Working with multiple microarrays

Vincent Carey and Robert Gentleman

Statistics and Genomics - Lecture 2

Department of Biostatistics
Harvard School of Public Health

January 23-25, 2002

Outline

- Recap of Lecture 1
- High-level view of data structures
- Exploratory data analysis: expression density diagnostics
- Gene selection/filtering concepts and tools
- Biological caveats

Recap of Lecture 1, Part I

- genome biology and the central dogma of molecular biology:
  - cell function and cell pathology is explainable through understanding of gene expression
  - gene expression is measurable in terms of specific mRNA abundance

Recap of Lecture 1, Part II

When microarrays are used to measure mRNA abundance, statistical modeling is needed to eliminate systematic variations within and between arrays.

Microarrays measure abundance of mRNA for prespecified sequences. These sequences may be related to genes, they may or may not be *correct*. For example minor mutations may not be detected.

Oligomer arrays in brief

- Affymetrix chips have short artificially synthesized nucleotide sequences of length 25.
- Probe pairs are constructed with perfect match and mismatch signals. A mismatch is created by inverting the 13th nucleotide to its complementary base pair.
- Approximately 20 probe pairs are used to represent each EST; $>$ 10000 EST's/chip
- A raw expression value for gene g in sample k is computed as some function $f_{gk} = f_{gk}(PM_{gk}, MM_{gk})$ over the probe sets for that gene.
- Normalization (see lecture 4) and other processing must be carried out prior to any other analyses. The reasons are similar to those given for cDNA arrays.

Collections of oligomer arrays

After conormalization of expression profiles on G genes in n samples, we have a $G \times n$ expression matrix

For inference about the roles of genes in disease processes, multiple covariates will be required (e.g., age, disease-defining phenotype, environmental exposures, treatment history).

These must be associated with the expression level data.

Examples of experiments

- A cohort study. We have n patients and for each of them we have obtained data on K covariates and estimated expression levels on G genes.
- A designed experiment. For each of two different cell lines expression level data on G genes is available at four time points under two sets of experimental conditions (with two replicates at each time point).
- A longitudinal study. A cohort of n patients are followed over l time points. For each patient K covariates and expression level data on G genes are available.

In each case one would like to identify sets of genes that have associations with particular variables (either patient covariates or experimental conditions).

Typically G is between 10,000 and 30,000, n ranges from three or four to a few hundred and K takes on similar values to n.

One might also be interested in the effects of some genes adjusting for different covariates but that problem, while more complex, is in principle similar to the problem of selecting genes with expression levels that are associated with levels of a single variable.

Clearly there are other questions of interest and different approaches that will be appropriate, but we will focus simply on a ranking of genes in order of *interestingness*. Selection is then easily done by choosing those genes with the highest ranks.

Some sample question:

- In an study of breast cancer in women can we detect genes that have expression levels that correlate with lymph node status (whether there is metastasis detected in the proximal lymph nodes).
- In the time course experiment can we detect genes that have a particular pattern of expression over the experimental conditions?
- In the longitudinal study can we detect genes whose expression level is related to changes in the patients clinical status?

Given that there are thousands of tests/regressions to be run, one cannot hope to interact with them all. So we need automatic procedures and we are likely to miss some things.

There has been very little attention paid to test diagnostics. Since we are carrying out a great many tests with very little scrutiny is seems like it would be prudent to incorporate some testing diagnostics in the process.

Not only do we carry out the test but we should also attempt to determine whether it is appropriate. Some of these issues will be addressed subsequently.

Using collections of arrays in R

A data structure called `exprSet` has been defined in the `Biobase` package to provide coordinated access to expression levels and phenotype data.

For an example based on the Golub ALL/AML discrimination data, the following setup may be used:

```
library(Biobase)
library(golubEsets)
data(golubMerge)
```

Using collections of arrays in R..

now `show(golubMerge)` yields

```
Expression Set (exprSet) with
7129 genes
   72 samples
      phenoData object with 11 variables
                  and 72 cases
    varLabels
      Samples: Samples
      ALL.AML: lymphocytic vs myelogenous
      BM.PB: bone marrow vs periph blood
      T.B.cell: T.B.cell
      Gender: Gender  ....
      Source: institution
```

Using collections of arrays in R

Familiar subscripting operations work as expected for a $G \times K$ matrix:

`show(golubMerge[1:4,])` yields

```
Expression Set (exprSet) with
   4 genes
   72 samples
       phenoData object with 11 variables
                   and 72 cases
    varLabels
      Samples: Samples
      ALL.AML: lymphocytic vs myelogenous
                 ...
```

Using collections of arrays in R

```
ALL.samps <-
    golubMerge$ALL.AML == "ALL"
show( golubMerge[ , ALL.samps] )
```

yields restriction to ALL cases.

```
Expression Set (exprSet) with
   7129 genes
   47 samples
       phenoData object with 11 variables
                  and 47 cases
    varLabels
      Samples: Samples
      ALL.AML: lymphocytic vs myelogenous
         ...
```

Using exprSets in R

Convenience functions for working with exprSets include

- `exprs()`: retrieve the numerical $G \times K$ matrix of expression values
- `phenoData()`: retrieve the phenotype data
- `geneNames()`: retrieve the gene identifiers

```
set.seed(123)
geneNames(golubMerge)[sample(1:100,size=5)]
```

```
[1] "AFFX-BioC-5_st" "AFFX-DapX-5_at"
[2] "AB000468_at" "AB001325_at"
[5] "AB002380_at"
```

Annotation: tip of the iceberg

These are examples of Affymetrix identifiers:

```
[1] "AFFX-BioC-5_st" "AFFX-DapX-5_at"
```

The `annotate` package defines mappings between different nomenclature systems.

If the translation from Affymetrix to GenBank has been loaded into `HGu95togenBank` then we can translate as follows:

```
get("AFFX-BioC-5_st", env=HGu95togenBank)
[1] "J04423"
```

For the Affymetrix identifier, `AFFX-BioC-5_st` the corresponding GenBank identifier is `J04423`.

Annotation: tip of the iceberg

GenBank can now be interrogated with this accession number to learn about this gene.

Simply point your web browser to
`http://www.ncbi.nih.gov/Genbank/`
and type in `J04423`.

```
1: J04423
E.coli 7,8-diamino-pelargonic acid (bioA), biotin synthetase
(bioB), 7-keto-8-amino-pelargonic acid synthetase (bioF), bioC
protein, and dethiobiotin synthetase (bioD), complete cds
gi|145422|gb|J04423.1|ECOBIO[145422]
```

Annotation: tip of the iceberg

However, this process is rather slow and prone to error. Instead we provide a facility in `annotate` that will allow you to produce web pages with the links to the appropriate web sites built in.

The function `ll.htmlpage` has been developed for LocusLink identifiers.

`http://www.ncbi.nlm.nih.gov/LocusLink/`

Recap thus far

- Collections of conormalized arrays are available for exploration on WWW or in Bioconductor
- The core is a $G \times K$ numerical array of expression values
- `exprSets` coordinate this array with K records of arbitrary phenotype data, and convenience functions exist for subscripting, e.g., by phenotypic condition
- `annotate` library performs mappings between various nomenclature systems to facilitate interpretation

Coming up

- non–specific filtering: thresholds on variation or magnitude
- determining which genes are important
- covariate–dependent filtering
- gene clustering: identifying collections of genes with common expression patterns
- expression density diagnostics: tools for evaluating the diversity of expression distributions in cohorts

Data reduction

Microarrays generally contain probes for thousands of genes.

Not all genes are expressed in all tissue types.

In order to ease the computational burden (and reduce the chance of spurious results) it is useful to remove those genes that have little or no variation in the samples being analysed.

Since no reference is made to any other variables or experimental conditions we refer to this as non–specific filtering.

Non–specific filtering

After normalization and estimation of the expression levels some anomalies may exist. For example there are sometimes estimated expression levels that are negative or ones that are much too high.

For example in Golub (1999) they applied the following transformations to the data:

1. thresholding: floor of 100 and ceiling of 16,000;
2. exclusion of genes with $\max / \min \leq 5$ or $(\max - \min) \leq 500$, where max and min refer respectively to the maximum and minimum intensities for a particular gene across the mRNA samples;
3. take a base 10 logarithmic transformation.

Non-specific filtering

Other examples of non-specific filtering:

- require a proportion of the samples to have expression level greater than some constant A.
- require the samples to exhibit inter–quartile range (IQR) larger than some specified constant.
- the gap filter: either a gap (jump) of at least A_1 units, in the central portion of the data or an IQR of size A_2 units.

The gap filter

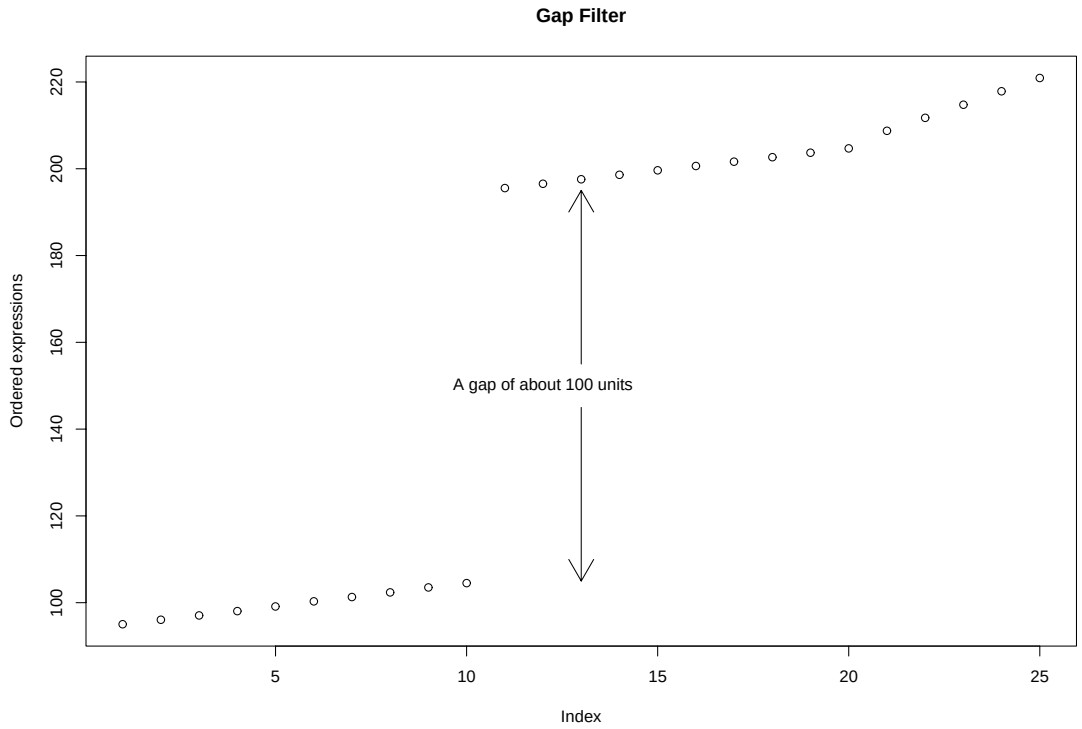


Figure 1: The gap filter.

Differentially expressed genes

Most analyses of microarray data have been directed towards the identification of genes with expression levels that are associated with a covariate or response of interest.

- ▷ *Qualitative covariates or factors*
  e.g. treatment, cell type, tumor class.
- ▷ *Quantitative covariates*
  e.g. dose, time.
- ▷ *Responses*
  e.g. survival, cholesterol level, weight.
- ▷ *Any combination of the above.*

Gene selection with covariates

First we consider the history of gene selection.

Many of the earliest papers used *fold change* as a method of selection. Fold change is simply the ratio of the expression level in one group to that in a second group.

Using fold change allowed investigators to compare two groups. It did not pay any attention to statistical variation.

Gene selection with covariates

A rather obvious extension is to use a statistical test and rank genes according to their p–values.

Tests include: t–test, ANOVA, the Cox model and virtually any other test.

Other methods include the use of ROC curves.

These methods take account of variation and allow us to select genes for further consideration on the basis of any statistical test.

Gene selection with covariates

It seems that one method that can be used to enhance the selection of genes is to use multiple tests. Genes can then be selected either on the basis of their maximum ranking or minimum ranking over the tests employed.

The choice of test or sequence of tests to be applied must be tailored to

- the contrast to be evaluated (location shift, quantile shift...)
- the distributions of gene expression levels in covariate-dependent strata.

Gene selection by subsample polling

A simple but seemingly effective method of selecting genes for further analysis goes as follows:

- select a method for ranking genes (such as p–value from some test).
- select subsamples of the data (examples include leave–one–out or bootstrap)
- for each subsample rank (or select) genes according to the method above
- select genes by simple majority poll over the subsamples.

Exploratory data analysis tools

It is customary to perform diagnostic tests for even the simplest two-sample comparisons.

The scale of the comparison problem with microarray data hinders usual approaches (plotting, outlier testing, interactive quality assessment).

We will consider how to investigate the distributions of gene-specific expression levels as preparation for good choice of statistical test.

Expression density diagnostics

Assume availability of conormalized arrays from well-defined cohorts.

Target question: For what values of $g \in 1, \ldots, 12625$ does expression of gene g differ in an interesting way between cases and controls?

Let superscripts D and $\bar{D}$ denote cases and controls resp.

Response: Choose the best method for testing

$$H_g : F_g^D = F_g^{\bar{D}}$$

or some special case of interest.

Complications for testing

- *multiplicity*: computational and inferential problems of thousands of tests possibly with data filtering
- *outliers*: measurement or recording error vs. true pathology
- *non-standard conditions*: detection limits high and low
- *mixtures*:
  - If the case-defining phenotype of interest is imprecisely characterized, F_g^D may be a mixture.
  - If there is population stratification $F_g^{\bar{D}}$ may be a mixture.

Is there a problem?

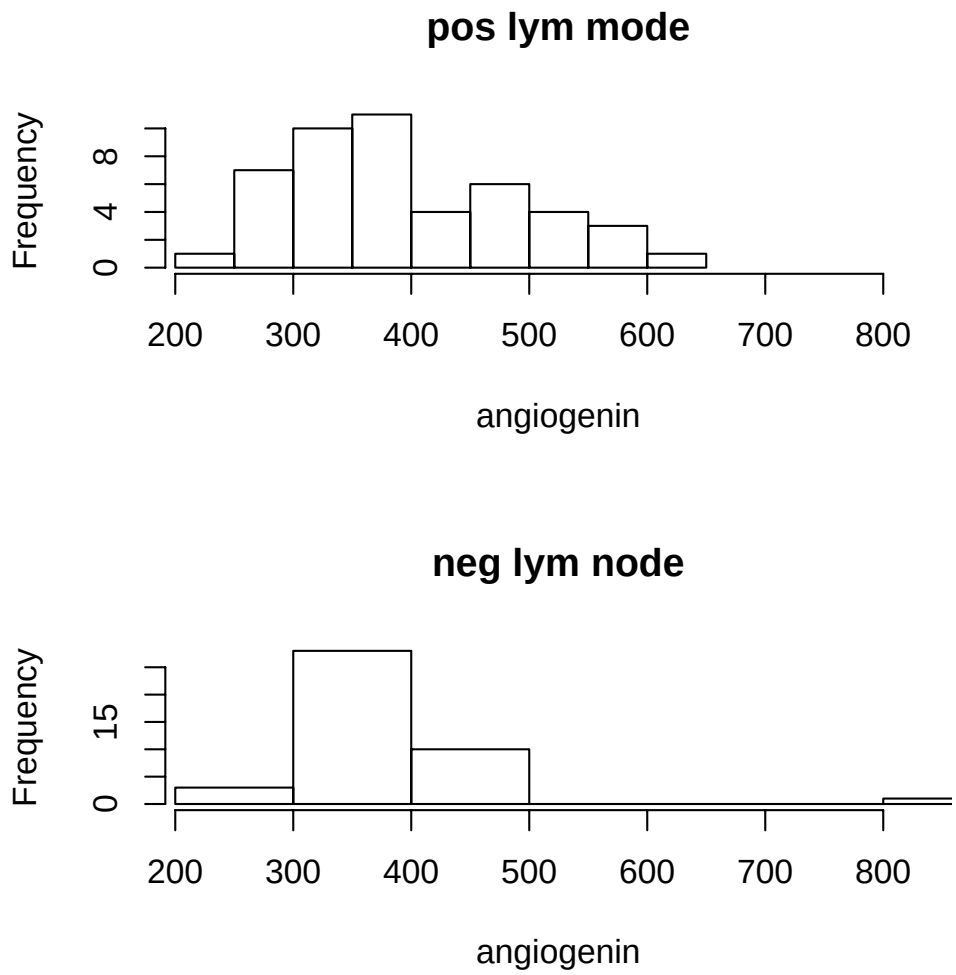
Two subquestions emerge:

- How varied are the 'shapes' of $F_g^{\bar{D}}$, $g = 1, \ldots, 12625$?
- Are shape disparities between F_g^D and $F_g^{\bar{D}}$ really likely to complicate detection of contrasts of real clinical significance?

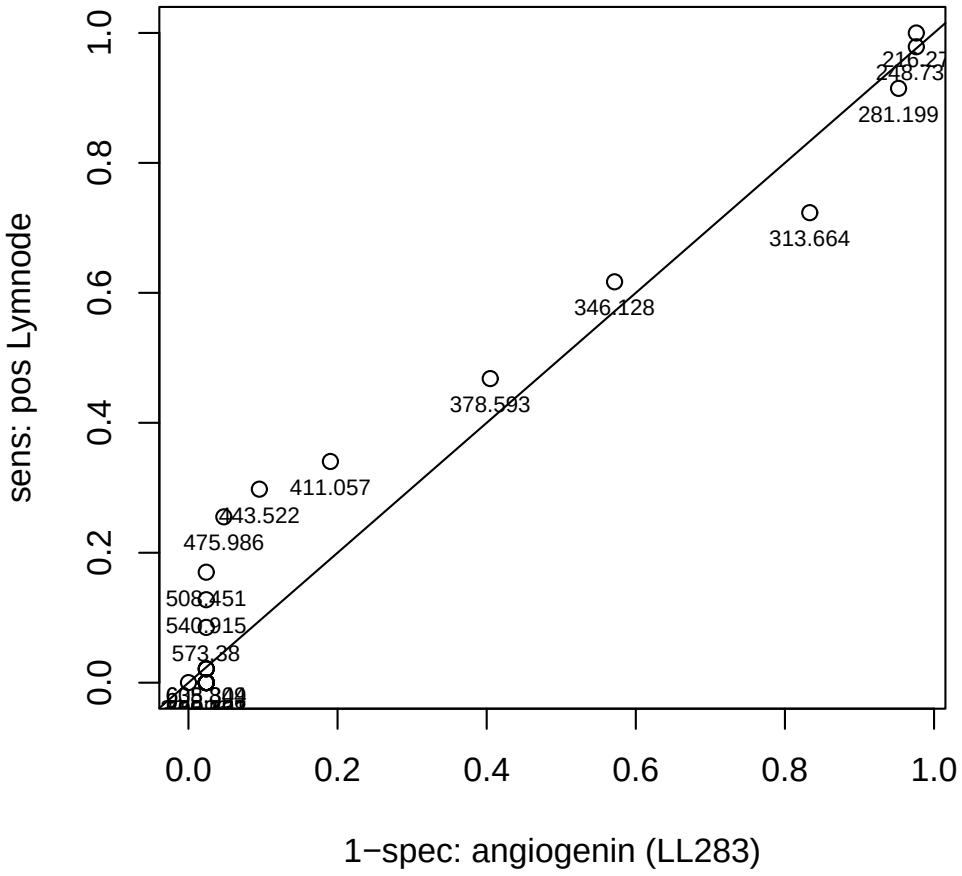
We need to be able to look to find out. This implies a need for high-throughput EDA tools.

If expression distribution diversity exists, adaptive testing or transformation should improve the discovery process.

47 metBC, 42 nonmetBC compared on 1103_at, $p_t > .20$



ROC curve contrasting LN+, LN-



Comments on the example

- move to ROC motivated by Pepe, Anderson et al.
- they argue that location shifts are not the only contrasts of interest
- a useful screening test might be developed on the basis of a modestly sensitive expression criterion that is quite specific
- use other functionals on $(F_g^D, F_g^{\bar{D}})$ to measure contrast, e.g. ROC(t) or pAUC-ROC

Assessment of diversity in $F_g^{\bar{D}}$

Notation: $Y_{gj}^{raw}, j = 1, \ldots, N^{\bar{D}}$ is the normalized expression level on gene g in subject j

$Y_{gj}^{raw} \sim F_g^{\bar{D}}$

There is considerable variation in location and scale after normalization, so we robustly center and rescale.

$$Y_{gj} = \frac{Y_{gj}^{raw} - \text{med } Y_g^{raw}}{\text{MAD } Y_g^{raw}}$$

High throughput EDA

- sort distributional shapes into a manageable number of classes, and 'unclassifiable'
  - 'graph based' classification
  - test based classification
- interpret class diversity, examine 'unclassifiable' genes as feasible
- evaluate, gene by gene, the concordance of shapes between cohorts
- provide guidance on choice of tests.

Graph-based classification

The empirical cdf of $Y_g^{\bar{D}}$ is regarded as an $N^{\bar{D}}$-dimensional multivariate datum for exploratory reasoning about $F_g^{\bar{D}}$.

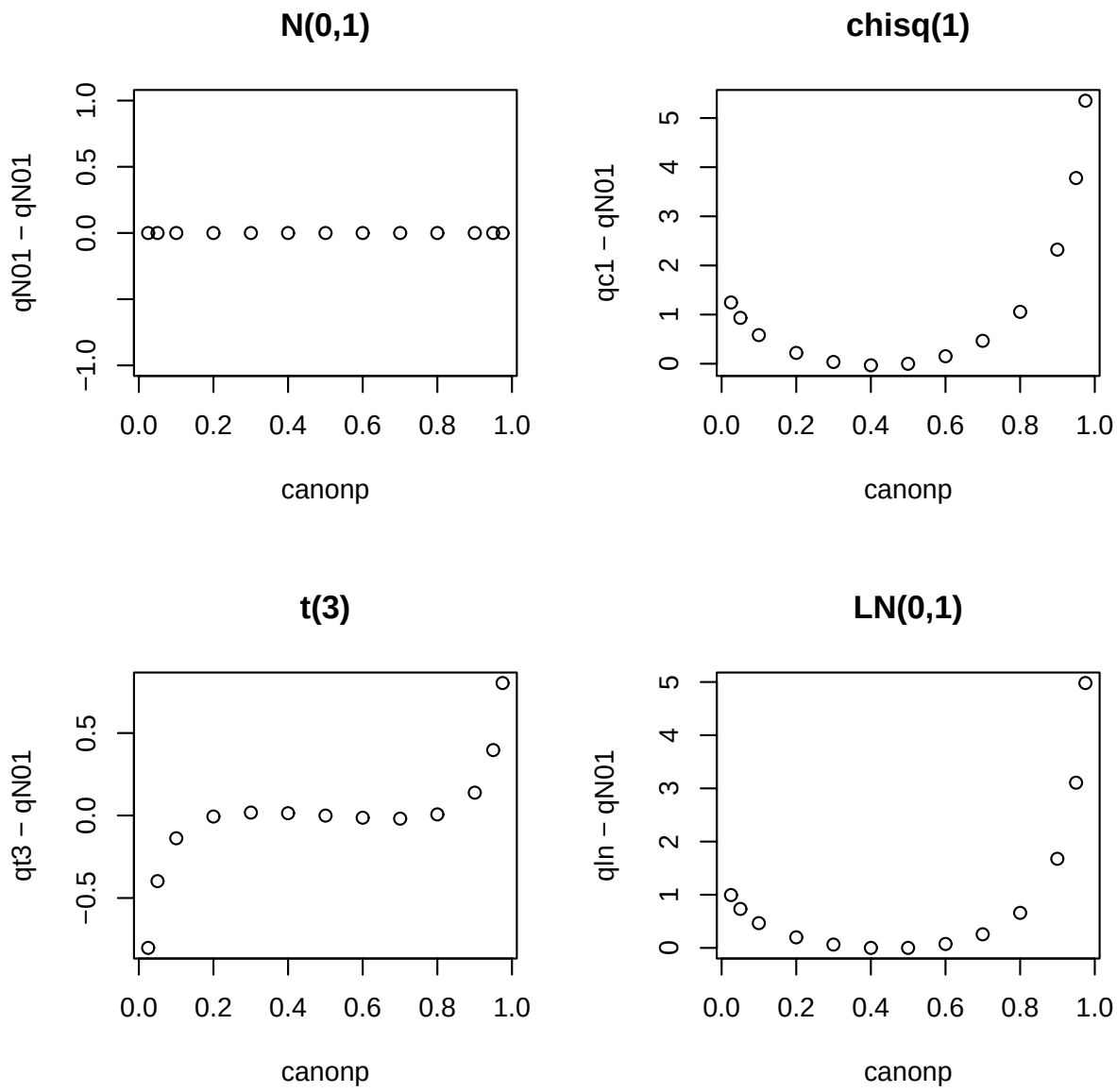
To simplify visualization, we focus on the Q-Q normal transformation, or the Q-Q normal-difference transformation in which the locus $y = 0$ corresponds to standard Gaussian

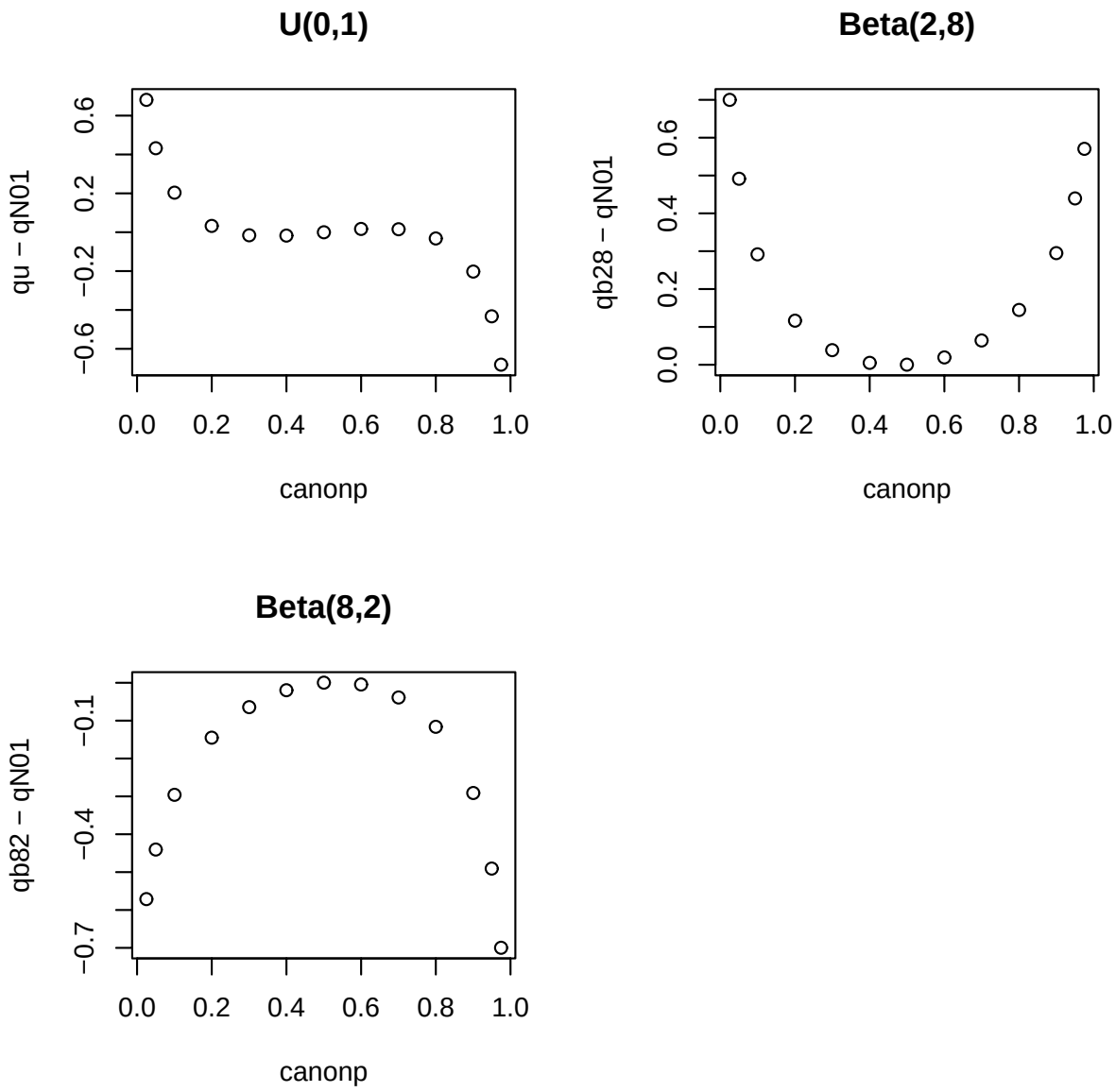
Visualization strategies for distributions

We compute the Q-Q normal difference (QQND) plot as a tool for either

- supervised sorting of distributional shapes (visually guided grouping of genes with similar QQND plots)
- unsupervised sorting by data-driven algorithms

examples follow


N(0,1)
qN01 – qN01
canonp
chisq(1)
qc1 – qN01
canonp
t(3)
qt3 – qN01
canonp
LN(0,1)
qln – qN01
canonp


U(0,1)
qu − qN01
−0.6
−0.2
0.2
0.6
0.0
0.2
0.4
0.6
0.8
1.0
canonp
Beta(2,8)
qb28 − qN01
0.0
0.2
0.4
0.6
0.0
0.2
0.4
0.6
0.8
1.0
canonp
Beta(8,2)
qb82 − qN01
−0.7
−0.4
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
canonp

Graph-based algorithm

- setup

  a) identify a set of R reference distributions Ref
  $= \{\Phi, t_3, \chi^2_1, \text{LNORM}(0,1), \text{specific mixtures}, \ldots\}$

  b) simulate m_r $N^{\bar{D}}$-dimensional representatives from
  $\text{Ref}_r \in \text{Ref}$, and recenter/scale each sample to median 0 and unit MAD

- for each g, use k-nearest neighbor classification with parameters k (number of neighbors to be polled) and l (number of assents required) to associate Y_g with an element of Ref, or "unclassifiable"

Options for the graph-based algorithm

Aspects of k-NN

- $m_r = m$ all r, some modest common number of representatives, require $l > 1$ assents – choice of k
- $m_r = 1$ all r, use a theoretically defined representative and only use the closest to classify
- m_r proportional to the prior probability of distributional shape r in the cohort

Another classification approach: fit a neural net to the representatives and use it to predict class membership for the gene expression data

Test based algorithm

- setup
  a) identify a set of K reference distributions Ref $= \{\Phi, t_3, \chi^2_1, \text{LNORM}(0,1), \text{specific mixtures}, \ldots\}$
  b) determine corrections for application of K-S tests to samples that have been centered/rescaled to zero median and unit mad
- for each $g \in 1..G$ and each $r \in 1..R$, compute corrected K-S tests of $H_o : Y_g^{\bar{D}} \sim \text{Ref}_r$ and associate Y_g with an element of Ref, or "unclassifiable", based on maximum p

Demonstration in R

setup of a small test matrix of diverse distributions

```
# 6 x 20 x 50 test problem
test <- matrix(NA,nr=120,nc=50)
test[1:20,] <- rnorm(1000)
test[21:40,] <- rt(1000,3)
test[41:60,] <- rbeta(1000,2,8)
test[61:80,] <- rmixnorm(750,250,0,1,4,1)
test[81:100,] <- runif(1000)
test[101:120,] <- rlnorm(1000)
test[121:140,] <- rchisq(1000,1)
```

Graph-based classification results

```
rescale.test <- t(apply(test, 1, centerScale))
fq.test <- fq.matrows(rescale.test)
out <- knn( train=fq.ref, cl=row.names(fq.ref),
      test=fq.test, k=10, l=2)
        method: multiCand
      INFERRED (% in 50 sims)
GIVEN|b28 csq1 ln mix1 n01 t3  u
-----------------------------------
  b28 |75    0  0   10  10  0  5
 csq1 | 0   65 35    0   0  0  0
   ln | 5   35 55    0   0  0  0
 mix1 | 5    0  0   90   0  0  0
  n01 |10    0  0    0  65 10 10
   t3 | 5    0  0    0  20 65  0
    u | 5    0  0    0   5  0 90
```

Graph-based classification result 3

method: nnet (size=6)

INFERRED (% in 50 sims)

GIVEN	b28	csq1	ln	mix1	n01	t3	u
b28	40	5	15	15	20	0	5
csq1	0	85	10	0	0	0	0
ln	5	80	15	0	0	0	0
mix1	5	0	0	90	0	0	0
n01	10	0	0	0	60	20	10
t3	5	0	5	0	15	75	0
u	10	0	0	0	5	0	85

Test-based classification result

```
  method: test-based (max p|p>.1)
      INFERRED (% in 50 sims)
GIVEN|b28 csq1 ln mix1 n01 t3  u
-----|---------------------------
  b28  55    0  5   10  15 10  5
 csq1   0   75 15    0   0  0  0
   ln  10   10 70   10   0  0  0
 mix1  15    0 20   65   0  0  0
  n01  15    0  0    0  30 40 10
   t3  10    0  0    5  20 60  0
    u  20    0  0    0  20  5 55
```

Summary of distribution classification methods

- none of the methods attempted dominates for all cases, but test-based seems least effective
- choice of method may depend upon target contrast
- need to enrich the reference set and do more work on calibrating models
- scope of reference set needs to be keyed to sample size

Shape diversity in practice

row is shape of expression distribution among AML, column is shape among ALL, in Golub full dataset

e1AML \ e1ALL	b28	b82	csq1	ln	mx1	mx2	n01	t3	u
b28	222	18	15	100	38	2	85	118	9
b82	12	46	0	4	0	7	32	36	3
csq1	10	1	29	27	0	0	4	8	1
ln	84	5	31	81	14	0	23	53	4
mix1	58	8	7	14	11	1	22	26	3
mix2	11	25	0	1	2	5	20	25	2
n01	43	20	1	10	6	3	44	51	2
t3	87	46	2	20	13	9	88	127	6
u	86	29	2	22	9	8	63	50	6

Shape diversity in practice

Preceding table indicates that the distribution of expression levels for a given gene measured in tissue from ALL patients is often of different *form* than the distribution of expression levels for the same gene measured in tissue from AML patients

For example, for 118 genes, the AML tissue-based expression levels have the shape of Beta(2,8), while the ALL tissue-based expression levels have the shape of t_3.

For a total of 33 genes, the distribution of one tissue's expression levels has a Gaussian shape, while the distribution of the expression levels in the other tissue has a lognormal(0,1) shape

Concordance of knn/nnet

classes chosen by knnMulti vs nnet on golub AML data

e1AML \ e2AML	b28	b82	csq1	ln	mix1	mix2	n01	t3	u
b28	246	0	0	9	1	0	9	30	0
b82	0	88	0	0	0	5	1	0	0
csq1	0	0	44	27	1	0	0	0	0
ln	1	0	64	135	0	0	0	12	0
mix1	5	0	0	0	117	0	2	0	0
mix2	0	7	0	0	0	77	0	0	0
n01	0	8	0	0	0	0	40	2	0
t3	0	11	0	0	0	1	29	283	0
u	0	4	0	0	0	0	2	0	168

Power ramifications

rejection rate of 2-sided tests (500 sims)

- two samples of size 50 from given dists,
- rescaled to med 0 mad 1

test = Student's t

	n01	csq1	ln01	b28	mix1
n01	0.002	0.824	0.508	0.046	0.256
csq1	-	0.084	0.054	0.520	0.212
ln01	-	-	0.010	0.232	0.034
b28	-	-	-	0.010	0.060
mix1	-	-	-	-	0.008

fluctuations across shapes more important than absolute magnitude

Power ramifications

rejection rate of 2-sided tests

- two samples of size 50 from given dists,
- rescaled to med 0 mad 1

test = Wilcoxon

	n01	csq1	ln01	b28	mix1
n01	0	0.118	0.026	0.000	0.004
csq1	-	0.010	0.000	0.040	0.000
ln01	-	-	0.000	0.014	0.004
b28	-	-	-	0.000	0.000
mix1	-	-	-	-	0.000

Power ramifications

rejection rate of 2-sided tests

- two samples of size 50 from given dists,
- rescaled to med 0 mad 1, then separate by .18

test = student's t

	n01	csq1	ln01	b28	mix1
n01	0.06	0.964	0.844	0.354	0.712
csq1	-	0.104	0.042	0.256	0.098
ln01	-	-	0.024	0.042	0.006
b28	-	-	-	0.078	0.312
mix1	-	-	-	-	0.048

Power ramifications

rejection rate of 2-sided tests

- two samples of size 50 from given dists,
- rescaled to med 0 mad 1, then separate by .18

test = wilcoxon

	n01	csq1	ln01	b28	mix1
n01	0.016	0.670	0.386	0.126	0.260
csq1	-	0.122	0.022	0.002	0.000
ln01	-	-	0.030	0.000	0.002
b28	-	-	-	0.044	0.096
mix1	-	-	-	-	0.006

Power ramifications

rejection rate of 2-sided tests

- two samples of size 50 from given dists,
- rescaled to med 0 mad 1, then separate by .5

t-test

	n01	csq1	ln01	b28	mix1
n01	0.696	1.000	0.996	0.964	0.996
csq1	-	0.252	0.124	0.064	0.072
ln01	-	-	0.172	0.018	0.086
b28	-	-	-	0.708	0.910
mix1	-	-	-	-	0.472

Power ramifications

rejection rate of 2-sided tests

- two samples of size 50 from given dists,
- rescaled to med 0 mad 1, then separate by .5

`wilcoxon`

	n01	csq1	ln01	b28	mix1
n01	0.716	1.000	0.990	0.942	0.992
csq1	-	0.832	0.642	0.188	0.340
ln01	-	-	0.774	0.394	0.602
b28	-	-	-	0.806	0.944
mix1	-	-	-	-	0.728

sensitivity not uniform over shape contrasts, but not too bad!

Recap of EDA

- provided some tools for sorting gene-specific expression distributions within cohorts into classes (defined by a set of reference distributions)
- diversity of expression distributions within strata exists
- distributional shape may vary sharply within gene, between strata
- contrast sensitivity for a median shift depends on stratum-specific distributions and test used

Multiple testing problem

Simultaneously test G null hypotheses, one for each gene g

H_g : the expression level of gene g
is not associated with the covariate or response.

Microarray experiments monitor the expression levels of thousands of genes simultaneously $\Longrightarrow$ large multiplicity problem.

Refs. Dudoit *et al.* (2001), Efron *et al.* (2000), Golub *et al.* (1999), Tusher *et al.* (2001), Westfall *et al.* (2001).

Differentially expressed genes

Question. Identify the genes that have a different expression response according to covariate A.

Approach. Simultaneously test G null hypotheses, one for each gene g

H_g : no effect due to A on the expression response of gene g.

- Compute a paired t-statistic for each gene.
- Compute permutation p-values from the distribution of the test statistics for the valid permutations of the responses (expression values).
- Adjust for multiple hypothesis testing.

Multiple hypothesis testing

	# not rejected	# rejected	
# true null hypotheses	U	V	G_0
# non-true null hypotheses	T	S	G_1
	$G-R$	R	G

From Benjamini & Hochberg (1995).

Type I error rates

1. **Per-family error rate (PFER)**. The PFER is defined as the expected number of Type I errors, *i.e.*,

$$PFER = E(V).$$

2. **Per-comparison error rate (PCER)**. The PCER is defined as the expected value of (number of Type I errors/number of hypotheses), *i.e.*,

$$PCER = E(V)/G.$$

Type I error rates

3. **Family-wise error rate (FWER)**. The FWER is defined as the probability of at least one Type I error, *i.e.*,

$$FWER = p(V \geq 1).$$

4. **False discovery rate (FDR)**. The FDR of Benjamini & Hochberg (1995) is the expected proportion of Type I errors among the rejected hypotheses, *i.e.*,

$$FDR = E(Q),$$

where by definition

$$Q = \begin{cases} V/R, & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

Strong *vs.* weak control

N.B. All probabilities are **conditional** on which hypotheses are true.

Strong control refers to control of the Type I error rate under *any combination* of true and false hypotheses,
i.e., under $\cap_{g \in K} \mathrm{H}_g$ for any $K \subseteq \{1, \ldots, G\}$.

Weak control refers to control of the Type I error rate only when *all* the null hypotheses are true, *i.e.*, under the **complete null hypothesis** $\mathrm{H}_0^C = \cap_{g=1}^{G} \mathrm{H}_g$ with $G_0 = G$.

In general, weak control without any other safeguards is unsatisfactory.

Comparison of Type I error rates

In general, for a given multiple testing procedure,

$$PCER \leq FWER \leq PFER,$$

and

$$FDR \leq FWER,$$

with $FDR = FWER$ under the complete null.

p-value adjustment

If interest is in controlling the FWER, the **adjusted p-value** for hypothesis H_g is:

$$\tilde{p}_g = \inf \{\alpha : \mathrm{H}_g \text{ is rejected at FWER } \alpha\}.$$

Hypothesis H_g is rejected at FWER α if $\tilde{p}_g \leq \alpha$.

Adjusted p-values for other Type I error rates are defined similarly.

p-value adjustment

- The level of the test does not need to be determined in advance.
- Some multiple testing procedures are most conveniently described in terms of their adjusted p-values.
- Adjusted p-values can usually be easily estimated using resampling.
- For any given procedure, adjusted p-values provide a convenient way of relating the Type I error rate to the number of rejected hypotheses.
- Different multiple testing procedures can be readily compared based on the corresponding adjusted p-values.

Notation

For hypothesis H_g, $g = 1, \ldots, G$

observed test statistic: t_g

observed unadjusted p-value: p_g

Ordering of the observed absolute test statistics: $\{r_g\}_{g=1,\ldots,G}$

such that $|t_{r_1}| \geq |t_{r_2}| \geq \ldots \geq |t_{r_G}|$.

Ordering of the observed unadjusted p-values: $\{r_g\}_{g=1,\ldots,G}$

such that $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_G}$.

The corresponding random variables are denoted by upper case letters.

Control of the FWER

Bonferroni **single-step** adjusted p-values

$$\tilde{p}_g = \min\bigl(Gp_g, 1\bigr).$$

Holm (1979) **step-down** adjusted p-values

$$\tilde{p}_{r_g} = \max_{k=1,\ldots,g} \Bigl\{\min\bigl((G-k+1)\,p_{r_k}, 1\bigr)\Bigr\}.$$

Hochberg (1988) **step-up** adjusted p-values (Simes inequality)

$$\tilde{p}_{r_g} = \min_{k=g,\ldots,G} \Bigl\{\min\bigl((G-k+1)\,p_{r_k}, 1\bigr)\Bigr\}.$$

Control of the FWER

Westfall & Young (1993) **step-down minP** adjusted p-values

$$\tilde{p}_{r_g} = \max_{k=1,\ldots,g} \left\{ p\Big(\min_{l \in \{r_k,\ldots,r_G\}} P_l \leq p_{r_k} \mid \mathrm{H}_0^C \Big) \right\}.$$

Westfall & Young (1993) **step-down maxT** adjusted p-values

$$\tilde{p}_{r_g} = \max_{k=1,\ldots,g} \left\{ p\Big(\max_{l \in \{r_k,\ldots,r_G\}} |T_l| \geq |t_{r_k}| \mid \mathrm{H}_0^C \Big) \right\}.$$

Westfall & Young (1993) adjusted p-values

- Step-down procedures: successively smaller adjustments at each step.
- Take into account the *joint* distribution of the test statistics.
- Less conservative than Bonferroni, Holm, or Hochberg adjusted p-values.
- Can be estimated by resampling, but computer intensive, especially for minP.

Westfall & Young (1993) adjusted p-values
maxT *vs.* minP

- The maxT and minP adjusted p-values are the same when the test statistics are identically distributed.
- When the test statistics are not identically distributed, procedures based on maxT adjusted p-values can lead to unbalanced adjustments.
- maxT adjusted p-values are more tractable computationally than minP p-values.
- Procedures based on maxT adjusted p-values can be more powerful in "small n, large G" situations.

Control of the FDR

Benjamini & Hochberg (1995): step-up procedure which controls the FDR under some dependency structures

$$\tilde{p}_{r_g} = \min_{k=g,\ldots,G} \left\{ \min\left(\frac{G}{k}\, p_{r_k}, 1 \right) \right\}.$$

Benjamini & Yekutieli (2001): conservative step-up procedure which controls the FDR under general dependency structures

$$\tilde{p}_{r_g} = \min_{k=g,\ldots,G} \left\{ \min\left(\frac{G \sum_{g=1}^{G} 1/g}{k}\, p_{r_k}, 1 \right) \right\}.$$

Yekutieli & Benjamini (1999): resampling based adjusted p-values for controlling the FDR under certain types of dependency structures.

"Significance Analysis of Microarrays, SAM"

Order statistics: $T_{(1)} \geq \cdots \geq T_{(G)}$.
Permutation estimates of the expected values of the order statistics under the complete null: $\bar{t}_{(g)}$, $g = 1, \ldots, G$.

1. Efron *et al.* (2000). Reject $\mathrm{H}_{(g)}$ if $|t_{(g)} - \bar{t}_{(g)}| \geq \Delta$,
where Δ is chosen based on a permutation estimate of the PFER under the complete null.

Adjusted p-values (for PCER):
$\tilde{p}_{(g)} = \sum_{l=1}^{G} p\big(|T_{(l)} - \bar{t}_{(l)}| \geq |t_{(g)} - \bar{t}_{(g)}| \mid \mathrm{H}_0^C\big)/G.$

Only weak control of the PFER.

The adjusted p-values are not monotone in g, *i.e.*, in the test statistics.

"Significance Analysis of Microarrays, SAM"

2. Tusher *et al.* (2001). Reject H_g if $t_g \geq cut_{up}(\Delta)$ or $t_g \leq cut_{low}(\Delta)$, where $cut_{low}(\Delta)$ and $cut_{up}(\Delta)$ are chosen from the Quantile-Quantile plot of $t_{(g)}$ *vs.* $\bar{t}_{(g)}$ and based on a permutation estimate of the PFER under the complete null.

Order statistics are not used in the computation of the PFER. It is thus controlled in the strong sense.

Discussion

- In multiple testing situations, there are several possible definitions of Type I error rates (FWER, PCER, or FDR). New proposals should be formulated precisely, within the standard statistical framework, to allow a better understanding of the properties of different procedures.
- Strong control of the Type I error rate is essential in the microarray context.
- Adjusted p-values provide flexible summaries of the results from a multiple testing procedure.

Discussion

- Substantial gains in power are obtained by taking into account the joint distribution of the test statistics (*e.g.* Westfall & Young (1993)).
- FDR controlling procedures are promising alternatives to more conservative FWER controlling procedures.
- More work is needed to develop procedures that take into account the joint distribution of the test statistics.
- Resampling methods are needed to estimate adjusted p-values for complex multivariate datasets.
- 2D-multiple testing problems: thousands of genes, several hypotheses for each gene.

Discussion

Rather than choosing a specific error rate to control:

1. Choose a number r of hypotheses to reject with which the researcher feels comfortable, and evaluate the adjusted p-values $\tilde{p}_{(r)}$ necessary to reach this number under various procedures and types of error control.
2. For a given level, find the number of hypotheses that would be rejected under one method, and give the level required to achieve that number under other methods.
3. Find the number of hypotheses that would be rejected using a procedure controlling FWER at a fixed level, and find how many others would be rejected using procedures controlling FDR and PCER at that level.

Back to gene selection

Earlier we presented methods that allowed you to select differentially expressed genes.

While the identification of differentially expressed genes is a useful practice the analyses of these data will require more complex methods.

In the next few slides we cover some of the reasons why differential expression per se is not sufficient.

Interesting genes: Ratios

The ratio of BAX (BCL2-associated X protein) to BAD (BCL2-antagonist of cell death) determines the cells fate.

If the level of BAD is larger than that of BAX then apoptosis (programmed cell death) is suppressed.

When the level of BAX is larger than that of BAD then apoptosis is promoted.

Interesting genes: Pathways

Most important biological activities are not the result of a single molecular activity.

They generally result from choreographed activities of multiple molecules. These activities and their constituents are called pathways.

To understand how organisms function we will have to understand the relevant pathways for that organism.

The role of microarray data in understanding biochemical pathways is not yet clear. However, we will probably need to examine the relative levels of many genes simultaneously.

Interesting genes: Mutations

The mutation of a single nucleotide can greatly affect the performance of the resulting mRNA.

Current microarray technology is unlikely to detect such mutations.

Thus, we may see *normal* levels of expression in patients that have the mutation and in those that do not.

It is possible to build microarrays to detect mutations and polymorphisms. Some exploration in this area is underway.

Data analysis

Data analysis is likely to be very problematic.

For example, either the over expression or mutation of *c-for* or *v-for* can cause cancer.
If we consider the data that might obtain, those with cancer can have an elevated expression level, a normal expression level (if the mutation is not detected), or a low expression level (if the mutation is such that we no longer identify the mRNA).

Data analysis: Tissue types

Different genes are expressed in different tissue types.

The analysis will need to account for this when the tissue that was assayed is not homogeneous.
The less homogeneous the tissue the more diverse we expect the outcome to be.

Data analysis: Proteomics

Switching the focus to proteomics seems unlikely to provide simple answers to all questions.
A simple example of what can go wrong there is given by p53. One role of p53 is the regulation of apoptosis.

Under some conditions elevated levels of the protein induce apoptosis.
A common mutations produces an inert version of the protein with a longer half–life than the functional version.

The defective version is detected by most assays and hence one will be lead to believe that there is an abundance of p53 when in fact there is a deficit of functioning p53.

Annotation

The investigation of biological questions requires the integration of many data sources. For example, expression, sequence, structure, pathway and so on.

For microarray data we typically have some identifier for each probe. That identifier must be translated to other systems.

The relationships may be applied either before or after gene selection. For example, we may want to carry out an analysis only on genes that are located on Chromosome 1. Or, after having selected the top ranking genes we might want to determine their chromosomal location.

Annotation

Our approach to providing these data is to supply a standard set of files representing the different translations for each chip that we are interested in. For example, there are a number of files in the `annotate` library that are associated with the Affymetrix 6800 chip. They are:

`hgu68Chrom`	`hgu68Symbol`	`hgu68bp3`	`hgu68cc3`	`hgu68mf1`
`hgu68Cyto`	`hgu68bp1`	`hgu68cc1`	`hgu68id`	`hgu68mf2`
`hgu68Name`	`hgu68bp2`	`hgu68cc2`	`hgu68ll`	`hgu68mf3`

Annotation

The information that they encode is given below

hgu68Chrom Chromosome location.

hgu68Cyto Cytoband location.

hgu68Name The long name of the gene (if available).

hgu68Symbol The symbol for that gene.

hgu68id The GenBank Accession number for the probe.

hgu68ll The LocusLink Accession number for the probe.

Annotation

The other 9 files are arranged in 3 groups; they encode information that has been curated and is in conformance with the Gene Ontology (GO). The GO is an attempt to standardize terminology for three areas of cellular function that are of interest. The are

bp Biological process.

cc Cellular component.

mf Molecular function.

Annotation

The GO provides a tree structure that groups terms into less and less specific groupings. Each gene is assigned a particular ontology. This is typically too specific and there will be almost as many specific ontologies as there are probes. However, by starting at a specific ontology and tracing upwards to the top 3 nodes in the tree we get some indication of the general process, component, function of the probe.

Annotation

Both GenBank and the LocusLink accession numbers are provided so users have direct access to those databases.
We will provide a means of easily linking to those databases through the construction of web–pages. (See `ll.htmlpage`).
This makes is easy to provide such information to biologists and other collaborators in a manner that they are able to deal with.

Post processing can also be carried out using R. It is important to realize that R can open http connections and download the contents of different web–pages internally.

This would let you, for example, obtain and collate abstracts or references to the journal articles related to the genes of interest.

Annotation: GO

The objective of GO `www.geneontology.org` is to provide a controlled vocabulary for the description of the molecular function, biological process and cellular component of a gene product. The material in this section is based on information provided on the GO website.

The terms defined in the vocabulary can then be used as attributes of different gene products. This will facilitate uniform queries across databases. GO provides only the vocabulary. It is up to others to allocate the terms to different gene products.

Annotation: GO

A gene product is a physical thing. It may be a protein or an RNA. Examples of gene products include alpha-globin and small ribosomal RNA.

Molecular function is what something does. It describes only what the gene product can do. Examples of broad functional terms are "enzyme," "transporter," or "ligand." Examples of narrower functional terms are "adenylate cyclase," or "Toll receptor ligand."

Annotation: GO

Biological process is a biological objective. A biological process is accomplished via one or more ordered assemblies of molecular functions. Usually there is some temporal aspect to it. Examples of broad biological process terms are *cell growth and maintenance*, or *signal transduction*.

A biological process is not the same as a pathway. The representation of a pathway is more complex. Biological processes generally consist of more than one step.

A cellular component is a component of a cell. It must be a part of a larger object.