Estimation Issues for High Throughput Data

Mark Reimers, PhD

Biostatistics, Virginia Commonwealth University

Despair about High-Throughput Data Analysis

- Many statisticians feel despair at high-throughput data – e.g. microarrays, RNA-Seq, GWAS, ...

… large errors

… much non-random error

… non-reproducible results

…many ad-hoc practices

What if Microarrays Worked?

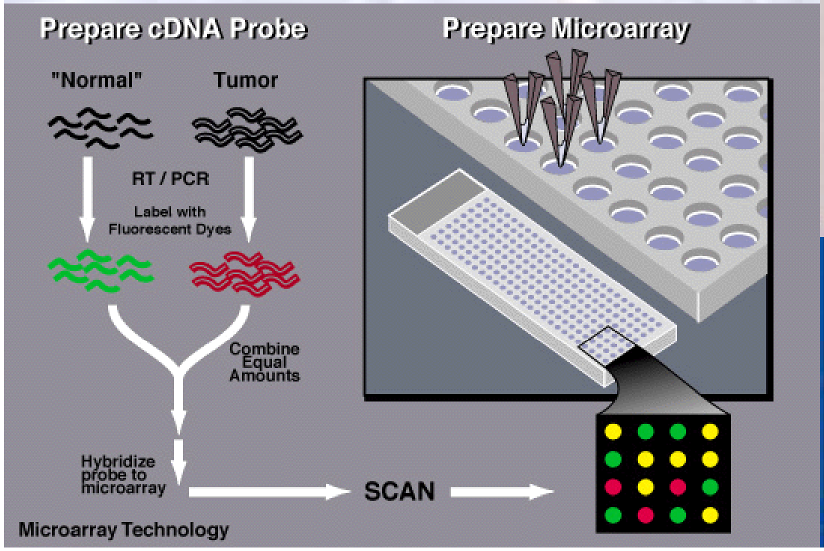
- Cheap assays with low overhead and rapid turn-around time suitable for population studies
- Could characterize response to a variety of pharmaceutical agents
- The Connectivity Map was funded ($1M) to do 100,000 arrays profiling effects of various small molecules on gene expression patterns
- The Genotype-Tissue Expression project, to find eQTLs in 50 tissues, will use arrays for genotyping, but not for expression profiling

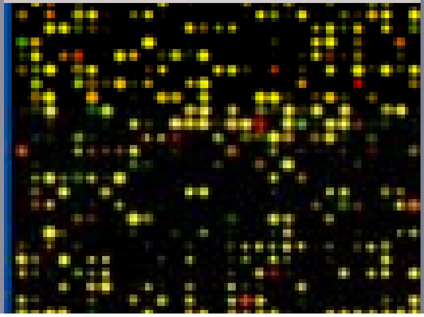
Outline

- The avalanche of high-throughput data
  - Genomic and proteomic technologies
  - Common characteristics of high-throughput data
  - Issues that torment us
- New strategies for estimation – how to effectively borrow information across measures
- An application to drug-transporter interaction

The Technologies: Genomics

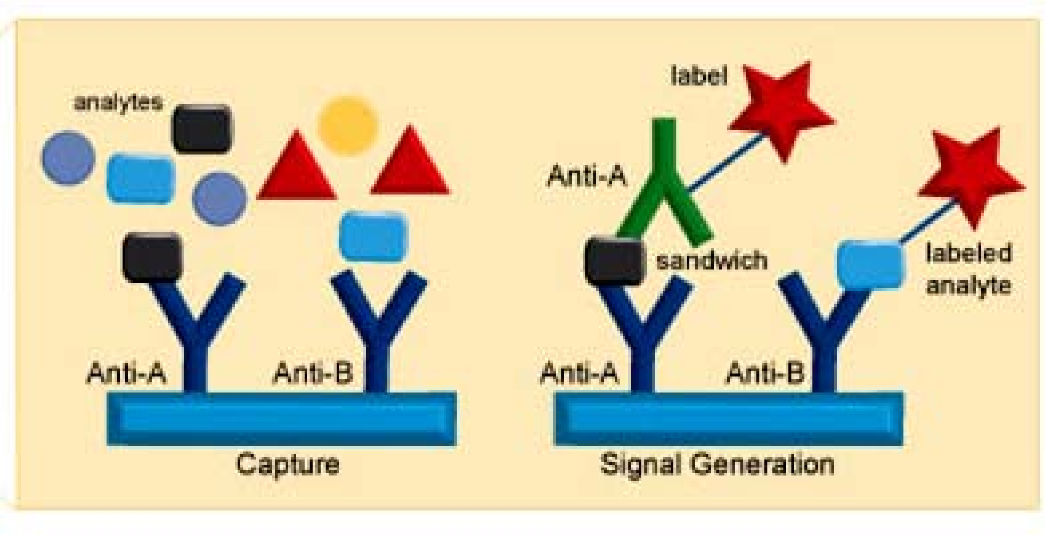
- Gene expression microarrays
- Microarrays for copy number, DNA methylation, genotypes, histone states, protein-DNA binding

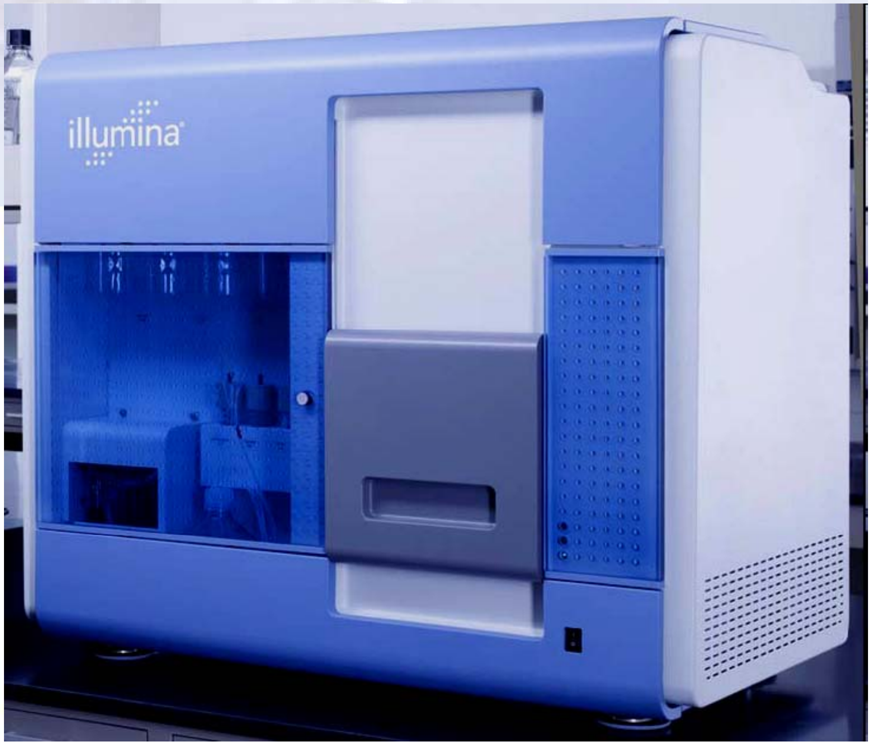




New Technologies

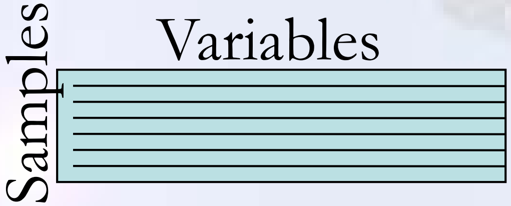
- New arrays to detect proteins, modified proteins, and antibodies
- QHTS – quantitative high-throughput sequencing





Characteristics of High-Throughput Data

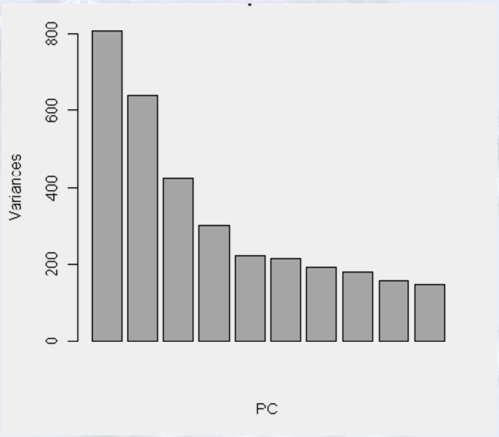
- ‘Wide’: P (# parameters) >> N (# samples)
  - – Genetics: P ~ 500,000; N ~ 1,000 - 5,000; P/N ~ 100 - 500
  - – Genomics: P ~ 50,000 ; N ~ 10 – 100; P/N ~ 500 - 5,000
  - – fMRI: P ~ 25,000 x 200; N ~ 20 – 50; P/N ~ 10,000
- Measures are parallel and of same type



High-Throughput Data is Correlated

- PCA of even highly variable genomic data sets represents 50% of the variance with 4 - 6 factors.
- Entropy has run wild; these cells have the most uncorrelated variation of any gene expression data set

Scree plot of PCA of gene expression profiles from 60 cancer cell lines



60 cell lines

Issues with High-Throughput Data

- Errors are more systematic than random
  - PCA of technical differences in many array data sets can represent 70-80% of ‘pure error’ variance with 2 factors
  - Estimates of parameters are correlated because the errors are correlated.
  - Correlations of estimates are often negative. Therefore the ‘positive dependence’ assumptions of the usual FDR theory are not satisfied.
  - Despite statisticians’ sermons, researchers don’t do randomized designs: comparison between samples is often confounded with differences in technical preparation

Will this always be so?

- Most high-throughput assays depend on several (often enzymatic or competing) processes; these occur in complex liquid mixtures
- Technologists are always pushing the envelope of what is technically possible using a delicately balanced measurement process.
- These assays are expensive, and hard to schedule randomly

How to Estimate Systematic Error

- The conventional strategy of analysis of covariance doesn’t help much here because…

  there are many known covariates for samples (dates of processing, batches of reagents, technician, …)

  …but …

  there are not many observations for each measure, hence few degrees of freedom to fit many covariates

Outline

- The avalanche of high-throughput data
- New strategies for estimation (‘normalization’) – how to effectively borrow information across measures
  - Estimating bias by non-parametric regression on technical characteristics of measures
  - Multivariate analysis of real or synthetic controls
  - Multivariate analysis of residuals from model
    - A new strategy: use benchmark studies
  - An effective hybrid approach
- An application to drug-transporter interaction

m3

The Overall Model

- We want to infer biological effects
- Effects confounded by technical common factors
- Some factors are known; some are unknown
- Model:

Coefficients of interest

$$\vec{y} \sim x^{\mathrm{T}} B + \alpha^{\mathrm{T}} A + \lambda^{\mathrm{T}} \nu + \varepsilon$$

known predictors
of interest

unknown common
factors biasing errors
“unknown unknowns”

common error factors
affecting outcomes
in known ways
“known unknowns”

Slide 13

m3 You are using the standard convention here,
but using the microarray convention later on
and the symbols are not quite consistent

which of these are known and which are unknown?

mreimers, 9/23/2010

Strategies

- *Strategy 1:* Use known technical characteristics of measures to identify those other measures most likely to be informative about biases in any particular measure *(known unknowns)*
- *Strategy 2:* Do PCA of controls across samples to infer common factors of bias and regress other measures on those factors *(unknown …)*
- *Strategy 3a):* Infer covariates by PCA of correlation structure of residuals from model (Leek & Storey) *(unknown unknowns)*
- *Strategy 3b) Use PCA of residuals to infer null space of 'technical' variation for technology*

Strategy 1:
Within sample regression of residuals on technical measures

Regressing Bias on Probe Characteristics

- Basic Idea:

1. Most technical variation between chips is caused by a few (unknown) systematic measurement factors
2. Probes with similar technical characteristics (thermodynamic characteristics, typical level of saturation, etc.) are biased by similar amounts by these factors in the assay

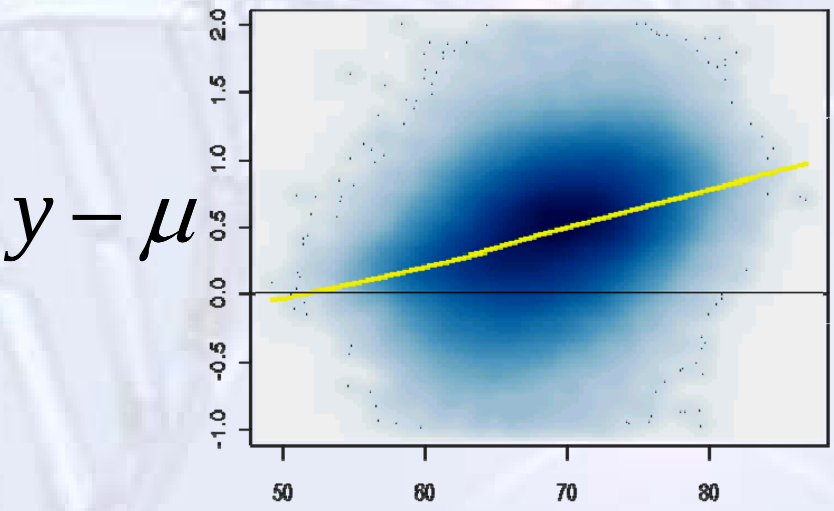
- Then technical variables are predictors of bias
- We can treat real biological differences as ‘noise’ in order to estimate bias from unknown factors.
- Model is:

$$\vec{y} \sim \vec{\mu} + x^{\mathrm{T}} B + A\alpha + \varepsilon$$

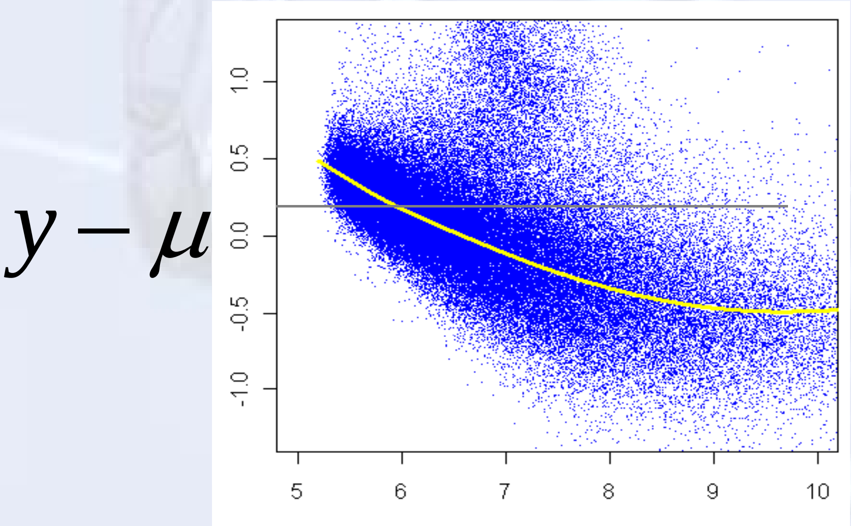
Indexing Biases by Technical Characteristics

- Thermodynamics can be indexed by equilibrium ‘melting’ (annealing) temperature T_m
- Saturation can be indexed roughly by average intensity (across all)
- Plots show deviations y-μ for two chips from (Cheung *et al*, ***Nature***, 2005) against two proxies for technical characteristics
- Yellow loess curves track trend

Deviations of log intensity from mean plotted against T_m



Deviations of log intensity from mean plotted against mean



Algorithm

1. Construct deviations from component-wise average (reference) profile
2. Identify which technical variables have the most effect
3. Estimate induced bias on any chip by regressing deviations from average on technical variables
4. Subtract estimated bias from individual estimates

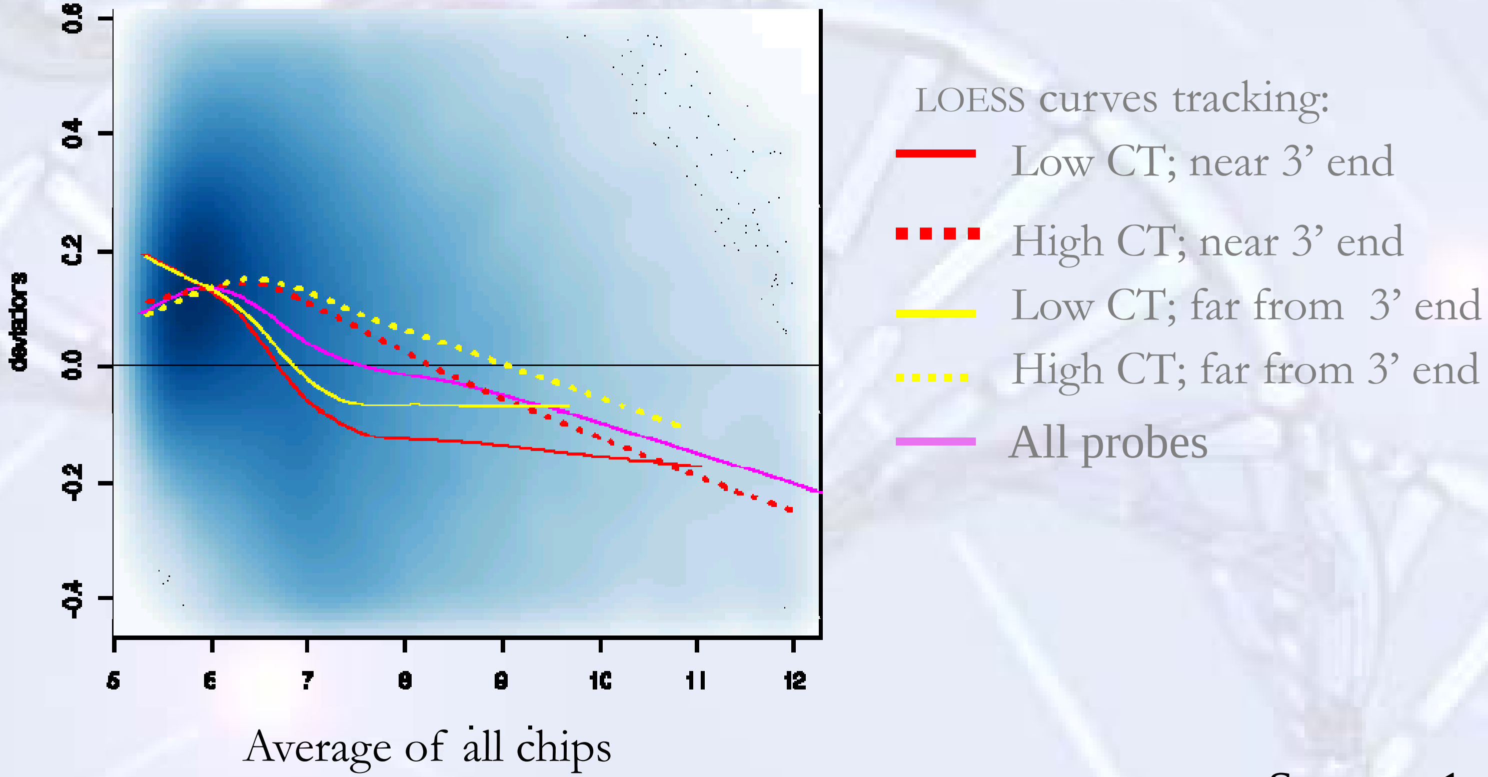
m5

Nonlinear Fits are Much Better

- Non-linear, non-additive interactions are usual
- Local regression better than linear model

Deviations of chip GSM 25410 from average of all chips in study

Overall downward trend (apparent loss of expression) at higher values of average intensity



Strategy 1

Slide 19

m5 maybe an easier graphic to understand effects of interactions

mreimers, 9/23/2010

Strategy 2:
Multivariate analysis of controls

Inferring covariates from Controls

- Variation in most measures reflect both biological and technical variability
- Variation of measures with little or no biological variability reflect mostly technical variability
- Variation of negative controls (e.g. random probes) reflects only technical variation
- The same common factors driving errors in these controls may drive errors in other measures

West’s Method

- (West, 2009) does PCA on 100 most stable genes across samples to infer two unknown covariates across samples
- Regresses all other genes on those inferred technical covariates; then subtracts estimates from measures
- Finds significantly sharper results when testing for differential expression between tumor classes

Negative Controls Predict Variation in Values of Low-Intensity Probes

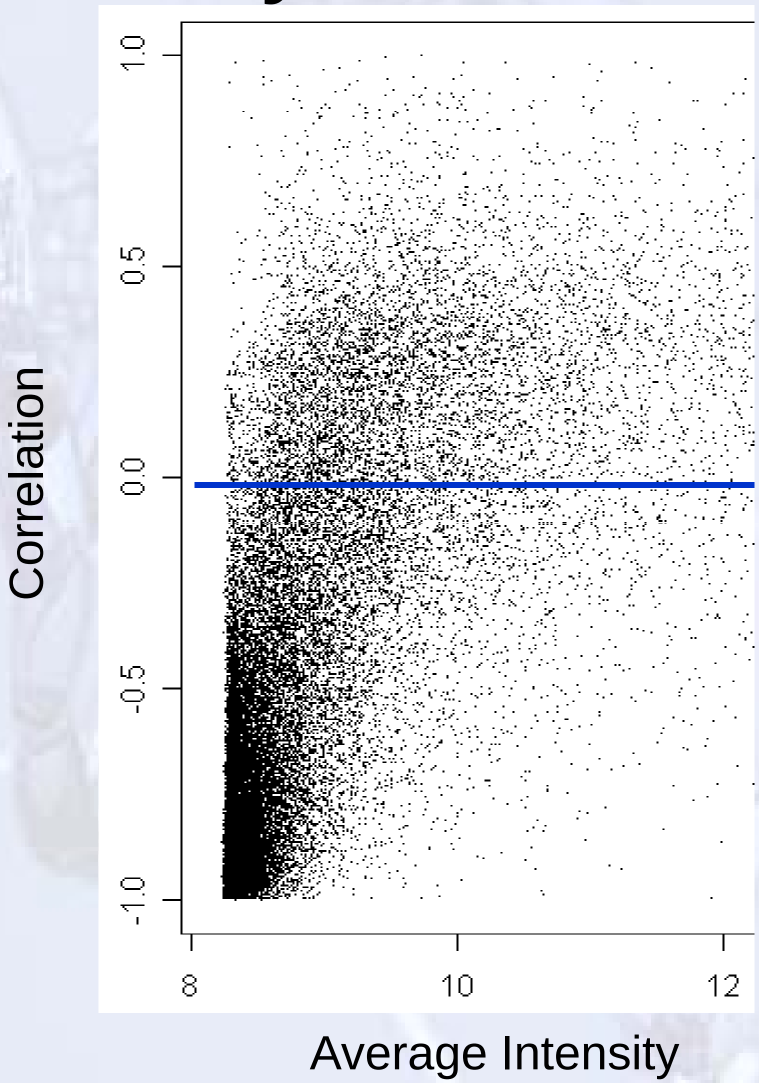
- MAQC Agilent data:

  4 samples x 5 replicates

  300 ‘negative’ controls

  (probes matching RNA that isn’t there)

- PC 1 of negative controls has very high correlations with most measures of genes of low mean intensity (many not expressed)



Synthetic Controls

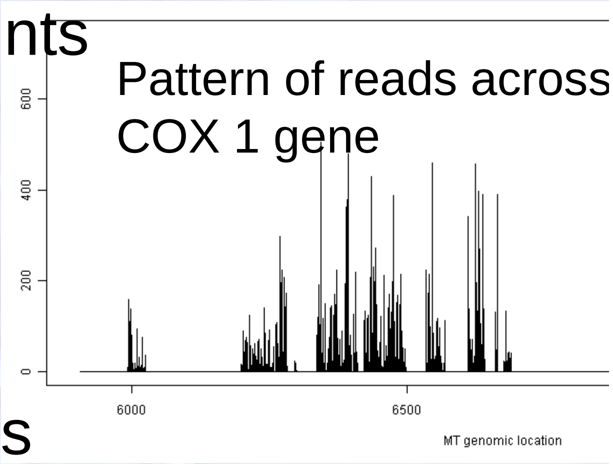
- Many HT technologies don’t have controls
- However often several measures may reflect the same underlying real object but differ in ways that may reflect different measurement biases
  - For example reads mapping to mitochondrial genes
  - The differences between these measures should be constant … but often are not
  - Tracking such differences could detect common factors affecting all measures
- I call these differences between real measures ‘synthetic controls’

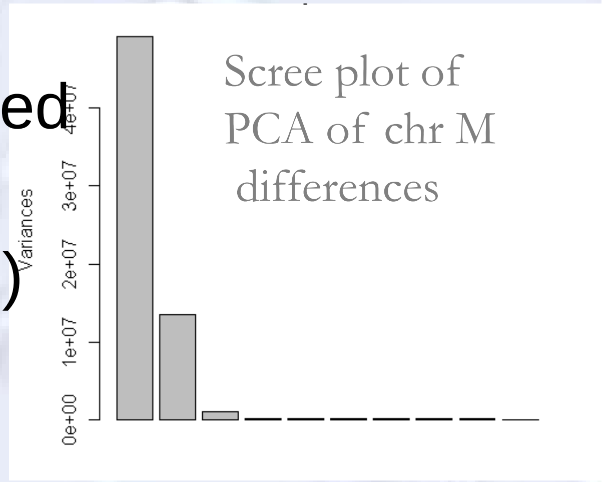
High-Throughput Sequencing Data

- HTS methods count the number of sequences matching each possible genomic position
- These ‘digital’ data are not supposed to suffer from technical biases
- Data: 5 replicates each of kidney and liver samples, normalized by total counts
  - Marioni et al, ***Genome Research***, 2008
- Puzzle: most house-keeping genes appear less expressed in harder-working liver tissue

Synthetic Controls Identify Problems

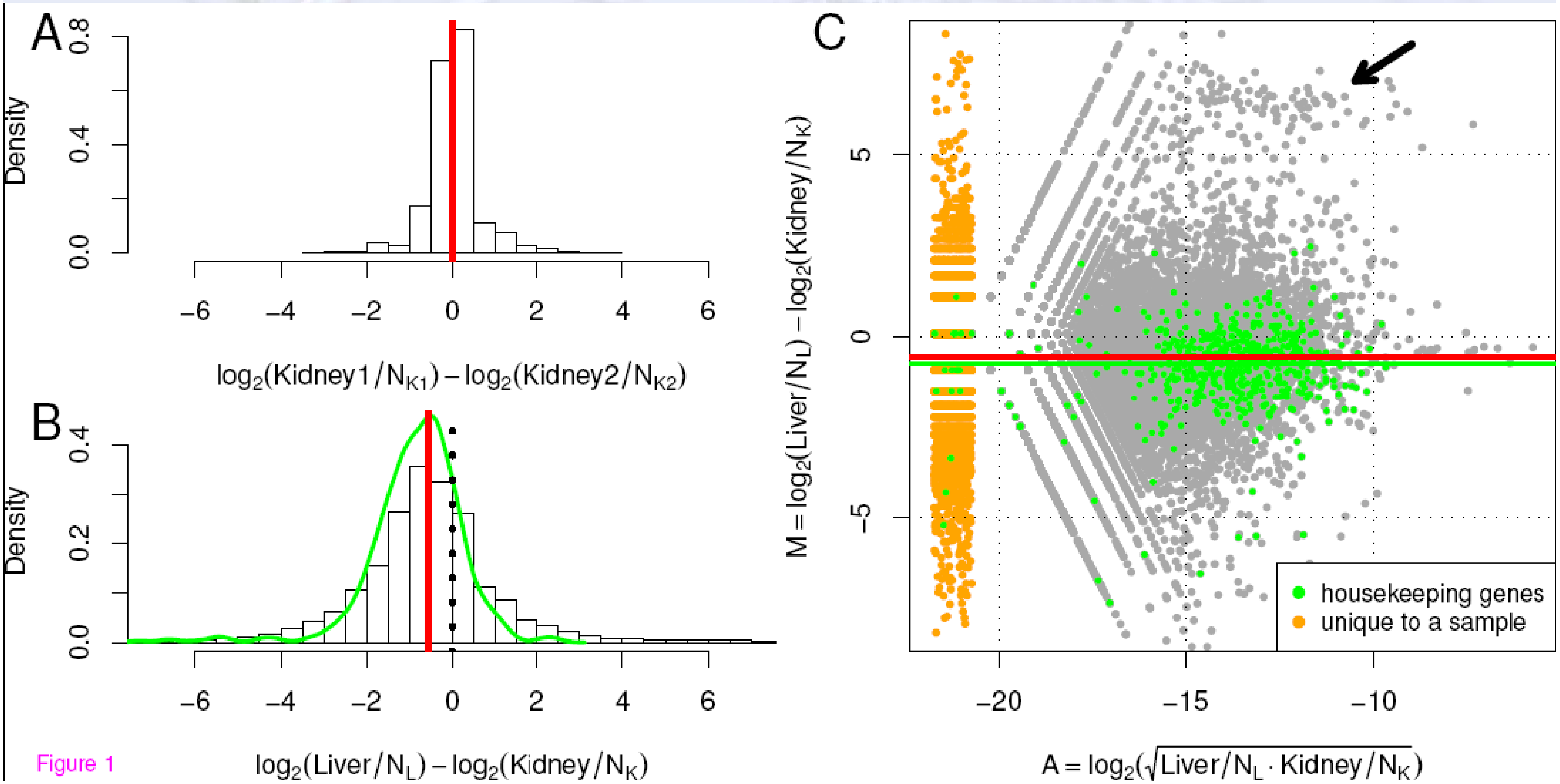
- I examined differences in sequence counts starting at various positions within the same highly expressed (mitochondrial) genes (synthetic controls)
- Two PC’s summarize almost all differences in read count patterns across 10 samples (at lower right)
- BUT strongest differences are confounded with tissue identity
- Robinson et al (***Genome Biology***, 2010) demonstrated this bias (based on other considerations) in the data set


Pattern of reads across COX 1 gene


Scree plot of PCA of chr M differences

Bias in RNA-Seq Data

from (Robinson & Oshlak, Genome Biology, 2010)



Strategy 3
Multivariate analysis of residuals or replicates

Inferred (Surrogate) Covariates

- Surrogate variable analysis (SVA)
  - Leek and Storey, ***PLoS Genetics***, 2007
- Motivation: many unmodeled (and unknown) factors affect the measures
- Even if known, most experiments don’t have sufficient d.f. to estimate their effects
- Idea: often the effects of several factors are somewhat correlated (on all probes)
- They show how to infer a manageable set of ersatz (surrogate) covariates that predict almost the same variation

The Leek-Storey Model

- There are factors f_1, …, f_K, which affect all genes via linear combinations of fixed $g_1\,(f_1)$, …, $g_K\,(f_K)$.
- The bias (systematic error) for gene i in array j is:

$$\sum_{k=1}^{K} \gamma_{ik}\, g_k(f_{kj}) = \sum_{k=1}^{K} \gamma_{ik}\, \tilde{g}_{kj}$$

- An additive representation with common functions can be represented as a linear combination of transformed variables
- In terms of my notation

$$\vec{y} \sim x^{\mathrm{T}} B + \lambda^{\mathrm{T}} \nu + \varepsilon$$

m2

How to Infer the Covariates

- Given observations ***Y***$_{L\ \mathrm{x}\ N}$ and predictors ***X***$_{L\ \mathrm{x}\ N}$,
  - (e.g. ***X*** might record diagnosis and age in columns)
- Fit the following model:

$$y_{ij} = \mu_i + \sum_{l=1}^{L} \beta_{il} x_{lj} + r_{ij}$$

$$r_{ij} = \sum_{k=1}^{K} \lambda_{ik} v_{kj} + \varepsilon_{ij}$$

- The residual matrix ***R*** is approximated by ***R ~ UDV***^{T} using singular value decomposition with ***K*** non-trivial components
- The kth row of matrix ***V*** records the kth inferred (surrogate) covariate across the N samples:

Slide 31

m2 Needs a picture and explanation of terms

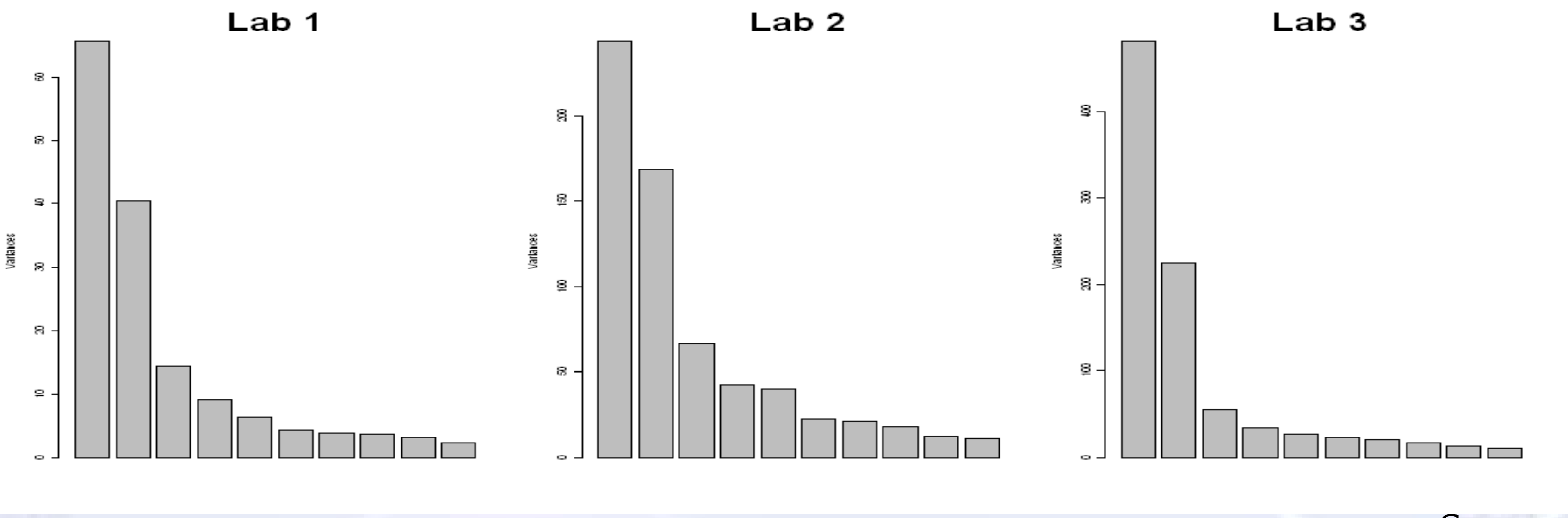
Explain: what is intuitive meaning behind low rank SVD?

mreimers, 9/23/2010

PCA of MAQC Residuals

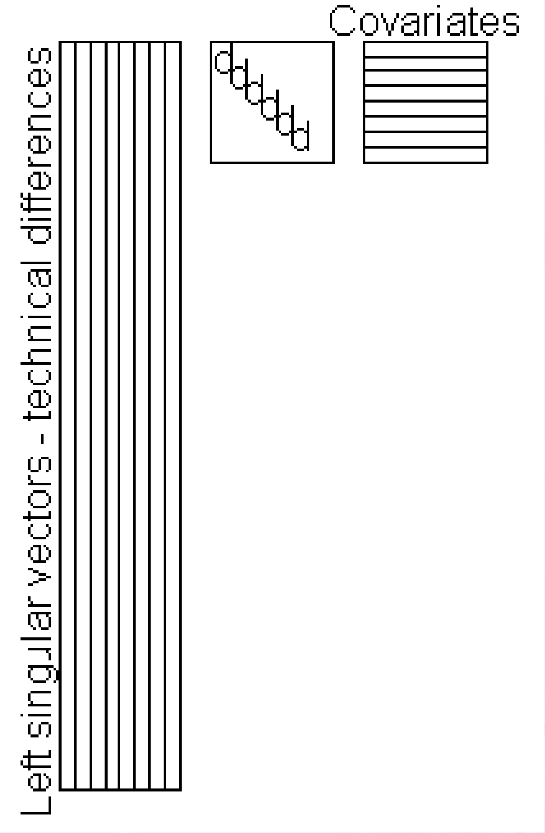
- Four samples: A: brain; B mixed tissue; C: 3:1 mixture of A & B; D 1:3 mixture of A & B
- Each sample hybridized five times in each of three labs

Scree plot of replicate PCA for Agilent 44K 1 color MAQC data set (3 sets of 4x5 reps)



Left- and Right-Handed Approaches Using SVD

- Right singular vectors represent surrogate variables
- Left singular vectors represent basis for subspace of purely technical variation
- Hypothesis: Technical errors are similar across labs
- Implication: one can ‘learn’ typical patterns of technical variation for each technology from one set of replicates



Strategy 3b

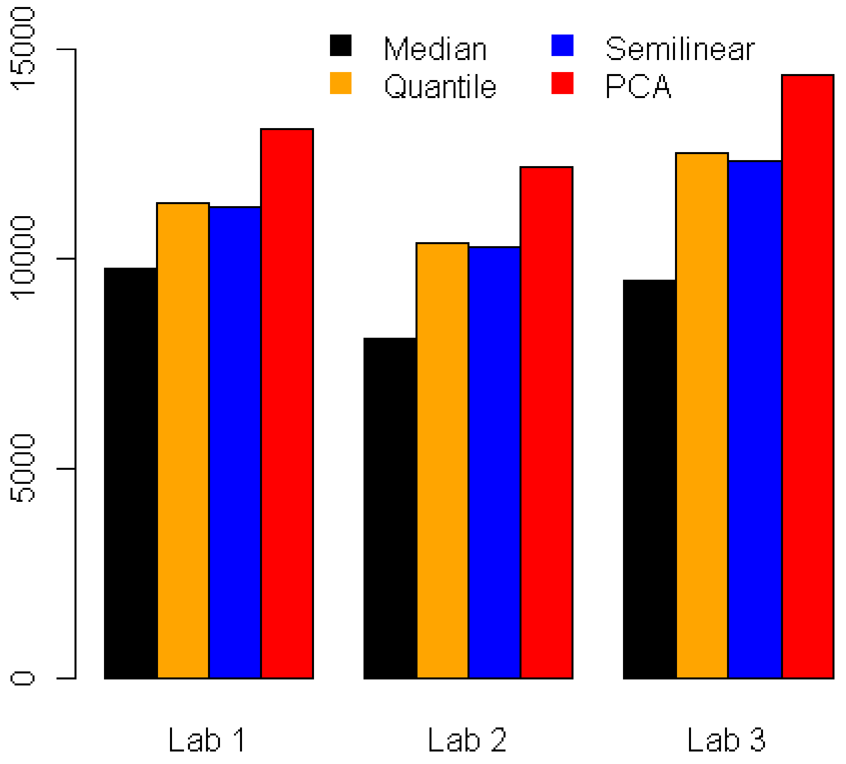
Algorithm

- Consider sets of technical replicates of the same samples, with only technical differences within sets
- PCA of replicates identifies major components
- Algorithm:
  - Construct technical differences from mean of each set
  - Robust PCA of differences
    - Outliers can be handled by simple winsorization
  - Find differences of each array from common mean of all arrays in experiment
  - Project each array’s difference onto K PC’s (K small)
  - Subtract projection (typically 50% of variance)
- Leverage points in regression are also winsorized

Results on MAQC Data

- Using 2 PC's (left singular vectors) from 4 groups of 5 replicates
- Somewhat more genes detected as differentially expressed across samples
- ~40% of variance within experiment explained by 2 PC's

Number of F-scores greater than 7



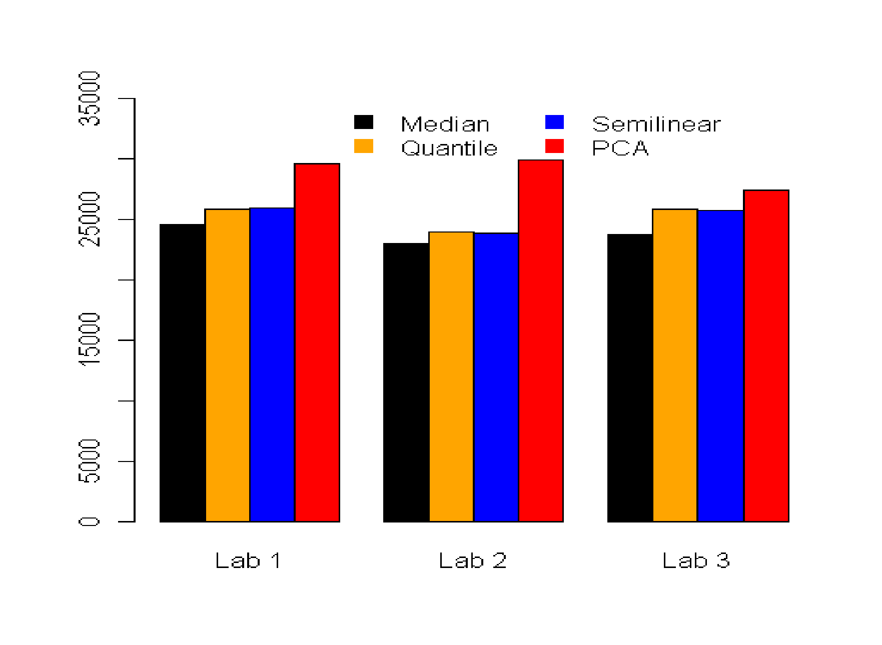
Strategy 1a

Improvement for MAQC Data: CDF of R^2 Measures Linearity

- Samples C & D are mixtures of A & B
- Expression measures in C & D should be linear combinations of those in A & B

R^2 measures linearity

PCA normalization Improves numbers of R^2 significant at $p < .005$



Can We Learn from Others' Errors?

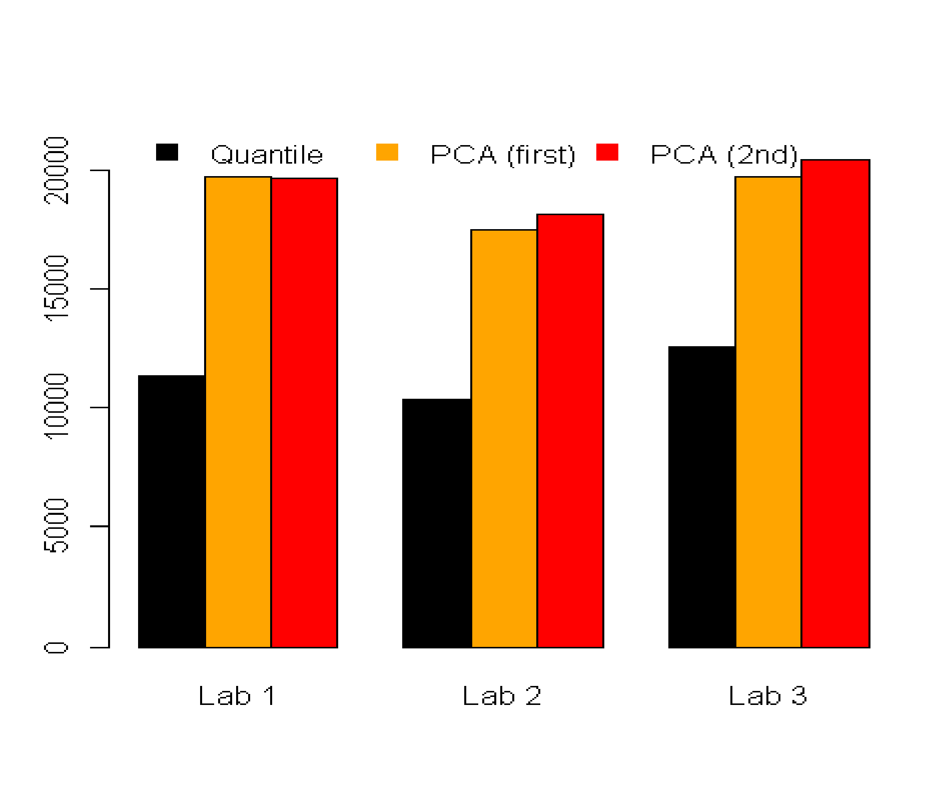
- Results in MAQC data are equivalent to consistent regression on inferred covariates
- Left singular vectors reflect characteristics of measures (probes) rather than of experiment
- Perhaps these characteristics reflect the technology and are stable across experiments
- Test: do SVD on residuals from one lab, and use those to normalize other labs following left singular vector procedure – "Benchmark PCA"

Strategy 1b

Results of Benchmark PCA on MAQC Data

- Using each lab’s PC’s to normalize the other two labs
- Five PC’s (left singular vectors) used
- Proportion of variance explained > 50%
  - – 5/40,000 expected if taking a ‘random’ subspace

Number of F-scores greater than 7



Strategy 1b

Integrating strategies and issues for the future

Combining Methods

- Use non-parametric surface most effectively to correct non-linear variation
  - spatial- and intensity-dependent biases
  - These effects often not well captured by PCA
  - Spatial inhomogeneities, saturation and cross-hybridization are nonlinear
  - LOESS adequate for 3 or 4 predictors
- Follow with PCA of residuals from technical differences
  - For CGH arrays a plausible proxies for technical differences are differences between signals from neighboring probes

Results with Combined Methods

- Expression arrays: Affy Focus, U133A
  - Improve ratio of differences / errors by 2X - 3X
- ChIP-chip: Nimblegen 380K:
  - Improve S/N by 2 X
- Array CGH: Agilent 44K and Nimblegen 380K:
  - Improve S/N by 2 X – 3 X (MS ready)

Application to Drug Discovery

- SLC genes are a large family of sodium-coupled importers – most uncharacterized
- Some sub-families import small organic molecules
- We would like to know which SLC’s may import various small molecule drugs
- Tumors over-expressing such SLC’s may be more sensitive to drugs they import

Using Correlations with NCI 60

- NCI 60 cell lines have largest collection of public cytotoxic drug-response data
- NCI 60 cell lines express varying amounts of many SLC importers
- If a particular SLC imports a particular drug then we expected a high correlation between the SLC expression pattern and the drug’s GI50 pattern across the NCI 60
- Some qPCR data available for a few SLC’s
- Microarray data available on all - but microarray measures are poor!

Results

- After normalization array data for many transporters exhibit statistically significant correlations with a small subset of drugs
- Examples:
  - Expression of SLC6A14 (known amino acid transporter) 66% correlated with GI50 for Urea
  - Expression of SLC43A3 (completely uncharacterized) is 70% correlated with GI50 of 2-Naphthacenecarboxamide

Implications and Challenges

- Systematic error is a major issue for the new high-throughput technologies – *including* the ‘digital’ technologies (HTS), which have ‘analog’ assay preparation steps
- We need a wide range of approaches for estimating biases in high-dimensional responses
- We can make significant improvements to current best practices to make microarrays (and HTS) substantially more accurate

Acknowledgements

- VCU
  - Tobias Guennel
- NCI
  - Michael Gottesman
  - William Reinhold
  - Jean-Pierre Gillet
  - John Weinstein (now at MD Anderson)
- UCSF
  - Pouya Khankhanian
- Weill-Cornell
  - Ari Melnick
  - Maria Figueroa
- Einstein (AECoM)
  - John Greally
- Roche-Nimblegen
  - Rebecca Selzer