Differential Gene Expression

Biol4230 Tues, March 27, 2018

Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

- The basics
  - the central dogma
  - all cells have the same genome
  - but tissues express very different genes
  - why?
- Measuring mRNA expression
  - the very old days: translation and hybridization
  - hybridization to SAGE to microarrays
  - RNA-seq
- Differential gene expression
  - abundance vs expression differences

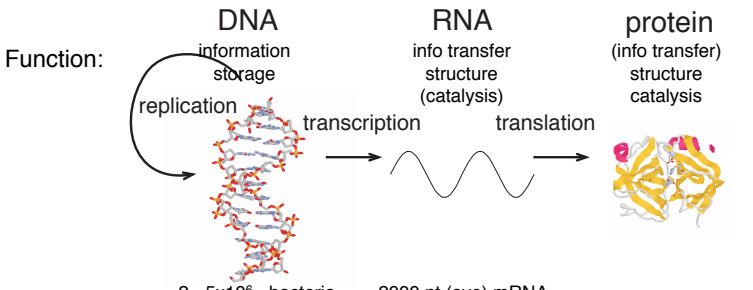
To learn more:

1. Pevsner, Chapter 8 pp. 296-323
2. Recombinant DNA, Chapter 13

3. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282,** 699–705 (1998)
4. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10,** 57–63 (2009)
5. Lovén, J. *et al.* Revisiting global gene expression analysis. *Cell* **151,** 476–482 (2012).

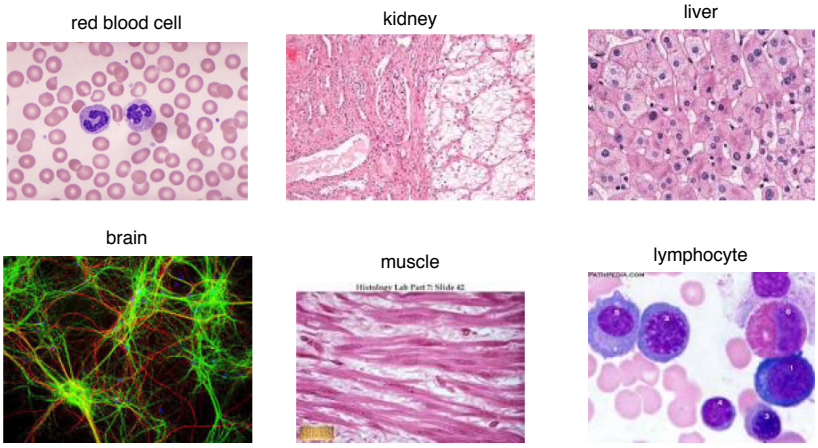
The Central Dogma of Molecular Biology
Molecules for Information transfer, storage, and function

Genetics / Bioinformatics
Template-driven synthesis



	DNA	RNA	protein
Size range:	2 - 5x10^6 - bacteria 18 x 10^6 - yeast 3x10^9 - mammals	2000 nt (ave) mRNA 2000, 5000nt rRNA 100 - 150nt tRNA, 5S	40 kb (400 aa) ave 50 - 30,000 aa
Abundance: (copies/cell)	1 (haploid) 2 (diploid)	>95% rRNA (structural) <5% mRNA < 1 – 100,000+	structural proteins enzymes signaling molecules

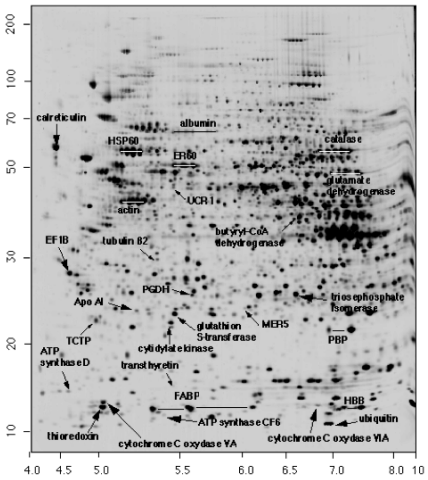
Cells in different tissues are different



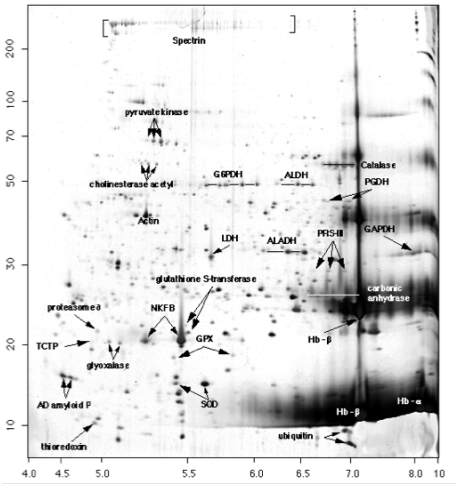
because they express different proteins from different mRNAs

Cells in different tissues are different

human HEPG2 (liver cell line)



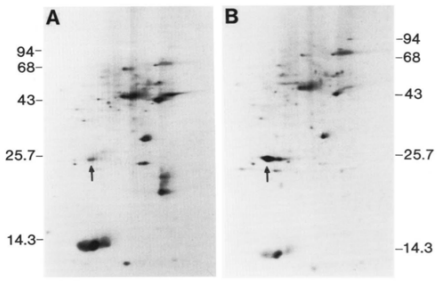
human red blood cell



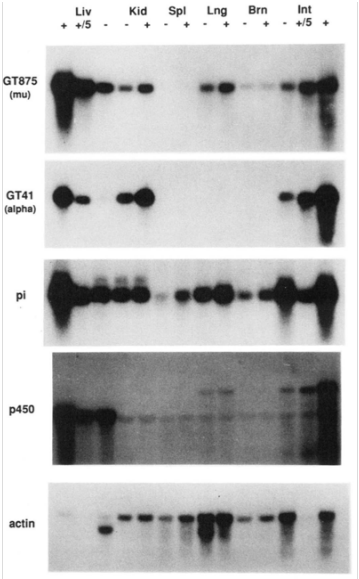
because they express different proteins from different mRNAs

induction of detoxification gene mRNAs

Liver Proteins

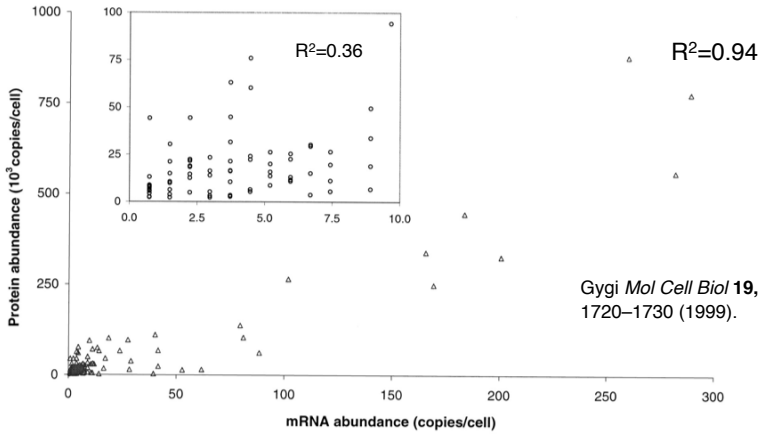


Pearson, W. R. et al *J Biol Chem* **258,** 2052–2062 (1983).



Pearson, W. R. et al. *J Biol Chem* **263,** 13324–13332 (1988).

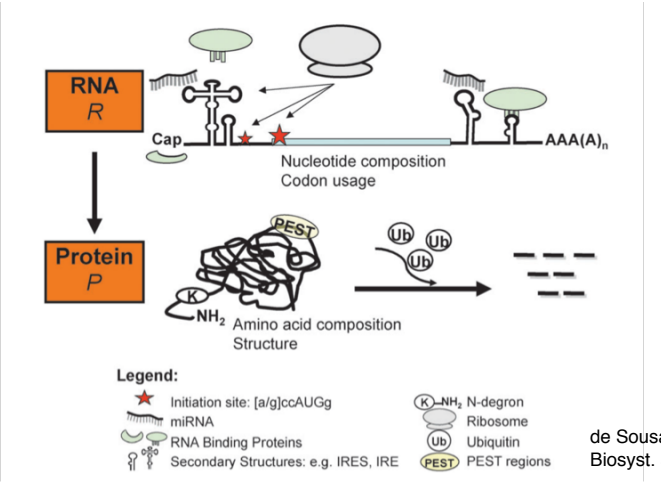
Protein abundance and RNA abundance



Gygi *Mol Cell Biol* **19,** 1720–1730 (1999).

FIG. 5. Correlation between protein and mRNA levels for 106 genes in yeast growing at log phase with glucose as a carbon source. mRNA and protein levels were calculated as described in Materials and Methods. The data represent a population of genes with protein expression levels visible by silver staining on a 2D gel chosen to include the entire range of molecular weights, isoelectric focusing points, and staining intensities. The inset shows the low-end portion of the main figure. It contains 69% of the original data set. The Pearson product moment correlation for the entire data set was 0.935. The correlation for the inset containing 73 proteins (69%) was only 0.356.

What determines protein levels?

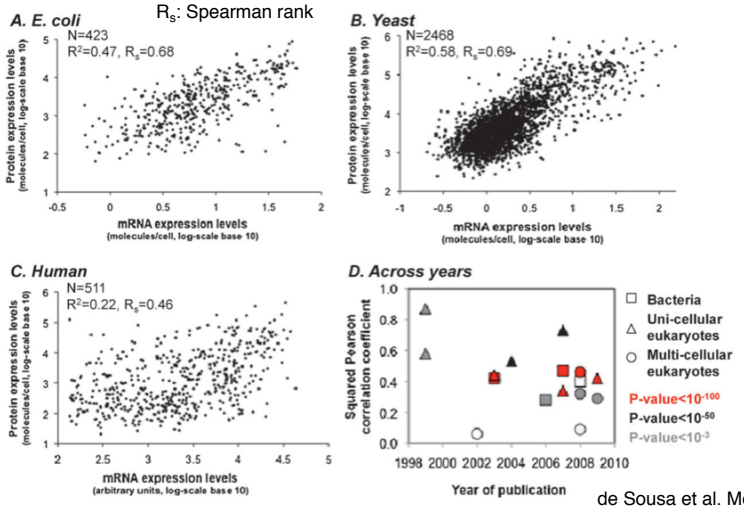


de Sousa et al. Mol. Biosyst. 5, 1512–1526.

from mRNA levels to promoter activity (transcription factor binding sites)

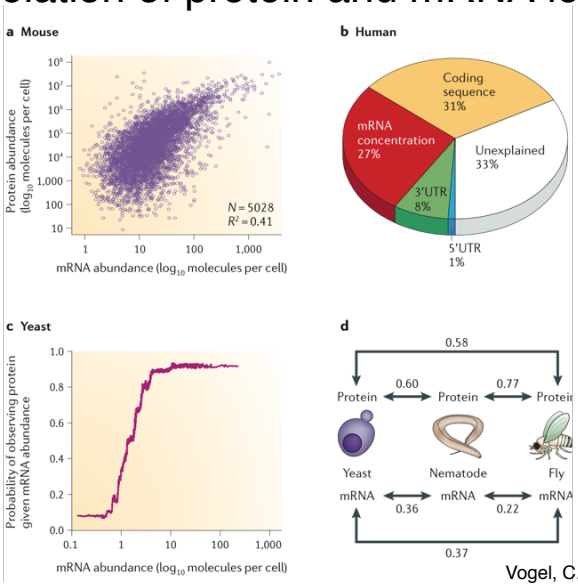
Correlation of protein and mRNA levels

R^2 : Pearson
R_s: Spearman rank



de Sousa et al. Mol. Biosyst. 5, 1512–1526.

Correlation of protein and mRNA levels



Vogel, C. & Marcotte, E. M.*Nat Rev Genet* **13,** 227–232 (2012).

Correlation of protein and mRNA levels

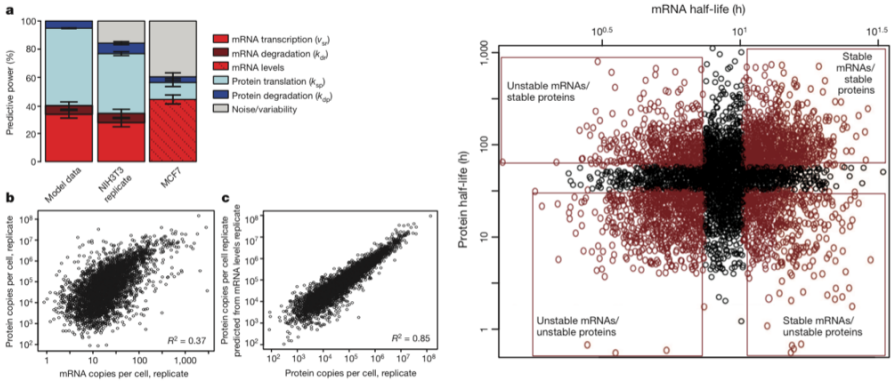
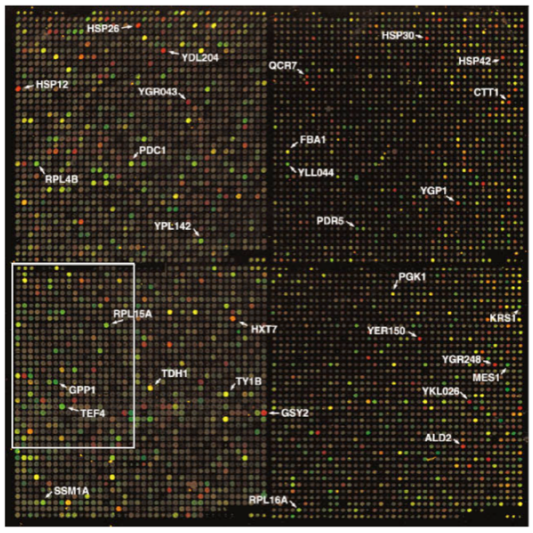


Figure 4 - Impact of different rates and rate constants on protein abundance. a, Protein levels are best explained by translation rates, followed by transcription rates. mRNA and protein stability is less important (left bar). b, In the replicate experiment mRNA levels explained 37% of protein levels in NIH3T3 cells (middle bar in a). c, The model explains 85% of variance in protein levels from measured mRNA levels (middle bar in a). The mouse fibroblast model has some predictive power for human orthologous genes in MCF7 cells (right bar in a). Error bars show 95% confidence intervals estimated by bootstrapping.

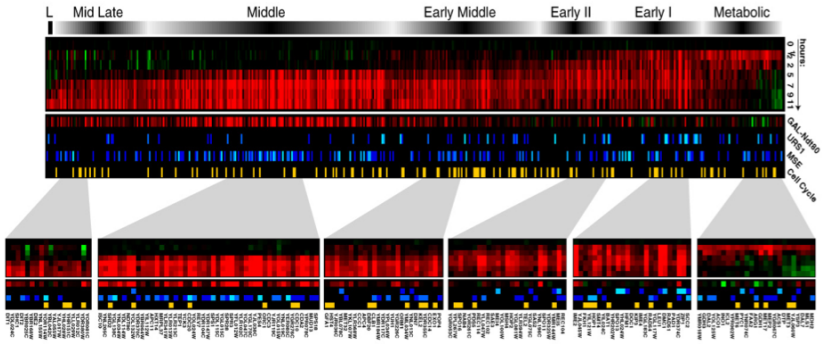
Schwanhäusser, B. *et al. Nature* **473,** 337–342 (2011).

The yeast genome on a chip



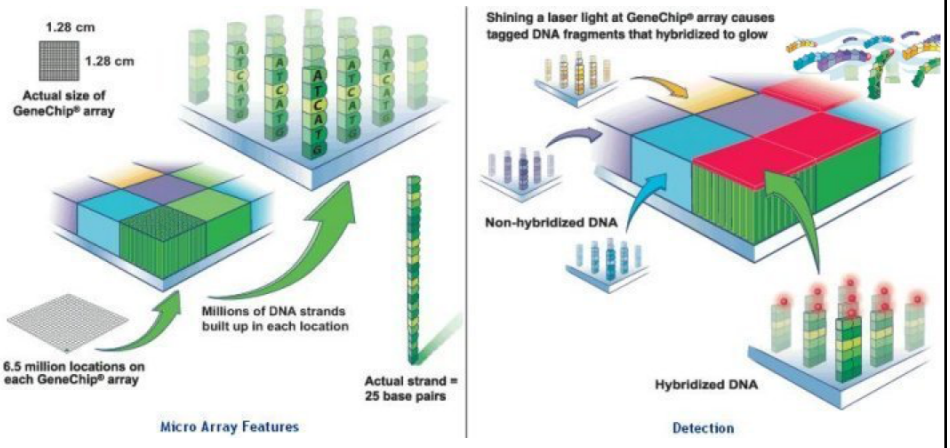
DeRisi et al *Science* **278,** 680–686 (1997).

Yeast genes induced during sporulation



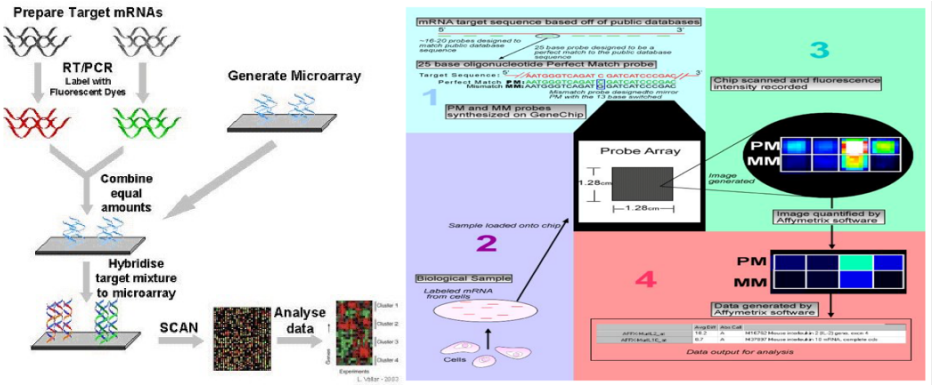
Chu, S. *et al. Science* **282,** 699–705 (1998).

Affymetrix GeneChip Overview



http://universe-review.ca/I11-50-microarray.jpg

Affymetrix GeneChip Overview



www.microarray.lu/images/overview_1.jpg

master.bioconductor.org/help/course-materials/2009/SeattleApr09/AffyAtoZ/AffymetrixAtoZSlides.pdf

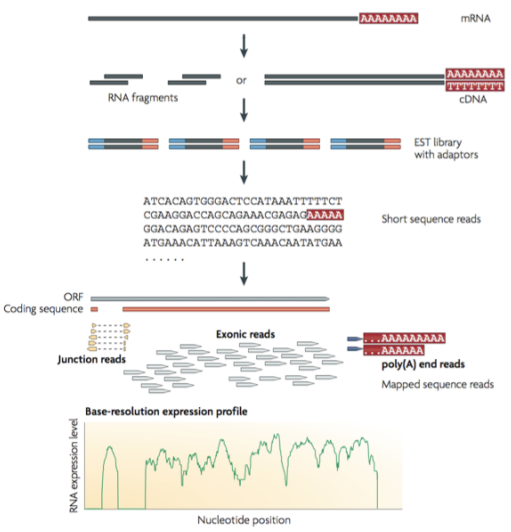
mRNA expression – accounting for differences

- Goals:
  - to quantify differences in mRNA abundance
  - ? to quantify amounts of mRNA in cell
- Data:
  - number of cells
  - amount of RNA (amount of mRNA?)
- Processes:
  - make cDNA from mRNA (equal efficiency?)
  - hybridize cDNA to oligonucleotides (GC hybridization differences)
  - does every probe set capture cDNA equally efficiently?
  - saturation (too much RNA/cDNA)
  - detection (too little RNA/cDNA)

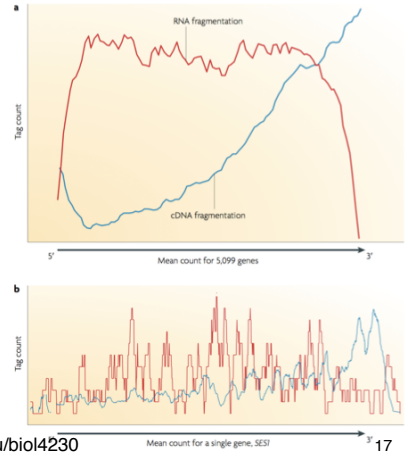
linearity

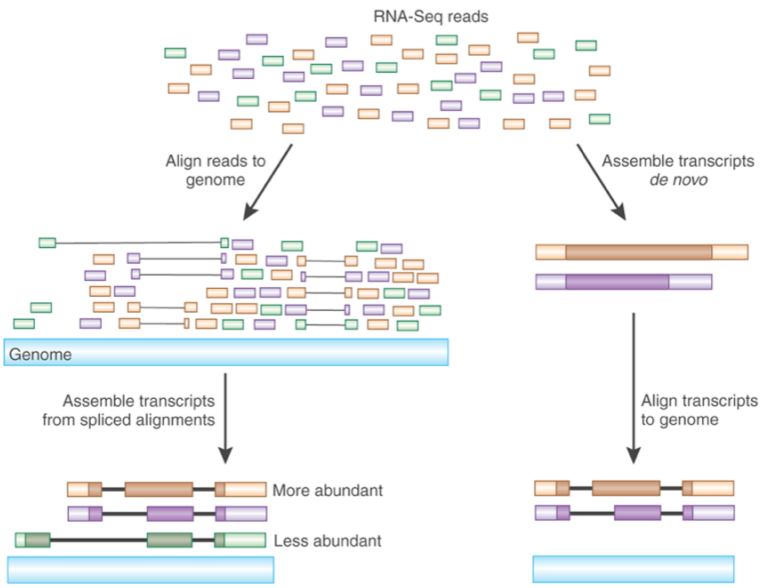
dynamic range

RNA-seq: digital RNA abundance

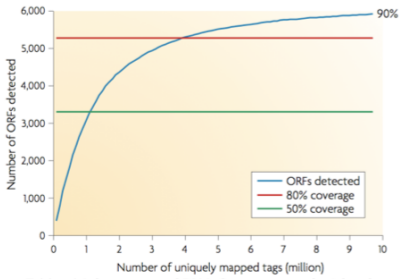


Wang et al. *Nat Rev Genet* **10,** 57–63 (2009).





Haas & Zody *Nat Biotechnol* **28,** 421–423 (2010).



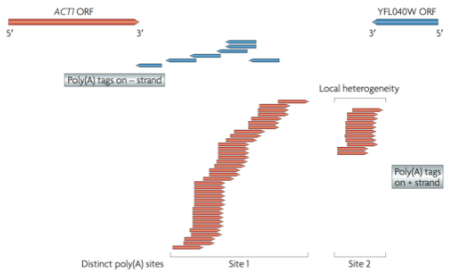


Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Wang, 2009

mRNA expression – accounting for differences

- Goals:
  - to quantify differences in mRNA abundance
  - ? to quantify amounts of mRNA in cell
- Data:
  - number of cells
  - amount of RNA (amount of mRNA?)
- Processes:

linearity
  - make cDNA from mRNA (equal efficiency?)
  - PCR DNA for sequencing

dynamic range
  - saturation (too much RNA/cDNA)
  - detection (too little RNA/cDNA)

Microarrays vs RNAseq

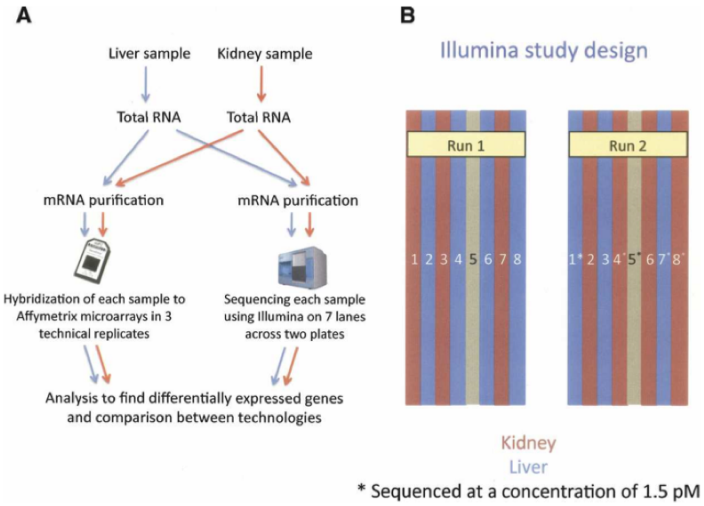


Figure 1. Graphical representation of the study design. (*A*) Summary of the experimental design. (*B*) The lanes in which each sample was sequenced across the two runs. In each run, the control sample was sequenced in lane *5*. Samples were sequenced at two concentrations: 1.5 pM (indicated by an asterisk) and 3 pM (no asterisk).

Marioni et al. Genome Res. **18,** 1509–1517 (2008).

Microarrays vs RNAseq

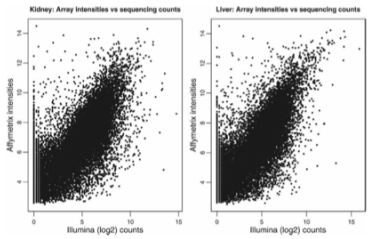


Figure 3. Comparing counts from Illumina sequencing with normalized intensities from the array, for kidney (*left*) and liver (*right*). In each panel, the average ($\log_2$) counts for each gene are plotted on the *X*-axis, and the corresponding normalized intensities from the array are shown on the *Y*-axis. To avoid taking the log of 0, we added 1 to each of the average counts prior to taking logs.

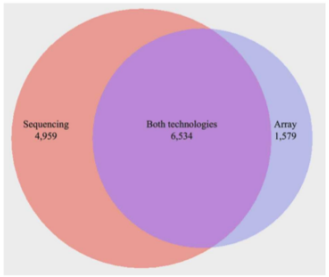


Figure 5. A Venn diagram summarizing the overlap between genes called as differentially expressed from the (*left* circle) sequence data and from the (*right* circle) array. The number of genes called by both technologies is indicated by the overlap between the two circles.

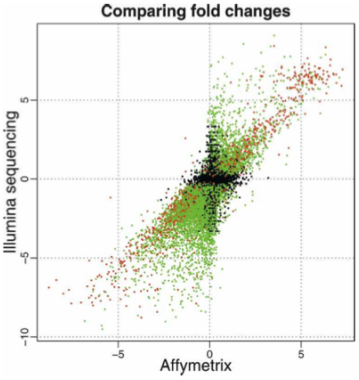


Figure 4. Comparison of estimated $\log_2$ fold changes (liver/kidney) from Illumina (*Y*-axis) and Affymetrix (*X*-axis). We consider only genes that were interrogated using both platforms and genes where the mean number of counts across lanes was greater than 0 for both the liver and kidney samples. (Red and green dots) Genes called as differentially expressed based on the Illumina sequencing data at an FDR of 0.1%, with a mean number of counts greater than (red) or less than (green) 250 reads in both tissues. (Black dots) Genes not called as differentially expressed based on the Illumina sequencing data. The set of differentially expressed genes that show the strongest correlation between the two technologies seems to be those that are mapped to by many reads (red), while the correlation is weaker for differentially expressed genes mapped to by fewer reads (green).

Marioni et al. Genome Res. **18,** 1509–1517 (2008).

How to compare relative mRNA expression?

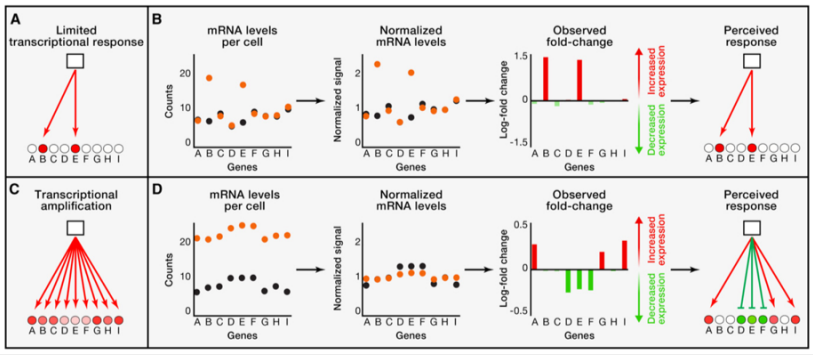


Figure 1. Normalization and Interpretation of Expression Data

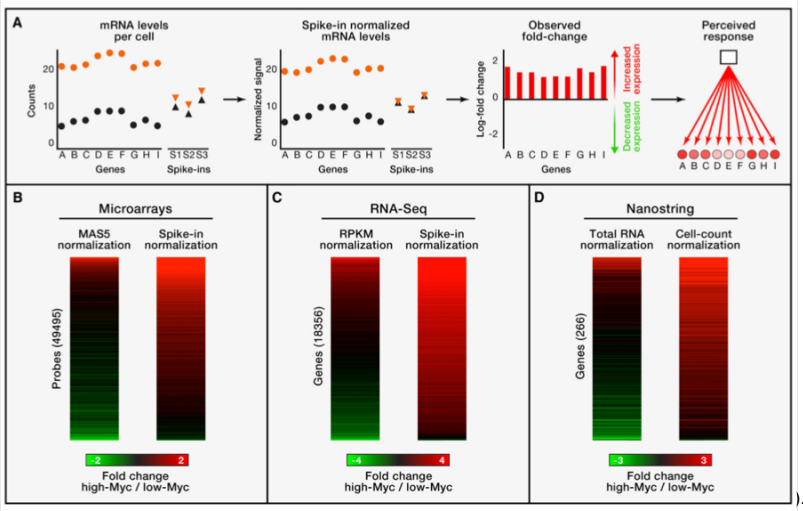
Lovén, J. *et al. Cell* **151,** 476–482 (2012).

How to compare relative mRNA expression?

Figure 1. Normalization and Interpretation of Expression Data (A) Schematic representation of pattern of change in gene expression when levels of total RNA in the two cells are similar. The square box represents a perturbation such as increased expression of a gene regulator or a change in environment or cell state. Red arrows point to target genes affected by the perturbation, which are represented as circles. Red shading of circles indicates relative transcriptional increase. (B) Schematic representation of microarray normalization when the overall levels of mRNA per cell are not changing in two conditions. Relative mRNA levels for nine different genes (A-I) are indicated along the y axis for condition 1 (black) and condition 2 (orange). The panels, from left to right, depict the actual relationship between mRNA levels for the two conditions; the effect of median normalization; the calculated fold-changes based on median normalization, with increased expression represented by red bars above the midline and decreased expression represented by green bars below the midline; and the perceived transcriptional response of a limited transcriptional increase in gene expression. (C) Schematic representation of pattern of change in gene expression when levels of total RNA in the two cells is different such as in transcriptional amplification, where most genes are expressed at higher levels. The square box represents a perturbation such as increased expression of a gene regulator or a change in environment or cell state. Red arrows point to target genes affected by the perturbation, which are represented as circles. Red shading of circles indicates relative transcriptional increase. (D) Schematic representation of microarray normalization when the overall levels of mRNA per cell are increased in one condition compared to another. Relative mRNA levels for nine different genes (A–I) are indicated along the y axis for condition 1 (black) and condition 2 (orange). The panels, from left to right, depict the actual relationship between mRNA levels for the two conditions; the effect of median normalization; the calculated fold changes based on median normalization, with increased expression represented by red bars above the midline and decreased expression represented by green bars below the midline; and the perceived transcriptional response following transcriptional amplification of gene expression.

Lovén, J. *et al. Cell* **151,** 476–482 (2012).

How to compare relative mRNA expression?



).

Figure 2. Spike-In Controls, Normalized to Cell Number, Enable Accurate Interpretation of Transcriptional Changes

Lovén, J. *et al. Cell* **151,** 476–482 (2012).

How to compare relative mRNA expression?

Figure 2. Spike-In Controls, Normalized to Cell Number, Enable Accurate Interpretation of Transcriptional Changes (A) Schematic representation of microarray normalization when the total level of mRNA per cell is different as in transcriptional amplification, but spike-in RNAs are used as standards for normalization. mRNA levels are indicated along the y axis for condition 1 (black) and condition 2 (orange); individual genes are represented along the x axis. Spike-in standards in the mRNA for condition 1 are represented by black triangles and spike-in standards in the mRNA for condition 2 are represented by orange triangles (S1–S3). The panels, from left to right, depict the actual relationship between mRNA levels for the two conditions; the effect of normalization using the spike-in standards; the resulting fold changes from condition 1 and condition 2, where increased expression is represented by red bars above the midline; and the perceived transcriptional response following transcriptional amplification of gene expression normalized with spike-in RNAs. (B) Heatmap showing the results of different normalization methods on the interpretation of microarray data. The data represent fold change of expression in high- Myc versus low-Myc cells. Each line represents data for individual probes on the microarray. Red indicates increased expression in high-Myc versus low-Myc cells. Green indicates decreased expression in high-Myc versus low-Myc cells. Black indicates no change in expression. Left: data using a standard microarray normalization method (MAS5). Right: the same data, now renormalized by using spike-in standards. (C) Heatmap showing the results of different normalization methods on the interpretation of RNA-sequencing data. The data represent fold change of expression in high-Myc versus low-Myc cells. Each line represents data for an individual gene. Red indicates increased expression in high-Myc versus low-Myc cells. Green indicates decreased expression in high-Myc versus low-Myc cells. Black indicates no change in expression. Left: data using a standard sequencing normalization (reads per kilobase of exon model per million mapped reads). Right: the same data, now renormalized by using spike-in standards. (D) Heatmap showing the results of different sample preparation methods on the interpretation of digital quantification data. The data represent fold change of counts of mRNA molecules in high-Myc versus low-Myc cells. Each line represents data for an individual gene. Red indicates increased expression in high-Myc versus low-Myc cells. Green indicates decreased expression in high-Myc versus low-Myc cells. Black indicates no change in expression. Left: the results if the quantification is performed with equal amounts of total RNA for the high-Myc versus low-Myc cells. Right: the results if the quantification is performed with RNA from equal numbers of high-Myc and low-Myc cells.

Lovén, J. *et al. Cell* **151,** 476–482 (2012).

Differential gene expression

- mRNA levels affect protein levels
  - no mRNA, no protein
  - little mRNA, sometimes lots of protein (long half-life)
  - lots of mRNA, often lots of protein
- RNA abundance:
  - most RNA is ribosomal RNA (rRNA)
  - 10 – 50 mRNA species account for >90% of mRNA abundance
  - sensitive methods detect < 1 molecule/cell (but not with single cells)
- which changes matter?
  - fold differences
    - 100X, from 1:100 molecules/cell?
    - 5X, from 50,000 to 250,000 molecules/cell?
  - mostly high abundance? mostly low abundance?