Recent development in microarray data analysis

Guan-Hua Huang
Institute of Statistics
National Chiao Tung University

Gene expression microarray

- The overwhelming majority of results rely on measures of relative expression -- genes are reported to be differentially expressed
- Has not yet led to big advances in diagnosis or treatment
- The main reason:
  - Probe characteristics can cloud the relationship between observed intensity and actual expression
  - Although this “probe effect” is large, it is also very consistent across different hybridizations
  - Relative measures of expression are substantially more useful than absolute ones.

A gene expression bar code for microarray data
(Zilliox & Irizarry. *Nature Methods* 2007)

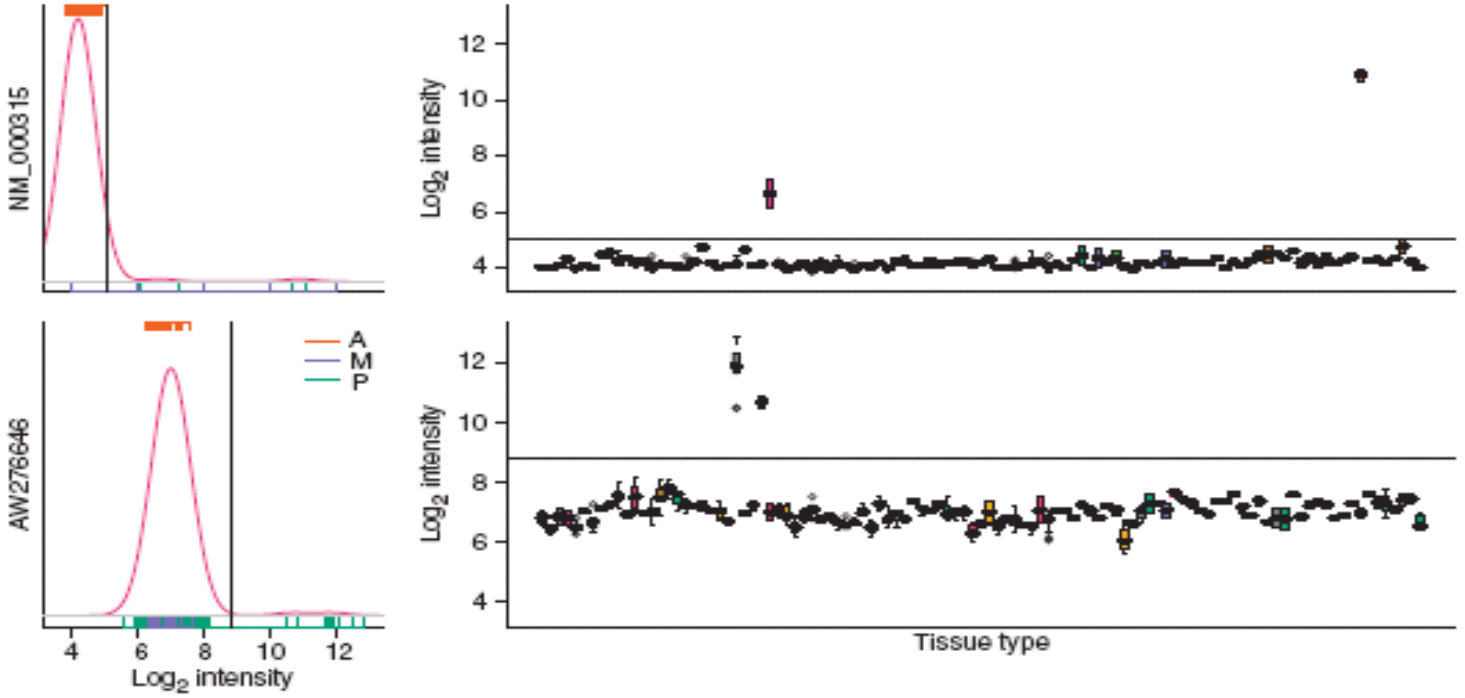
- Accurately demarcate expressed from unexpressed genes
- Select cutoff points for each platform and for each gene
- Use the vast amount of publicly available data sets (GEO, ArrayExpress) to select cutoff points
- Found that the probe effects are not large enough to change the expressed/unexpressed calls that form the bar code, making this new procedure robust to the lab/batch effects.

A gene expression bar code: for Affymetrix HGU133A chips

1. Obtain all the control samples for which the raw data (CEL files) were available from GEO and ArrayExpress
2. All raw data were preprocessed using RMA.
3. For each gene, select the cutoff point for expressed/unexpressed.
4. If a new sample is provided, simply compare the observed intensity to the determined cutoff point for each gene to determine its expressed/unexpressed – the gene expression bar code

Bar code cutoff point selection

- Any given gene will only be expressed in some tissues, multiple modes should be observed.
- The lowest intensity mode is due to a lack of expression.



Classification performance

Table 1 | Percentage accuracy comparison on independent data sets

GEO identifier	Data type	PAM (% correct)	Bar code (% correct)
GSE5388	Cortex	100	100
GSE2395	Respiratory system epithelia	0	100
GSE2665	Lymph node/tonsil	35	95
GSE1561	Breast tumor	69	100
GSE2603	Breast tumor	77	90
GSE6344	Kidney: normal versus cancer	100	100

PAM versus the bar code approach in six randomly selected data sets not included in the original database. The data described in **Supplementary Table 1** were used to train the prediction algorithms. GEO, Gene Expression Omnibus.

OPEN ACCESS Freely available online PLoS GENETICS

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer[1,2], Szabolcs Szelinger[1], Margot Redman[1], David Duggan[1], Waibhav Tembe[1], Jill Muehling[1], John V. Pearson[1], Dietrich A. Stephan[1], Stanley F. Nelson[2], David W. Craig[1]*

1 Translational Genomics Research Institute (TGen), Phoenix, Arizona, United States of America, 2 University of California Los Angeles, Los Angeles, California, United States of America

- Describe a framework for accurately and robustly resolving whether individuals are in a complex genomic DNA mixture using high-density SNP genotyping microarrays.

Determination criteria - relative differences

- Use raw allele intensity measures to estimate allele frequency, not the qualitative genotype
- The distance measure

$$D(Y_{ij}) = |Y_{ij} - Pop_j| - |Y_{ij} - M_j|$$

Y_{ij} : the allele frequency estimate for the individual *i* and SNP *j*

M_j : the allele frequency of the mixture at SNP *j*

Pop_j : the reference population's allele frequency

Hypothesis testing

- H_0 : the individual is not in the mixture
- H_1 : the individual is in the mixture
- Under H_0, $D(Y_{ij}) \leqq 0$
- Under H_1, $D(Y_{ij}) > 0$
- Test statistic : one sample t test

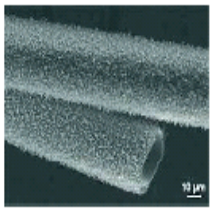
$$\frac{mean(D)}{sd(D)/\sqrt{n}} \overset{H_0}{\sim} Normal(0,1)$$

Bar code vs. resolving complex mixtures

- To overcome microarray probe effects
- Bar code – for each gene to determine its expressed/unexpressed
- Resolving complex mixtures – for each gene to calculate the difference between the individual and the reference relative to the difference between the individual and the mixture

Possible research topics

- ALE strata for subdividing a microarray dataset and analyze each stratum individually with the best performing methods
- Use public available datasets (GEO, ArrayExpress) to generate the “norm” for microarray analysis
- Use public available datasets (GEO, ArrayExpress) and bar code idea to simulate “real” microarray data



STRONG STUFF
Colossal carbon tubes take the strain.
www.nature.com/news

The death of microarrays?

High-throughput gene sequencing seems to be stealing a march on microarrays. **Heidi Ledford** looks at a genome technology facing intense competition.

- Tailor-made, small-market arrays to suit more specific research needs
- Improvements designed to drive prices down and expand into clinical diagnostics.
- creating arrays that can be used to isolate specific regions of the genome for sequencing -- ‘capture arrays’