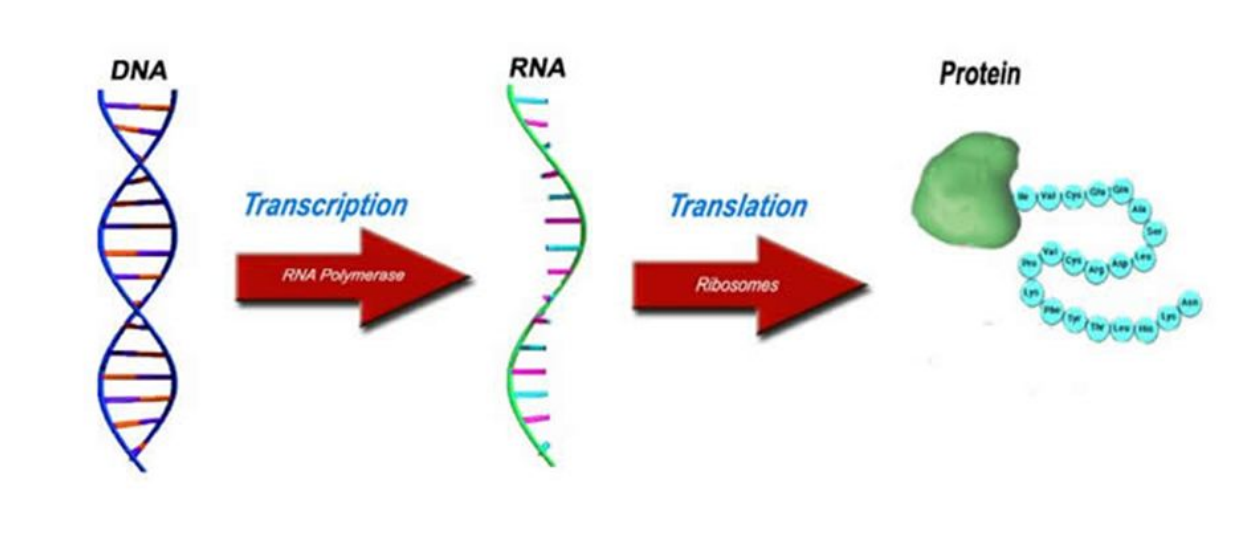
BF528 - Microarrays

Acknowledgement

This lecture was, in part, designed to be consistent with lecture material from Johns Hopkins University. Please see the link below for more information.

http://www.ams.jhu.edu/~dan/550.435/notes/COURSENOTES435.pdf

Central Dogma of Biology



Microarrays are a tool that help figure out **why**, **when**, and **to what extent** genes are transcribed

Definitions

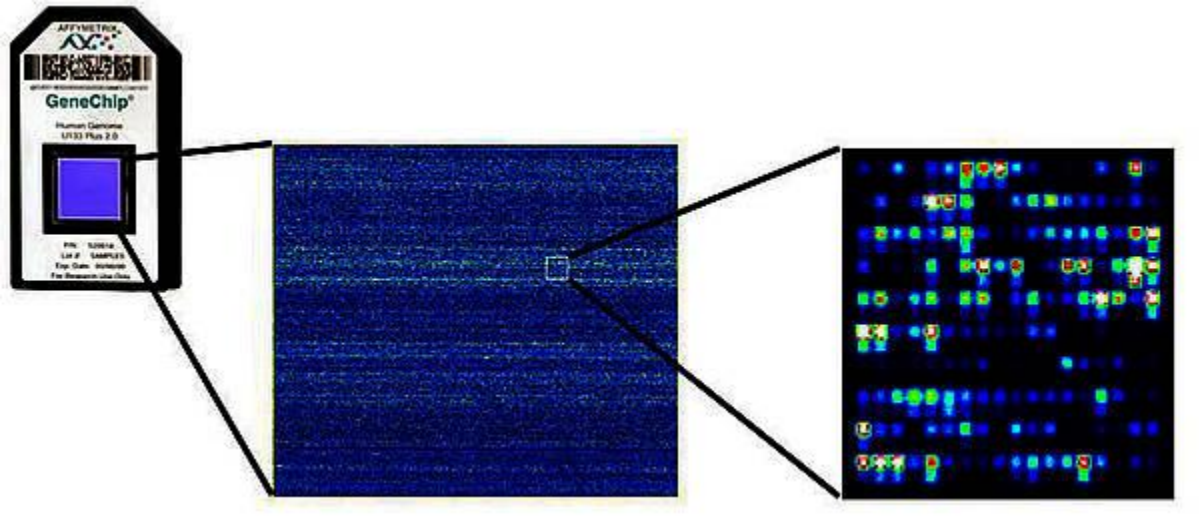
- Gene expression == mRNA abundance
- Increased expression == **induced**, or **up-regulated**
- Decreased expression == **down-regulated**
- Difference in expression == **differential expression**

Research Questions

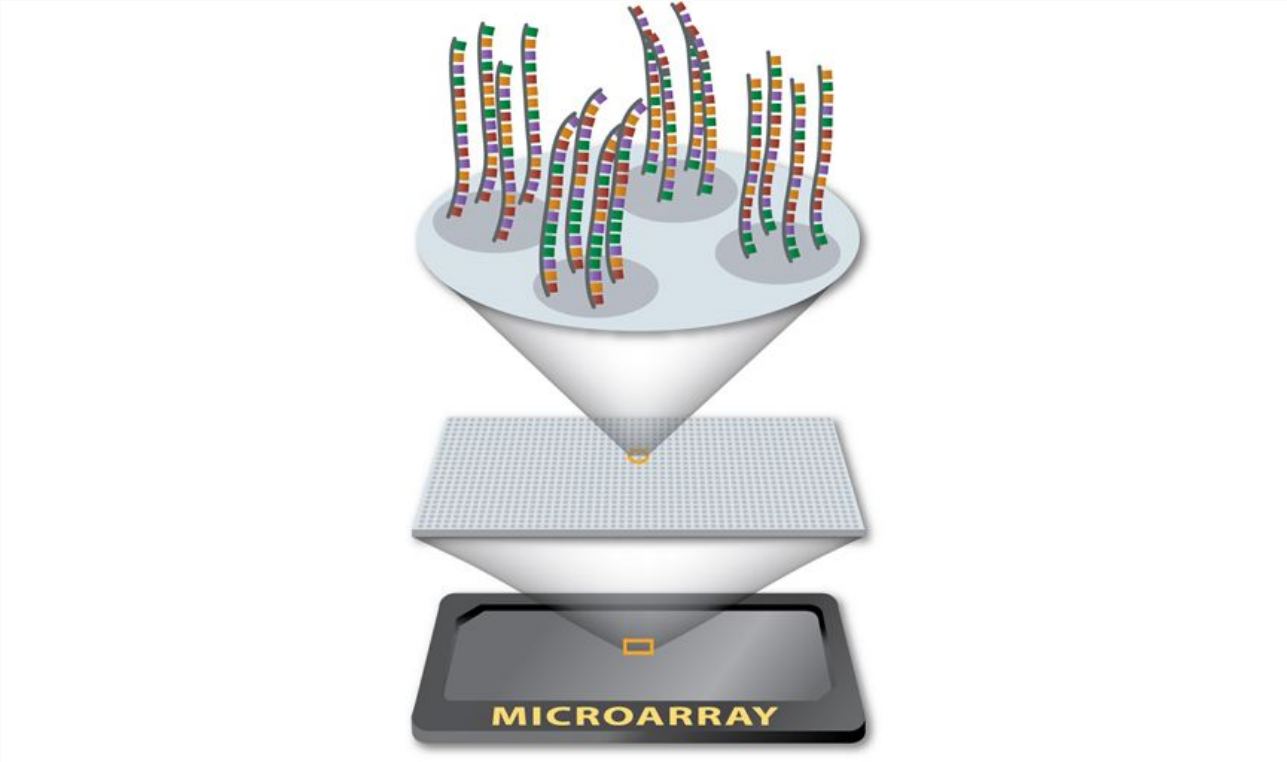
- Which genes are expressed in a given context?
- Which genes are expressed differently between different contexts?
- Which genes' expression correlate with a variable of interest?
- What do the (differentially) expressed genes tell us about the biological processes of the system?

What Is A Microarray?

“A collection of microscopic DNA spots attached to a solid surface”



What Is A Microarray?



Microarray Experiments

Experimental questions:

- Disease vs Normal tissue
- Changes in expression over time
- Before and after drug treatment
- Before and after toxin exposure
- Different tumor subtypes

Data driven questions:

- Which genes are differentially expressed?
- Which genes are co-expressed?
- Given gene expression, can we predict a condition?

Impact of Microarrays

~100k publications in PubMed

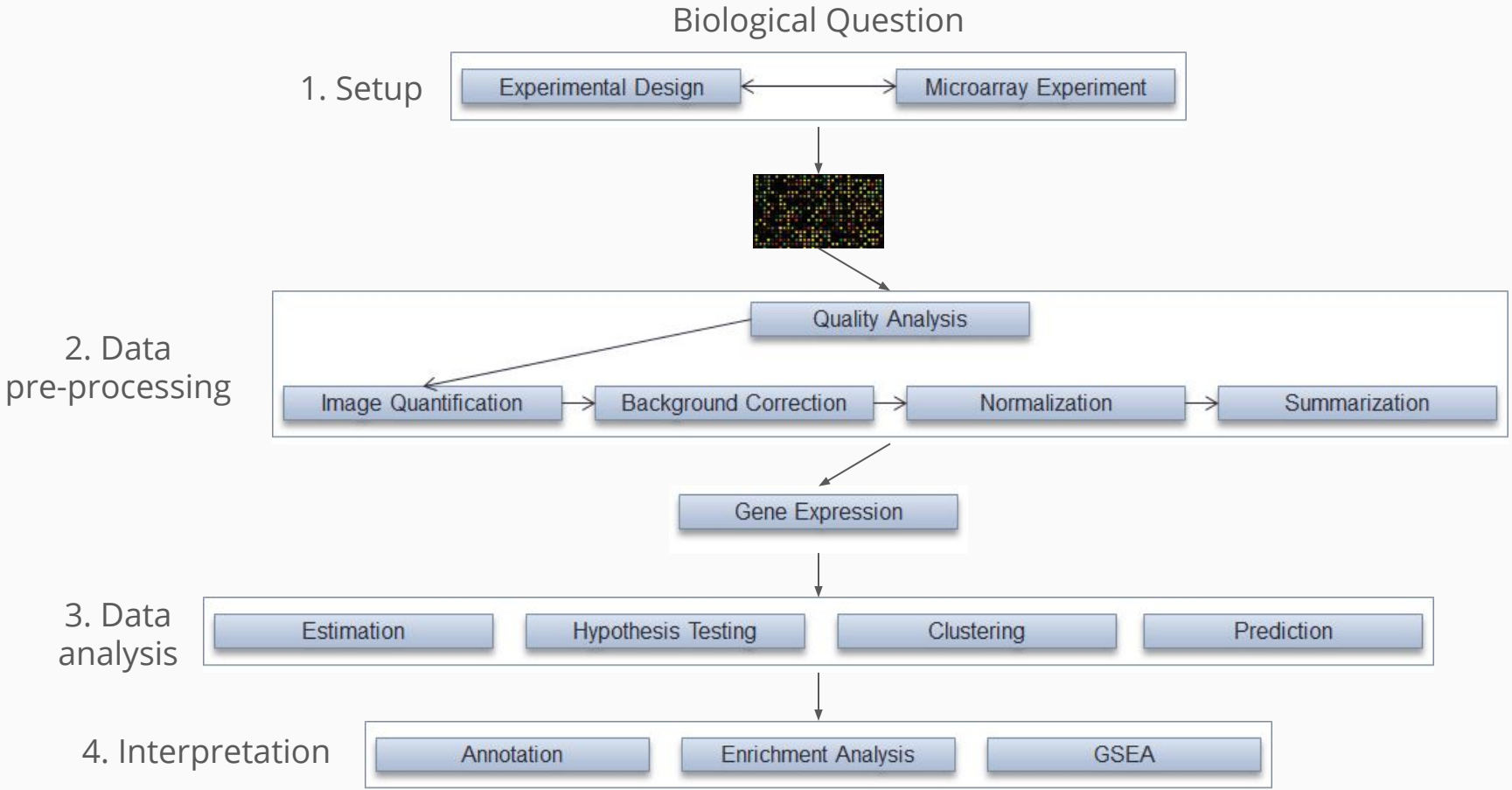
Breast cancer prognosis tool

Breast cancer predicts chemotherapy response

Colon cancer assay predicts risk of recurrence

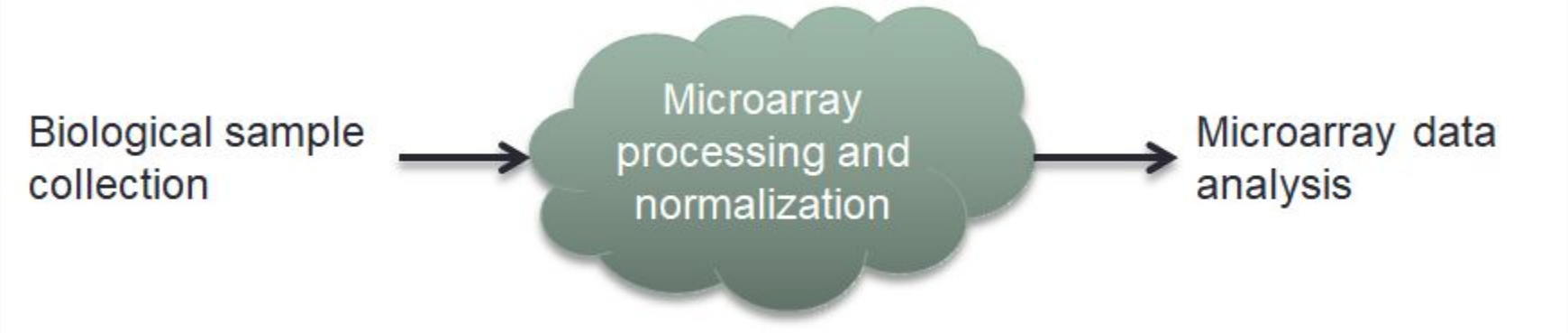
Prostate cancer assay predicts risk of progression

Microarray Study Design



Two Types of Microarrays

- Spotted arrays
- **High density oligonucleotide arrays**

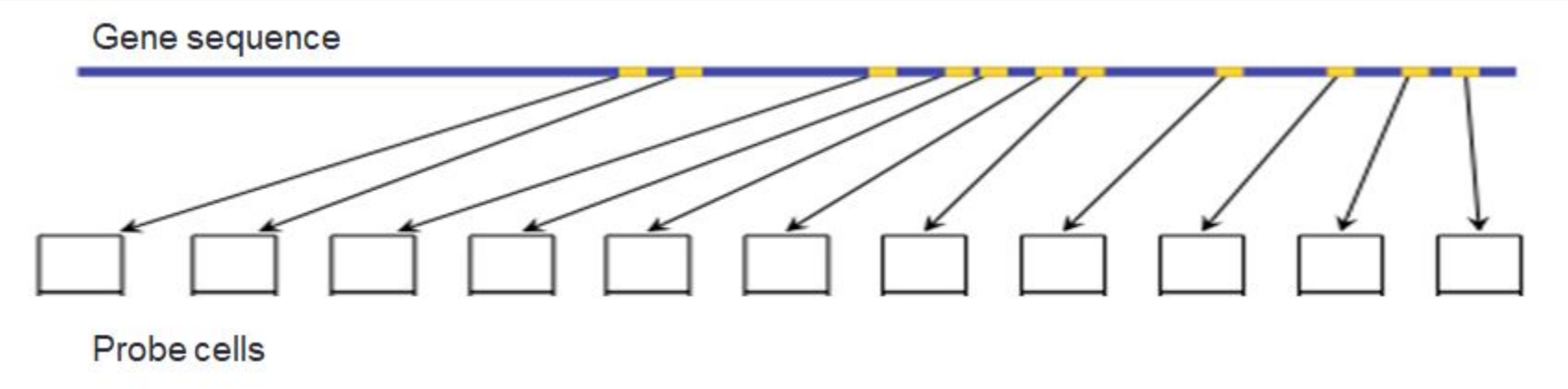


Oligo Arrays (in situ)

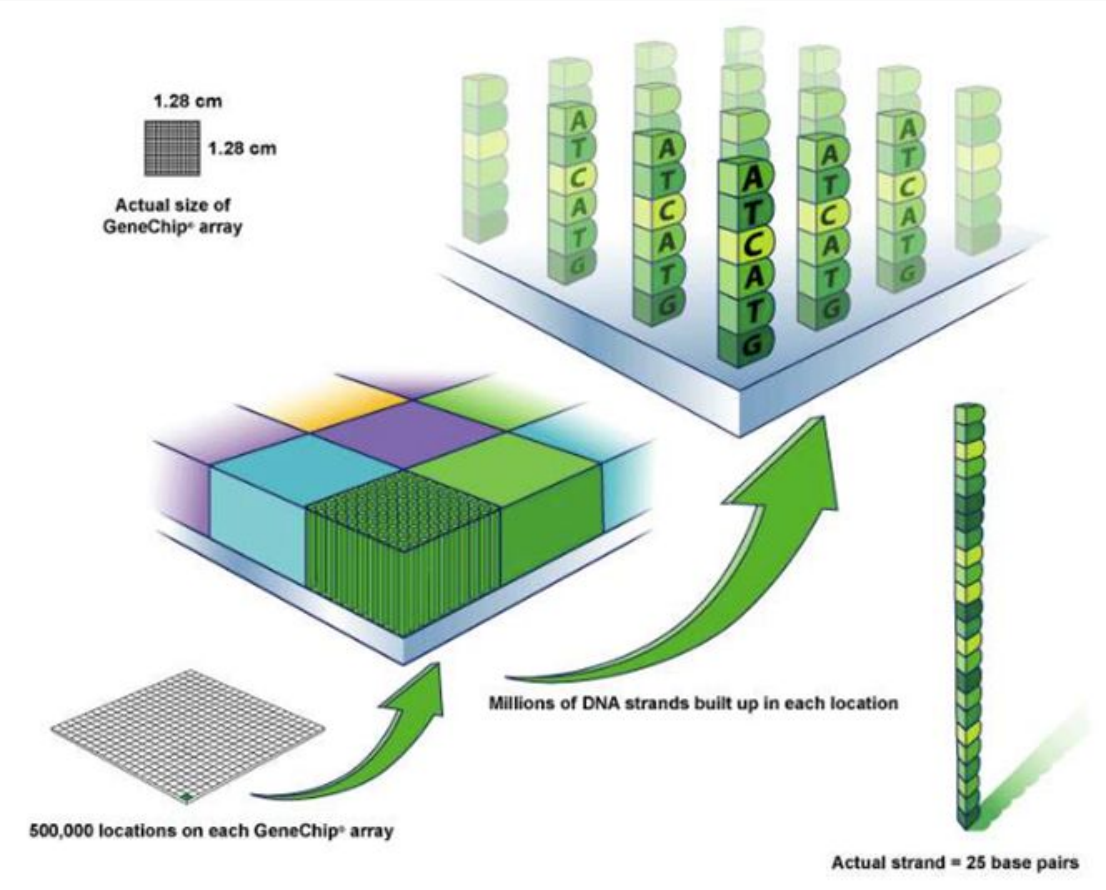
- Also called *Affy arrays* (orig. Affymetrix)
- **Probe** - a unique, short cDNA sequence
- Probes synthesized on the chip
- Terminology:
  - **probe** - an individual 25nt sequence
  - **probe cell** - millions of copies of a probe placed together
  - **probe set** - a set of unique probes for a given gene

Oligo Arrays (in situ)

- Each gene is represented by a set of probes
- Probe sets
  - Unique to each gene
  - Multiple probe cells tiled across exons

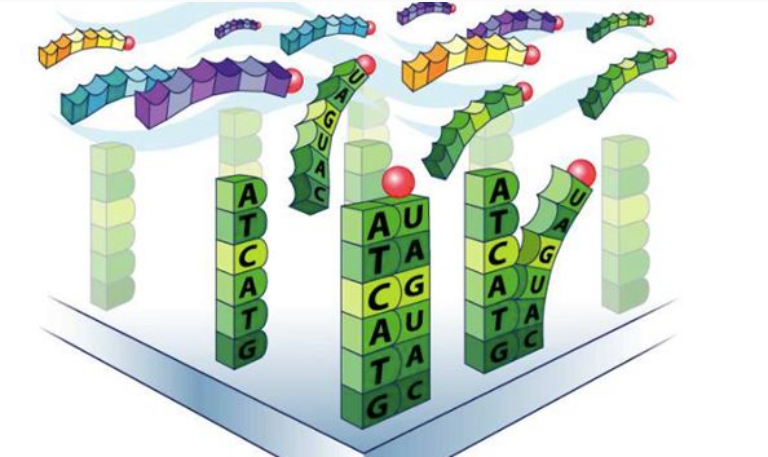


Oligo Arrays (in situ)



Oligo Arrays - Labeling & Scanning

- Sample cDNA tagged w/ fluorescent marker
- Sample washed over flowcell
- Unbound cDNA washed away
- Luminescence quantified in a scanner

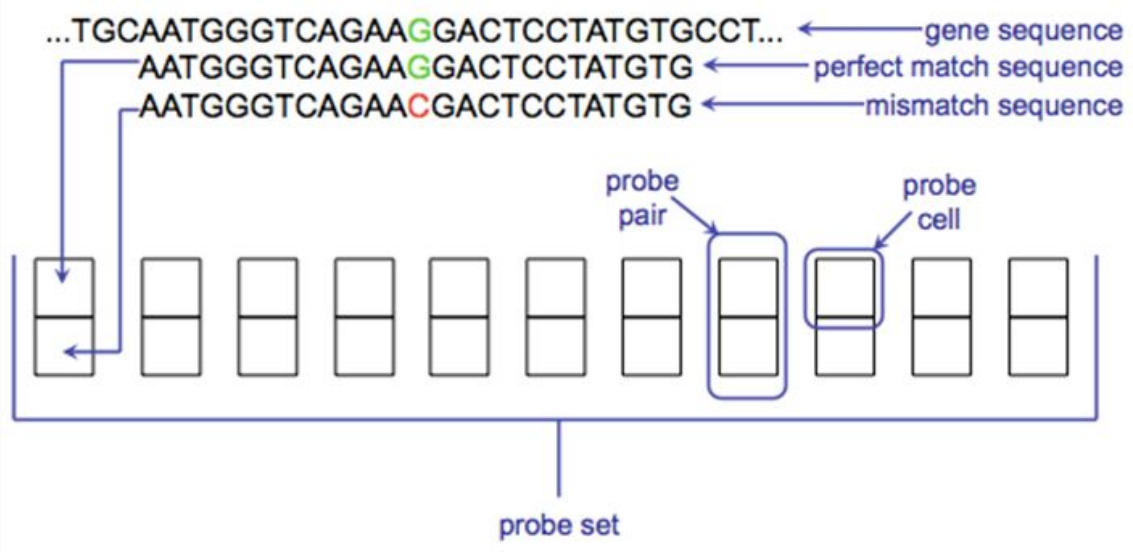


Oligo Arrays - Older Technologies

- Affymetrix U133A & B
- Probe cells grouped together, led to biases and artifacts

Oligo Arrays - Older Technologies

- Used *perfect match* and *mismatch probes*
- Mismatch probes contain...mismatch in probe sequence to account for non-specific binding



Oligo Arrays - Technology Update

- U133 Affy arrays designed against 2001 draft human genome
- Mismatch probes didn’t work well
- Genome reference improved, more genes needed to be profiled
- Gene ST array created and current

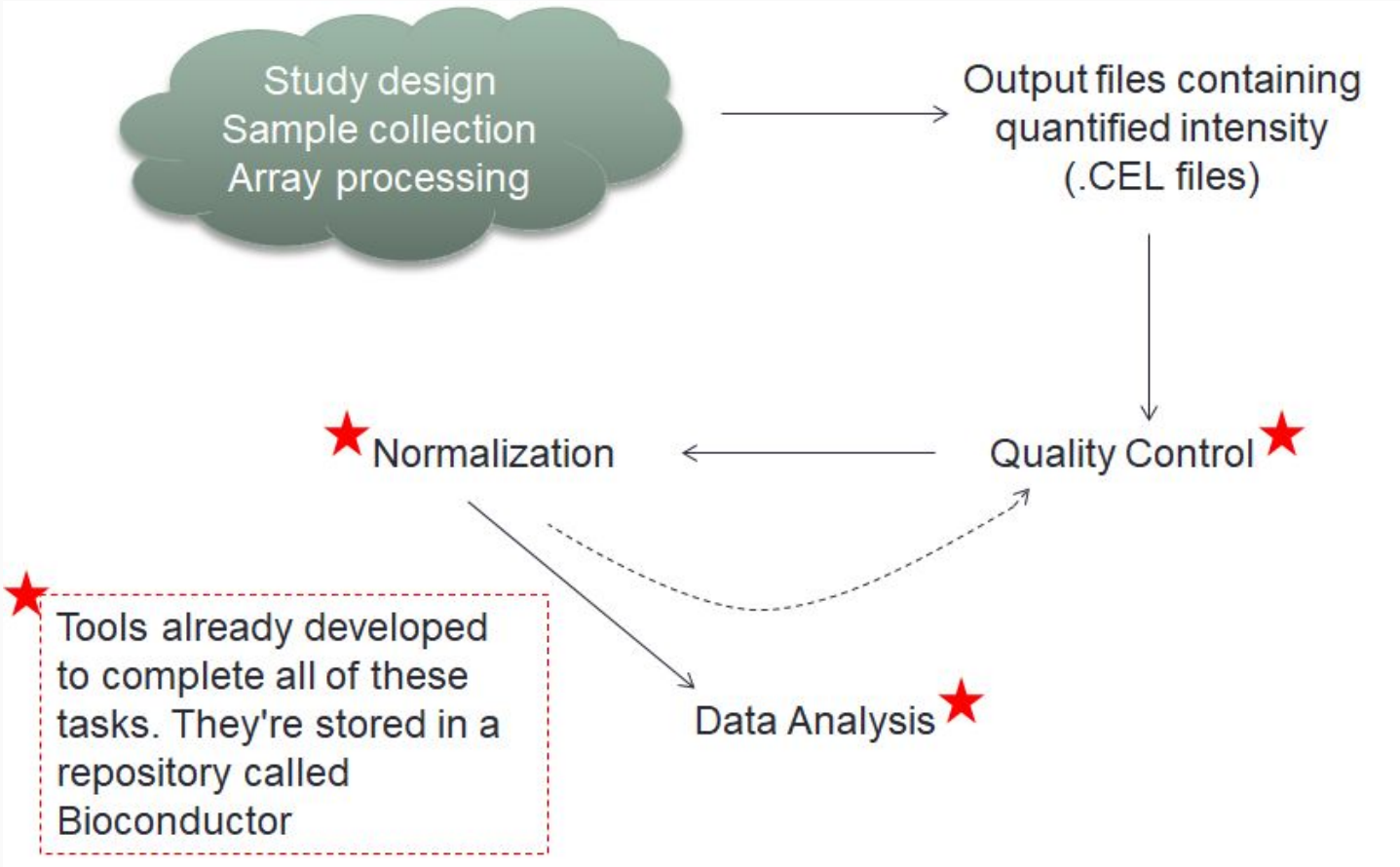
Oligo Arrays - Current Technology

Several chips:

- 3’ IVT - gene expression, measures 3’ UTR
- Gene - gene expression, tiled sequence
- Exon - DNA sequence, exon sequences only
- Tiling - DNA sequence, tiled across genome

Microarray Analysis Methods

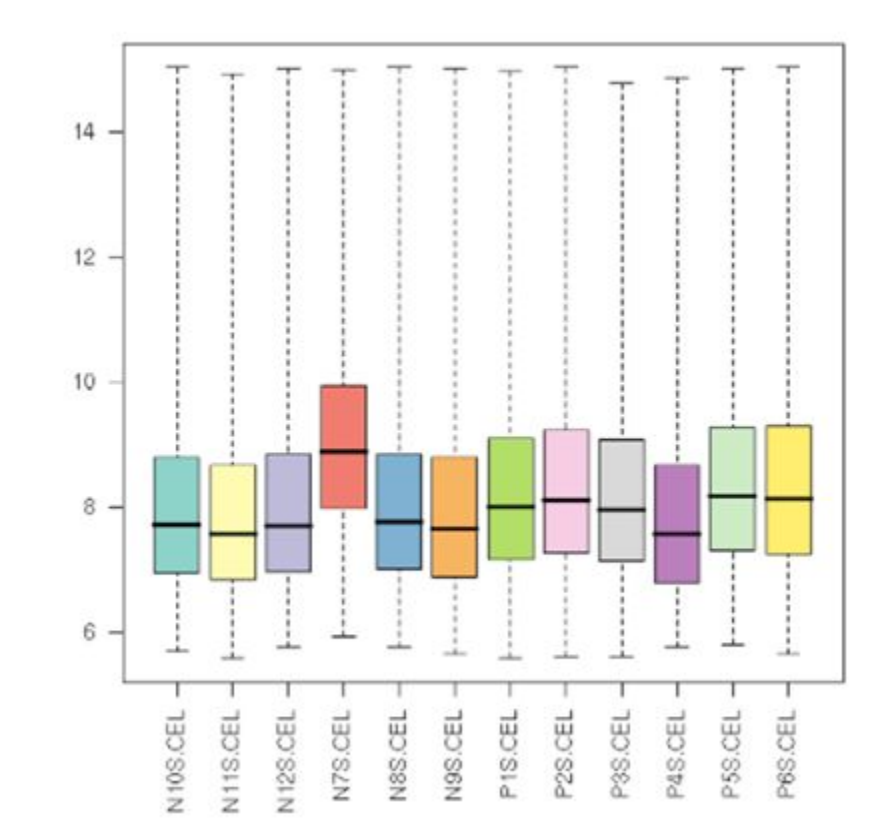
Typical Processing Steps



Normalization

- Statistically adjust a set of expression matrices so they are comparable
- Includes:
  - **Background correction** - remove artifacts/noise
  - **Data normalization** - adjust probe distributions across arrays to ensure comparability
  - **Probe summarization** - combine probe intensities within probeset to gene-level

Normalization



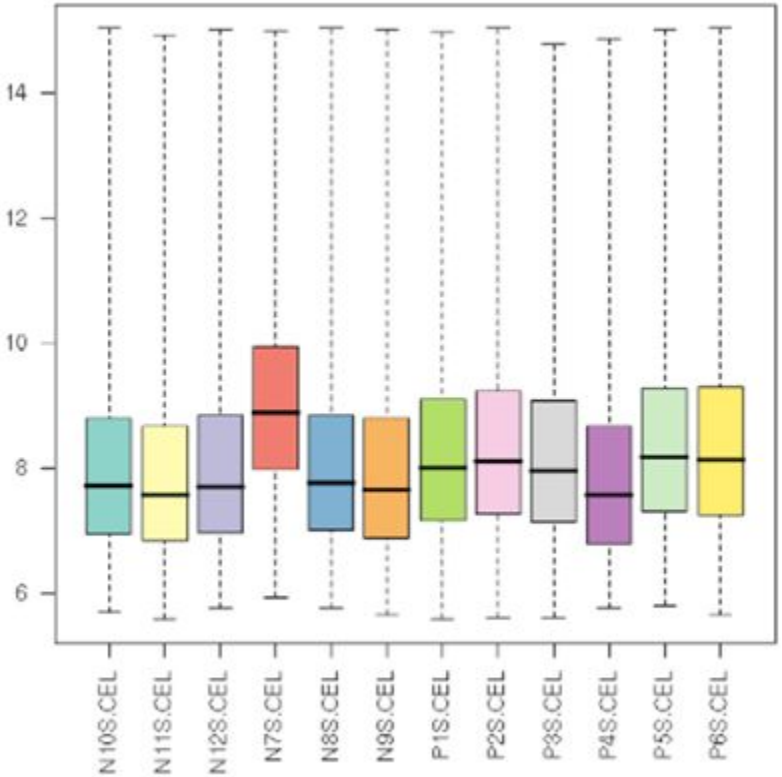
Old Normalization Method: MAS5

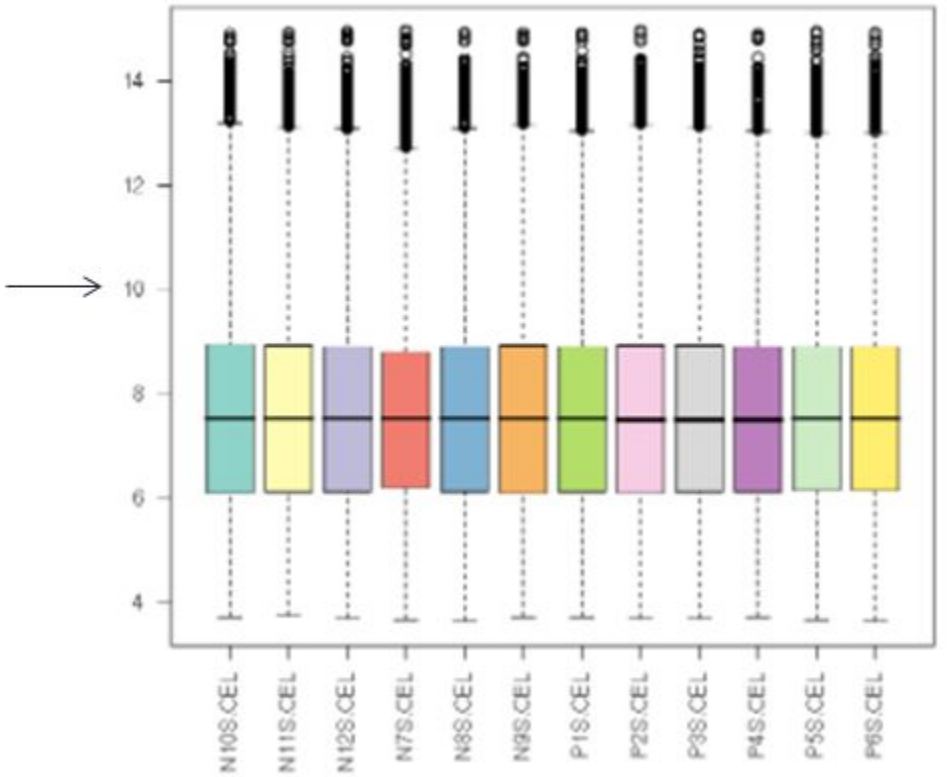
- Measured value = Noise + Probe Effects + Signal
- Background correction
- Intensity adjustment
  - Examine perfect/mismatch probe ratio
  - Calculate % present
  - AUC
- The good - usable with single chip, p-value for gene expression
- The bad - uses mismatch probes, very complicated

State of the Art Normalization: RMA

- Robust Multi-array Average (RMA)
- **Background correction** - kernel density estimation based
- **Normalization** - quantile normalization across batch of arrays
- **Summarization** - linear additive model controlling for probe effects, gene expression, and error

RMA Normalization

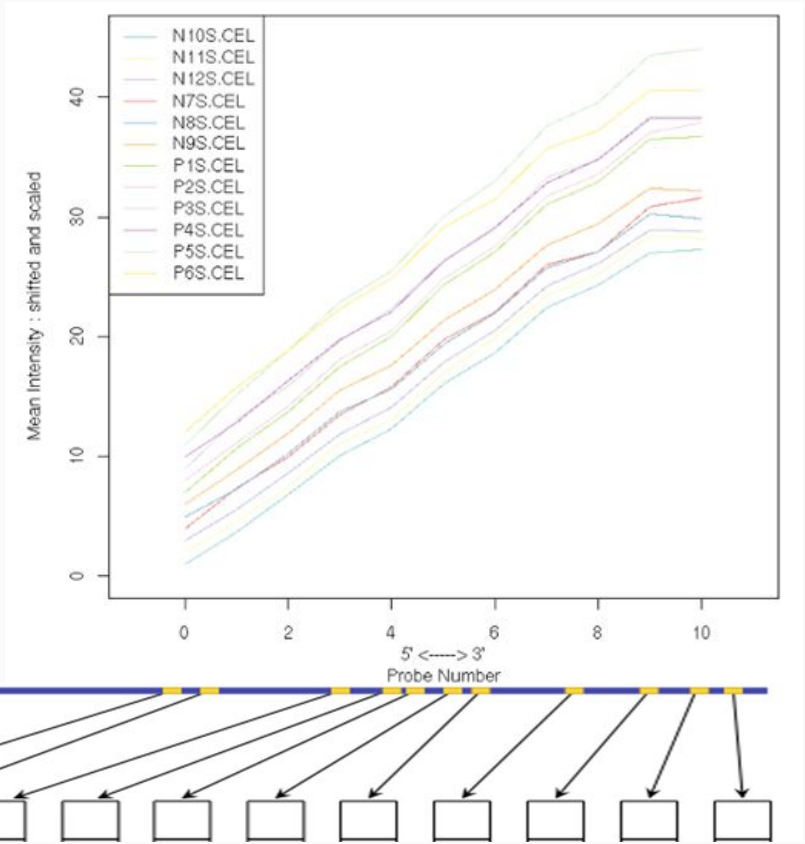




http://www3.it.nuigalway.ie/agolden/bioconductor/version1/MicroArrayAnalysis.pdf

Quality Control - RNA Quality

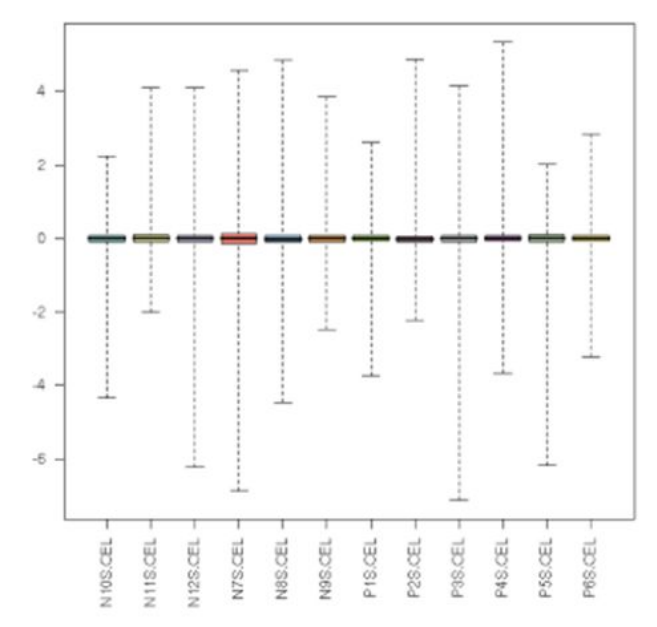
- RIN - RNA Integrity Number
- Ranges 0-9
- RNA Degradation Plot
- RNA degrades 5’ -> 3’
- Probe intensity tends to increase accordingly
- Slope >2 may indicate poor RNA quality



Quality Control - RLE

- **R**elative **L**og **E**xpression
- Subtract median intensity across all arrays from each probe
- Median RLE != 0 -> number of up/down genes not the same
- Large IQR means many genes are differentially expressed

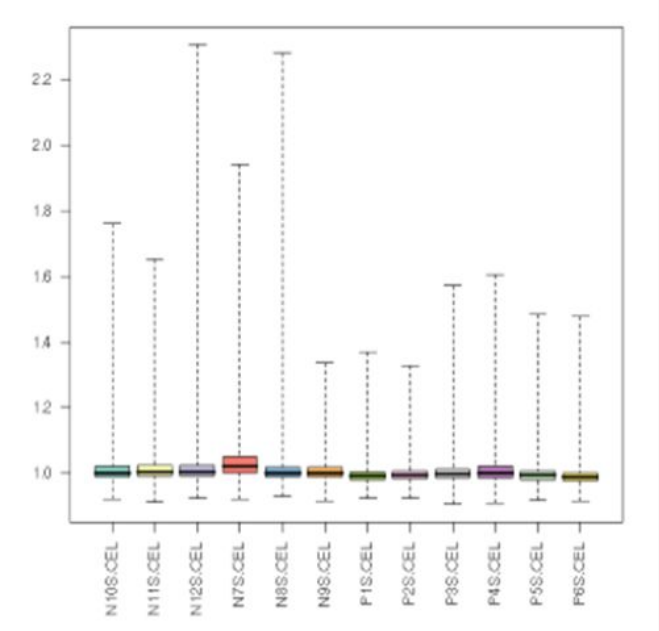
$$\mathrm{RLE}(\hat{\theta}_{i,j}) = \hat{\theta}_{i,j} - \mathrm{median}_j(\hat{\theta}_{i,j})$$



Quality Control - NUSE

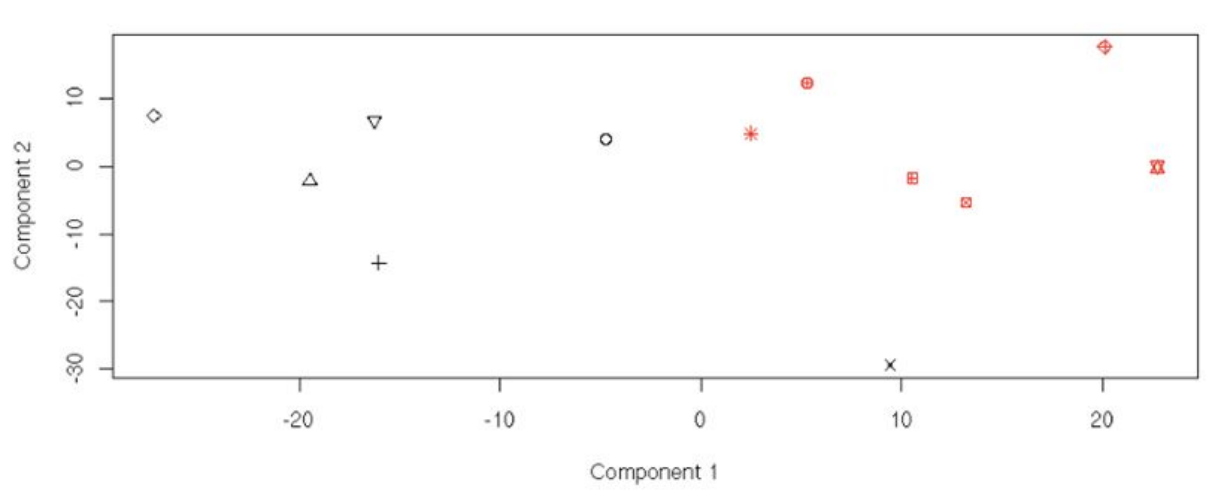
- **N**ormalized **U**nscaled **S**tandard **E**rror
- Arrays w/ large IQR or median
  NUSE > 1 may be poor quality

$$\mathrm{NUSE}(\theta_{gi}) = \frac{\mathrm{SE}(\theta_{gi})}{\mathrm{med}_i(\mathrm{SE}(\theta_{gi}))}$$



Quality Control - PCA

- Principal Component Analysis
- Data dimensionality reduction technique
- Identifies “directions of variance”



Quality Control Rules of Thumb

- RLE, NUSE, RNA Degradation, and PCA
- No single metric is indicative of poor quality
- Flagged samples should be examined before omission
- Make decisions based on multiple criteria
- Rule of thumb - filter if all three:

  Median RLE > 0.10, Median NUSE > 1.05, RIN < 4.0

Correcting for Batch Effects

- Batch effects - variation in data due to technical effects:
  - Arrays processed on different days
  - The technician who ran the arrays
  - Sample collection site
  - Variance in reagent type/quality
- Confounding - when batch effects are correlated with condition of interest
- ComBat - R package to correct batch effects

Correcting for Batch Effects