Biochem 3BP3

Gene Expression Analysis

Week of Nov 8, 2021

Gene Expression Analysis

- Gene Expression Analysis used primarily for two purposes:
  - Determine genetic underpinnings of observed phenotype – experimental
  - Annotation of genomes – development of gene models (intron/exon)
- The complete set of messenger RNAs in a cell is called the “Transcriptome”
- Transcriptomes are dynamic while genomes are static
  - the transcriptome will vary among cells, organs, life-stages, over time
  - many transcriptome libraries are needed for a full sampling
  - cell sorting technologies + NGS = single cell transcriptomics
- Transcriptome analysis is both genomic and statistical in methodology

Gene Expression Analysis

- For most studies, transcriptome = mRNA
  - rRNA is ignored as it dominates the transcriptome and would overwhelm library construction & sequencing
  - mRNA is sampled exclusively via isolation of polyadenylated RNA during library construction
  - transcriptome = mRNA = protein-coding portion of the genome
- transcriptome ≠ proteome
- Increasingly, studies sample:
  - total mRNA
  - ribosome-bound mRNA (i.e. actively translated)
  - miRNA (i.e. post-transcriptional regulation)

Transcriptomes can be sampled and measured in two ways:

INDIRECT = probe based technologies = microarrays

DIRECT = sequencing technologies = RNA-Seq

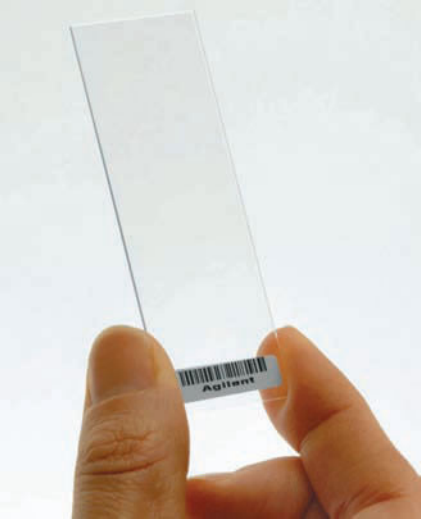
Direct Sampling

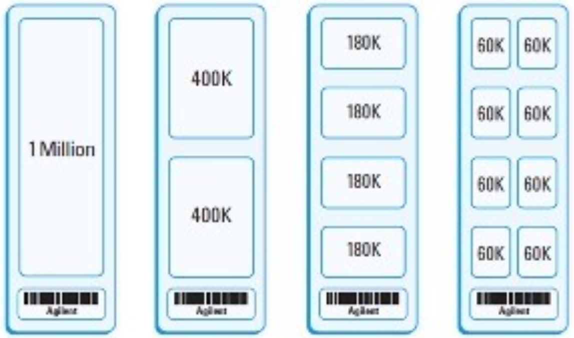
1. Construction and Sanger sequencing of cDNA libraries
   - long reads but only a small number of them (1,000s)
   - cDNA libraries normalized so all mRNA were equally abundant – sequencing was about sampling a broad diversity of mRNA but not about relative abundance
2. 1990s – Serial Analysis of Gene Expression (SAGE)
   - Higher volume isolation of 14 bp or 21 bp tags from mRNAs
   - Tags mapped to a known genome sequence (required!)
   - 10,000s
   - Relative abundance
3. RNA-Seq (focus of next week)
   - NGS of mRNA libraries
   - 1,000,000s
   - Gene identification and relative abundance data

Indirect Sampling

- Hybridization of mRNA samples to a set of pre-defined DNA probes
- **Probes designed based on known gene sequences – not a complete sampling of the transcriptome**
- More abundant mRNA molecules will bind more frequently to probes = higher signal (usually fluorescent); signal is used a proxy for relative abundance
- Original “microarrays” involved robotic spotting or *in situ* synthesis of probes on glass slides
- NGS was expected to make microarray technology obsolete but new BeadChip approaches have improved accuracy and very competitive cost
- For example, human muscle biopsy transcriptomics @ McMaster:
  - RNA-Seq – $540 / sample – 25 million mRNA sequenced
  - HT-12 BeadChip – $178 / sample – 47,000 probes

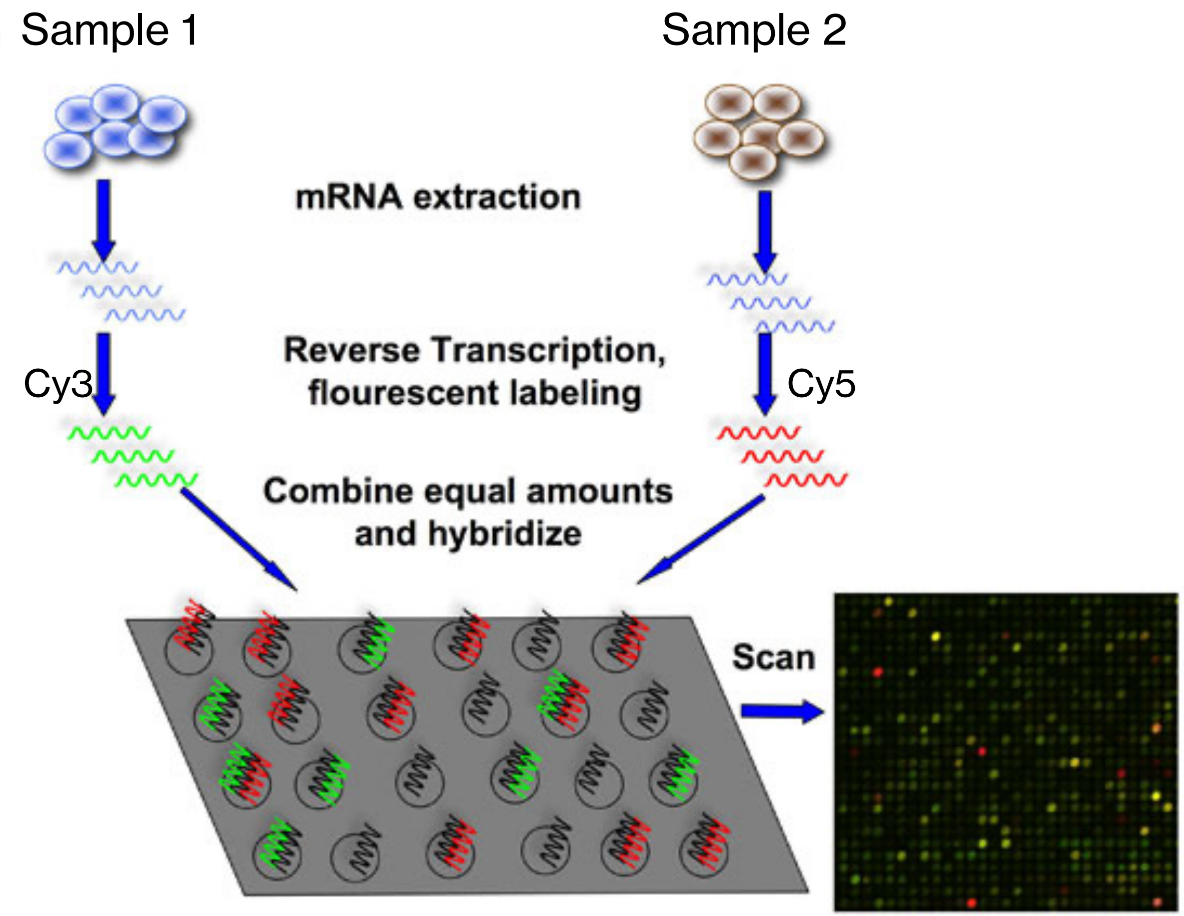
Glass Slide Microarrays



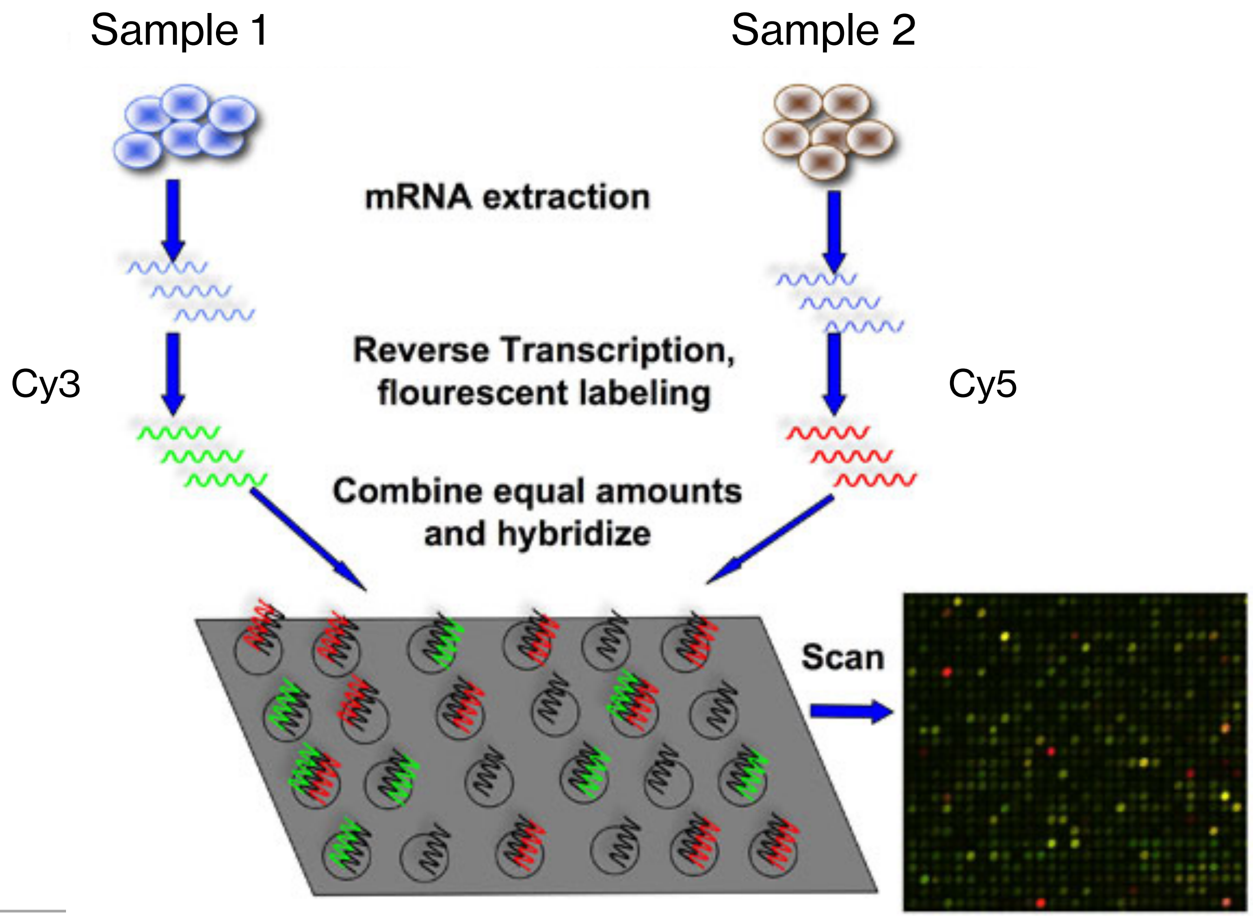


- High density probe arrays – many formats

Microarrays – Two-Dye Methods

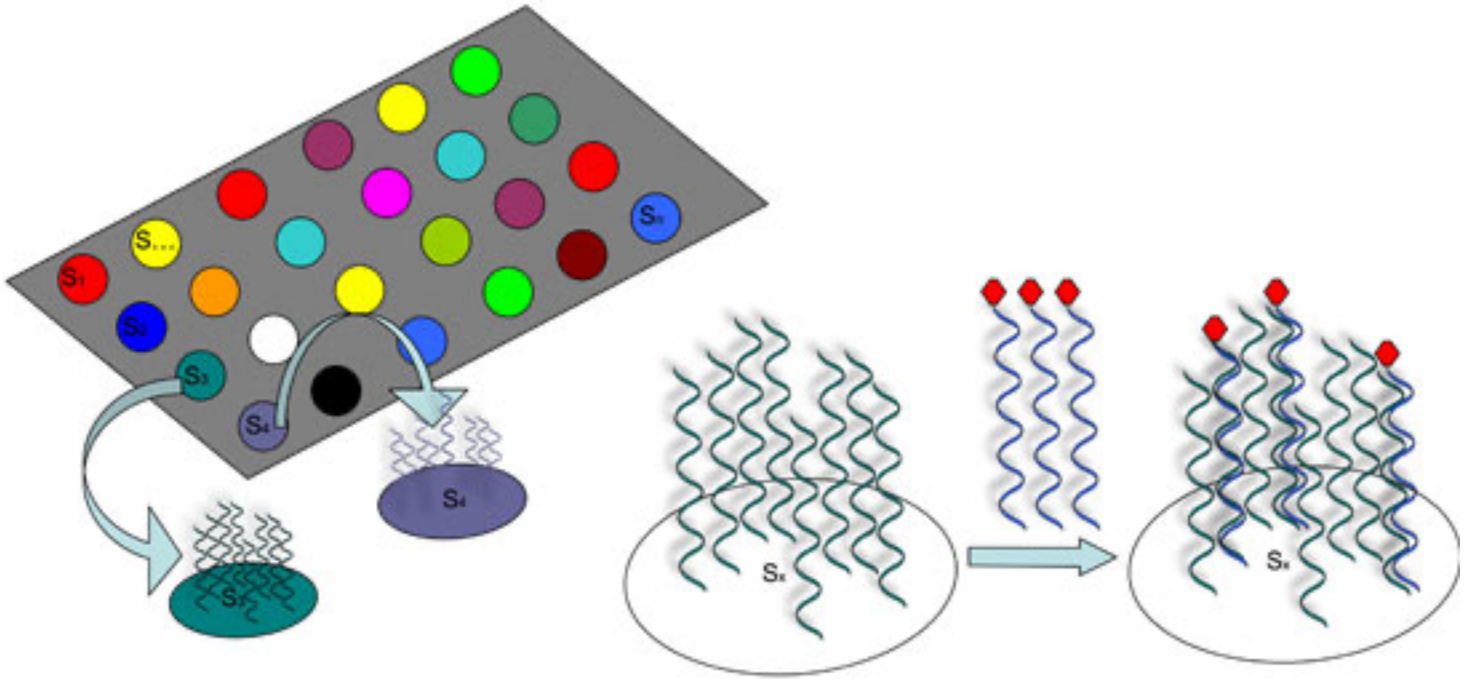


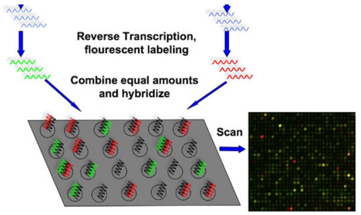
Microarrays – Two-Dye Methods



Cy3 – green, 532 nm
Cy5 – red, 635 nm

Microarrays – Two-Dye Methods



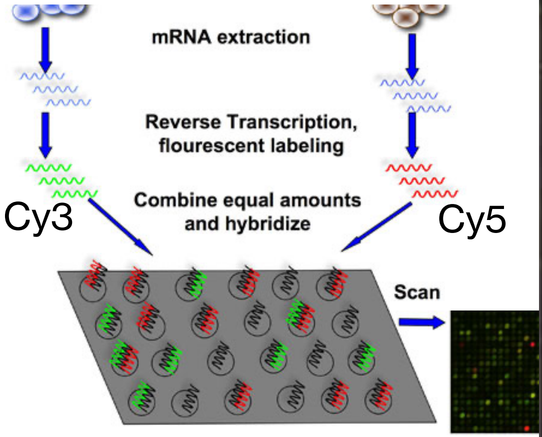

Reverse Transcription, flourescent labeling

Combine equal amounts and hybridize

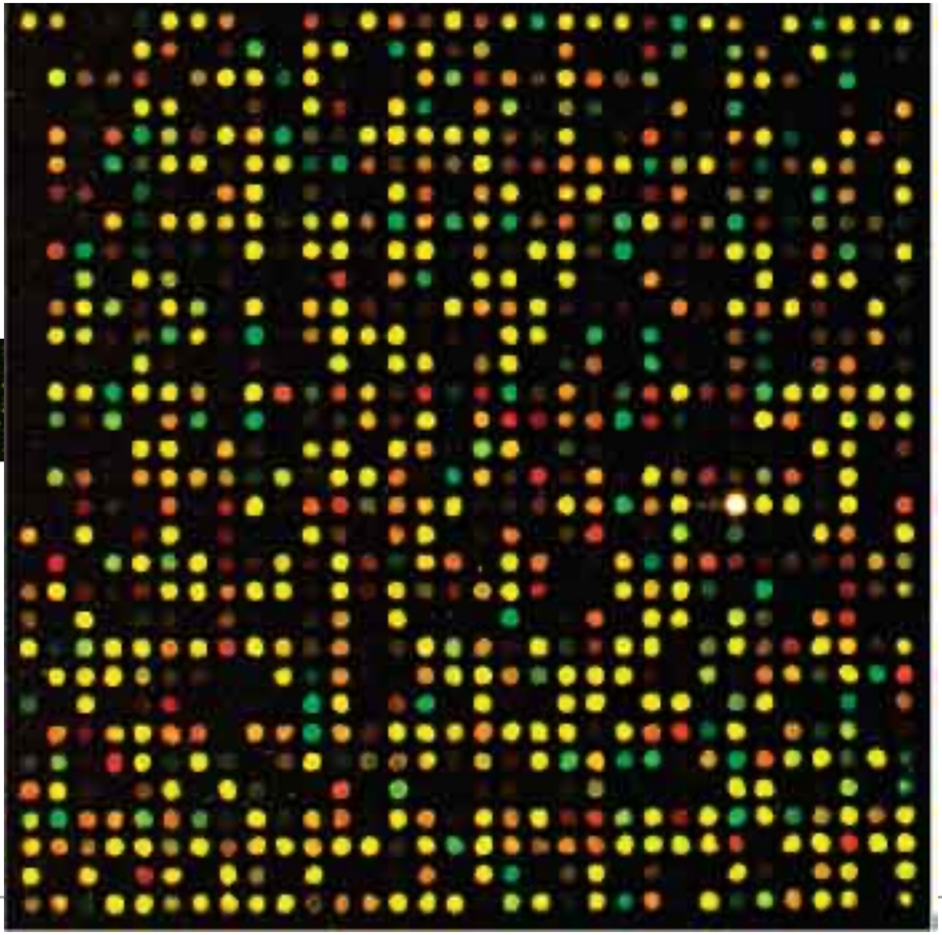
Scan

- Each spot is a collection of thousands of identical probes
- The more mRNA that bind to a probe within a spot, the brighter the overall signal for that spot

Microarrays – Two-Dye Methods

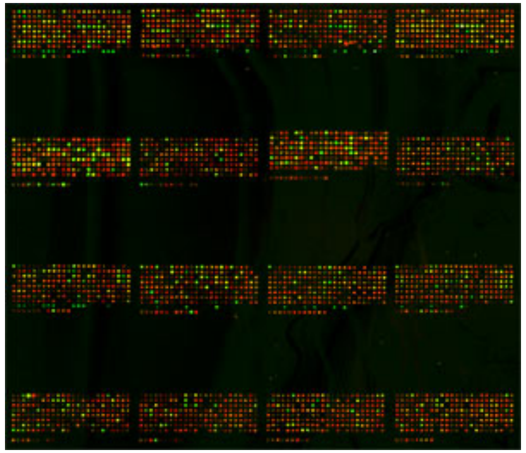

mRNA extraction
Reverse Transcription, flourescent labeling
Cy3
Combine equal amounts and hybridize
Cy5
Scan

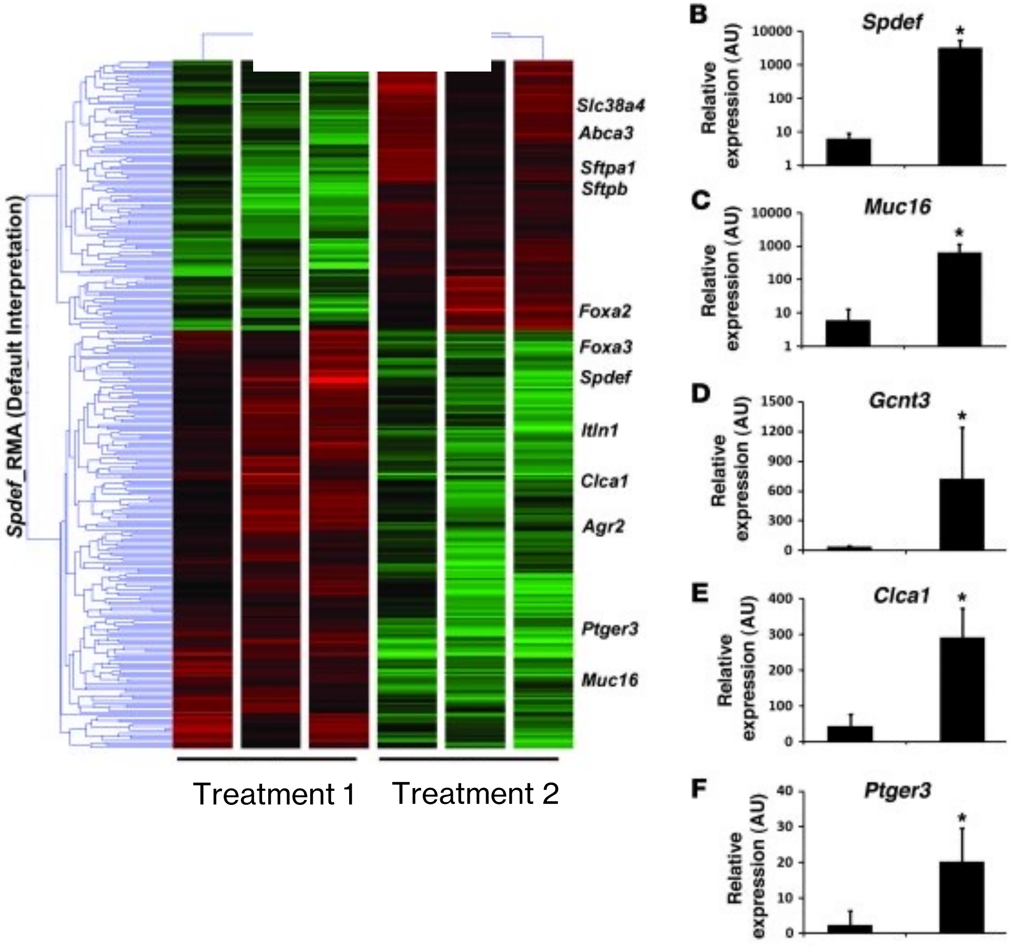
Cy3 – green, 532 nm
Cy5 – red, 635 nm



Equal expression	🟡
High expression of red	🔴
High expression of green	🟢
No expression	⚫

Microarrays – Two-Dye Methods



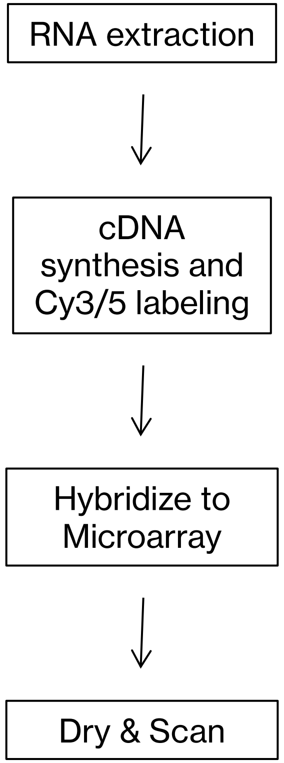


Fold change between the two samples is calculated

GREEN – downregulated

RED - upregulated

Bench work



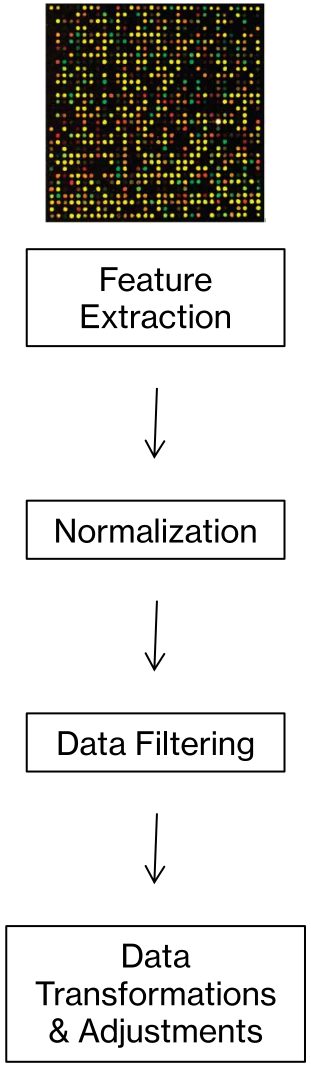
RNA extraction

cDNA synthesis and Cy3/5 labeling

Hybridize to Microarray

Dry & Scan

Data cleanup



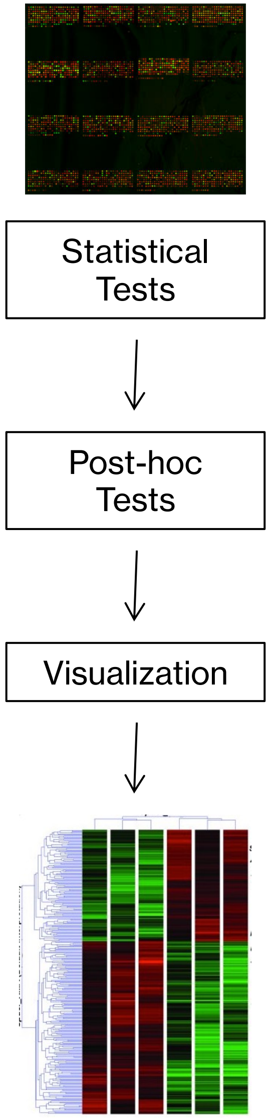
Feature Extraction

Normalization

Data Filtering

Data Transformations & Adjustments

Analysis



Statistical Tests

Post-hoc Tests

Visualization

Bench work

RNA extraction

↓

cDNA synthesis and Cy3/5 labeling

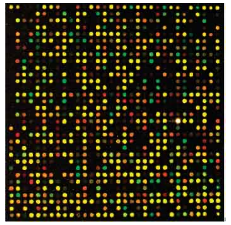
↓

Hybridize to Microarray

↓

Dry & Scan

Data cleanup



Feature Extraction

↓

Normalization

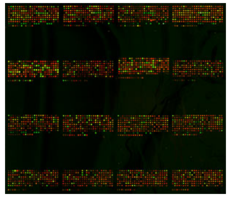
↓

Data Filtering

↓

Data Transformations & Adjustments

Analysis



Statistical Tests

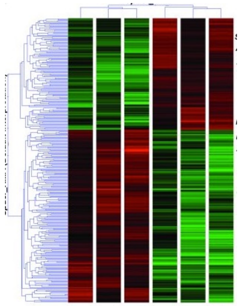
↓

Post-hoc Tests

↓

Visualization

↓



Feature Extraction

- Vendor Software
- Spot finding
- Outlier detection
- Measure feature intensity (Cy3 and/or Cy5)
- Background Subtraction or Detrending
- Commercial microarrays make extensive use of internal control probes

Two-Dye Approaches - Normalization

- Human error and imprecision of tools leads to slightly different loading of Cy3 labeled cDNA and the Cy5 labeled cDNA on the microarray
- Normalization attempts to factor out this technical variation
  - Normalize within microarrays (Cy3 versus Cy5 load)
  - Normalize among microarrays (each microarray will have slightly different loadings)
- Lowess normalization most commonly used (see Flash Update)

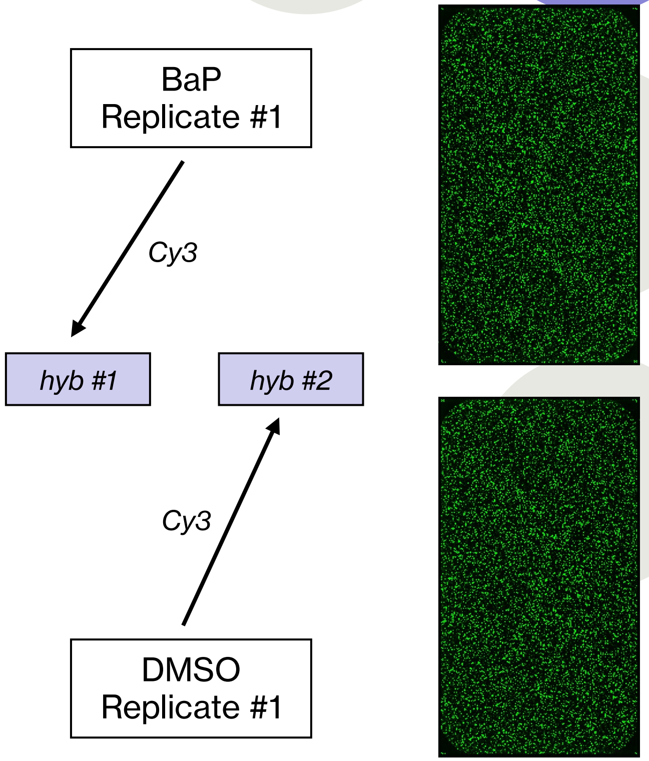
Feature Extraction

- Vendor Software
- Spot finding
- Outlier detection
- Measure feature intensity (Cy3 and/or Cy5)
- Background Subtraction or Detrending
- Commercial microarrays make extensive use of internal control probes

Two-Dye Approaches - Normalization

Single Sample Approaches

- Single sample approaches best for high quality microarrays & BeadChip technologies
- Each sample is hybridized to its own microarray
- You rarely see two-dye approaches anymore
- Only uses Cy3 and thus saves money on dye usage
- Cy3 is also more stable to laboratory ozone
- Avoids difficult statistical properties of fold change estimates
- Measurement is the intensity of Cy3 for each sample – how bright is the green?



BaP Replicate #1

Cy3

hyb #1

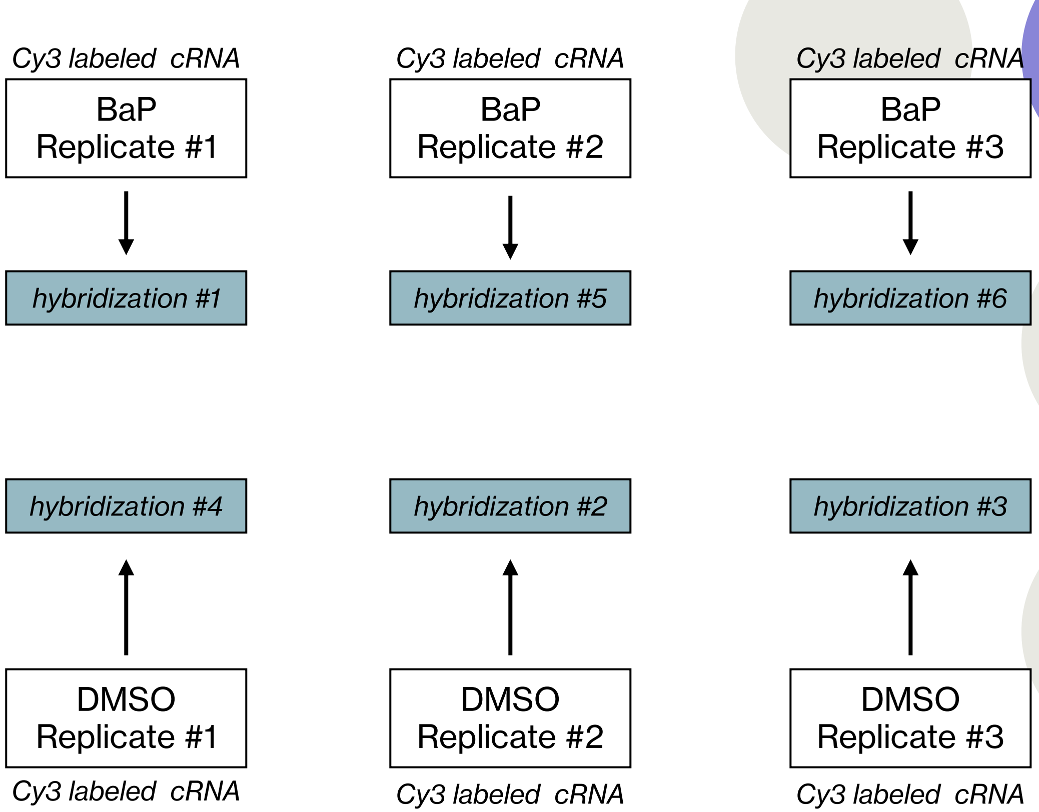
hyb #2

Cy3

DMSO Replicate #1

Normalization

- Each sample is hybridized to its own microarray
- Dye normalization is needed to factor out microarray loading error
- Non-linear scaling method based on rank invariant probes (see lab)
- Result is reliable estimates of relative transcript abundance not fold change estimates



Cy3 labeled cRNA
BaP Replicate #1
hybridization #1

hybridization #4
DMSO Replicate #1
Cy3 labeled cRNA

Cy3 labeled cRNA
BaP Replicate #2
hybridization #5

hybridization #2
DMSO Replicate #2
Cy3 labeled cRNA

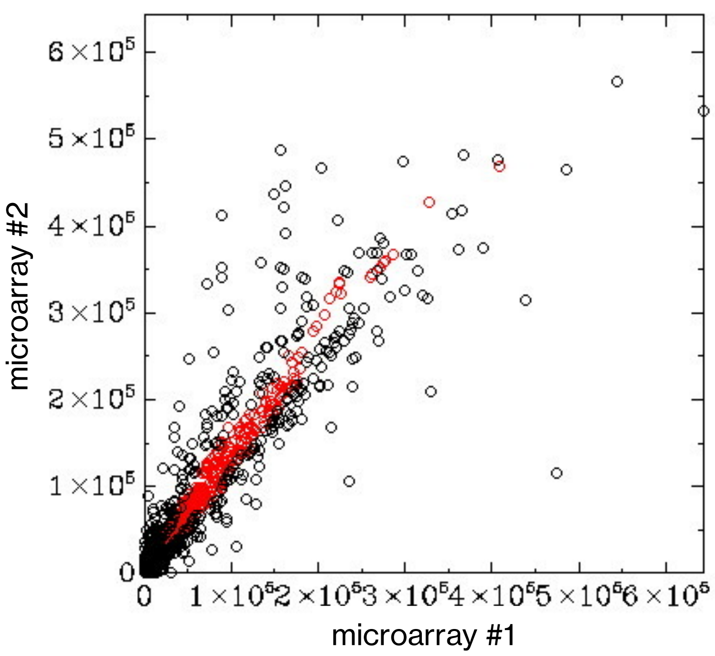
Cy3 labeled cRNA
BaP Replicate #3
hybridization #6

hybridization #3
DMSO Replicate #3
Cy3 labeled cRNA

Normalization

- Each sample is hybridized to its own microarray
- Dye normalization is needed to factor out microarray loading error
- Non-linear scaling method based on rank invariant probes (see lab)
- Result is reliable estimates of relative transcript abundance not fold change estimates

pre-normalization

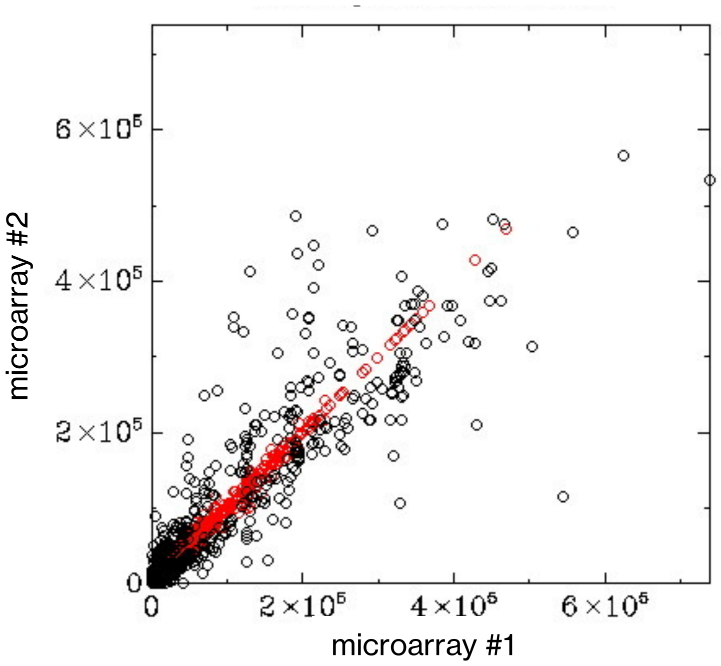


Cy3 intensity between two samples. Putative housekeeping genes in red – expected to be in equal abundance in both samples. Non-linear normalization needed based on plot above.

Normalization

- Each sample is hybridized to its own microarray
- Dye normalization is needed to factor out microarray loading error
- Non-linear scaling method based on rank invariant probes (see lab)
- Result is reliable estimates of relative transcript abundance not fold change estimates

pre-normalization



Cy3 intensity between two samples. Putative housekeeping genes in red – expected to be in equal abundance in both samples. Non-linear normalization needed based on plot above.

Bench work

RNA extraction

↓

cDNA synthesis and Cy3/5 labeling

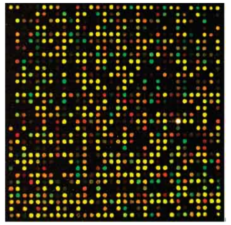
↓

Hybridize to Microarray

↓

Dry & Scan

Data cleanup



Feature Extraction

↓

Normalization

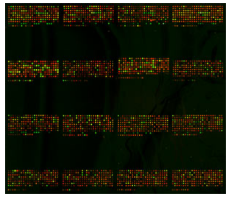
↓

Data Filtering

↓

Data Transformations & Adjustments

Analysis



Statistical Tests

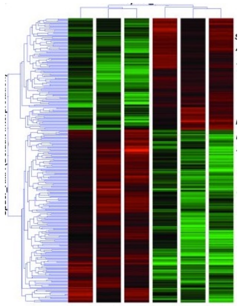
↓

Post-hoc Tests

↓

Visualization

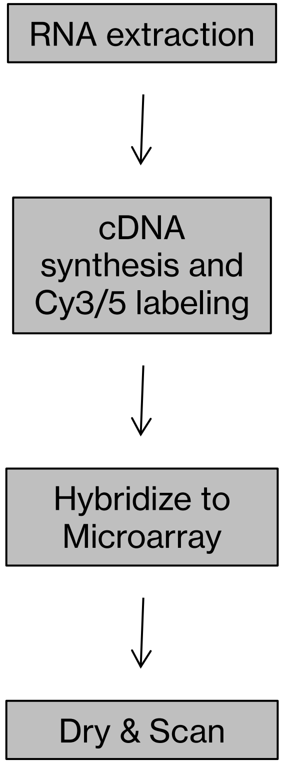
↓



Filtering

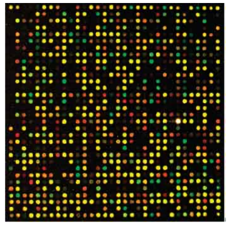
- Saturated probes – Cy3 signal at it’s maximum – all probes hybridized
- Probes not above background
  - within 2.6 x standard error of background
- Non-uniform probes
- Poorly replicated probes

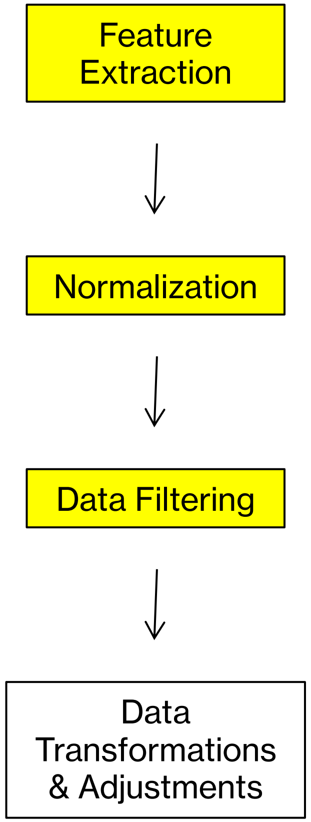
Bench work



RNA extraction ↓ cDNA synthesis and Cy3/5 labeling ↓ Hybridize to Microarray ↓ Dry & Scan

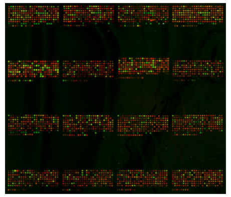
Data cleanup

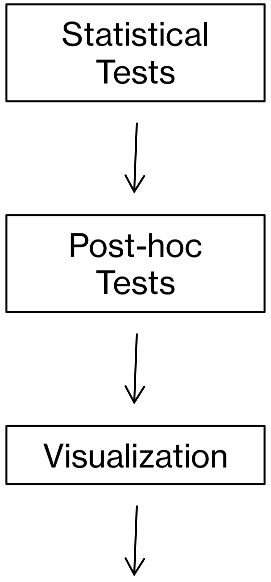




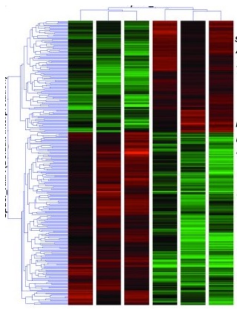
Feature Extraction ↓ Normalization ↓ Data Filtering ↓ Data Transformations & Adjustments

Analysis

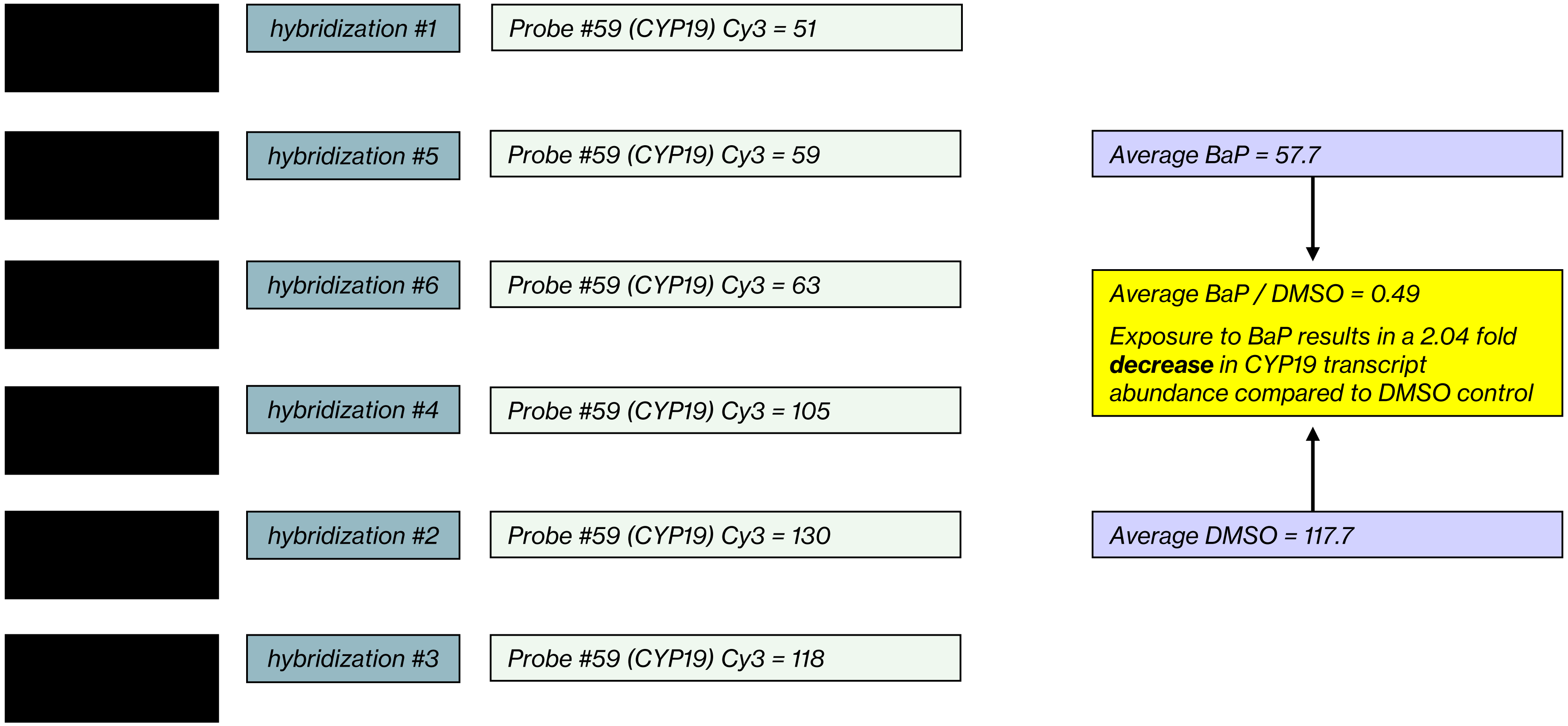




Statistical Tests ↓ Post-hoc Tests ↓ Visualization ↓



Experimental Design



[redacted] hybridization #1 — Probe #59 (CYP19) Cy3 = 51

[redacted] hybridization #5 — Probe #59 (CYP19) Cy3 = 59

[redacted] hybridization #6 — Probe #59 (CYP19) Cy3 = 63

[redacted] hybridization #4 — Probe #59 (CYP19) Cy3 = 105

[redacted] hybridization #2 — Probe #59 (CYP19) Cy3 = 130

[redacted] hybridization #3 — Probe #59 (CYP19) Cy3 = 118

Average BaP = 57.7

Average BaP / DMSO = 0.49

Exposure to BaP results in a 2.04 fold ***decrease*** *in CYP19 transcript abundance compared to DMSO control*

Average DMSO = 117.7

- Normalized values are quantitative and have normal statistical properties
- Mean, standard error, and ANOVA calculations follow normal formulas

Data Transformations

- Log transformation
  - Correction for fold change data in two dye experiments
    - 2 fold increase = 2.0
    - no change = 1.0
    - 2 fold decrease = 0.5
  - Reduce mean and variance relationships in one dye experiments, reduce Type I and Type II error
- Median centering
  - Focus analysis upon variation in the data, not magnitude
  - Important so analysis is not biased toward most abundant transcripts

Bench work

RNA extraction

↓

cDNA synthesis and Cy3/5 labeling

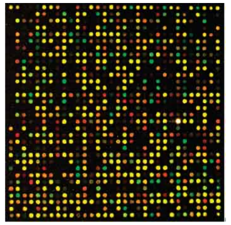
↓

Hybridize to Microarray

↓

Dry & Scan

Data cleanup



Feature Extraction

↓

Normalization

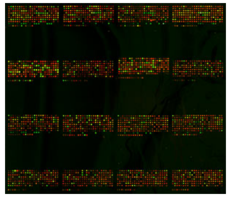
↓

Data Filtering

↓

Data Transformations & Adjustments

Analysis



Statistical Tests

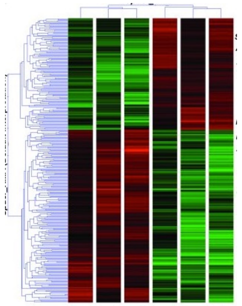
↓

Post-hoc Tests

↓

Visualization

↓



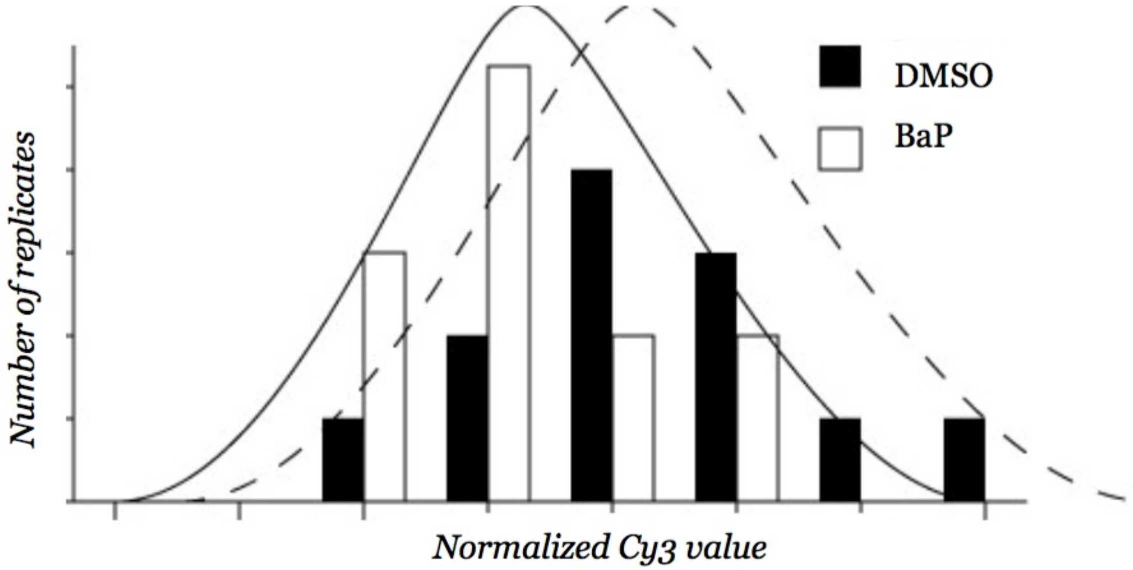
Statistical Significance - Replication

- Microarrays use familiar statistical tests such at t-test and ANOVA
- Increased replication within treatments increases statistical power among treatments
- Glass slide microarrays are expensive so many experiments only use triplication
- BeadChip microarrays are cheaper and often use higher replication

Average BaP / DMSO = 0.49

Exposure to BaP results in a 2.04 fold ***decrease*** *in CYP19 transcript abundance compared to DMSO control*

t-test with p<0.05



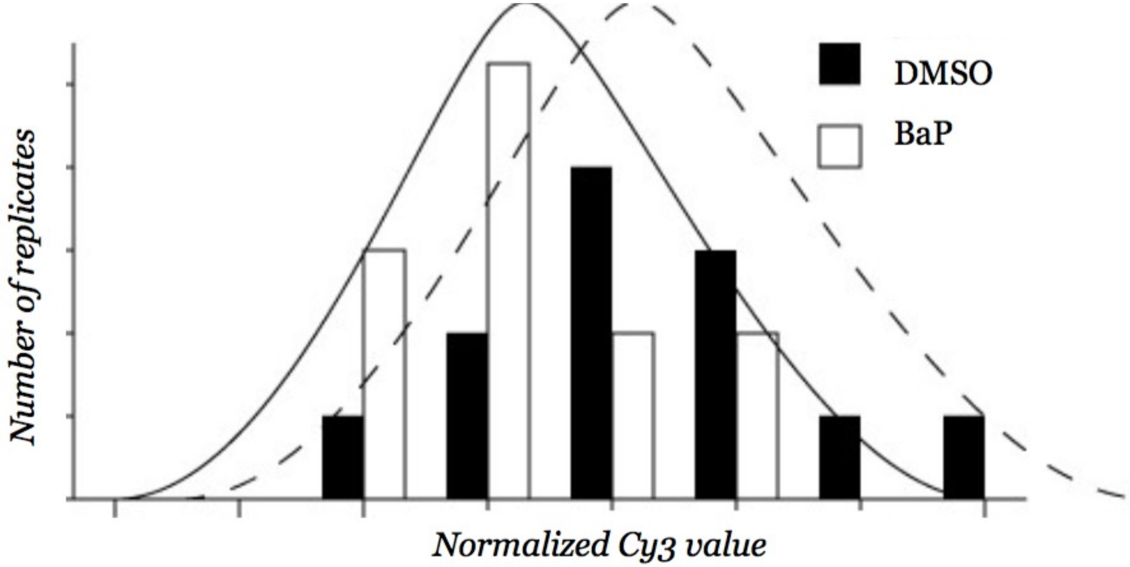
Statistical Significance - Replication

- Type I error
  - false positive or false discovery
  - controlled by the significance level of the test (α), e.g. α=0.05
- Type II error
  - false negative
  - has value β but often actual value is unknown
  - related to Power of the test (1 – β) and experimental replication

Average BaP / DMSO = 0.49

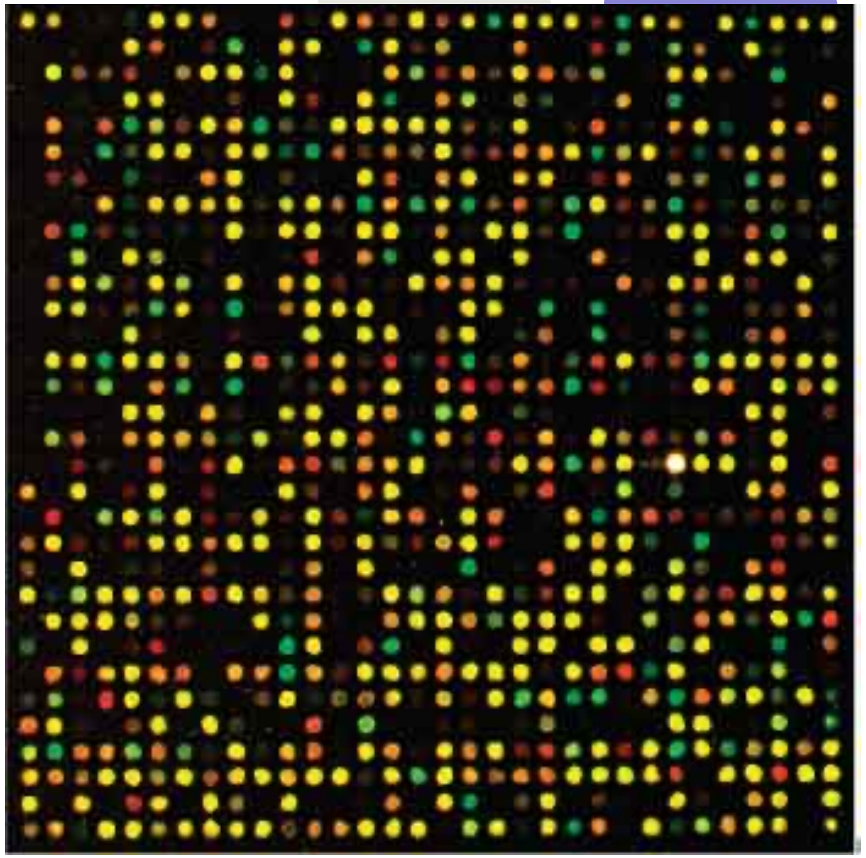
*Exposure to BaP results in a 2.04 fold **decrease** in CYP19 transcript abundance compared to DMSO control*

t-test with p<0.05



Statistical Significance - Error

- Microarrays perform multiple tests, inflating false positives
- Traditional statistics use the Bonferroni correction α/n where n = # tests, but this is often not appropriate for microarrays
- Microarrays methods often instead estimate the false discovery rate (FDR) and use arbitrary cut-offs
- Microarray methods often permutate appropriate FDR test distributions

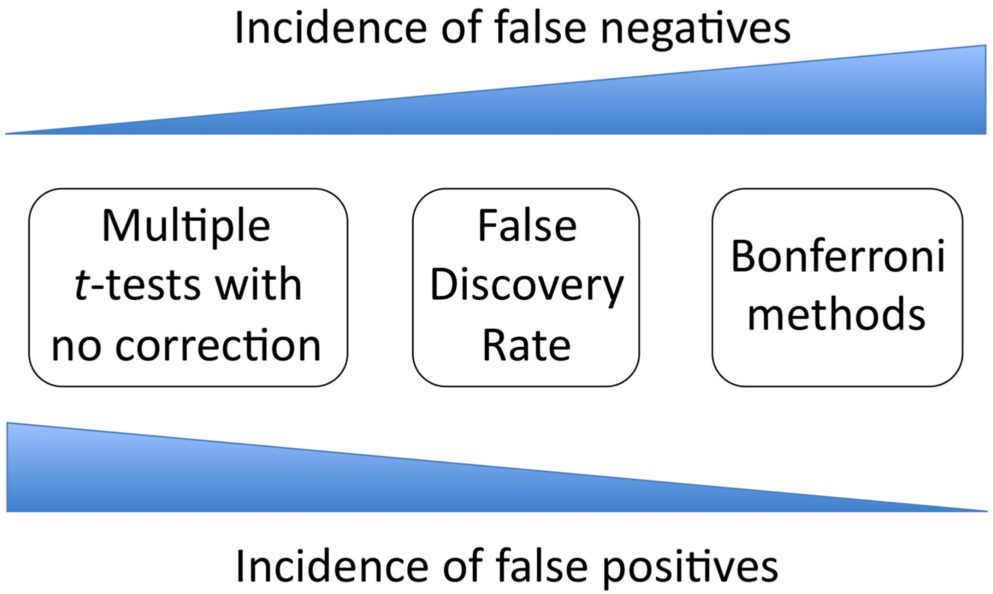


Agilent zebrafish microarray has 43,803 probes

Each spot measures Cy3 for a single probe (i.e. gene) and each spot undergoes it's own significance test. Thus there are 43,803 tests performed on the same mRNA sample – the tests are not independent!

Statistical Significance - Error

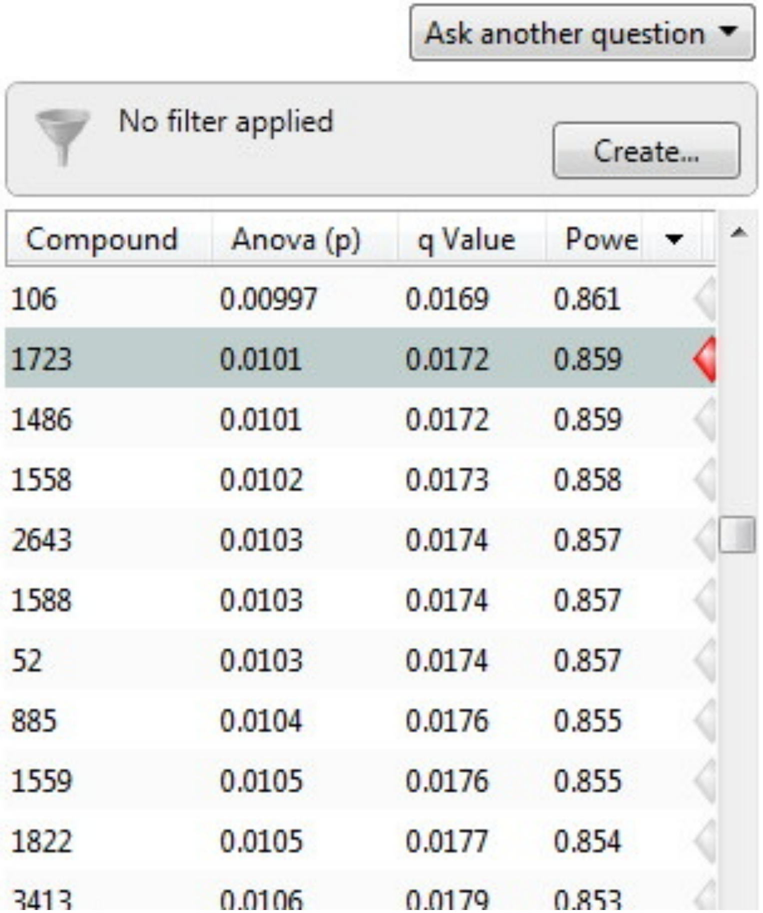
- The more t-tests you run on the same mRNA sample, the greater the chance of obtaining a statistically significant result through chance sampling
- Microarrays thus have the unavoidable presence of false-positive findings (Type I errors)
- Traditional corrections (i.e. Bonferroni) only work for small number of multiple tests; FDR permutations are a compromise for microarrays to use reasonable Type I and Type II error
- **Q-values are the name given to the adjusted p-values found using an optimized FDR approach,** http://www.nonlinear.com/support/progenesis/comet/faq/v2.0/pq-values.aspx


Incidence of false negatives
Multiple *t*-tests with no correction
False Discovery Rate
Bonferroni methods
Incidence of false positives

http://www.wormbook.org/chapters/www_statisticalanalysis/statisticalanalysis.html

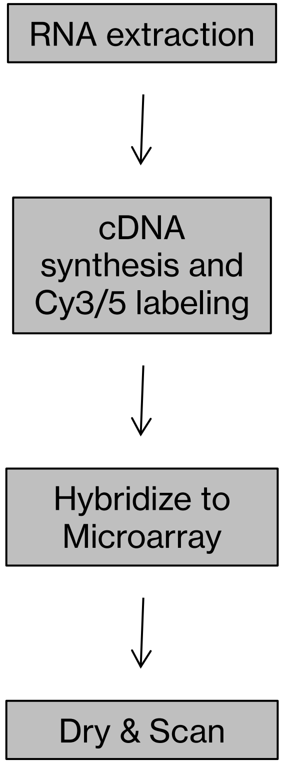
Statistical Significance – Metabolomics Example

- Order the results by q-value, Compound #1723 is the 800th compound in the list of 3516 total compounds
- Compound #1723 has p=0.0101 and q=0.0172
- A p-value of 0.0101 implies a 1.01% chance of false positives in the experiment. 0.0101 x 3516 = 35.51 false positives in the top 800 hits
- A q-value of 0.0172 implies only 1.72% of the top 800 compounds are false positives = 800 * 0.0172 = 13.76 false positives
- p-values are biased by 3516 multiple tests, q-values correct for this effect



Compound	Anova (p)	q Value	Powe
106	0.00997	0.0169	0.861
1723	0.0101	0.0172	0.859
1486	0.0101	0.0172	0.859
1558	0.0102	0.0173	0.858
2643	0.0103	0.0174	0.857
1588	0.0103	0.0174	0.857
52	0.0103	0.0174	0.857
885	0.0104	0.0176	0.855
1559	0.0105	0.0176	0.855
1822	0.0105	0.0177	0.854
3413	0.0106	0.0179	0.853

Bench work



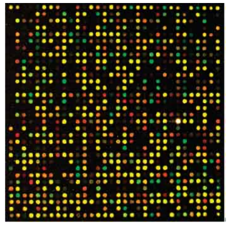
RNA extraction

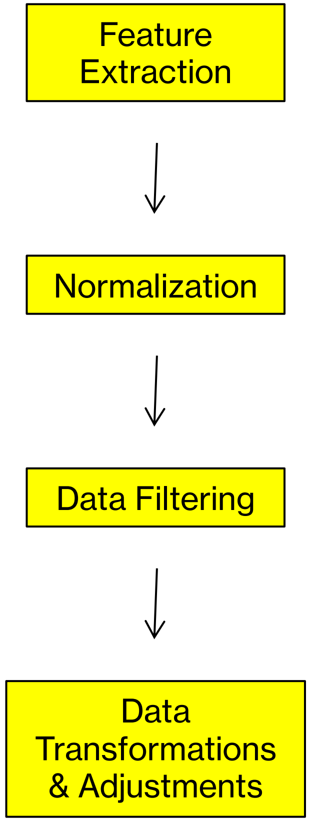
cDNA synthesis and Cy3/5 labeling

Hybridize to Microarray

Dry & Scan

Data cleanup





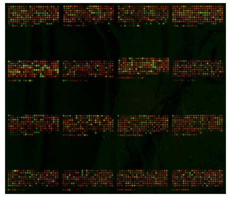
Feature Extraction

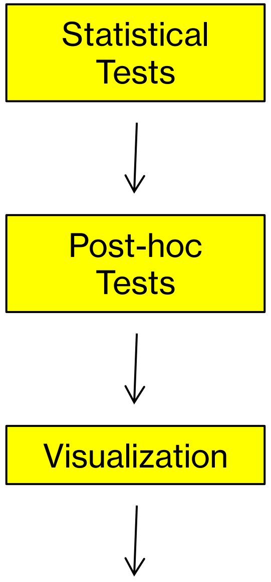
Normalization

Data Filtering

Data Transformations & Adjustments

Analysis

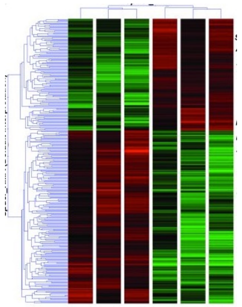




Statistical Tests

Post-hoc Tests

Visualization

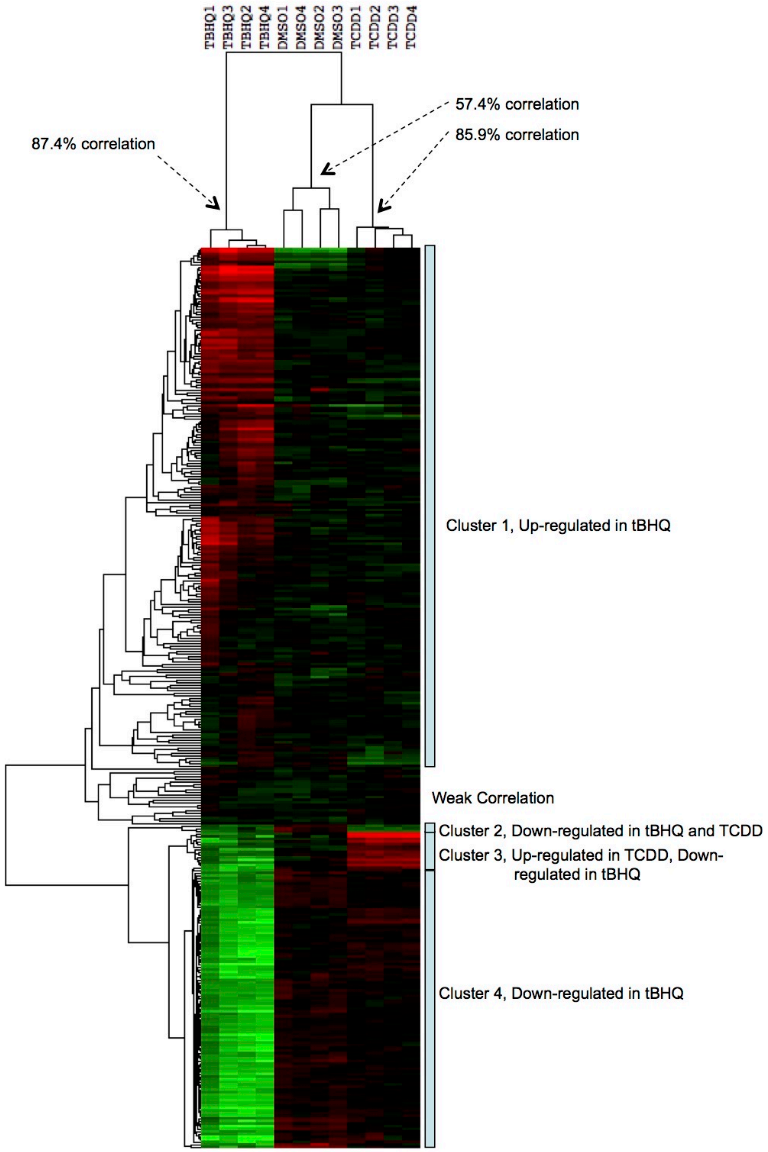


Visualization

- Heat Maps
  - variety of clustering algorithms
  - clusters of co-variation may indicate common regulatory pathways
  - particularly good for time series data

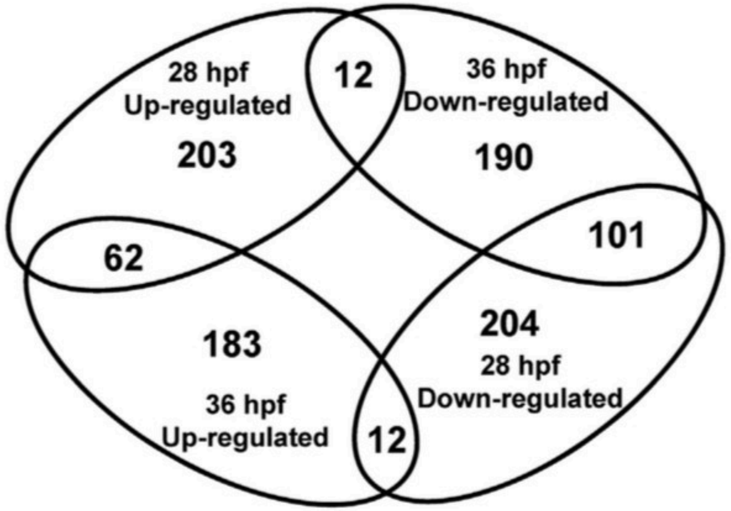
Visualization

- Cluster by genes (rows) and / or treatments (columns)
- Clusters are hierarchical
- Smaller clusters have higher correlation in expression patterns
- Median centre to downplay abundance and focus analysis on patterns



Visualization

- Heat Maps
  - variety of clustering algorithms
  - clusters of co-variation may indicate common regulatory pathways
  - particularly good for time series data
- VENN diagrams
  - excellent for visualizing shared response and designing follow-up experiments

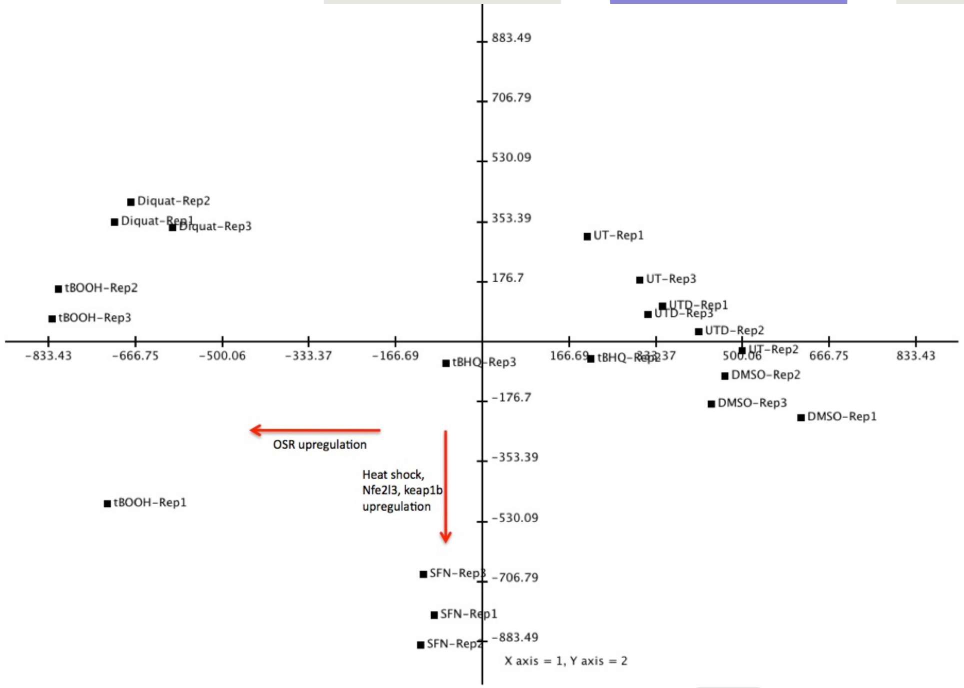


28 hpf (eGFP → dnMTF-1)
277 probes up-regulated and 317 probes down-regulated

36 hpf (eGFP → dnMTF-1)
257 probes up-regulated and 303 probes down-regulated

Visualization

- Heat Maps
  - variety of clustering algorithms
  - clusters of co-variation may indicate common regulatory pathways
  - particularly good for time series data
- VENN diagrams
  - excellent for visualizing shared response and designing follow-up experiments
- Principal Component Analysis (PCA)
  - powerful tool for trend analysis
  - look at the genes involved in the trends – what are they?
  - use DAVID for Gene Ontology term enrichment analysis for genes involved in trends

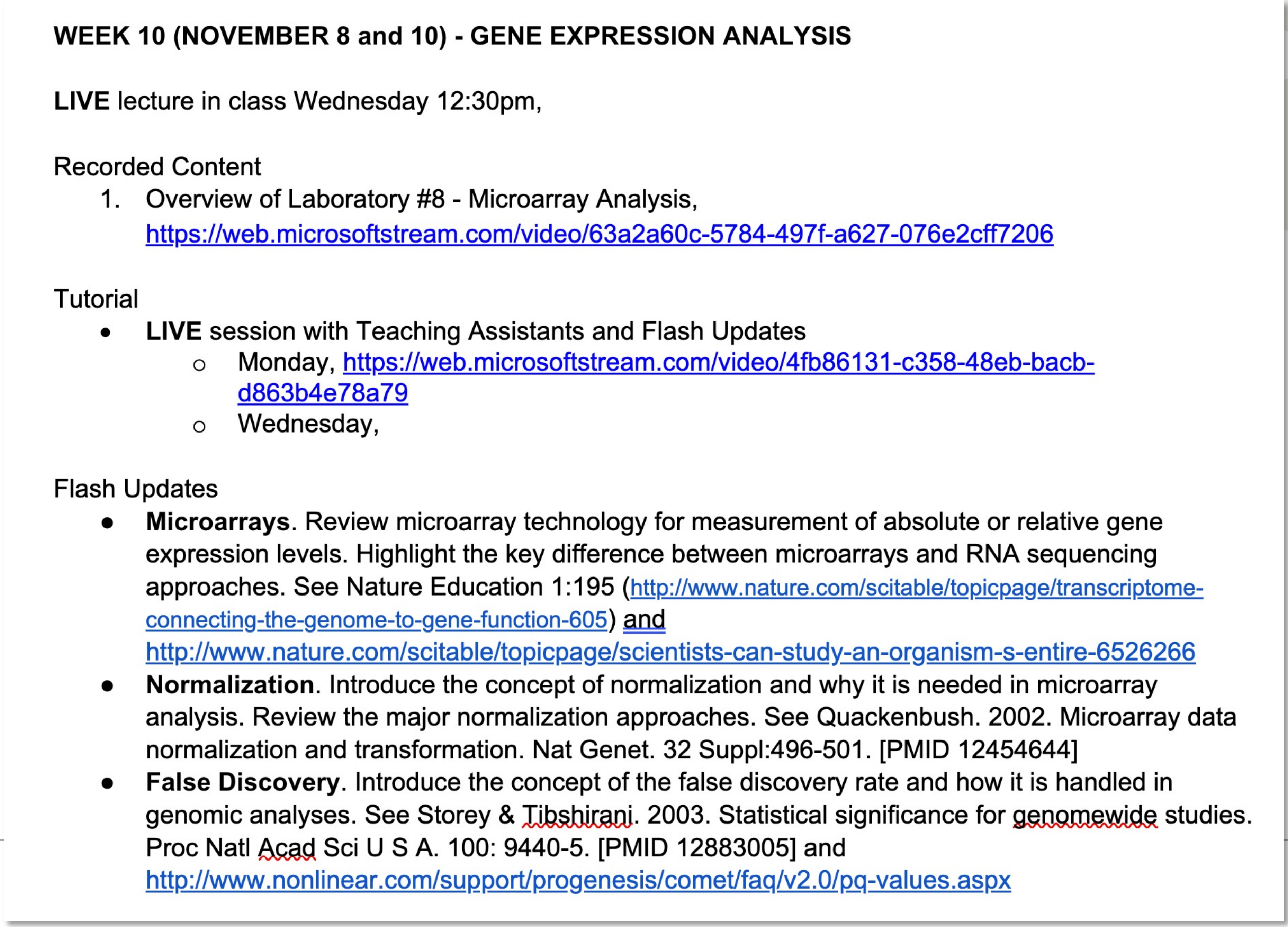


Conclusions

- Design of the experiment and the microarray probes critical
  - How many probes?
  - Quality of probes for each gene?
  - Coverage of the transcriptome and target genes?
  - Coverage of the genome and hypothetical genes?
  - Multiple probes for genes?
- Design and analysis involve trade-offs
  - Level of replication, experiment cost, **statistical power**
  - Replication costs: BeadChip < microarray < RNA-Seq
  - Subtle data may require permissive analyses
    - Allow higher false discovery rate?
    - Downplay variance (e.g. rank product methods)?
- Visualization and post-hoc tests needed when large numbers of probes are significant

Conclusions

- Independent verification often required
    - Quantitative PCR verification of selected results
    - Gene knockdown or over-expression experiments
- Trust and Reliability
    - Robust experimental design, microarray probes, and statistical analyses engender trust of overall results
    - Subtle results, high variation, or downplaying of error require more extensive independent verification of overall results
- Interpretation of overall results can be difficult
    - Fold change versus biological relevance
    - Poorly understood genes
    - Prediction of underlying biological processes
        - GO enrichment? Interactome analyses? KEGG?



WEEK 10 (NOVEMBER 8 and 10) - GENE EXPRESSION ANALYSIS

LIVE lecture in class Wednesday 12:30pm,

Recorded Content

1. Overview of Laboratory #8 - Microarray Analysis, https://web.microsoftstream.com/video/63a2a60c-5784-497f-a627-076e2cff7206

Tutorial

- **LIVE** session with Teaching Assistants and Flash Updates
    - Monday, https://web.microsoftstream.com/video/4fb86131-c358-48eb-bacb-d863b4e78a79
    - Wednesday,

Flash Updates

- **Microarrays**. Review microarray technology for measurement of absolute or relative gene expression levels. Highlight the key difference between microarrays and RNA sequencing approaches. See Nature Education 1:195 (http://www.nature.com/scitable/topicpage/transcriptome-connecting-the-genome-to-gene-function-605) and http://www.nature.com/scitable/topicpage/scientists-can-study-an-organism-s-entire-6526266
- **Normalization**. Introduce the concept of normalization and why it is needed in microarray analysis. Review the major normalization approaches. See Quackenbush. 2002. Microarray data normalization and transformation. Nat Genet. 32 Suppl:496-501. [PMID 12454644]
- **False Discovery**. Introduce the concept of the false discovery rate and how it is handled in genomic analyses. See Storey & Tibshirani. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 100: 9440-5. [PMID 12883005] and http://www.nonlinear.com/support/progenesis/comet/faq/v2.0/pq-values.aspx