Experiments with microarrays: generalities and some local experiences

Terry Speed
Genetics & Bioinformatics

WEHI Postgraduate Teaching Course
June 23, 2003
The basic cDNA and GeneChip technologies

cDNA-microarrays

probe preparation
- cDNA collection
- insert amplification by PCR
  - vector specific primers
  - gene specific primers
- printing
- coupling
- denaturing

ratio Cy5/Cy3

hybridization

target preparation
- Cy3 or Cy5 labeled cDNA
- modified oligoDT
- total RNA
- cells/tissue

high-density oligonucleotide arrays

mRNA reference sequence

perfect match mismatch

in situ synthesis by photolithography

ratio array 1/array 2

fragmented

biotin labeled cRNA (generated using T7 promoter)

double-stranded cDNA (generated using T7 primer)

total RNA

cells/tissue
Recent technology developments

Long (60-75 bp) oligonucleotide instead of cDNA probes (immobilized substrate) on glass slides, hybridized to a cDNA target (mobile substrate):
- data quality better than standard cDNA slide
- still some other issues

Short (25 bp) or longer (40+ bp) oligonucleotide probes synthesized in situ using a maskless photolithography technology (digital light processor)
- faster to make and cheaper than Affymetrix chips
- still some other issues
With microarray data, always log$_2$
Instead of plotting one colour (chip) against the other: rotate

\[ \log_2 R \text{ vs } \log_2 G \]

\[ M = \log_2 \frac{R}{G} \text{ vs } A = \log_2 \sqrt{RG} \]
MA-plots

Extensively in later slides, we use of scatter plots of

\[ M = \log \text{ratio} = \log \frac{f}{g} = \log f - \log g \]

against the corresponding

\[ A = \log \text{geom mean} = \log \sqrt[2]{fg} = \frac{[\log f + \log g]}{2}. \]

Usually, \( f = \text{Cy 5} \) intensity for a spot on a cDNA slide, and \( g = \text{Cy3} \) for the same spot on the same slide. At other times, \( f \) is an expression measure of a gene from one chip, and \( g \) the same thing from another chip.
Overview of today’s overview

- Purpose
- Planning
- Pre-processing
- Some local studies
- Validation and follow-up
Purpose
Some uses of microarrays

To detect gene expression (i.e. mRNA)
To compare gene expression levels
To map genes contributing to phenotypes
To map deleted or duplicated regions
To identify genes modifying other genes

and more… …
Planning
Issues that affect the planning of microarray experiments

Scientific
Aim of the experiment
Specific questions and priorities between them.
How will the experiments answer the questions posed?

Practical (Logistic)
Types of mRNA samples:
reference, control, treatment 1, cell line tissue sample, ...
Amount of mRNA available.
Number of slides/chips available for experiment.

Other Information
The experimental process prior to hybridization:
sample isolation, mRNA extraction, amplification or pooling, labelling.

Controls planned:
positive, negative, ratio, etc.

Verification method:
Northern, RT-PCR, QRT-PCR, in situ hybridization, etc.
Decisions to be made (mainly cDNA arrays)

Layout of the array
Which cDNA sequences to print, and where to print them?
- Library
- Controls
- Spatial arrangement

Allocation of samples to the slides
Hybridizations planned
- Treatment vs control, or vs common reference or..

Replication
- Number of replicates of each hybridization
- Use of dye swaps in replication
- Different types of replicates (technical, biological)

Other considerations
- Physical limitations: the number of slides and the amount of material
- Extensibility - linking

Carrying out the experiment
- Reagents (e.g. slide/chip batches), equipment e.g. (scanners), people to use
- Order of hybridizations
Recall the general statistical principles of design laid down by R A Fisher in the context of agricultural experiments, but which have been found to be of value elsewhere in science and technology (industry, engineering, social science, other biological and medical science, etc):

RANDOMIZATION
REPLICATION
LOCAL CONTROL
Graphical representation of cDNA hybridizations

(a) A → 5 → B

(b) A ← C ← B
Natural design choice (cDNA)

Case 1: Meaningful biological control (C)

**Samples:** Liver tissue from four mice treated by cholesterol modifying drugs.

**Question 1:** Genes that respond differently between the T and the C.

**Question 2:** Genes that responded similarly across two or more treatments relative to control.

Case 2: Use of a reference sample (more later)

**Samples:** Different tumor samples.

**Question:** To discover tumor subtypes.
Where there is choice: direct or indirect?

Two samples: KO vs. WT, or mutant vs. WT, write T vs C.

Direct

\[
\text{T} \quad \text{C}
\]

average (log (T/C))

Indirect

\[
\text{T} \quad \text{Ref} \quad \text{C}
\]

log (T / Ref) – log (C / Ref)
Time series (time-course) experiments

Possible designs:
1) All sample vs common pooled reference
2) All sample vs time 0
3) Direct hybridization between times.

Diagram:
- T1 → T2 → T3 → T4 → T5 → T6 → T7
- Ref
- Relations:
  - T1 vs T7
  - T2 vs T6
  - T3 vs T5
  - T4 vs T6
  - T5 vs T7
  - T1 vs Ref
  - T2 vs Ref
  - T3 vs Ref
  - T4 vs Ref
  - T5 vs Ref
  - T6 vs Ref
  - T7 vs Ref

Legend:
- Dotted yellow arrows: Pooled reference
- Green dashed arrows: Compare to T1
- Solid blue arrows: t vs t+1
- Solid red arrows: t vs t+2
- Solid purple arrows: t vs t+3
## Some cDNA design choices with time series experiments

<table>
<thead>
<tr>
<th>N=3</th>
<th></th>
<th>t vs t+1</th>
<th>t vs t+2</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>T1T2</td>
<td>T2T3</td>
<td>T3T4</td>
</tr>
<tr>
<td>A)</td>
<td>T1 as common reference</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
</tr>
<tr>
<td>B)</td>
<td>Direct Hybridization</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N=4</th>
<th></th>
<th>t vs t+1</th>
<th>t vs t+2</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>T1T2</td>
<td>T2T3</td>
<td>T3T4</td>
</tr>
<tr>
<td>C)</td>
<td>Common reference</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
</tr>
<tr>
<td>D)</td>
<td>T1 as common ref + more</td>
<td>.67</td>
<td>.67</td>
<td>1.67</td>
</tr>
<tr>
<td></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
</tr>
<tr>
<td>E)</td>
<td>Direct hybridization choice 1</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
</tr>
<tr>
<td></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
</tr>
<tr>
<td>F)</td>
<td>Direct hybridization choice 2</td>
<td>1</td>
<td>.75</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
<td><img src="T1arrowT2T3T4" alt="Diagram" /></td>
</tr>
</tbody>
</table>
Use of common reference cDNA

With more than 4 times, it makes good sense to use a common reference design, unless other considerations over-rule this (e.g. wt vs mutant time course). In other situations, e.g. with tumor samples, use of a common reference is often the design of choice (again, other issues may suggest the contrary, e.g. *matched* tumor vs normal pairs). We get *extensibility*, and the ability to make *pairwise comparisons*.

What should we use as the common reference? If possible, mRNA of scientific interest, e.g. a control or wt, or time zero. Often a *pool* is used: it could be a *general* pool, e.g. Strategene, or it could be a *specific* one, e.g. of mRNA from all the times in a time course. With mice it might be *mouse-specific* or a pool of *all times from all mice*. Many factors are relevant.
More complex designs: 
2 factors, each with 2 levels

Example 1: Suppose we wish to study the joint effect of two drugs, A and B. 
4 possible treatment combinations:
   C: No treatment
   A: drug A only.
   B: drug B only.
   A.B: both drug A and B.

Example 2: Our interest is in comparing two strain of mice (mutant and wild-type) at two different times, postnatal and adult. 
4 possible samples (labelling as above):
   C: WT at postnatal
   A: WT at adult (effect of time only)
   B: MT at postnatal (effect of the mutation only)
   A.B: MT at adult (effect of both time and the mutation).
A biologist’s approach

\[ y_1 = \log \left( \frac{A}{C} \right) = a \]
\[ y_2 = \log \left( \frac{B}{C} \right) = b \]
\[ y_3 = \log \left( \frac{AB}{C} \right) = a + b + ab \]

Estimate \( ab \) by \( y_3 - y_2 - y_1 \) (or averages of such terms).
A statistician’s approach

Different ways of estimating parameters.

e.g. B effect.

\[
1 = (\mu + b) - (\mu) = b \\
2 - 5 = ((\mu + a) - (\mu)) -((\mu + a)-(\mu + b)) = (a) - (a + b) = b
\]

cDNA: many design options, some better than others.

Affymetrix: just hybridize all four samples to the chip.
Summary

In general there will be features specific to your scientific and practical context, and a mix of indirect and direct may be the best choice for you.

We (Bioinformatics) are always happy to discuss design options with would-be microarrayers.
Replication

Why replicate slides?

- Averages of replicates give better estimate of the log-ratios
- Essential to estimate the variance of log-ratios
  (Essential for valid statistical analysis)

Different types of replicates

- Technical replicates
  - Within slide and between slides
- Biological replicates
Technical replication - labelling

3 sets of self–self hybridizations: (cerebellum vs cerebellum)
Data 1 and Data 2 were labeled together and hybridized on two slides separately.
Data 3 were labeled separately.
Biological replicates: averaging helps

Apo A1 Data Set
Randomization

Not widely considered for this sort of experiment. Sometimes hard to implement, but sometimes very important. I mention two examples.

First, an experiment to determine the efficacy of pooling liver mRNA across mice with Affymetrix chips. It didn’t give us a clear answer because of inattention to design issues. In part this was my fault. Technical replicate variation turned out to be larger than biological variation.
What went on?

How could biological variability be less than technical variability? Sadly, here’s how.

<table>
<thead>
<tr>
<th>Chip</th>
<th>Hybridization date</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1, M2, M3</td>
<td>3/28/02</td>
</tr>
<tr>
<td>All <em>same</em> day, also all <em>same</em> chip lot</td>
<td></td>
</tr>
<tr>
<td>M123(pool, rep 1)</td>
<td>3/28/02</td>
</tr>
<tr>
<td>M123(pool, rep 2)</td>
<td>4/4/02</td>
</tr>
<tr>
<td>M123(pool, rep 3)</td>
<td>5/22/02</td>
</tr>
<tr>
<td>All <em>different</em> days and all <em>different</em> chip lots</td>
<td></td>
</tr>
</tbody>
</table>
Another example?

In a recent article in *The Lancet*, Petricoin *et al* (2002) demonstrated that patterns in SELDI-TOF mass spec data can distinguish between serum samples from healthy women and those from women with ovarian cancer, even when the cancers are at early stages. This set off a flurry of activity.

More recently, Baggerly *et al* (2003, submitted) obtained *perfect separation* of the same classes of samples using the same data, but using only features selected from the *noise* region of the spectra.

It seems as though the original investigators processed all the cancer patients as a group, and all the healthy patients as a group. This allows artifacts to determine group differences.
Use of dye-swap replicates

These two hybridizations:

\[ \text{T} \quad \text{C} \quad \text{T} \quad \text{C} \]

are usually to be preferred to these:

\[ \text{T} \quad \text{C} \quad \text{T} \quad \text{C} \]

Why?
Pre-processing
Normalization

Why?
To correct for systematic differences between samples on the same slide, or between slides/chips, which do not represent true biological variation between mRNA samples.

How do we know it is necessary?
By examining replicate and/or self-self hybridizations, where no true differential expression is occurring.

We find biases which vary with overall spot intensity, location on the array, dye, plate origin, pins, scanner, scanning parameters,....
cDNA: self-self hybridizations

False color overlay
Boxplots within pin-groups
MA-plots + lowess smooth
A series of non self-self hybridizations

From the NCI60 data set

Early Goodman lab, UC Berkeley

Early Ngai lab, UC Berkeley

Early PMCRI, Melbourne Australia
What is to be done?

The challenge is to adjust or correct the raw data in the hope of removing these systematic effects, bearing in mind that

• the distinction between systematic and other effects is often far from clear, and
• it will usually be quite hard to demonstrate that we are improving and not worsening the “true” signal, as
• determining just how much adjustment is desirable is quite hard.

We do what we can, seeking validation wherever possible.
Normalization: methods

a) Normalization based on a global adjustment

\[ \log_2 R/G \rightarrow \log_2 R/G - c = \log_2 \frac{R}{(kG)} \]

Choices for \( k \) or \( c = \log_2 k \) are \( c = \) median or mean of log ratios for a particular gene set (e.g. housekeeping genes).

Or, total intensity normalization, where \( k = \sum R_i / \sum G_i \).
b) Intensity-dependent normalization.

Here we run a line through the middle of the MA-plot, shifting the $M$ value of the pair $(A,M)$ by $c=c(A)$, i.e.

$$\log_2 \frac{R}{G} \rightarrow \log_2 \frac{R}{G} - c(A) = \log_2 \frac{R}{(k(A)G)}.$$ 

One estimate of $c(A)$ is made using the lowess function of Cleveland (1979): LOcally WEighted Scatterplot Smoothing.

c) Spatial normalization. Frequently necessary. Details omitted
Which spots to use for normalization?

The lowess lines can be run through many different sets of points, and each strategy has its own implicit set of assumptions justifying its applicability.

For example, we can justify the use of a global lowess approach by supposing that, when stratified by mRNA abundance, a) only a minority of genes are expected to be differentially expressed, or b) any differential expression is as likely to be up-regulation as down-regulation.

The use of other sets of genes, e.g. control or housekeeping genes, involve similar assumptions. Ultimately we must judge on the accuracy and precision of our results.
Global scale, global lowess, pin-group lowess; spatial plot after, smooth histograms of M after
MA-plots of Affymetrix chip pairs: before normalization

Log ratios v log geom means for chips 1,2; 1,3,..., 5,6.
MA-plots of chip pairs: after quantile normalization
Summary of normalization

Reduces systematic (not random) effects
Makes it possible to compare several arrays

Use logratios (MA-plots)
Lowess normalization (dye bias)
MSP titration series
Pin-group location normalization
Pin-group scale normalization
Between slide scale normalization

More? Use controls (e.g. Lucidea scorecard)
Normalization introduces more variability
Outliers (bad spots) are handled with replication
Some local studies
Use of microarrays to map genes

Here we summarize an attempt, so far unsuccessful, to use microarrays to map genes contributing to a complex trait (generally a hard task). This approach has worked in the past.

The trait: resistance/susceptibility to *Leishmania major*, a murine analogue of human *Leishmania*.

Previous work: QTL mapping identified two loci, *lmr1/2* (*lmr* = leishmania resistance). Several generations of selection and crossing yielded *reciprocal compound congenic lines* denoted by *B6.lmr1/2* and *B/c.lmr1/2*.
Our players: BALB/c and C57BL/6

One susceptible, the other resistant to *Leishmania major*
Compound congenics

B/c.lmr1/2:
BALB/c background
Imr1 and Imr2 from C57BL/6
Predict: more resistant than BALB/c

B6.lmr1/2:
C57BL/6 background
Imr1 and Imr2 from BALB/c
Predict: more susceptible than C57BL/6
Microarray data to be analysed

Analysis strategy: find genes on each side of the line with high interaction (note $2 \times 2$ factorial structure, actually $2 \times 2 \times 2$). Select common genes from the congenic interval.
How does cytokine signalling intensity affect biological outcome? A pilot experiment

Expression analysis of $\gamma^{-/-}$ vs SOCS-1$^{-/-} \gamma^{-/-}$ livers, following IFN$\gamma$ injection.
The pilot experiment, ctd

Does IFN\(\gamma\) differentially affect gene expression in IFN\(\gamma\)^{-/-} and SOCS1^{-/-} IFN\(\gamma\)^{-/-} livers? Another 2\times2 factorial.
Results from the pilot experiment

Detection of SOCS1 as expected (main effect)

~300 genes induced in response to IFNγ (main effect)

Magnitude of gene expression generally greater in the absence of SOCS1 (interaction!)

Results encouraging enough to go on to the full experiment.
Time-course study of IFN\(_{\gamma}\) signal transduction

Two different genotypes (\(\gamma^{-/-}\) and S1\(^{-/-}\) \(\gamma^{-/-}\))

9-point IFN\(_{\gamma}\) injection time-course:

Three Controls

One non-injected.

PBS injections (4h and 12h).

Total = 24 mice.
Processing the time-course experiment

24 arrays

Scan arrays

24 arrays $\times$ 12,000 genes $\times$ 16 probes/gene = $4.6 \times 10^6$ data points

Normalisation
Background adjustment
Weighted average of probe set

24 arrays $\times$ 12,000 expression values

= 12,000 gene expression profiles, 24 points each
Cluster Analysis

• 218 differentially expressed genes were clustered according to their expression profiles using a program called GeneCluster.

• GeneCluster is used by first choosing the number of different clusters of genes that you think you might find in your experiment.

• GeneCluster then finds and displays the most common clusters of expression profiles. Each gene is forced to belong to a cluster.
Output from GeneCluster (49 clusters)

Trend 1: 145 genes

Trend 2: 43 genes
Validation
A common story: select genes, follow up

Issues: how should we select genes? Using fold-change, a t-statistic, or something else? We can give guidance here, especially on the last option.

Some slides from an on-going study of Hamish Scott, courtesy of Joelle Michaud. I won’t even attempt to describe the science here. Let’s just say that they did many hybridizations, seeking differentially expressed genes.
Ranking Genes

Genes ranked using B-statistic which estimates the odds that each gene is differentially expressed.

B-statistic gives useful compromise between average log-fold change and Student’s t-statistic for each gene.
TF-Hela

MA-plots summarizing replicate hybridizations.
RT-PCR Confirmations of PTGs

<table>
<thead>
<tr>
<th>Gene</th>
<th>M</th>
<th>EGFP</th>
<th>RUNX1</th>
<th>RUNX1CBFb</th>
<th>EGFP</th>
<th>RUNX1</th>
<th>RUNX1CBFb</th>
<th>EGFP</th>
<th>RUNX1</th>
<th>RUNX1CBFb</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGFBP3</td>
<td>M-1.1</td>
<td>[1:200]</td>
<td>[1:500]</td>
<td>[1:750]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD55/DAF</td>
<td>M-0.7</td>
<td>[1:50]</td>
<td>[1:250]</td>
<td>[1:500]</td>
<td>[1:750]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARHB</td>
<td>M-1.3</td>
<td>[1:200]</td>
<td>[1:500]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOP</td>
<td>M0.7</td>
<td>[1:50]</td>
<td>[1:250]</td>
<td>[1:100]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
E. Winzeler’s custom *P. falciparum* chip

Probe design based on genomic sequence, not gene models (e.g. UniGene clusters)

Chip contains *P. falciparum* genes as well as control genes from other species (human, mouse, bacterial, etc.)

Number of probes per gene depends on gene length (can have a few, or hundreds!)

Three genotypes: wildtype (3D7), knockout strain, variant of WT modified to behave similarly to KO

Each hybridized to three different chips
**P. falciparum** (knockout vs. wildtype)

**normal q-q plot, KO vs. wildtype**

**M v. A, KO vs. wildtype**
Comparison between chip and QRT-PCR

<table>
<thead>
<tr>
<th>Gene</th>
<th>RMA fold change</th>
<th>QRT-PCR fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.8</td>
<td>86.6</td>
</tr>
<tr>
<td>2</td>
<td>33.2</td>
<td>76.4</td>
</tr>
<tr>
<td>3</td>
<td>4.2</td>
<td>~$10^5$ (knocked out)</td>
</tr>
<tr>
<td>4</td>
<td>3.2</td>
<td>1.8</td>
</tr>
<tr>
<td>5</td>
<td>2.7</td>
<td>2.1</td>
</tr>
<tr>
<td>6</td>
<td>1.4</td>
<td>No change</td>
</tr>
<tr>
<td>7</td>
<td>2.0</td>
<td>1.6</td>
</tr>
<tr>
<td>8</td>
<td>2.8</td>
<td>No change</td>
</tr>
<tr>
<td>9</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>10</td>
<td>1.4</td>
<td>2.9</td>
</tr>
<tr>
<td>11</td>
<td>2.6</td>
<td>2.5</td>
</tr>
</tbody>
</table>
You might check expression using ISH

Expression in the mouse olfactory bulb (Berkeley data)
3-dimension reconstruction from *in-situ* data
Follow-up

A BIG topic: more bioinformatics, less statistics
Which clone (gene, probe) is that?

Sometimes not such a simple question.
Might need to re-sequence the clone, perhaps go back to an old UniGene build, and then forward to the latest genome annotation.
You might want to BLAST an Affymetrix probe sequence against the latest version of the genome, and pass results over to other tools or databases, including GO and KEGG.
A growing number of tools to help here.
Gene Ontology (GO) consortium: goals

1. To compile a comprehensive structured vocabulary of terms describing different elements of molecular biology that are shared among life forms.

2. To describe biological objects (in the model organism database of each contributing member) using these terms.

3. To provide tools for querying and manipulating these vocabularies.

4. To provide tools enabling curators to assign GO terms to biological objects.
Top Level
Molecular Function
Biological Process
Cellular Component

Annotation: #genes
Mouse   12,325
Human  21,521
Yeast  6,910
Drosophila  7,536
Or, use http://www.genome.ad.jp/kegg/kegg2.html in one way
Perhaps linking to http://www.genmapp.org/
Or another type, from KEGG
Other things you could do

Relate gene expression to chromosomal position of genes

Seek shared DNA sequence motifs in the upstream regions of co-regulated genes

Comply with MIAME

…and much else. But it’s past time to stop.
Acknowledgments

Statistics/bioinformatics
WEHI
Gordon Smyth
Matt Ritchie
Asa Wirapati
Ken Simpson
Natalie Thorne
Tim Beissbarth
UCSF
Jean Yee-Hwa Yang
U of Uppsala
Ingrid Lönnstedt

For data/studies
WEHI
Sam Wormald
Janine Stubbs
Hamish Scott
Joelle Michaud
LLNL
Colleen Elso
LBNL
Matt Callow
UCB
John Ngai et al
U of Oklahoma
Eun-Soo Han
MSP normalization compared to other methods

Orange: Schadt-Wong rank invariant set
Yellow: GAPDH, tubulin
Light blue: MSP pool / titration
Red line: lowess smooth