Microarray Preprocessing

by Stuart Andrews with help from Uri-David Akavia & Bo-Juen Chen

Gene Expression

- A key function of all living organisms is the expression of genes

Gene Expression

- mRNA is created from DNA

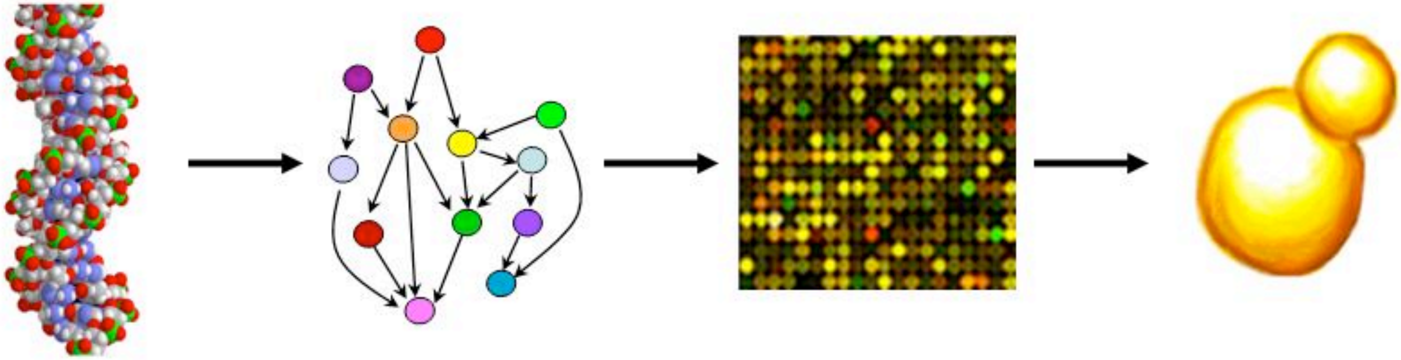
Gene Expression

- **protein is created from mRNA**

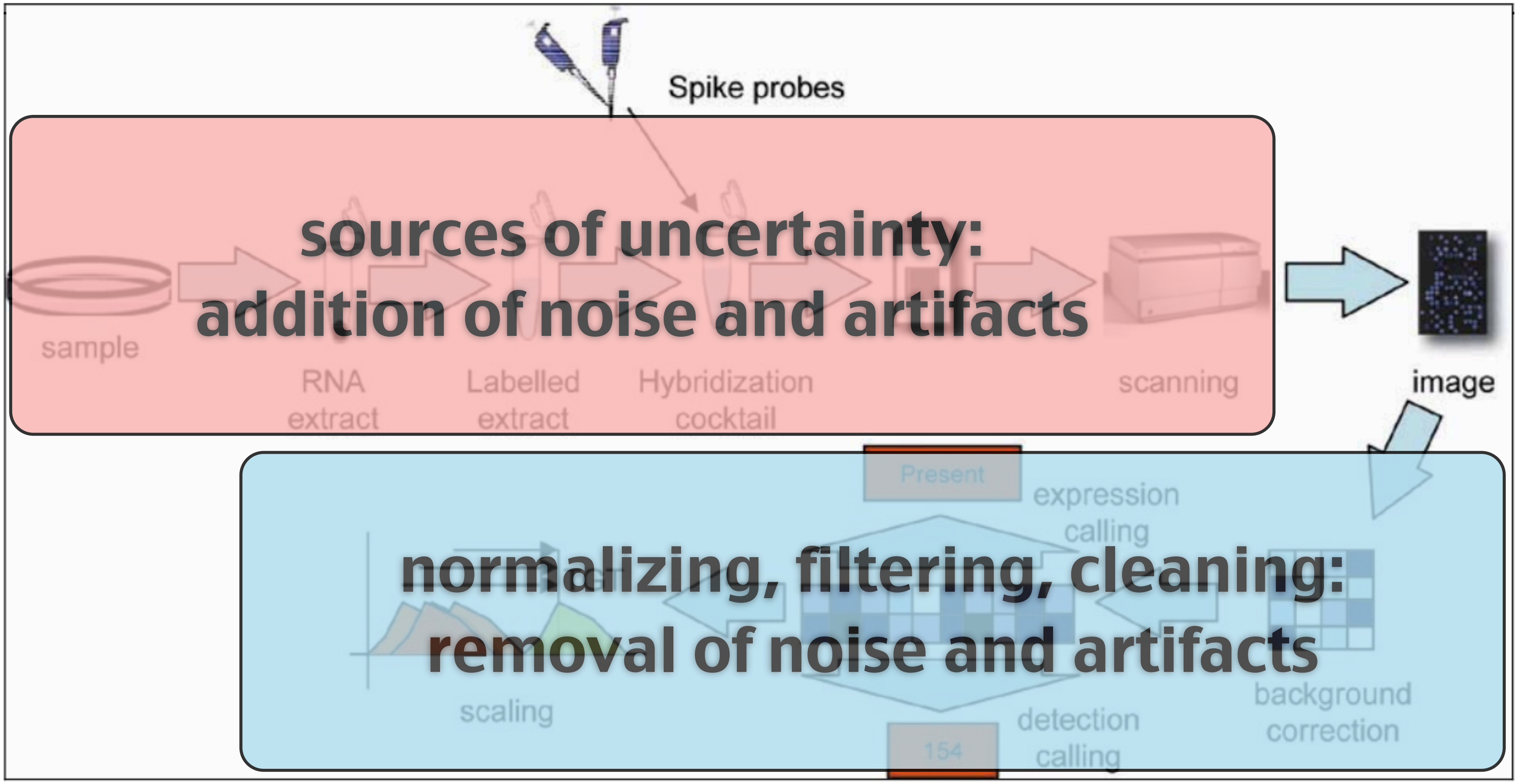
Microarrays

- High-throughput technology used to measure concentration of mRNA in a sample

Applications



Measurement Process



sources of uncertainty:
addition of noise and artifacts

normalizing, filtering, cleaning:
removal of noise and artifacts

Goal

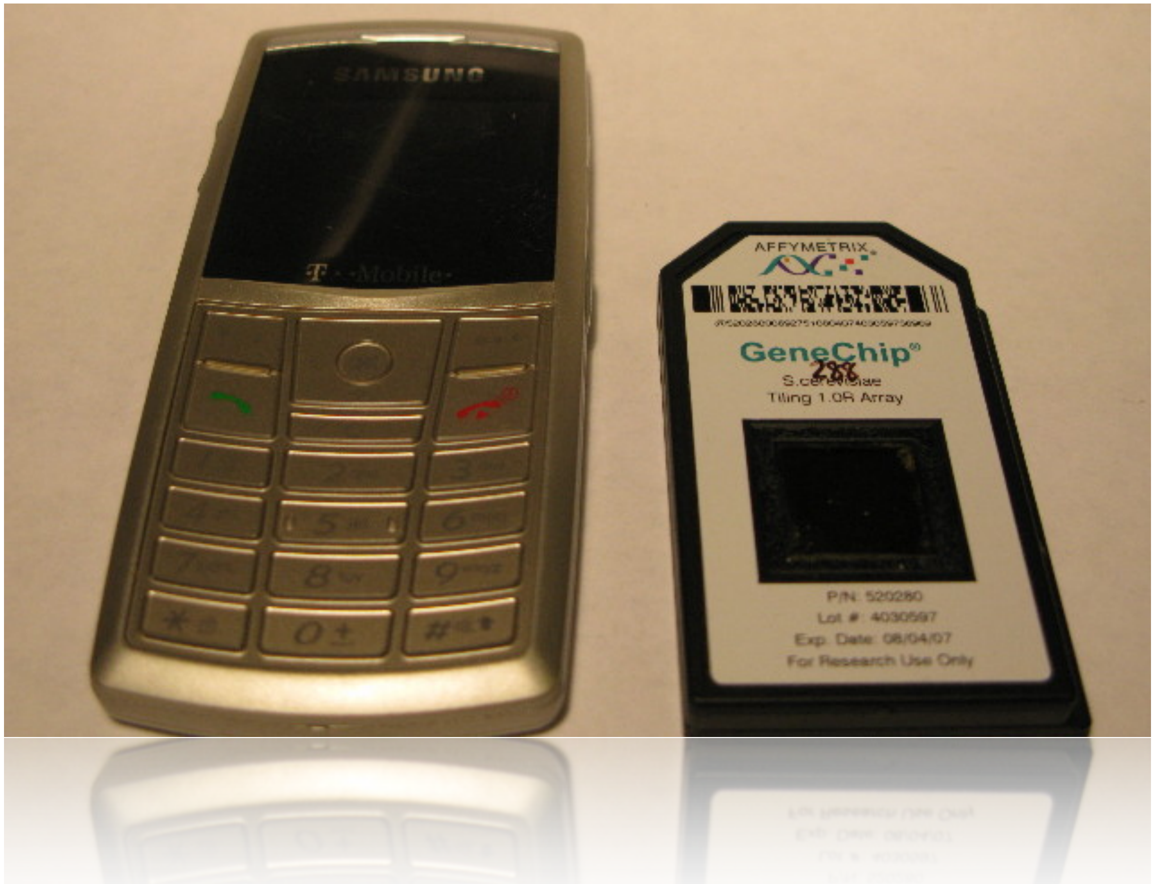
- Framework for preprocessing raw microarray data
- Produce reliable gene expressions
  - Jeff Settleman’s data

Strategy

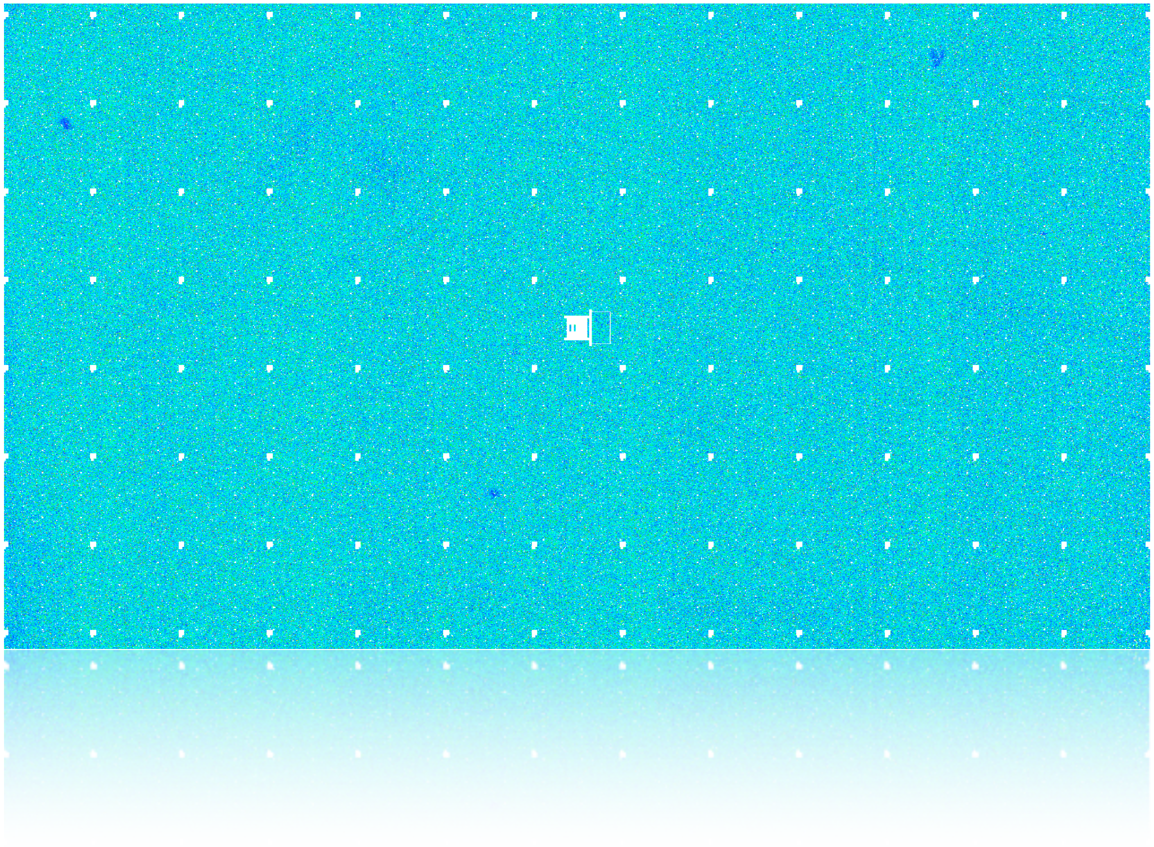
- Keep it simple, efficient, easy to modify, and easy to re-run
- Use visualizations to test, communicate, and confirm results

Overview

- Introduction to microarrays
- Probe normalization & summarization
- Batch effects
- Redundant probe set summarization
- From A to Z
- Summary


SAMSUNG
T-Mobile
AFFYMETRIX
GeneChip®
288
S.cerevisiae
Tiling 1.0R Array
P/N: 520280
Lot #: 4030597
Exp. Date: 08/04/07
For Research Use Only

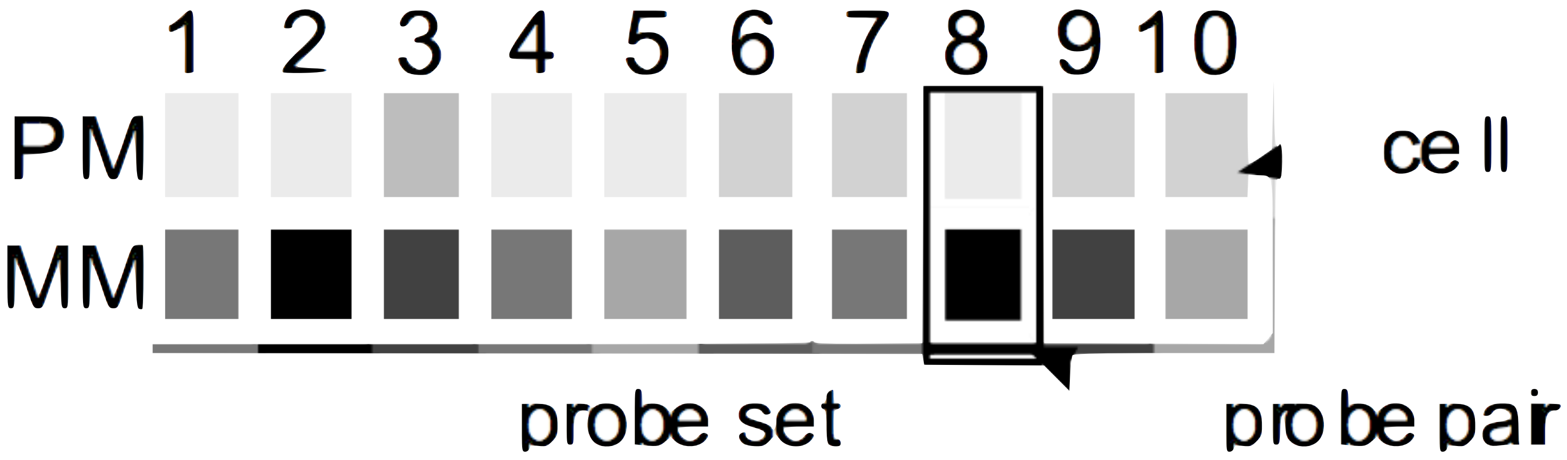
Micro



Array

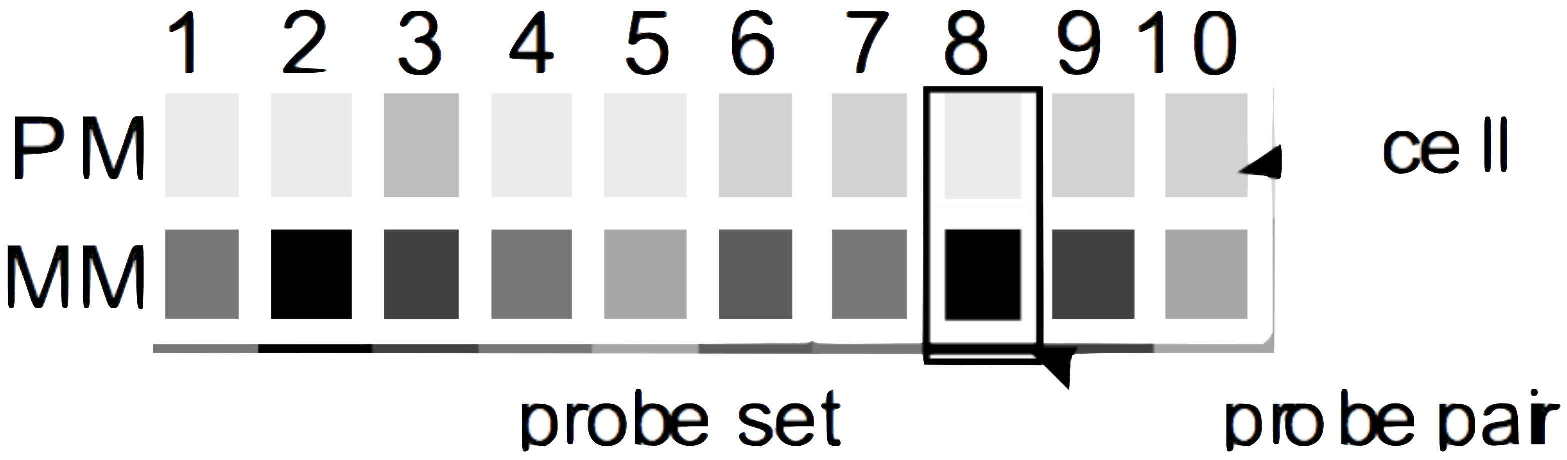
Probes, Probe Pairs, & Probe Sets

- Probes are short oligomers e.g. 25-mers



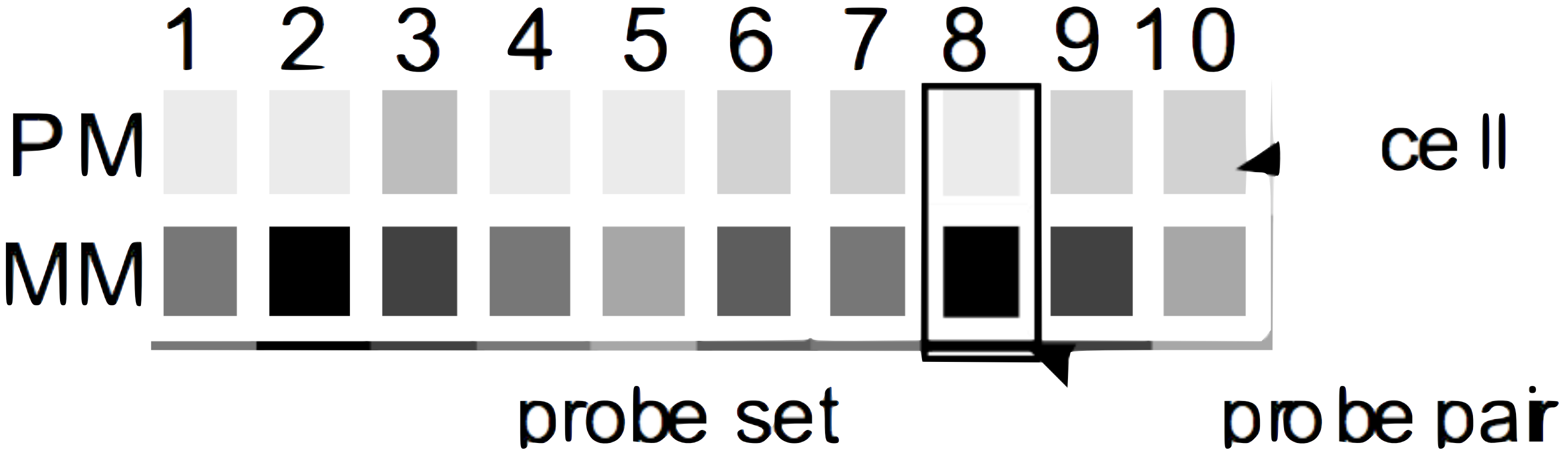
Probes, Probe Pairs, & Probe Sets

- **Perfect match & mismatch pairs**



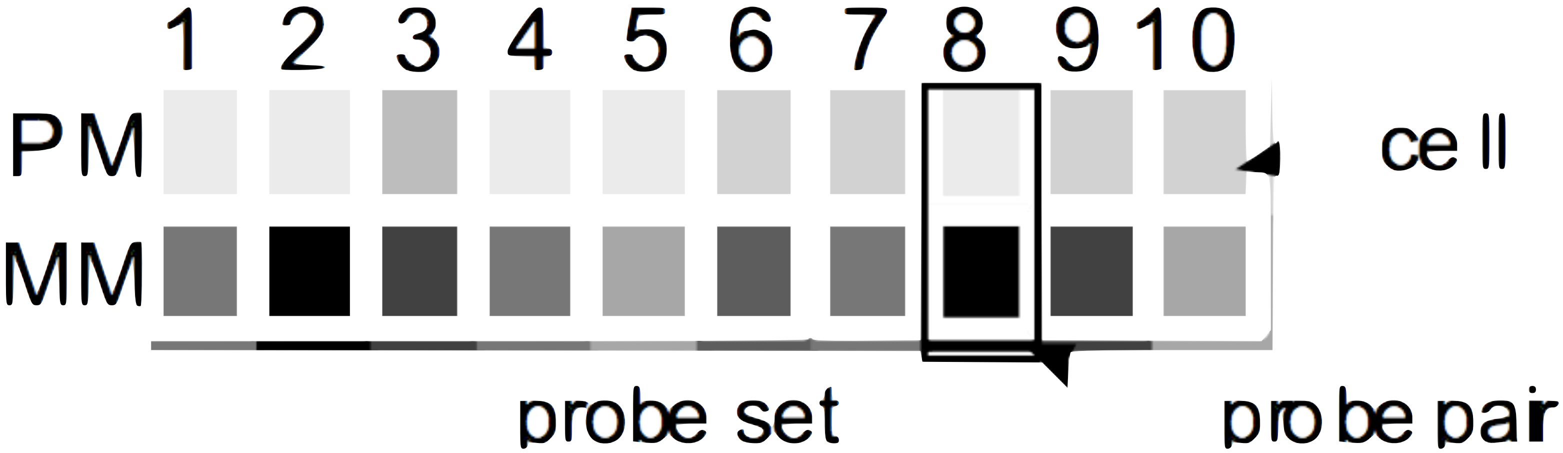
Probes, Probe Pairs, & Probe Sets

- 11–20 probe pairs per gene = a probe set



Probes, Probe Pairs, & Probe Sets

- **Many genes have redundant probe sets**



Chip Design

- Sample preparation & hybridization
  - Datasheet.pdf
- Probe sequence & gene auxiliary data
  - Annotation file.csv
- Probes locations & gene mappings
  - Library file.csv
- Intensity values
  - Sample.CEL

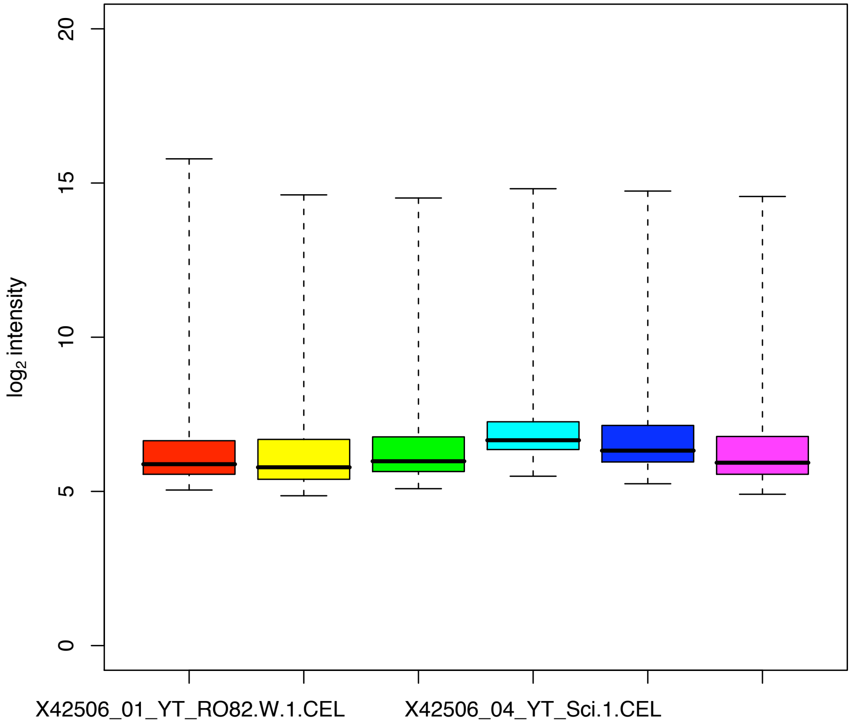
Filter Rule #1

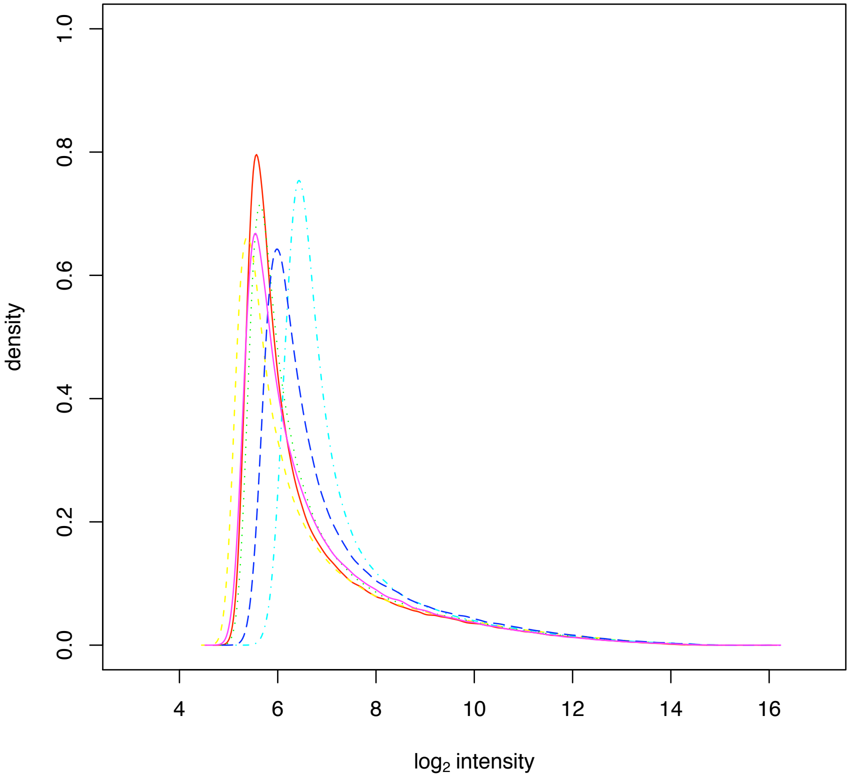
- **Remove control & garbage probes**
  - **Probes included for quality control**
  - **Probes discovered to encode incorrect sequence, as the Human genome is revised**
  - **Probes discovered to hybridize to antisense strand**
  - **Probes with annotation warnings, etc.**

Probe Normalization & Summarization

Inputs: Probe Intensities

Each color is a different chip (sample)





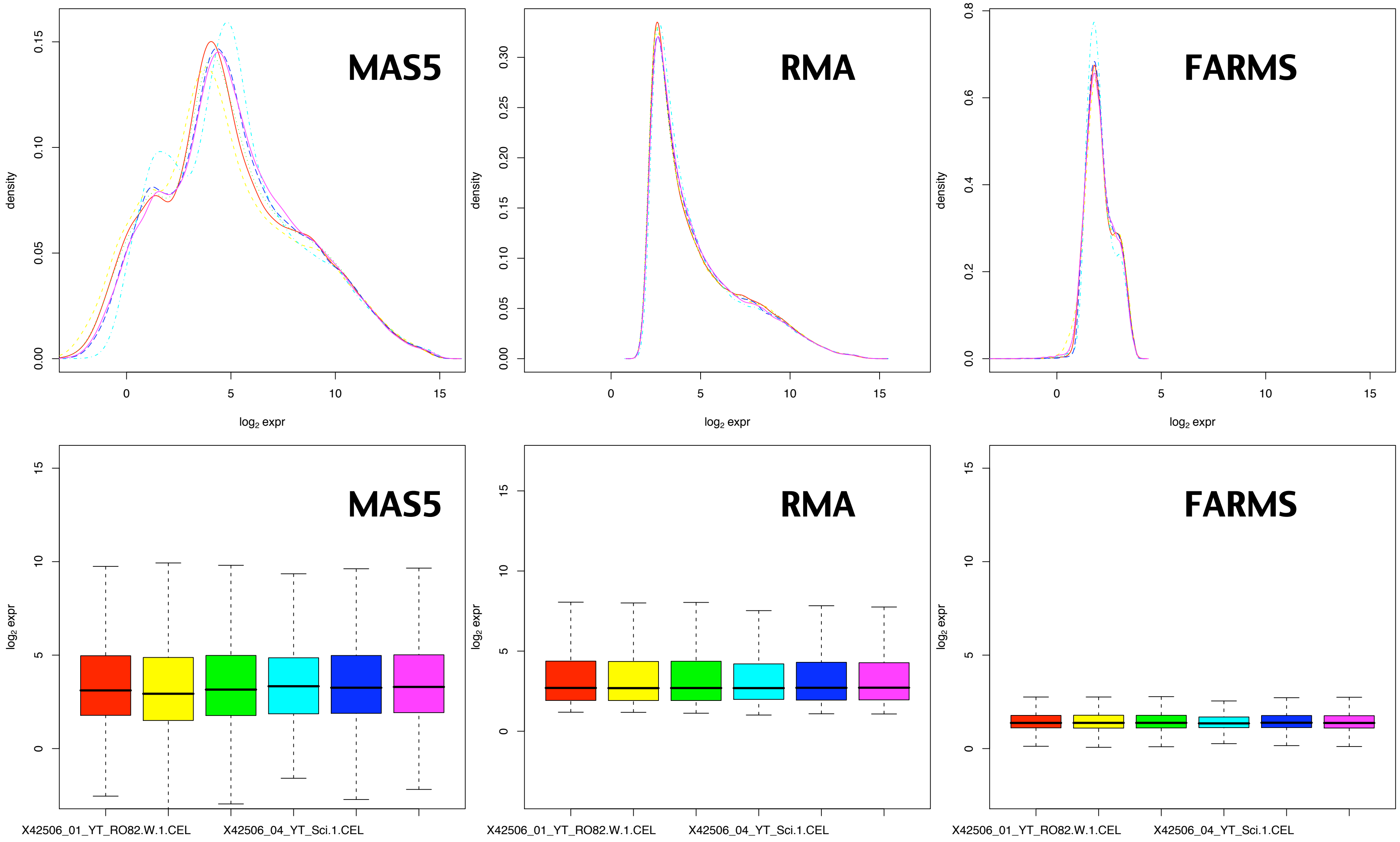
Outputs: Transcript Exprs

- Calculate a single expression value for a transcript from a set of probes
  - Remove background chip effects
  - Normalize intensity values
  - Summarize

Algorithms

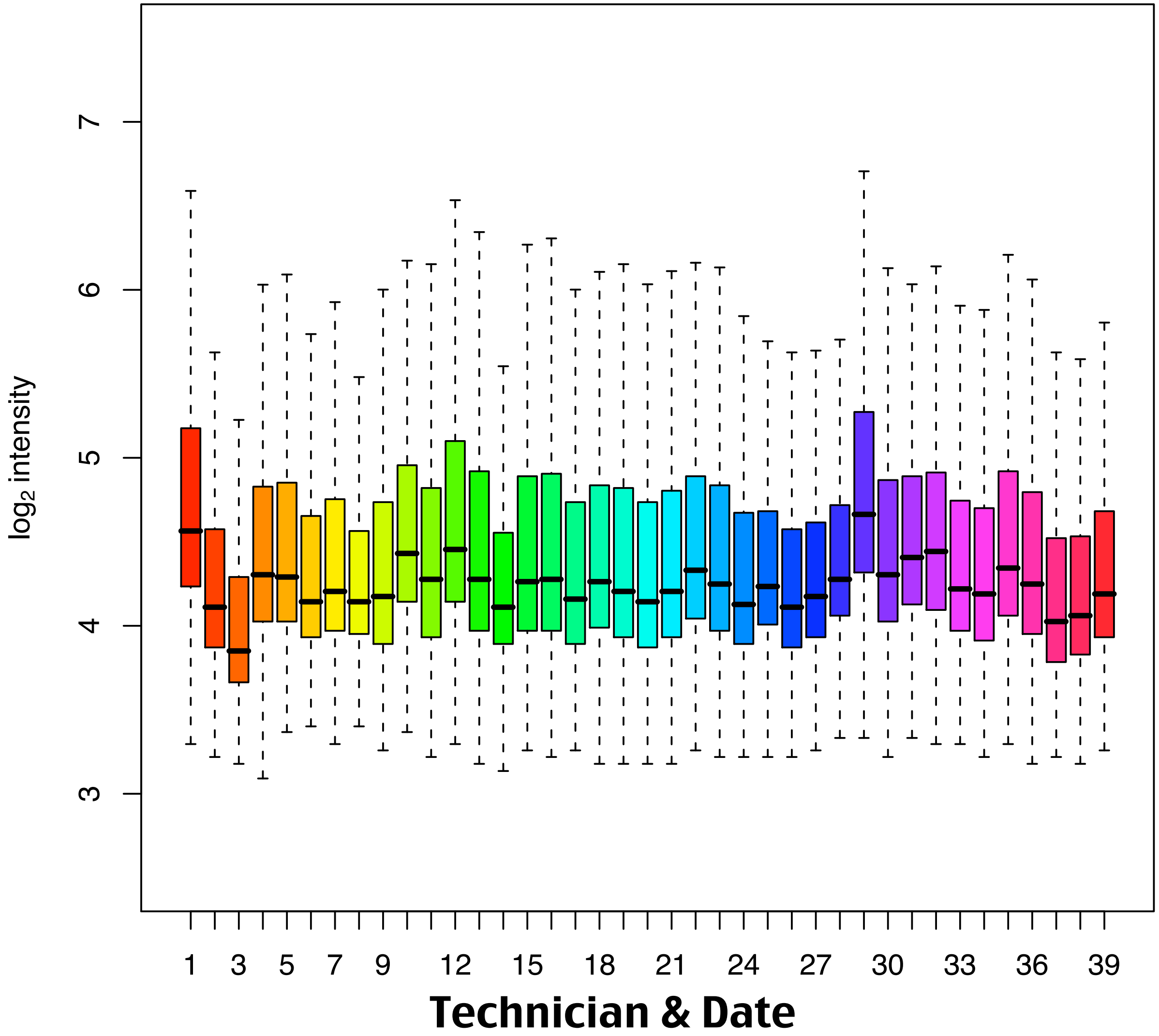
- MAS5
  - local average & PM-MM background correction
  - robust Tukey-biweight summarization
  - fixed scale
- RMA
  - normal-exponential foreground/background model
  - linear model of log expression
- Plier, MBEI, FARMS, many others ...

Log2 Probe Set Expressions



Batch Effects

Batches



Batch variations:

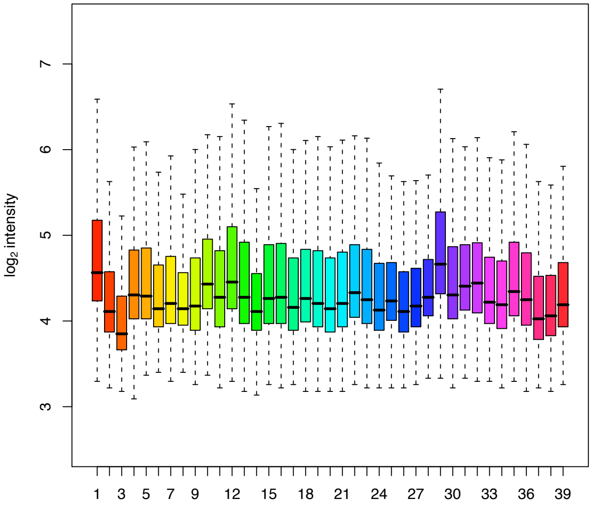
Biology,
or
artifacts?

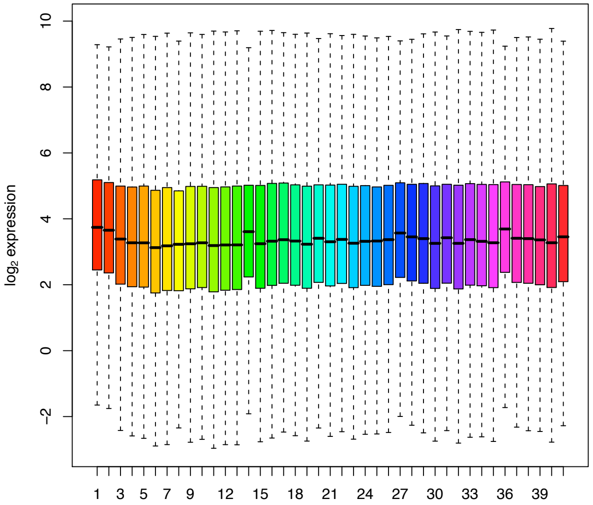
Batches

CEL Intensities

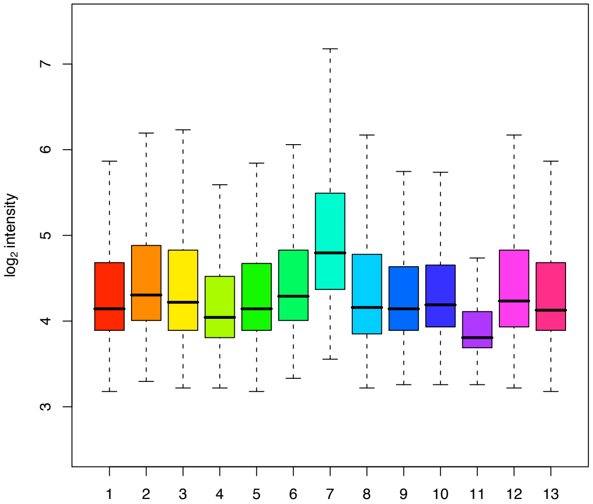
MAS5 Expressions

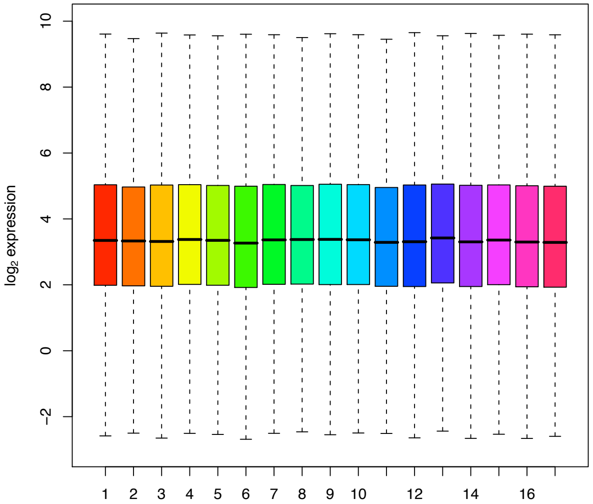
Technician and date





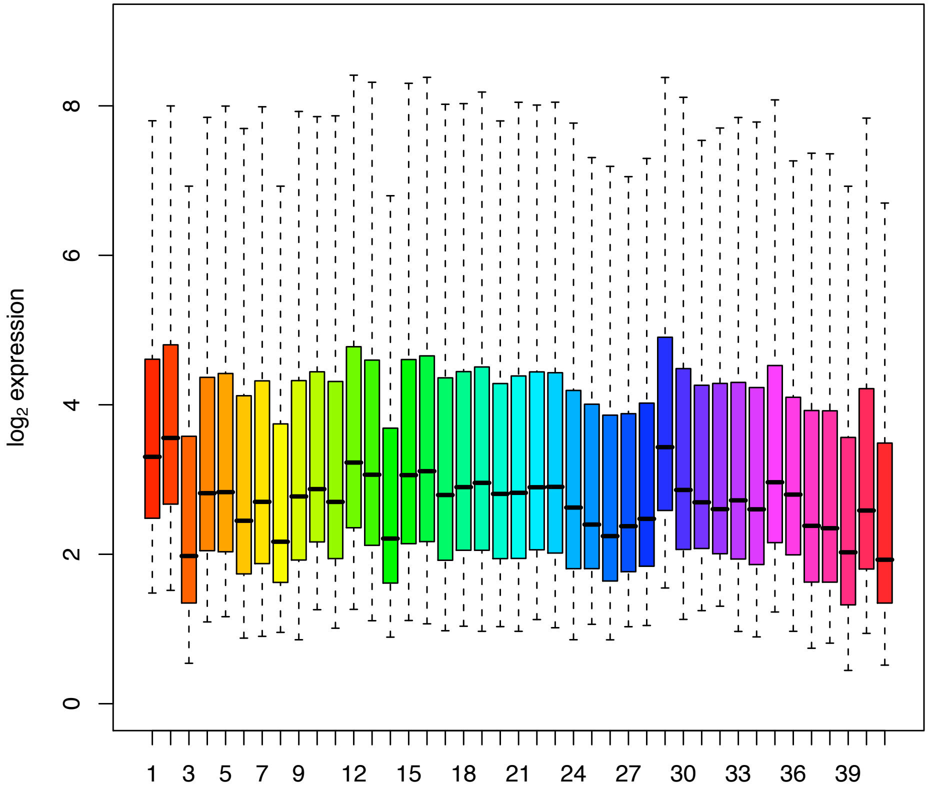
Tissue





Caveat

RMA expressions computed batch-by-batch



Problem:

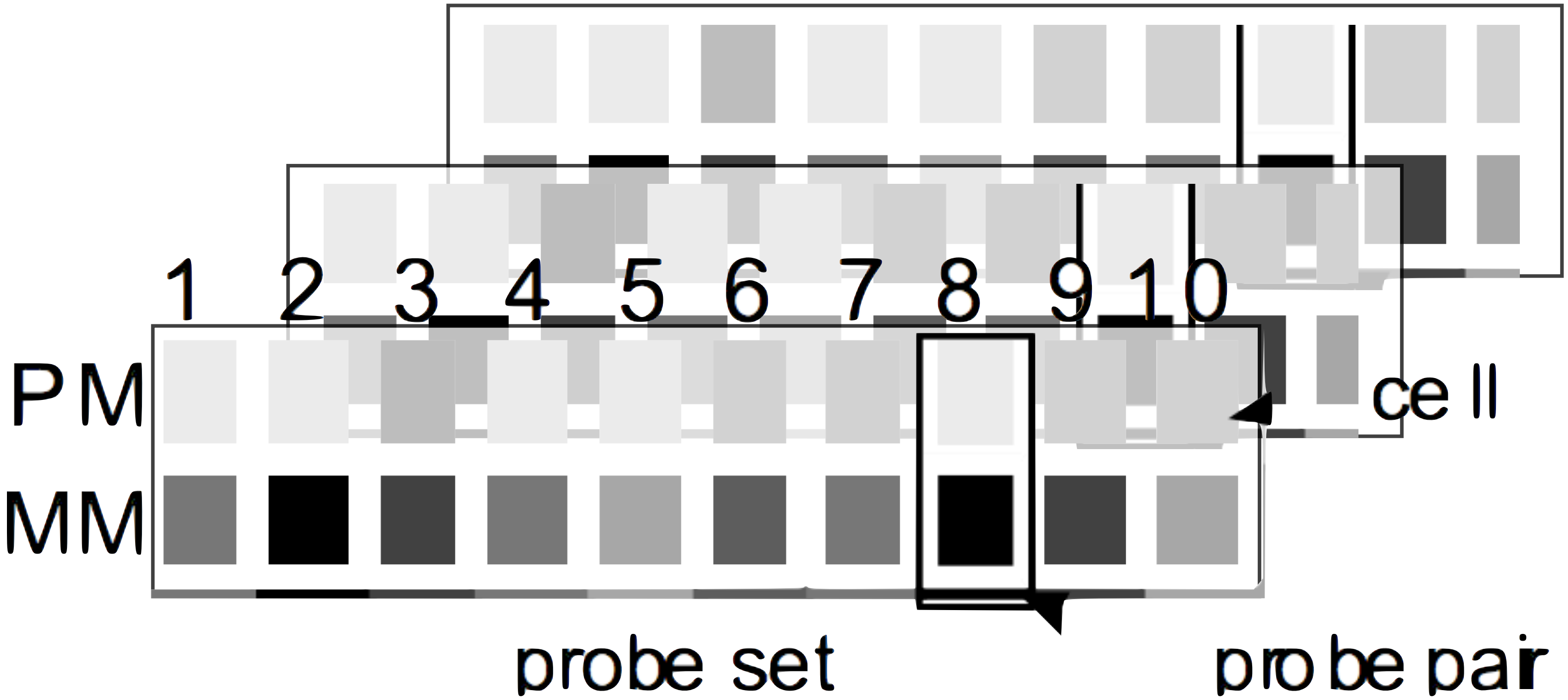
Batch variations are not removed by linear models when computed batch-by-batch

Solution:

Use global linear model with batch factors

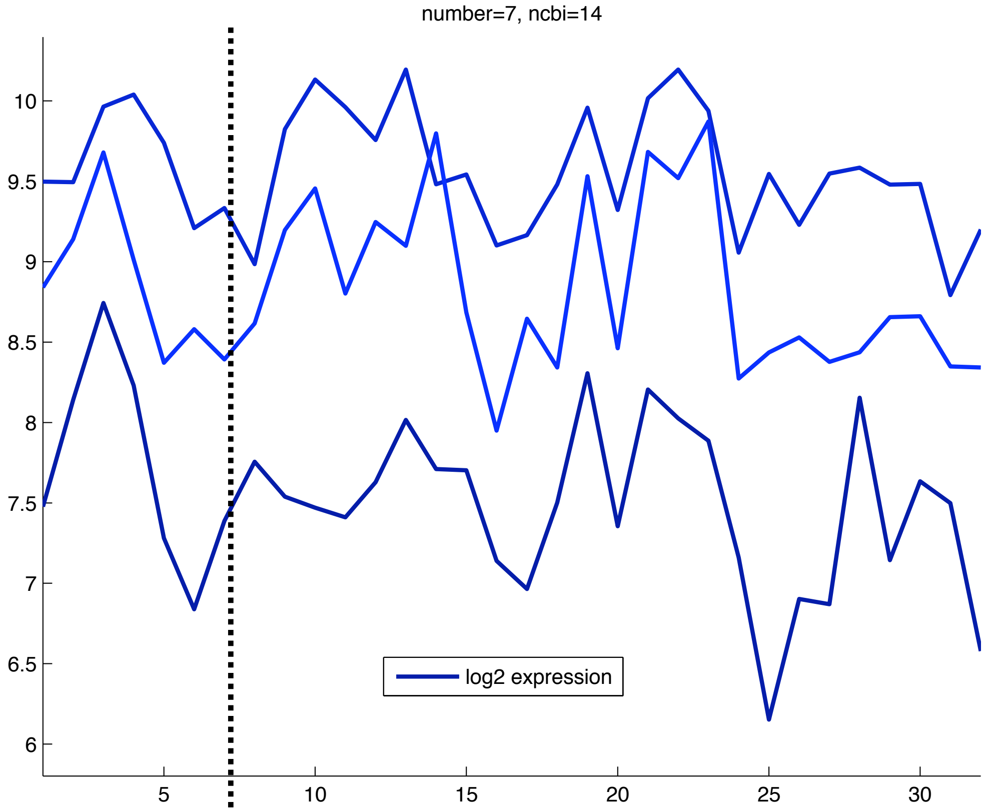
Batch Effects

- Conclusions
  - Small variations exist among the technician & date batches
  - Likely to be artifactual and not biologically meaningful
  - Normalization of batch MEAN AND VARIANCE is recommended



Redundant Probe Set Summarization

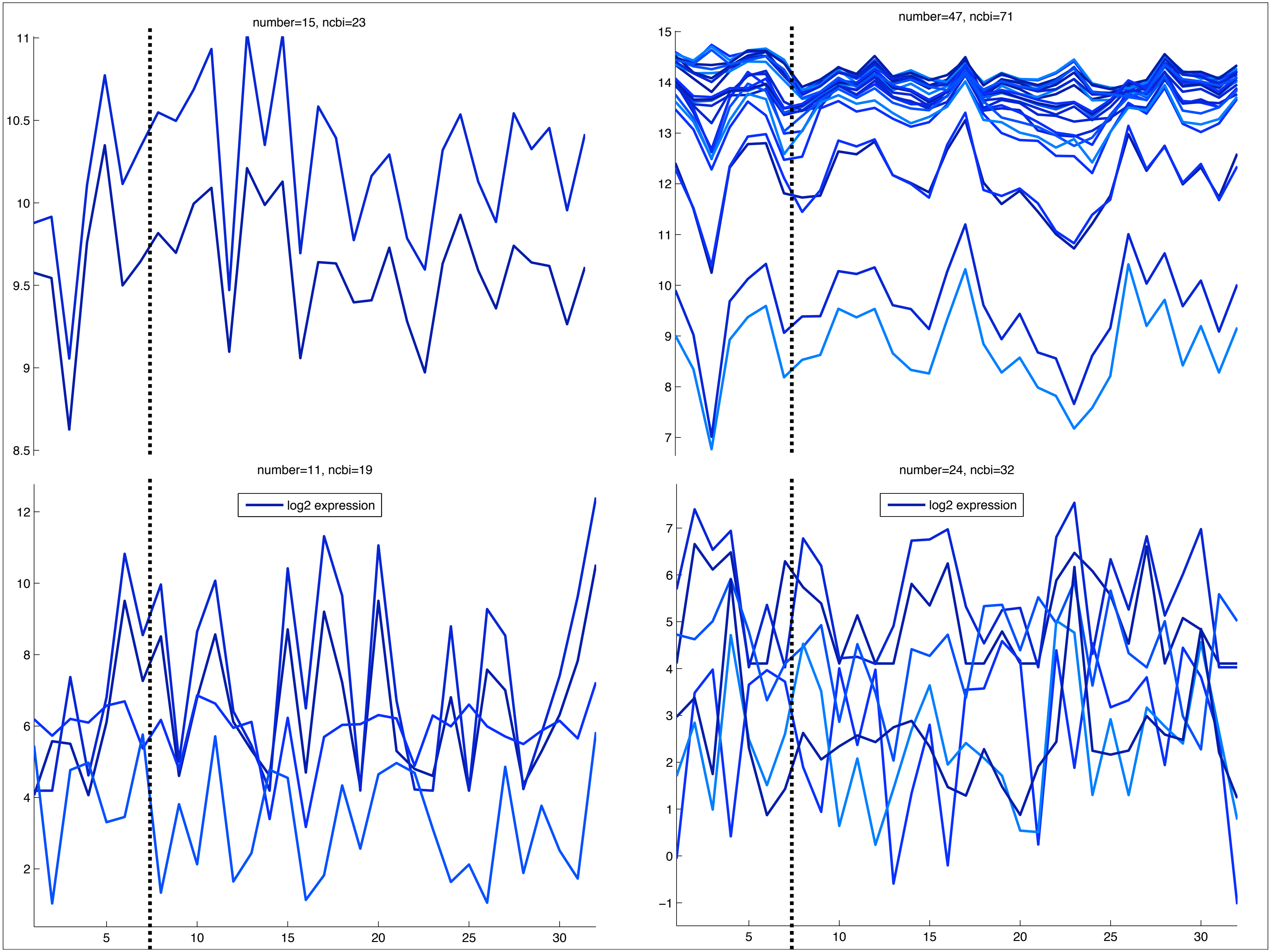
Redundant Profiles



NCBI 14
3 probe sets

Batch 1
(samples 1–7)

Batch 2
(samples 8–33)


number=15, ncbi=23
11
10.5
10
9.5
9
8.5
number=47, ncbi=71
15
14
13
12
11
10
9
8
7
number=11, ncbi=19
log2 expression
12
10
8
6
4
2
5
10
15
20
25
30
number=24, ncbi=32
log2 expression
7
6
5
4
3
2
1
0
−1
5
10
15
20
25
30

Redundant Profile Summarization

- Previous methods
  - Take average profile
  - Use most variant
  - Use biologically motivated outlier filtering

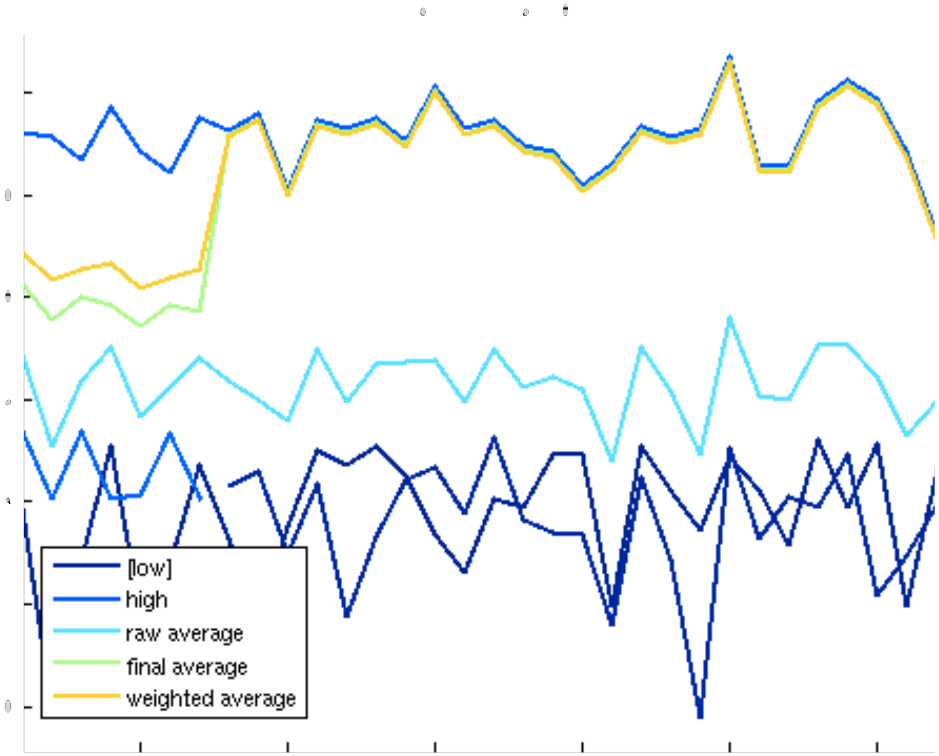
Filtering Example

- Profiles that are consistently low are indistinguishable from noise and should be removed

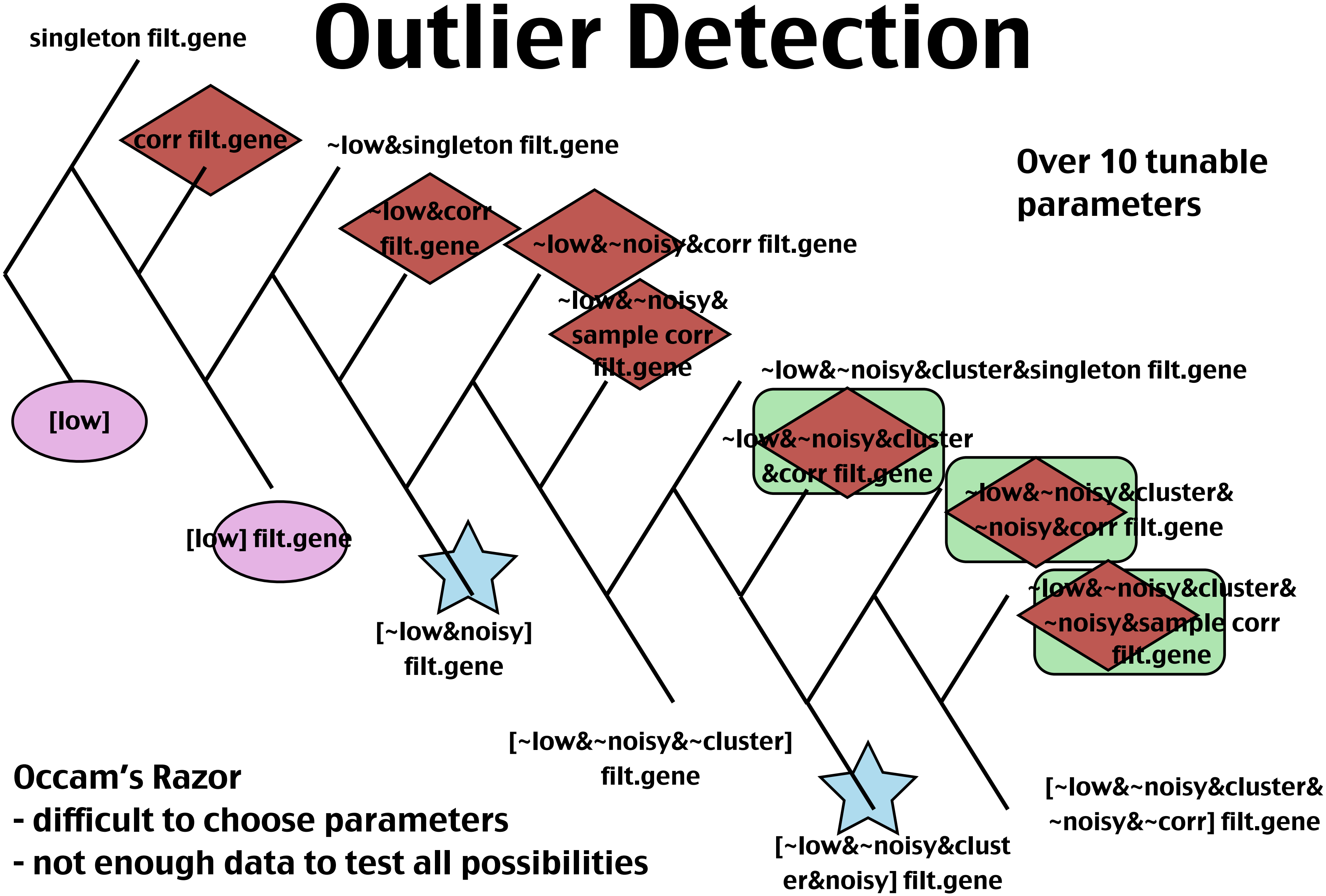
- Set inconsistenly low values to NaN
  - "Specs of Dust"

Filtering Example

- **Care must be taken to avoid the introduction of additional batch effects**



Outlier Detection


singleton filt.gene
corr filt.gene
~low&singleton filt.gene
~low&corr filt.gene
~low&~noisy&corr filt.gene
~low&~noisy& sample corr filt.gene
~low&~noisy&cluster&singleton filt.gene
~low&~noisy&cluster &corr filt.gene
~low&~noisy&cluster& ~noisy&corr filt.gene
~low&~noisy&cluster& ~noisy&sample corr filt.gene
[low]
[low] filt.gene
[~low&noisy] filt.gene
[~low&~noisy&~cluster] filt.gene
[~low&~noisy&clust er&noisy] filt.gene
[~low&~noisy&cluster& ~noisy&~corr] filt.gene

Over 10 tunable parameters

Occam's Razor

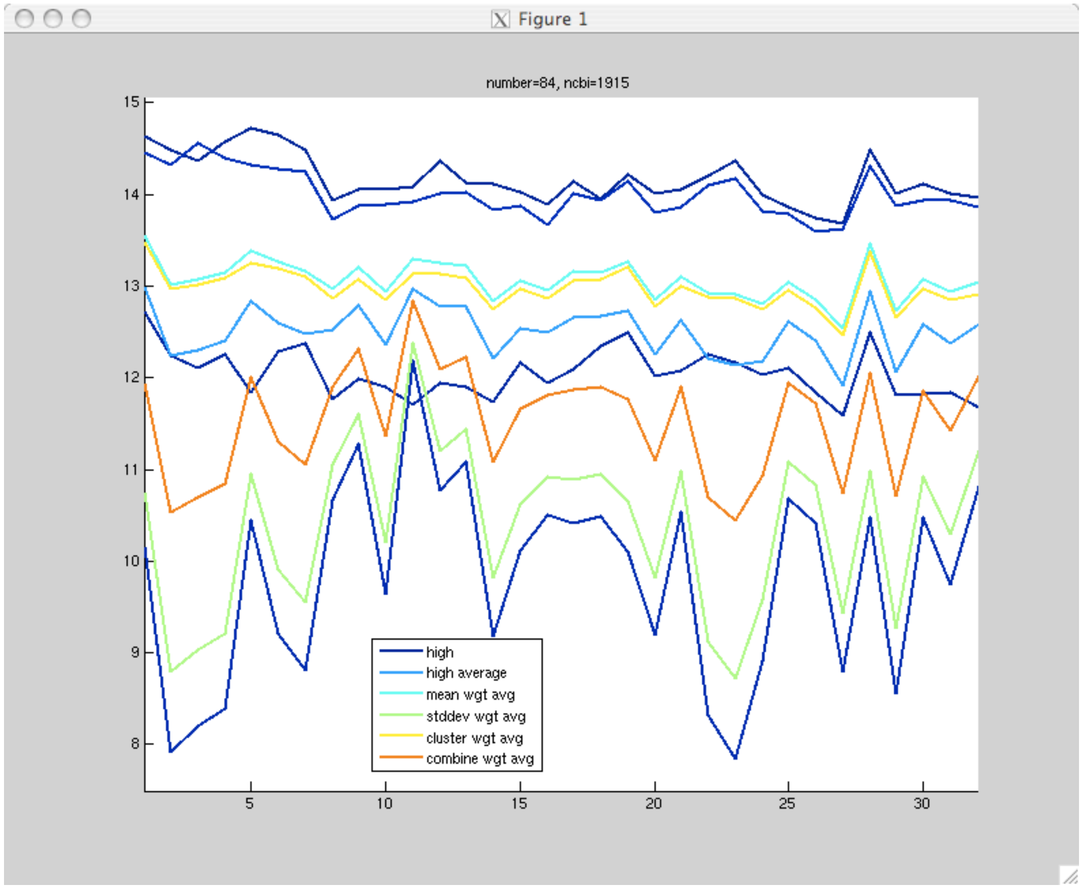
- difficult to choose parameters

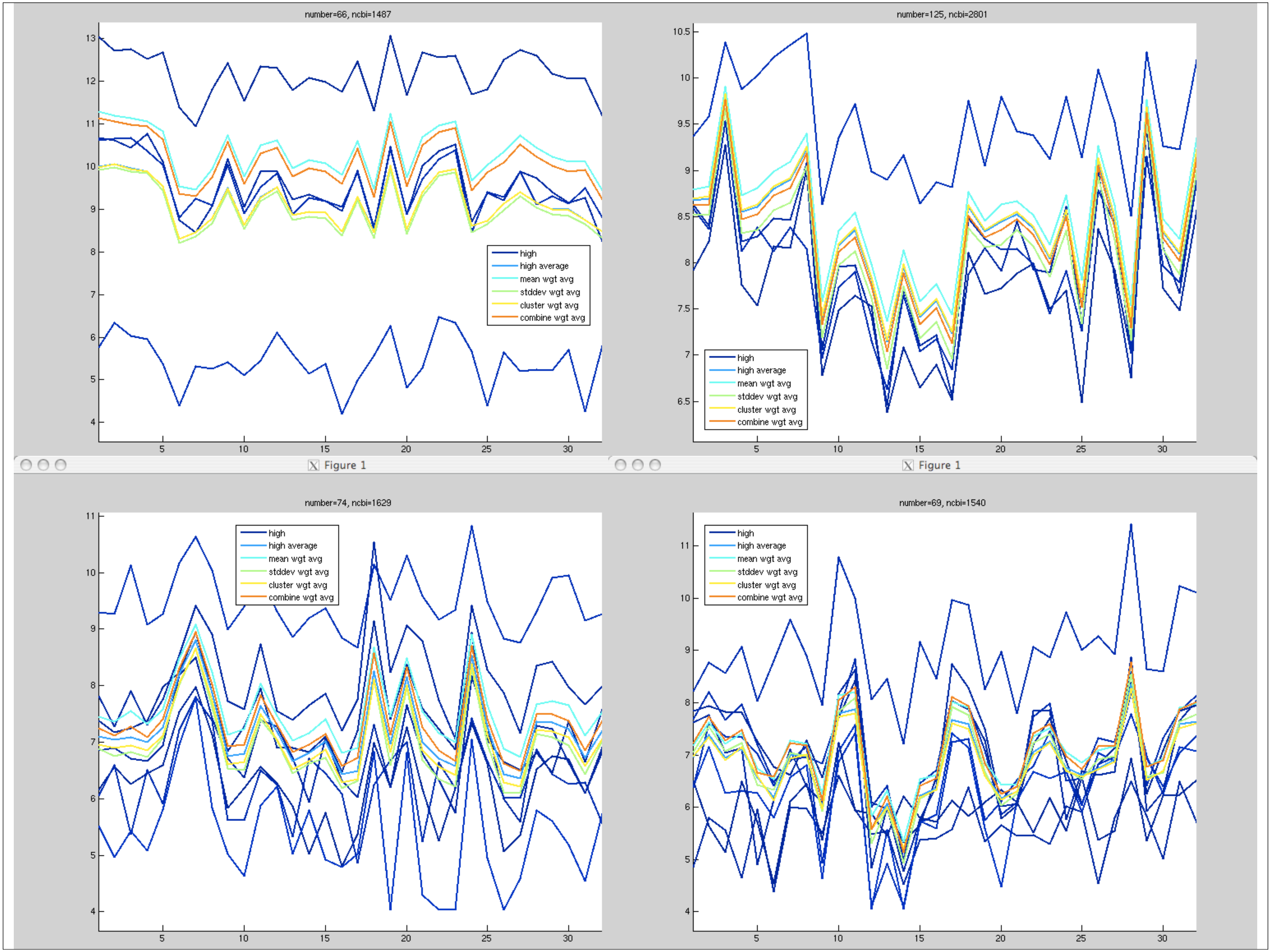
- not enough data to test all possibilities

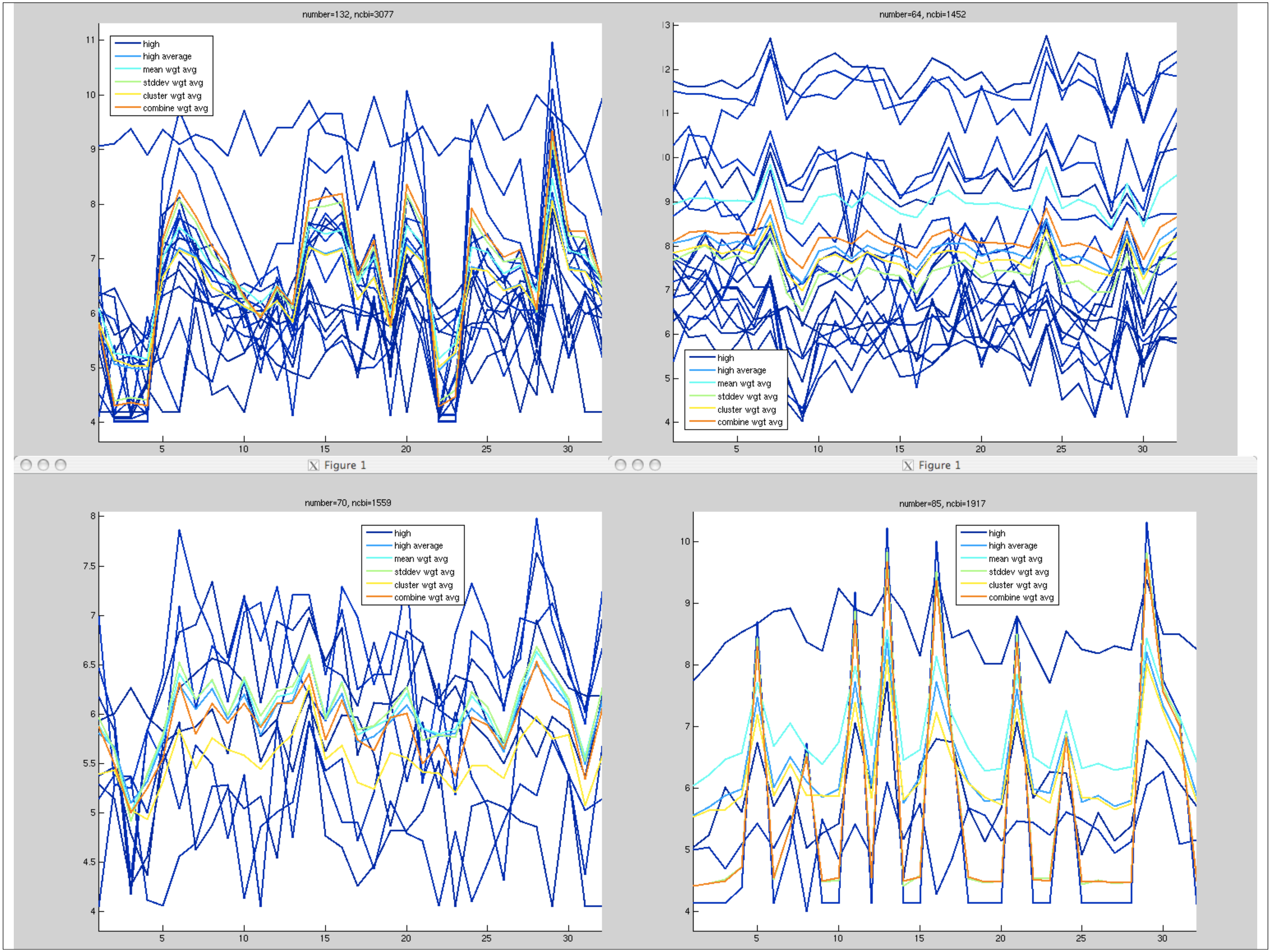
Alternative Solution

- Summarization by weighted median
  - Robust and biologically motivated
- Weights
  - Profile mean expression
  - Profile variance
  - Profile eigen-centrality
  - All weights combined

Weighted Median Summarization




number=66, ncbi=1487
13
12
11
10
9
8
7
6
5
4
5
10
15
20
25
30
high
high average
mean wgt avg
stddev wgt avg
cluster wgt avg
combine wgt avg
number=125, ncbi=2801
10.5
10
9.5
9
8.5
8
7.5
7
6.5
5
10
15
20
25
30
high
high average
mean wgt avg
stddev wgt avg
cluster wgt avg
combine wgt avg
Figure 1
Figure 1
number=74, ncbi=1629
11
10
9
8
7
6
5
4
5
10
15
20
25
30
high
high average
mean wgt avg
stddev wgt avg
cluster wgt avg
combine wgt avg
number=69, ncbi=1540
11
10
9
8
7
6
5
4
5
10
15
20
25
30
high
high average
mean wgt avg
stddev wgt avg
cluster wgt avg
combine wgt avg


number=132, ncbi=3077
high
high average
mean wgt avg
stddev wgt avg
cluster wgt avg
combine wgt avg
4
5
6
7
8
9
10
11
5
10
15
20
25
30
number=64, ncbi=1452
high
high average
mean wgt avg
stddev wgt avg
cluster wgt avg
combine wgt avg
4
5
6
7
8
9
10
11
12
13
5
10
15
20
25
30
Figure 1
Figure 1
number=70, ncbi=1559
high
high average
mean wgt avg
stddev wgt avg
cluster wgt avg
combine wgt avg
4
4.5
5
5.5
6
6.5
7
7.5
8
5
10
15
20
25
30
number=85, ncbi=1917
high
high average
mean wgt avg
stddev wgt avg
cluster wgt avg
combine wgt avg
4
5
6
7
8
9
10
5
10
15
20
25
30

From A to Z

1. Remove invalid probes
2. Normalize & summarize probes
3. Normalize batch mean & variance
4. Construct probe set profiles
5. Remove consistently noisy profiles
6. Compute weighted median of redundant profiles

   ... Validate and improve

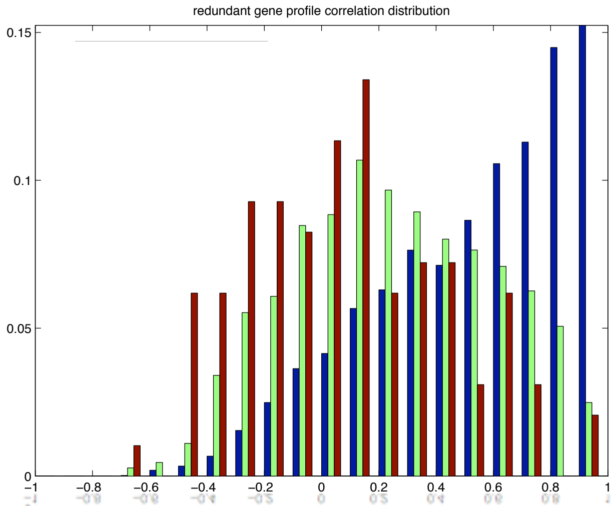
Summary

- Presented a framework for processing raw microarray data
- Batch effects identified and documented
- Introduced weighted median probe set summarization

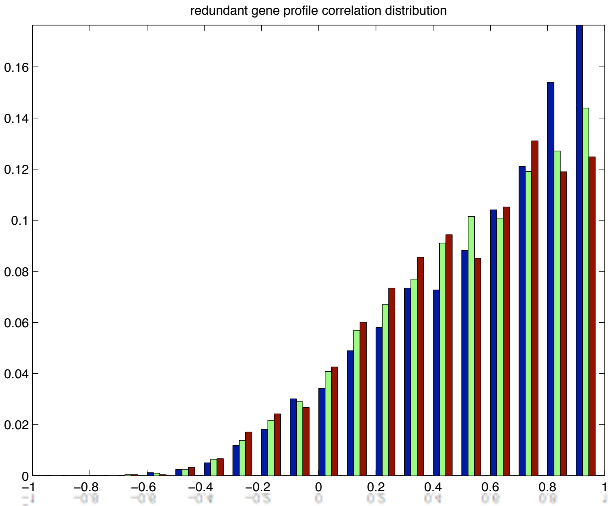
The End

For each NCBI gene, compute correlation matrix corr(i,j) for redundant profiles

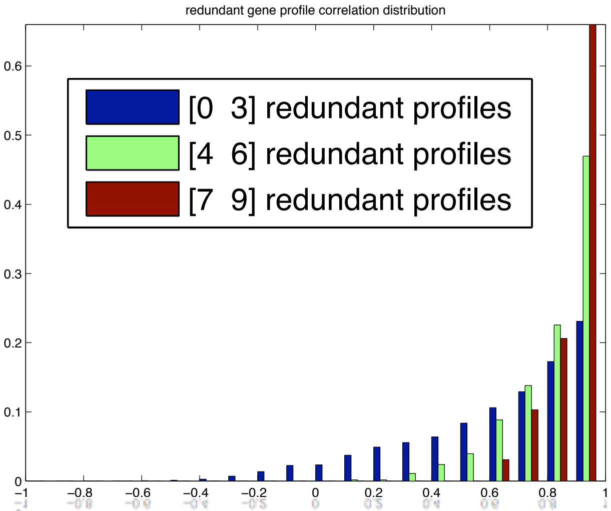
min corr(i,j)



all corr(i,j)



max corr(i,j)

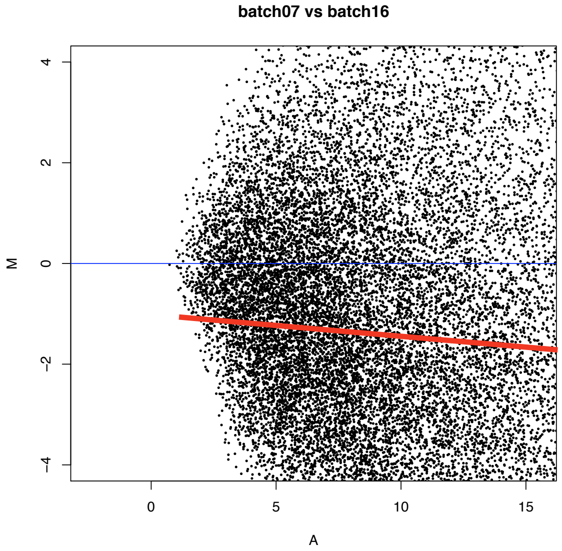


Redundant Profiles Correlations Distributions

Caveat 2

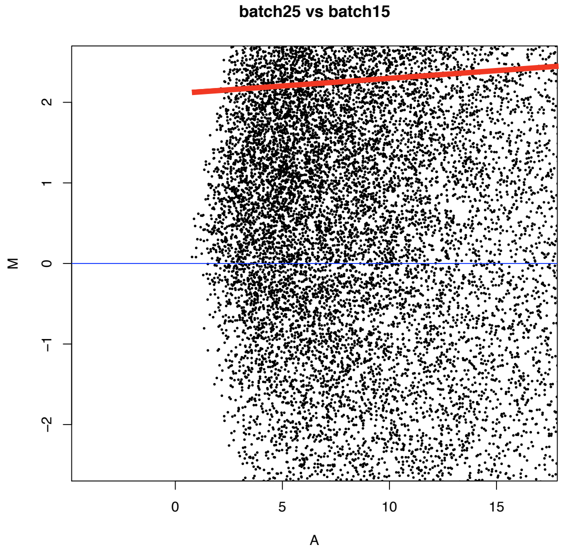
Points correspond to probesets, M = difference, A = average (of batch median expr)

MAS5 by tissue



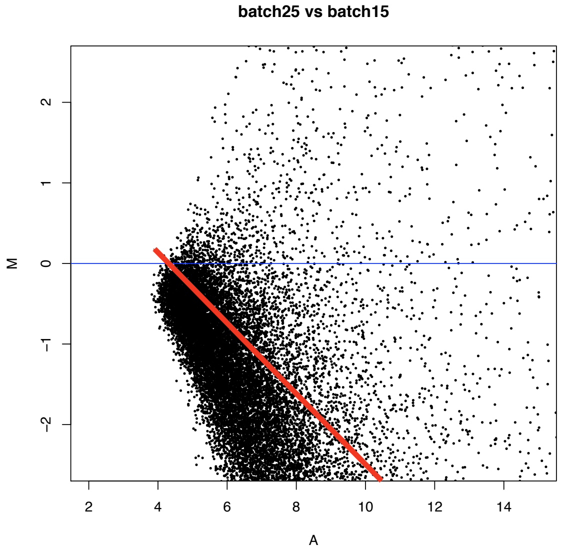
small variation

MAS5



moderate variation

RMA



large variation