Multiple testing and false discovery rates in mining gene expression data

Adapted from material by Yudi Pawitan

Alexander Ploner

Medical Epidemiology & Biostatistics
Karolinska Institutet

http://www.meb.ki.se/~aleplo/

June 16, 2007

Data mining

Definition 1

The nontrivial extraction of implicit, previously unknown, and potentially useful information from data

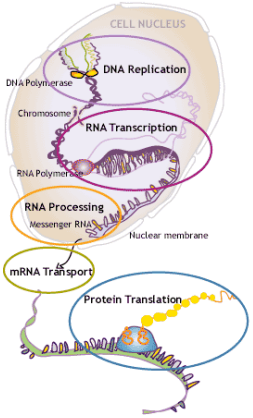
Definition 2

The science of extracting useful information from large data sets or databases

Or just staying sane

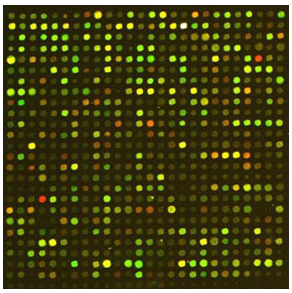
There have always been a considerable number of people who busy themselves examining the last thousand numbers which have appeared on a roulette wheel, in search of some repeating pattern.
Sadly enough, they have usually found it.

High-throughput cell biology



Source: http://nobelprize.org/medicine/educational/dna/

Gene expression microarrays



Example: Stockholm breast cancer study

- 159 breast cancer patients treated at Karolinska hospital
- Full clinical information: age, stage, grade, tumor size, survival etc.
- Gene expression measured on tumor tissue collected during surgery
- HGU133A Affymetrix chips, 22283 sequences, 11833 after filtering.
- Outcome: relapse or death due to breast cancer within five years

Which genes are differentially expressed between survivors and non-survivors?

Overview

Testing a single feature

1. Collect data $y_1, \ldots, y_n$ on a feature, together with some structure (e.g. grouping)
2. Parametrize the problem, formulate hypotheses:

   $H0: \Delta = 0$ (no difference) vs $H1: \Delta \neq 0$.

3. Compute test statistics $Z = z$. Reject $H0$ if

   $$|Z| > c$$

   where c is a critical value.

Issues

- Choice of test statistic (t-test, Wilcoxon, F-test, etc.)
- Significance level: $\alpha = P_0(|Z| > c)$
- Power or sensitivity: $P_1(|Z| > c)$
- Operating characteristics, sample size

The p-value perspective

- P-value = $P_0(|Z| > |z|)$
- Reject $H0$ if p-value $< \alpha$ – significance level?
- Distribution of p-value under $H0$?
- Distribution of p-value under $H1$?

Example: Breast Cancer – one gene

CDC2, important for cell cycle regulation

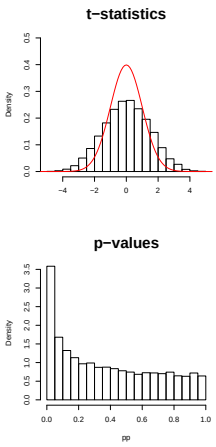
```
> t.test(CDC2 ~ status)

Welch Two Sample t-test
data:  CDC2 by status
t = -4.53, df = 115.1, p-value = 1.453e-05

alternative hypothesis: true difference in
    means is not equal to 0

95 percent confidence interval: -0.83 -0.32
sample estimates:
mean in group FALSE  mean in group TRUE
          6.856015            7.434217
```

Example: Breast Cancer – 11333 genes



Top-regulated genes:

Name	t-statistic	p-value
GPR56	4.96	2.0e-06
STAT5B	-4.99	2.0e-06
SNHG3	4.77	4.0e-06
CDKN1C	-4.76	4.0e-06
PDF	4.80	4.0e-06
MLF1IP	4.70	6.0e-06
H2AFZ	4.61	8.0e-06
MMRN2	-4.56	1.0e-05
C3orf63	-4.51	1.3e-05
ANKRD12	-4.48	1.5e-05
⋮		

p-values ≤ 0.05 = 2031 (18%)
p-values ≤ 0.01 = 799 (7%)

Problems with 1000s of simultaneous tests

Intuitively:

$$11,333 \times 0.05 = 566.65$$

Naive testing will generate

- false positives,
- false discoveries,
- irreproducible results.

Some error control required!

A hypothetical example

m =10,000 tests where we know the true state of nature

	Test decision		
True status	Accept H0	Reject H0	
H0	$U = 9,025$	$V = 475$	$m_0 = 9,500$
H1	$T = 100$	$S = 400$	$m_1 = 500$
Total	$W = 9,125$	$R = 875$	$m = 10,000$

Classical error types

- **False positive** (type-I error) rate: We observe

$$\alpha = \frac{V}{U+V} = \frac{475}{9500} = 5\%$$

- **False negative** (type II error) rate: We observe

$$\text{FNR} = \frac{T}{S+T} = \frac{100}{500} = 20\%$$

Nota bene: Sensitivity = 1-FNR = 80%

Generalized error types

- **Family-wise error rate**: Defined as

$$\text{FWER} = P(V > 0|H0) = P_0(V > 0)$$

- **False discovery** rate: We observe

$$\text{FDR} = \frac{V}{R} = \frac{475}{875} = 54\%$$

- **False non-discovery** rate: We observe

$$\text{FNDR} = \frac{T}{T+U} = \frac{100}{9125} = 1.1\%$$

Adjusted p-values controls FWER

- P-value is associated with $P_0(V > 0)$.
- Classically: adjustment of individual P-values to control FWER
  - Order the p-values $p_{(1)} \leq \cdots \leq p_{(m)}$
  - Compute adjusted p-values $\tilde{p}_{(1)} \leq \cdots \leq \tilde{p}_{(m)}$
  - Declare top k features significant if $\tilde{p}_{(k)} \leq \alpha$.

  Given suitable conditions, this guarantees $P_0(V > 0) \leq \alpha$.
- Single-step vs. stepwise procedures:
  - Single-step treat all p-values equally
  - Stepwise consider ranking of p-values
- Many, many variants: Bonferroni, Sidak, Hommel, Hochberg, Westfall-Young's minP, W-Y maxT etc.

Different kinds of error control

In reality, we have unknown subsets

$$\begin{aligned} M_0 &= \{i_1, \ldots i_{m_0}\} \subset \{1, \ldots m\} \\ M_1 &= \{1, \ldots m\} - M_0 \end{aligned}$$

of true and false hypotheses.

Control of the error level $P(V > 0) \leq \alpha$ can be

- exact: under true H_{M_0} (true null)
- weak: under $H_{\{1,\ldots m\}}$ (global null)
- strong: under all possible H_{M_0} (any null)

Bonferroni adjustment

The simplest and most conservative:

$$\tilde{p}_{(k)} = \min(m\, p_{(k)}, 1)$$

Typical proof of FWER control:

$$\begin{aligned} P_0(V > 0) &= P_0(\tilde{p}_k \leq \alpha \text{ for some } k) \\ &= P_0(p_k \leq \alpha/m \text{ for some } k) \\ &\leq \sum_k P_0(p_k \leq \alpha/m) = \alpha. \end{aligned}$$

Effect of dependence?

Westfall-Young's minP

Using the null distribution of minP = $\min(p_1, \ldots, p_m)$, we define

$$\tilde{p}_{(k)} = P_0(\min_{1 \leq l \leq m} P_l \leq p_{(k)}) = F(p_{(k)}),$$

under the global null H_{M_0}, with P_l the raw p-value of the l-th hypothesis.
Proof of FWER control:

$$\begin{aligned} P_0(V > 0) &= P_0(\tilde{p}_k \leq \alpha \text{ for some } k) \\ &= P_0(p_k \leq F^{-1}(\alpha) \text{ for some } k) \\ &\leq P_0(\text{minP} \leq F^{-1}(\alpha)) \\ &= \alpha \end{aligned}$$

Dependence? What is the null distribution of minP?

Holm's adjustment: A step-down Bonferroni

Briefly:

$$\tilde{p}_{(k)} = \max_{\ell=1,\ldots,k} \left(\min((m-\ell+1)p_\ell, 1)\right).$$

Better power than Bonferroni by:

1. Starting with smallest p-value:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)},$$

2. Performing sequential Bonferroni adjustment, adapting the number of hypotheses
3. Making sure to preserve order

Benjamini-Hochberg FDR as step-up p-value adjustment

The original B-H FDR was presented as a sequential adjustment of p-values:

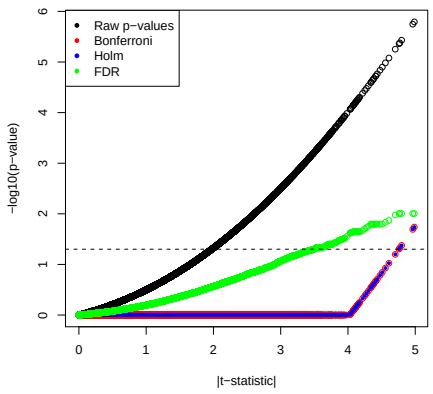
$$\tilde{p}_{(k)} = \min_{\ell=k,\ldots,m} \left\{\min\left(\frac{m}{\ell} p_{(\ell)}, 1\right)\right\}$$

This controls FWER weakly, but FDR strongly under

- independence,
- positive regression dependency,

Which is more interesting: control for FWER or FDR? Advantages and disadvantages?

Example: Breast Cancer



Example: Breast Cancer

Name	t-stat.	Raw p	Bonf.	Holm	FDR
STAT5B	-4.99	2.0e-06	0.018	0.018	0.010
GPR56	4.96	2.0e-06	0.020	0.020	0.010
PDF	4.80	4.0e-06	0.042	0.042	0.010
SNHG3	4.77	4.0e-06	0.047	0.047	0.010
CDKN1C	-4.76	4.0e-06	0.049	0.049	0.010
MLF1IP	4.70	6.0e-06	0.063	0.063	0.011
H2AFZ	4.61	8.0e-06	0.095	0.095	0.014
MMRN2	-4.56	1.0e-05	0.118	0.118	0.015
C3orf63	-4.51	1.3e-05	0.143	0.143	0.016
ANKRD12	-4.48	1.5e-05	0.165	0.165	0.016
⋮			⋮		
	# ≤ 0.05:	2,031	5	5	137
	# ≤ 0.01:	799	0	0	5

Benjamini & Hochberg 1995: FDR

Simplest idea:

$$\text{FDR} = E\left(\frac{V}{R}\right)$$

What to do when $R = 0$?
B-H: set FDR=0 if $R = 0$. Then:

$$E\left(\frac{V}{R}\right) = E\left(\frac{V}{R}\middle| R > 0\right) P(R > 0)$$

- What if $m = m_0$?
- Relationship with FWER?
- As $m \to \infty$ what happens to $P(R > 0)$?

Alternative: pFDR (Storey 2001)

Storey: Let FDR be undefined if $R = 0$. Use

$$\text{pFDR} = E\left(\frac{V}{R}\middle| R > 0\right),$$

the *positive* FDR.

Difference?

Alternative: extending FDR to $R = 0$

Look at the 2-by-2 table of conditional probabilities for *one* gene:

	Test decision		
True status	Accept H0	Reject H0	Total
nonDE (H0)	$1-\alpha$	$Pr(V_i = 1 \mid H0) = \alpha$	1
DE (H1)	$1-\beta$	$Pr(S_i = 1 \mid H1) = \beta$	1
Total		$P(R_i = 1) = \pi$	1

Probability of true nonDE genes $Pr(H0 = 1) = \pi_0$. The probability of rejection is a mixture:

$$Pr(R_i = 1) = \pi = \pi_0\alpha + (1-\pi_0)\beta.$$

Alternative: extending FDR to $R = 0$

For a collection of m features, assumed independent: (S, T, U, V) is multinomial. Given $R = r > 0$, the conditional dist. of (V, S) is binomial with size r and probability

$$p = \frac{\pi_0\alpha}{\pi}.$$

For $r > 0$ therefore

$$E\left(\frac{V}{R}\middle| R = r\right) = \frac{\pi_0\alpha}{\pi}$$

and

$$\begin{aligned} E\left(\frac{V}{R}\middle| R > 0\right) &= E_{R|R>0}\left\{E\left(\frac{V}{R}\middle| R = r, r > 0\right)\right\} \\ &= E_{R|R>0}\left(\frac{\pi_0\alpha}{\pi}\right) = \frac{\pi_0\alpha}{\pi}. \end{aligned}$$

Alternative: extending FDR to $R = 0$

Finally, define $E(V/R|R = 0) \equiv E(V/R|R = 1)$, then

$$\begin{aligned} E\left(\frac{V}{R}\right) &= E\left(\frac{V}{R}\middle| R > 0\right) P(R > 0) + \\ &\quad E\left(\frac{V}{R}\middle| R = 0\right) P(R = 0) \\ &= \frac{\pi_0\alpha}{\pi}. \end{aligned}$$

We shall use FDR in this sense.

Mixture model

We are interested in the association between a feature and an outcome variable, e.g. gene expression and survival.
Let H be the indicator of the hypothesis

$$\begin{aligned} H0 &\sim \text{no association (no DE)} \\ H1 &\sim \text{association (DE)} \end{aligned}$$

The proportion of null genes: $P(H0) = \pi_0$

Mixture model

Let Z be a test statistic and assume

$$\begin{aligned} Z|H0 &\sim F_0(z) \\ Z|H1 &\sim F_1(z) \end{aligned}$$

The observed statistics come from a mixture

$$F(z) = \pi_0 F_0(z) + (1-\pi_0)F_1(z).$$

For a critical value $c > 0$, reject H0 if $|z| > c$, so as a function of c:

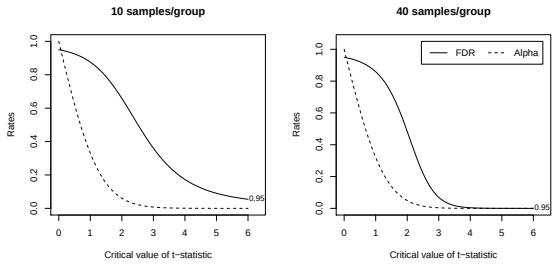
$$\begin{aligned} P(V_i = 1|H0) &= P_0(|Z| > c) = F_0(-c) + 1 - F_0(c) \\ P(R_i = 1) &= P(|Z| > c) = F(-c) + 1 - F(c) \\ \text{FDR}(c) &= \frac{\pi_0\{1 + F_0(-c) - F_0(c)\}}{1 + F(-c) - F(c)}. \end{aligned}$$

Mixture model: normal theory

Useful framework for theoretical understanding & experimentation:

- Comparing two groups, with n samples/group
- Data from normal distribution:
  - Group 1: $y \sim N(0, \sigma^2 = 1)$
  - Group 2, null genes: $y \sim N(0, \sigma^2 = 1)$
  - Group 2, regulated genes: $y \sim N(\pm D, \sigma^2 = 1)$ for DE genes
- E.g.: Proportion of DE = 5%, $D = 1$
- More general: vary σ and D as multiple of σ

Example: $\pi_0 = 0.95$ and $D = 1$



Estimation of FDR

For observed statistics $z_1, \ldots, z_m$, how can we estimate FDR(c)?
Given estimates of π_0, F_0 and F, compute empirical FDR as

$$\widehat{\text{FDR}}(c) = \frac{\widehat{\pi}_0\{1 + \widehat{F}_0(-c) - \widehat{F}_0(c)\}}{1 + \widehat{F}(-c) - \widehat{F}(c)}, \quad c > 0.$$

In practice, impose monotonicity in c by

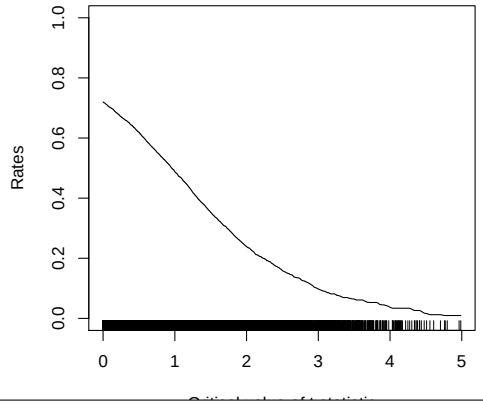
$$\widehat{\text{FDR}}(c) = \min_{0<z<c} \left\{ \frac{\widehat{\pi}_0\{1 + \widehat{F}_0(-z) - \widehat{F}_0(z)\}}{1 + \widehat{F}(-z) - \widehat{F}(z)} \right\}.$$

Multiple testing

Hypothesis testing
Multiple testing
Classical error control
Definitions of FDR
Mixture model
FDR Estimation
Estimating F_0
Estimating π_0
Local false discovery rate
Summary

Example: Breast cancer

Empirical FDR, $\pi_0 = 0.72$



Multiple testing

Hypothesis testing
Multiple testing
Classical error control
Definitions of FDR
Mixture model
FDR Estimation
Estimating F_0
Estimating π_0
Local false discovery rate
Summary

Estimation: P-value perspective

P-value from an observed z is

$$p = P_0(|Z| > |z|)$$

where the probability is computed under H_0.
Using p-values as test statistics with critical values α, we get:

$$\widehat{\text{FDR}}(\alpha) = \frac{\widehat{\pi}_0 \alpha}{\widehat{D}(\alpha)}, \quad 0 < \alpha < 1,$$

$\widehat{D}(\alpha)$ empirical distribution of p-values.

Multiple testing

Hypothesis testing
Multiple testing
Classical error control
Definitions of FDR
Mixture model
FDR Estimation
Estimating F_0
Estimating π_0
Local false discovery rate
Summary

Estimation: P-value perspective

From the data:

$$\widehat{D}(\alpha) = \frac{R(\alpha)}{m}$$

where $R(\alpha)$ is number of features with p-value $\leq \alpha$.
Therefore:

$$\widehat{\text{FDR}}(\alpha) = \frac{m\widehat{\pi}_0 \alpha}{R(\alpha)}$$

At $\alpha = p_{(k)}$ (observed p-values):

$$\widehat{D}(p_{(k)}) = k/m$$

Multiple testing

Hypothesis testing
Multiple testing
Classical error control
Definitions of FDR
Mixture model
FDR Estimation
Estimating F_0
Estimating π_0
Local false discovery rate
Summary

Estimation: P-value perspective

Together:

$$\widehat{\text{FDR}}(p_{(k)}) = \frac{m\widehat{\pi}_0 p_{(k)}}{k} \leq \frac{m p_{(k)}}{k}.$$

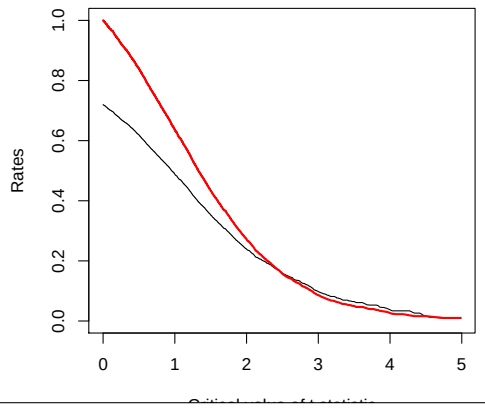
Monotone version:

$$\begin{aligned} \widehat{\text{FDR}}(p_{(k)}) &= \min_{\ell \geq k} \frac{m\widehat{\pi}_0 p_{(\ell)}}{\ell} \\ &\leq \min_{\ell \geq k} \frac{m p_{(\ell)}}{\ell} \quad = \text{BH adjustment} \end{aligned}$$

BH is conservative because it ignores π_0 (strong control of FDR).

Example: Breast cancer

Empirical FDR (black) vs B-H FDR (red)



The Key Theorem of FDR Estimation

Benjamini-Hochberg 1995, Storey 2002: essentially, the empirical FDR bounds the FDR on average, or:

$$\text{FDR}(\alpha) \leq \widehat{\text{FDR}}(\alpha)$$

Controlling $\widehat{\text{FDR}}(\alpha)$ makes sense, but it may not be possible to limit $\widehat{\text{FDR}}(\alpha)$ to an arbitrarily low value.

Multiple testing

Hypothesis testing
Multiple testing
Classical error control
Definitions of FDR
Mixture model
FDR Estimation
Estimating F_0
Estimating π_0
Local false discovery rate
Summary

Estimation: minor problem when $R = 0$

What is $\widehat{\text{FDR}}(\alpha)$ for $\alpha < p_{(1)}$ where $R(\alpha) = 0$?

Easiest: set to $\widehat{\text{FDR}}(p_{(1)})$

Alternative: set $R(\alpha) = 1$ and compute FDR using

$$Q = \frac{V}{\max(R, 1)}.$$

This creates negative bias, since

$$\begin{aligned} EQ &= P(R > 0) E\left(\frac{V}{\max(R,1)} \middle| R > 0 \right) + \\ &\quad P(R = 0) E\left(\frac{V}{\max(R,1)} \middle| R = 0 \right) \\ &= P(R > 0) \text{ FDR}. \end{aligned}$$

Multiple testing

Hypothesis testing
Multiple testing
Classical error control
Definitions of FDR
Mixture model
FDR Estimation
Estimating F_0
Estimating π_0
Local false discovery rate
Summary

Bias correction for minor problem?

We can apply a correction factor $1/P(R > 0)$, like Storey (2002):

$$\widehat{\text{pFDR}}(\alpha) = \frac{m\widehat{\pi}_0\alpha}{\max(R(\alpha), 1)P\{R(\alpha) > 0\}}$$

If m genes are independent, and α is probability of rejection for a null gene:

$$P(R > 0) \geq P(V > 0) = 1 - (1 - \alpha)^m$$

Multiple testing

Hypothesis testing
Multiple testing
Classical error control
Definitions of FDR
Mixture model
FDR Estimation
Estimating F_0
Estimating π_0
Local false discovery rate
Summary

Limited effect of bias correction

For large m and $\alpha \approx k/m$,

$$1 - (1 - \alpha)^m \approx 1 - e^{-k}.$$

For k=1 to 7:

```
0.632 0.865 0.950 0.982 0.993 0.998 0.999
```

The difference between $\widehat{\text{pFDR}}$ and $\widehat{\text{FDR}}$ only matters for the top 5 genes.

Multiple testing

Hypothesis testing
Multiple testing
Classical error control
Definitions of FDR
Mixture model
FDR Estimation
Estimating F_0
Estimating π_0
Local false discovery rate
Summary

Estimating F_0 through permutation

General procedure to generate null distribution when analytical solution may not exist. Old idea from R.A. Fisher in 1930s.

Example: comparison of two groups

group 1: $x_1, \ldots, x_{n_1}$

group 2: $y_1, \ldots, y_{n_2}$

Idea: under $H0$ of no difference, group labels can be permuted without changing the distribution of a test statistic Z.

Multiple testing

Hypothesis testing
Multiple testing
Classical error control
Definitions of FDR
Mixture model
FDR Estimation
Estimating F_0
Estimating π_0
Local false discovery rate
Summary

Estimating F_0: Permutation method

So:

- generate permuted data $x_1^*, \ldots, x_{n_1}^*, y_1^*, \ldots, y_{n_2}^*$
- compute Z^* from the permuted data
- repeat a large number of times p
- the collection of $z_1^*, \ldots, z_p^*$ is a sample from the null distribution of Z.

For example, for the observed z

$$\text{p-value} = \text{proportion of } |z^*|\text{'s} > |z|$$

Estimating F_0: Permutation method

For many genes: the collection might be gene-specific (as if done separately), or might be combined. Advantage of each?
Observed statistics from m genes: $z_1, \ldots, z_m$.
For $i = 1, \ldots, p$:

- Permute the data matrix
- Recompute statistics $z_{1i}, \ldots, z_{mi}$

The permutation p-value for gene g is

$$\text{p-value}_g = \text{proportion of } |z_{gi}|\text{'s over } i > |z|_g, \text{ for fixed } g.$$

Alternatively: mixing the genes

$$\text{p-value}_g = \text{proportion of } |z_{ji}|\text{'s over } i \text{ and } j > |z|_g.$$

Estimating F_0: Permutation method

What is the difference between the permutation and bootstrap methods?
How does the permutation deal with dependence between genes?
Other applications: how to permute?

- Paired data
- More than two groups
- Stratified data
- Censored survival data.

Estimating π_0

Traditionally based on the distribution of p-values: null hypotheses correspond to

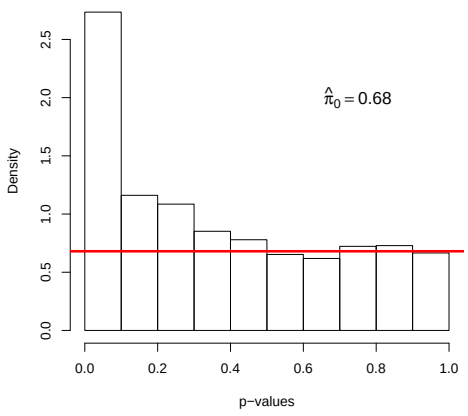
- large p-values,
- flat histogram.

Problems:

- correlated p-values,
- many features with small effects.

Many proposals – none entirely convincing
(e.g. Broberg 2005)

Example: a simple estimate for π_0



Local false discovery rate: fdr

For the mixture model: Let H be indicator of

$$\begin{aligned} H0 &\sim \text{no association, no DE} \\ H1 &\sim \text{association, DE} \end{aligned}$$

Let Z be a test statistic with conditional **densities**

$$\begin{aligned} Z|H0 &\sim f_0(z) \\ Z|H1 &\sim f_1(z) \end{aligned}$$

The observed statistics come from a mixture density

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z).$$

Local fdr: definition

What we want:

$$P(H0|Z = z).$$

This is the **local fdr**. Why?
Using Bayes's formula

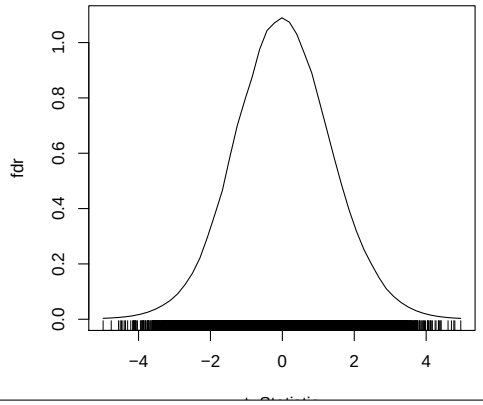
$$\begin{aligned} \text{fdr} &= P(H0|z) \\ &= \frac{P(H0)f(z|H0)}{f(z)} \\ &= \pi_0 \frac{f_0(z)}{f(z)}. \end{aligned}$$

Local fdr: estimation

- $f(z)$ from the observed $Z \sim f(z)$ with smoothing
- $f_0(z)$ from permutation
- π_0 as above (or not)

Example: Breast cancer

Local fdr (two-sided)



Local fdr vs global FDR

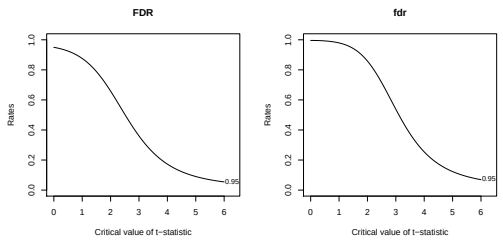
If we reject H0 for $|Z| > c$, then

$$\begin{aligned} \text{FDR}(c) &= \frac{\text{null genes with } |Z| > c}{\text{genes with } Z > z} \\ &= \pi_0 \frac{P(|Z| > c|H0)}{P(|Z| > c)} \end{aligned}$$

Compare with local fdr. Advantages and disadvantages?
Connection?

Example: Breast cancer

Global FDR vs local fdr



Local fdr vs global FDR

- global FDR: FDR(c) = 5% if we consider *all genes* with $|Z| \geq c$ significant
- global FDR does not apply to a gene, but to a collection of genes
- local fdr(c)= 0.05 applies to a gene with $Z = z$, it contributes 0.05 to the global FDR
- global FDR = Ave(local fdr)

Summary

1. Traditional multiplicity adjustments too conservative for high-dimensional data
2. FDR methods offer a good shot at controlled hypothesis discovery
3. FDR methods still under development
4. Ample software available

Further research

FDR methodology

- Proof of optimality? [Storey 2007, Perelman 2007]
- Correlated test statistics? [Pawitan 2006]

Applications

- Gene-gene correlation
- Genome-wide association studies
- Genomic-imbalances: copy-number variation, LOH map
- QTL mapping

People & Software

At Karolinska Institutet

Yudi Pawitan, Stefano Calza, Arief Gusnanto, Elena Perelman

At Institute Gustave Roussy, Paris

Stefan Michiels

At Genome Institute of Singapore

Lance Miller, Karaturi Krishna Murty

At `www.bioconductor.org`

`OCplus`, `multtest`