_at	8.6888	8.5025	8.5471	8.5412	8.5624	8.3073
1053_at	9.1558	9.1835	9.4294	9.2111	9.1204	9.2494
117_at	7.0700	7.0034	6.9047	9.0414	8.6382	9.2663
121_at	9.7174	9.7440	9.6120	9.7581	9.7422	9.7345
1255_g_at	4.2801	4.4669	4.2360	4.3700	4.4573	4.2979
1294_at	6.3556	6.2381	6.2053	6.4290	6.5074	6.2771
1316_at	6.5759	6.5330	6.4709	6.6636	6.6438	6.4688
1320_at	6.5497	6.5388	6.5410	6.6605	6.5987	6.7236
1405_i_at	4.3260	4.4640	4.1438	4.3462	4.3876	4.6849
1431_at	5.2191	5.2070	5.2657	5.2823	5.2522	5.1808
1438_at	7.0155	6.9359	6.9241	7.0248	7.0142	7.0971
1487_at	8.6361	8.4879	8.4498	8.4470	8.5311	8.4225
1494_f_at	7.3296	7.3901	7.0886	7.2648	7.6058	7.2949
1552256_a_at	10.6245	10.5235	10.6522	10.4205	10.2344	10.3144
1552257_a_at	10.3224	10.1749	10.1992	10.2464	10.2191	10.2405

Table by Bing Zhang

We “detect” differences even when none exist.

- A T-test yields a p-value below 0.05 for one in twenty tests when no difference exists.
- This occurs because p-values in random data are uniformly distributed.
- If you perform 1,000 T-tests, you should expect that 50 will be “significant” by random chance alone.

http://haohanw.blogspot.com/2013/12/ml-p-value-null-hypothesis-uniform.html

Multiple testing correction

- Protect against *any* errors: Bonferroni

$$Threshold = \frac{0.05}{NumTrials}$$

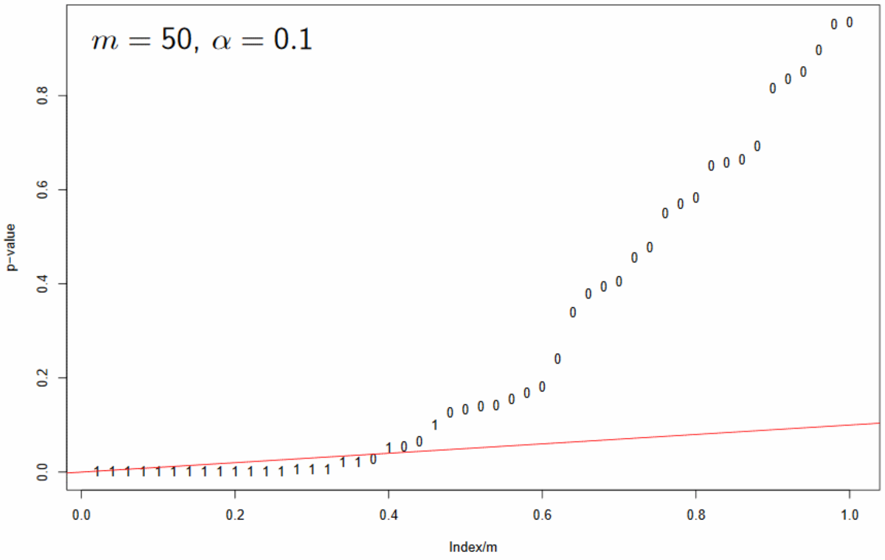
- Limit the *rate* of false positives: Benjamini-Hochberg

(first, sort p-values)

$$Threshold_1 = \frac{1}{NumTrials} * 0.05$$

$$Threshold_2 = \frac{2}{NumTrials} * 0.05$$

$$Threshold_3 = \frac{3}{NumTrials} * 0.05$$



Common multiple testing complaints

- "All my hits vanished after MTC."
  - Maybe the null distribution was correct!
  - Maybe your experiment was underpowered.
- "Dr. X published without MTC, so why must I?"
  - People publish bad statistics all the time. *Don't be one of them*.
- "I did five groups, so I just compared between each pair of groups."
  - Comparing between all pairs in five groups yields ten comparisons; you must correct!

$$\frac{n(n-1)}{2} = \frac{5 * 4}{2} = 10$$

Performing your experiment took caution; analyzing it well also requires caution.

"If you torture the data long enough, it will confess."

https://en.wikiquote.org/wiki/Ronald_Coase

Takeaway messages

- Microarrays are giving way to RNA-Seq for gene expression measurement due to increased flexibility.
- Statistical considerations are inextricable from bioinformatics processing.
- Visualizing data in biclusters and volcano plots is very common in gene expression studies.
- Systems biology can easily yield situations where multiple testing is a problem. Bonferroni and B-H FDR can help!