Outperforming the Optimal Discovery Procedure

Detecting differential expression in microarray data

Alexander Ploner[1] Elena Perelman[2] Stefano Calza[3] Yudi Pawitan[1]

[1]Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[2]iBusiness AB, Stockholm, Sweden

[3]Biomedical Sciences and Biotechnologies, University of Brescia, Italy

MCP 2007, July 9-11

Optimal Discovery Procedure (ODP)

Storey 2005; 2007

Consider $i = 1, \ldots m$ genes $\equiv$ hypotheses with

$$H_0^{(i)} : X_i \sim f_i \qquad \text{vs} \qquad H_1^{(i)} : X_i \sim g_i.$$

w.l.g.: H_0 for $i = 1, \ldots m_0$ and H_1 for $i = m_0 + 1, \ldots m$

Then the **optimal** common-cutoff procedure rejects $H_0^{(i)}$ for

$$S(\mathbf{x_i}) = \frac{g_{m_0+1}(\mathbf{x_i}) + \ldots + g_m(\mathbf{x_i})}{f_1(\mathbf{x_i}) + \ldots + f_{m_0}(\mathbf{x_i})} \geq \lambda$$

No procedure with the same number of expected false positives has more expected true positives.

Proof: Neyman-Pearson lemma $\square$

Estimating the optimal discovery procedure (EODP)

Storey et al. 2005; 2007

1. Estimate densities: parametric/normal, for 2-sample

$$\hat{f}_j(\mathbf{x_i}) = \phi(\mathbf{x_i}|\hat{\mu}_j, \hat{\sigma}^2_{j0}) \qquad \hat{g}_j(\mathbf{x_i}) = \phi(\mathbf{x_{i1}}|\hat{\mu}_{j1}, \hat{\sigma}^2_{j1})\phi(\mathbf{x_{i2}}|\hat{\mu}_{j1}, \hat{\sigma}^2_{j1})$$

2. Strong(er) control: modified statistic $S^*(\mathbf{x_i}) = 1 + S(\mathbf{x_i})$ as

$$S^*(\mathbf{x_i}) = \frac{f_1(\mathbf{x_i}) + \ldots + f_{m_0}(\mathbf{x_i}) + g_{m_0+1}(\mathbf{x_i}) + \ldots + g_m(\mathbf{x_i})}{f_1(\mathbf{x_i}) + \ldots + f_{m_0}(\mathbf{x_i})}$$

estimated as

$$\hat{S}^*(\mathbf{x_i}) = \frac{\sum_{j=1}^m \hat{g}_j(\mathbf{x_i})}{\sum_{j=1}^m w_j \hat{f}_j(\mathbf{x_i})}$$

E.g. $w_j \equiv 1$ with Bayesian argument or $w_j = 0/1$ based on prelim. p-values

3. Relate λ to actual FDR: conventional permutation/bootstrap approach for computing FDR(λ) as for any test statistic

Local fdr using standard errors (fdr2d)

Ploner et al. 2006

Efron 2001: test statistic Z addressing H_0 vs H_1 following

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

The local fdr is the posterior probability of H_0:

$$\text{fdr}(z) = P(H_0|Z = z) = \pi_0 \frac{f_0(z)}{f(z)}$$

This works for any vector-valued test statistic $\mathbf{Z}$, e.g. $\mathbf{Z} = (Z_1, Z_2)$:

$$\text{fdr2d}(z_1, z_2) = \pi_0 \frac{f_0(z_1, z_2)}{f(z_1, z_2)}.$$

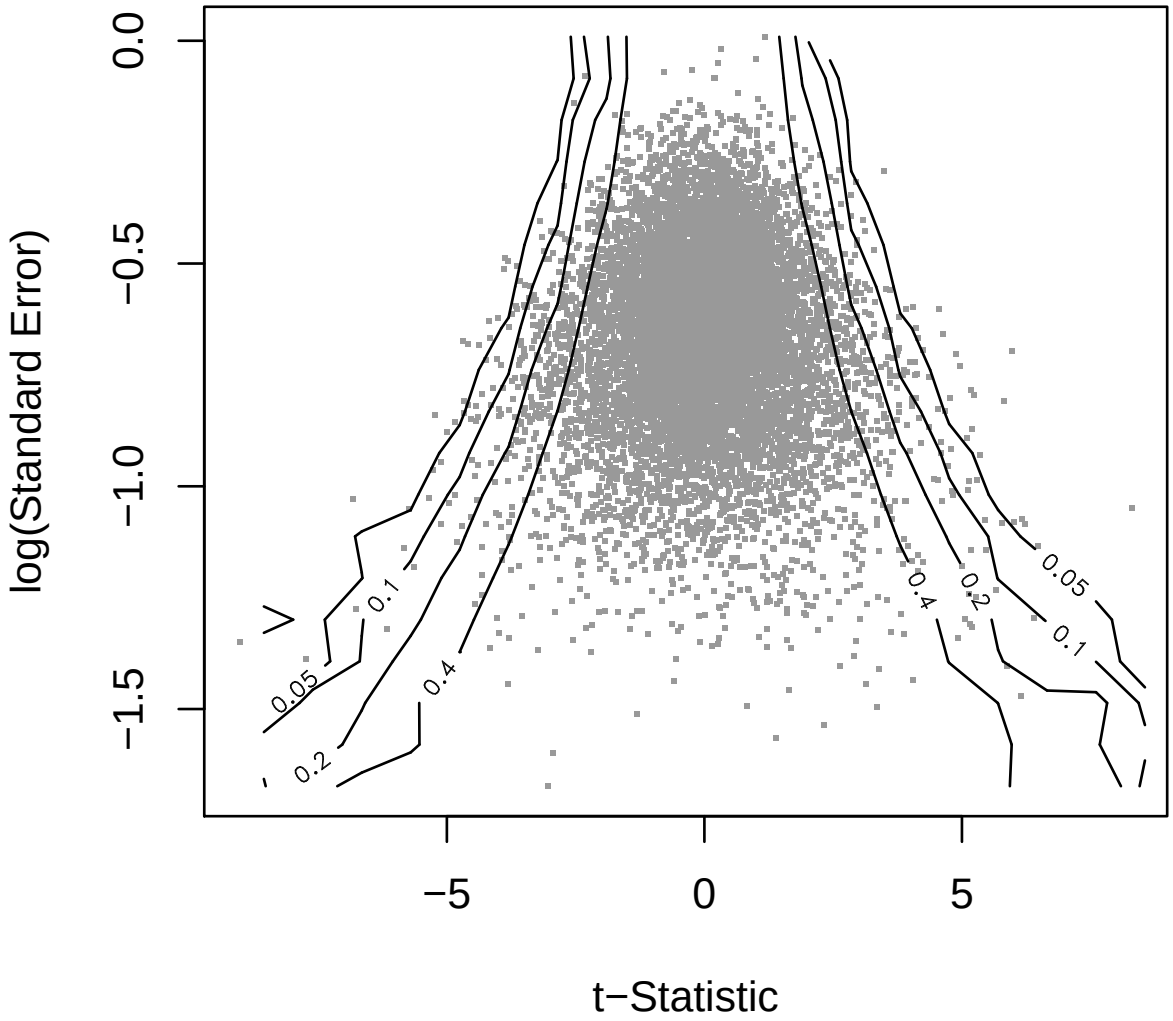
For the two-sample problem:

$$z_1^{(i)} = t_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{se_i} \qquad z_2^{(i)} = \log se_i$$

with conventional pooled standard error se_i.

Example: simulated data

$p = 10000$, $n_1 = n_2 = 7$, $\pi_0 = 0.80$, expression values $\sim N(0,1)$ and $N(\pm 1, 1)$

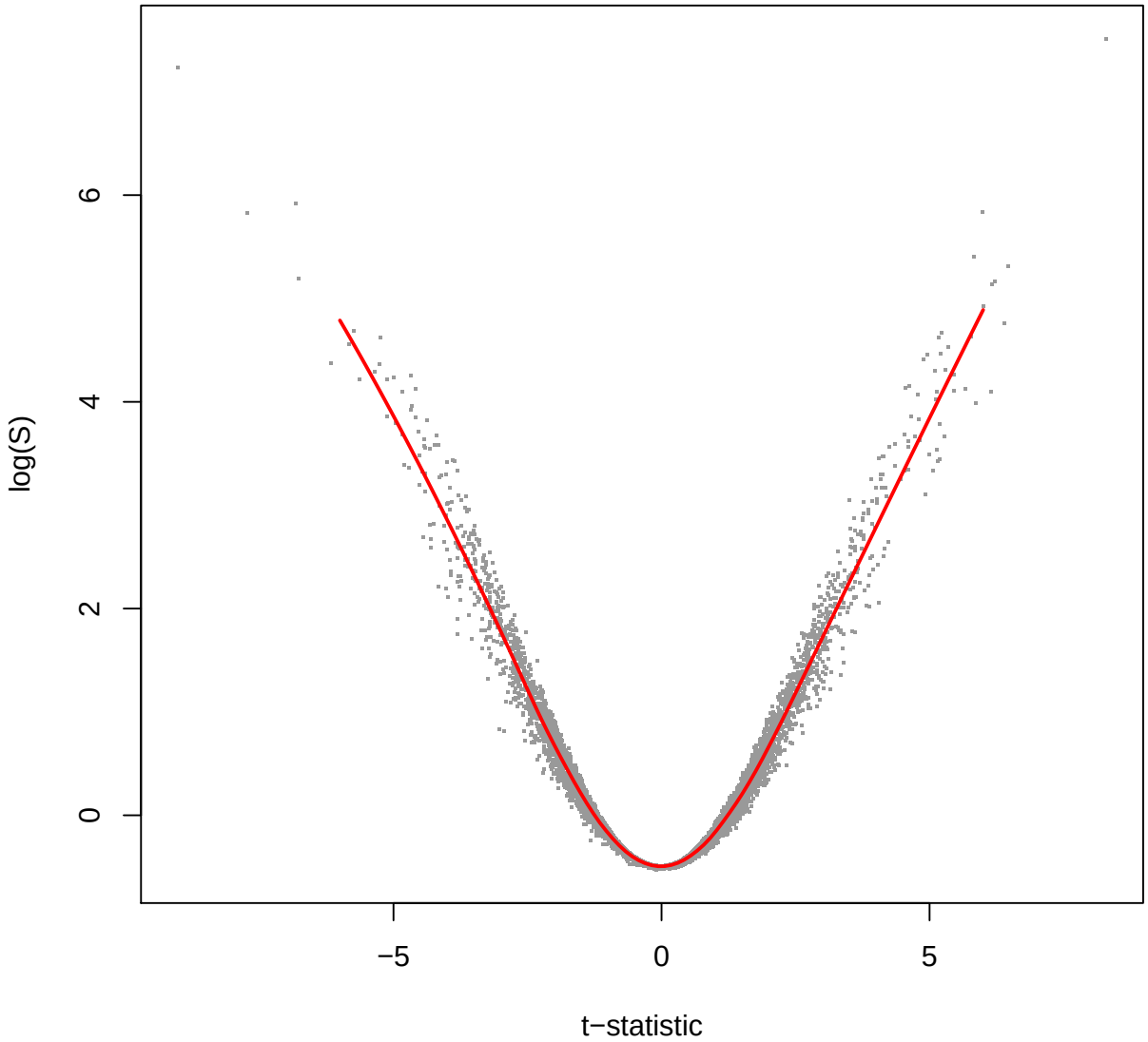


Comparing ODP and fdr2d

	ODP	**fdr2d**
Properties	Theoretical optimality	?
Assumptions	FDR-type	FDR-type
	Extra for optimality	
Computation	Expensive	Cheap
	Automatic	Smoothing parameters
Outperforms	Tusher et al. 2001 (SAM)	Tusher et al. 2001 (SAM – kind of)
	Efron et al. 2001 (fdr)	Efron et al. 2001 (fdr)
	Kerr et al. 2000	Smyth 2004
	Dudoit et al. 2002	
	Lönnstedt and Speed 2002	
	Cui et al. 2005	

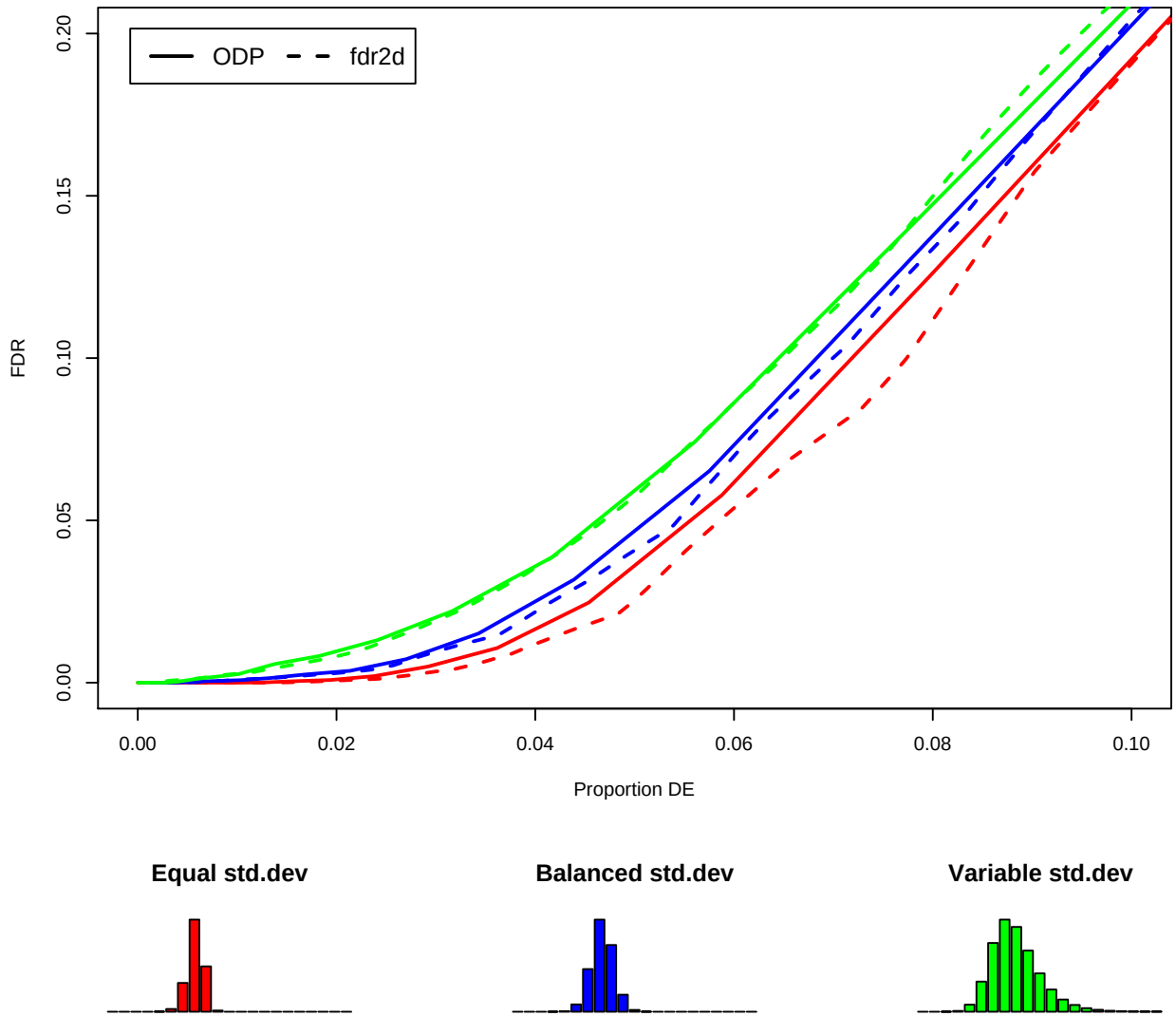
Comparing ODP and fdr2d: Test statistics

$p = 10000$, $n_1 = n_2 = 7$, $\pi_0 = 0.80$, expression values $\sim N(0, 1)$ and $N(\pm 1, 1)$



Comparing ODP and fdr2d: Performance

$p = 10000$, $n_1 = n_2 = 7$, $\pi_0 = 0.80$, expr. $\sim N(0, \sigma_i^2)$ and $N(\pm d \times \sigma_i, \sigma_i^2)$ (Smyth 2004)



S2d: Combining ODP and fdr2d

Perelman et al. 2007

fdr2*d* based on

$$\begin{aligned} Z_1 &= \log S(\mathbf{x_i}) \\ Z_2 &= \log se_i \end{aligned}$$

with conventional pooled standard error se_i for the difference of means.

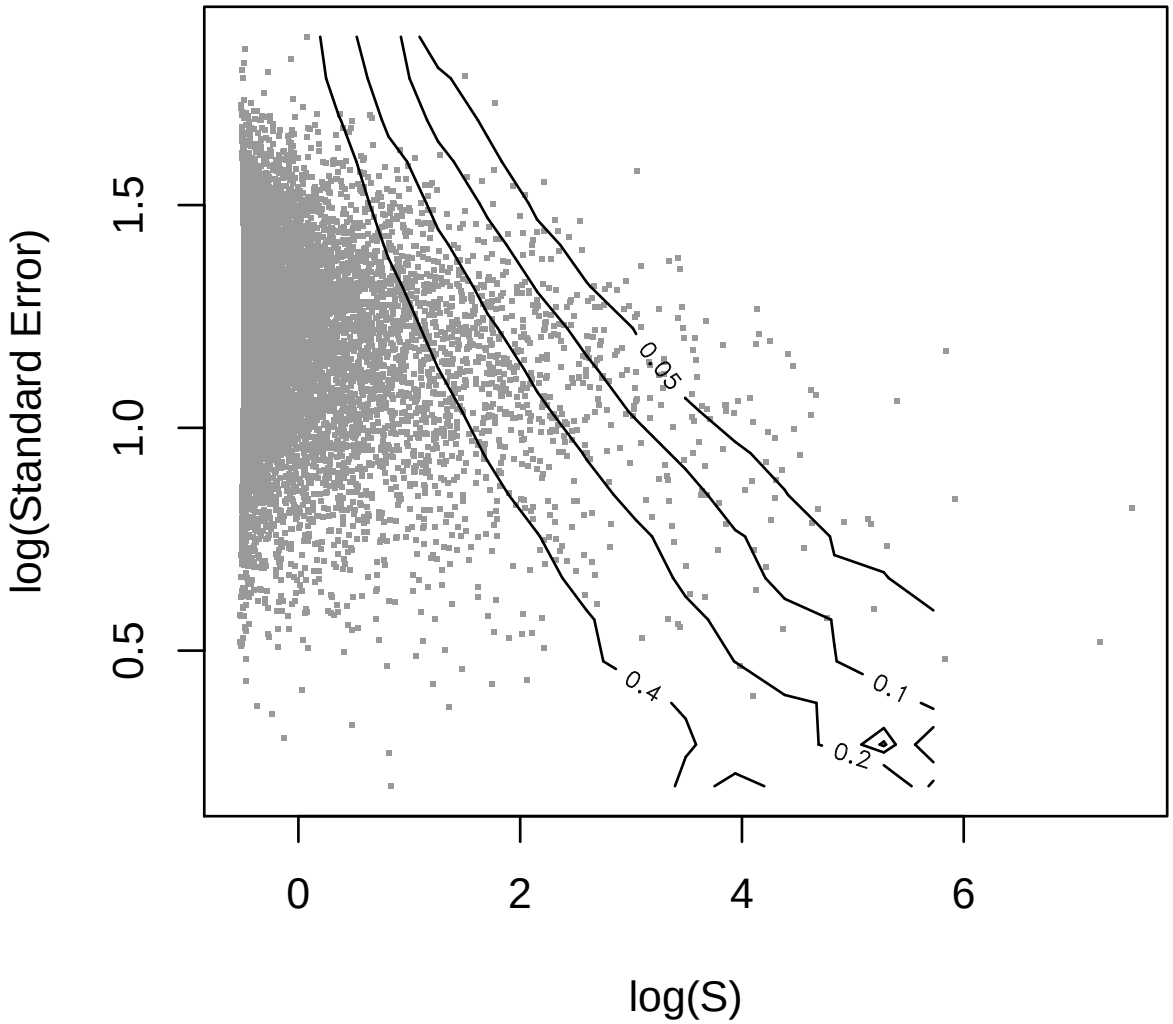
WTF?!

What's the motivation?

Maybe as idea: fdr2d transformation invariant – log(S) and t close

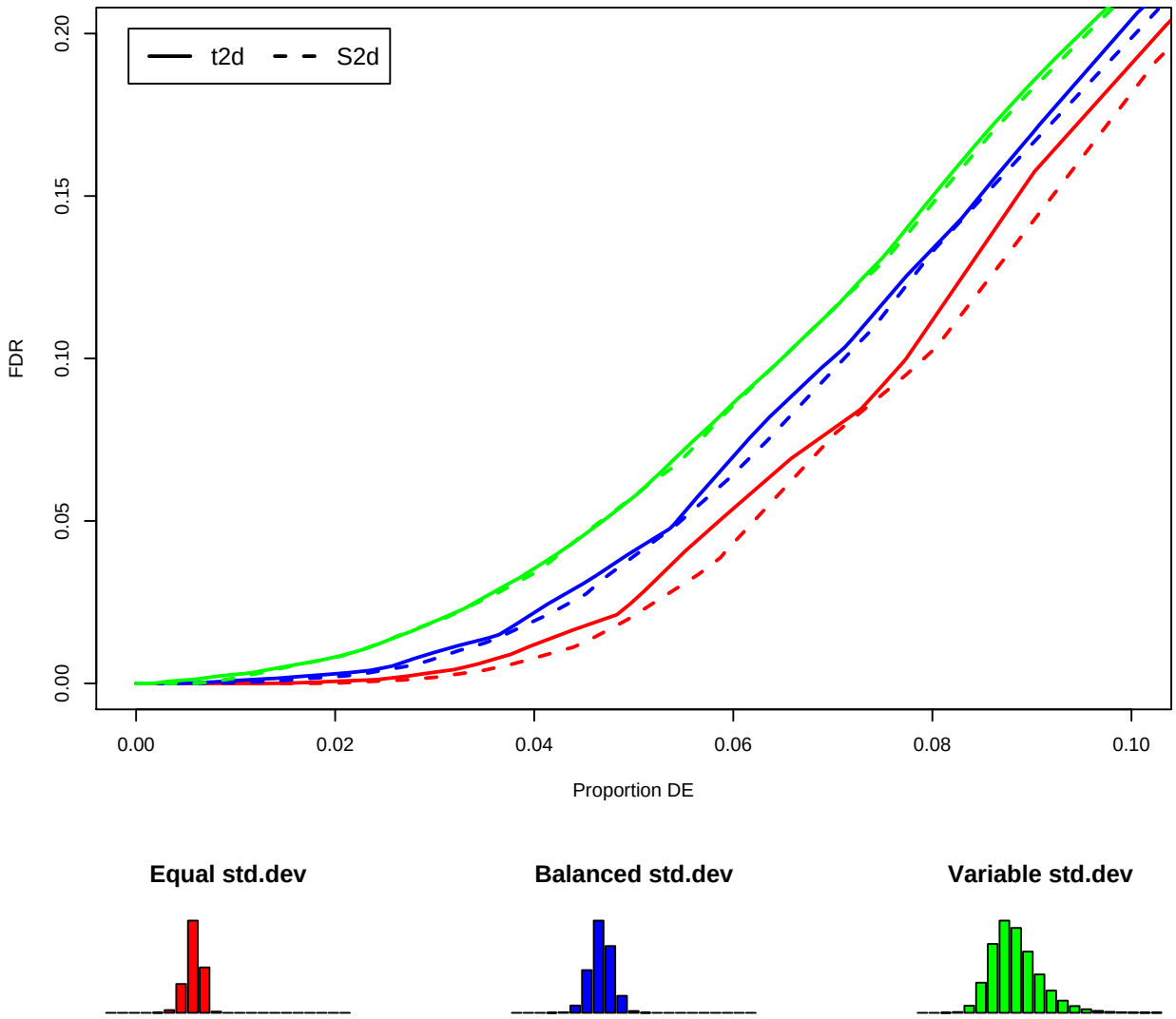
S2d: Proof of principle

$p = 10000$, $n_1 = n_2 = 7$, $\pi_0 = 0.80$, expression values $\sim N(0,1)$ and $N(\pm 1, 1)$



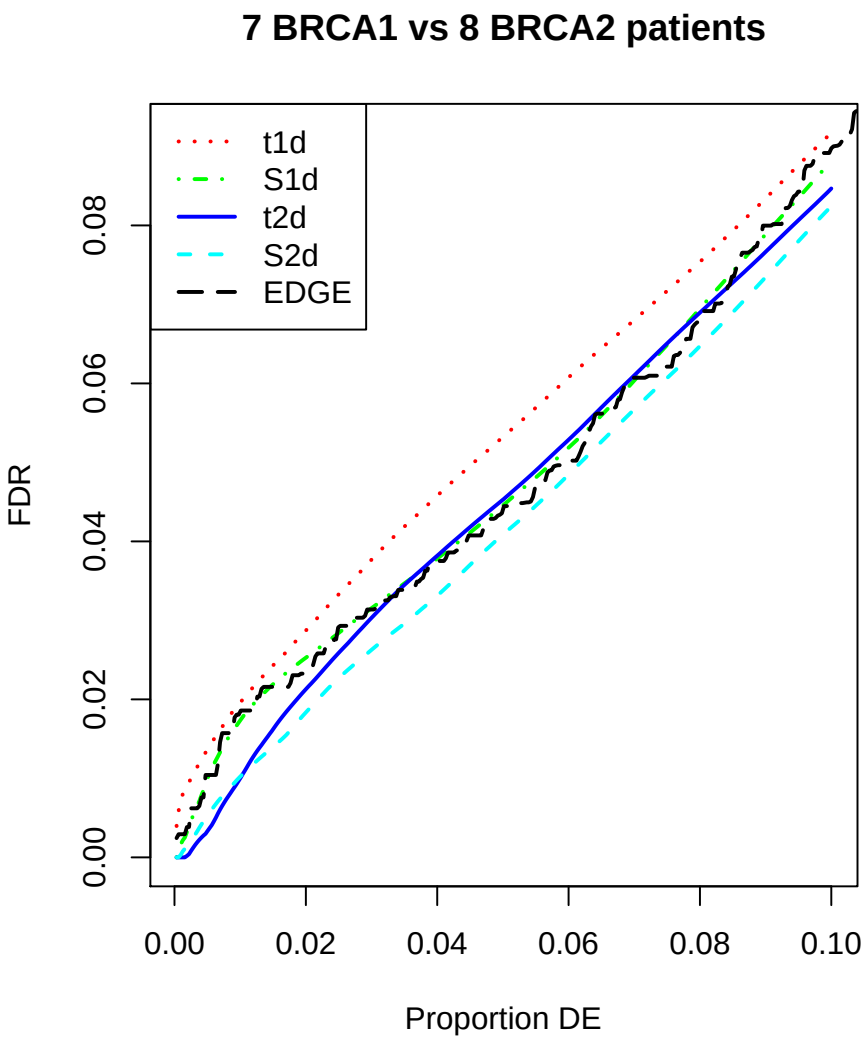
Comparing S2d and t2d for simulated data

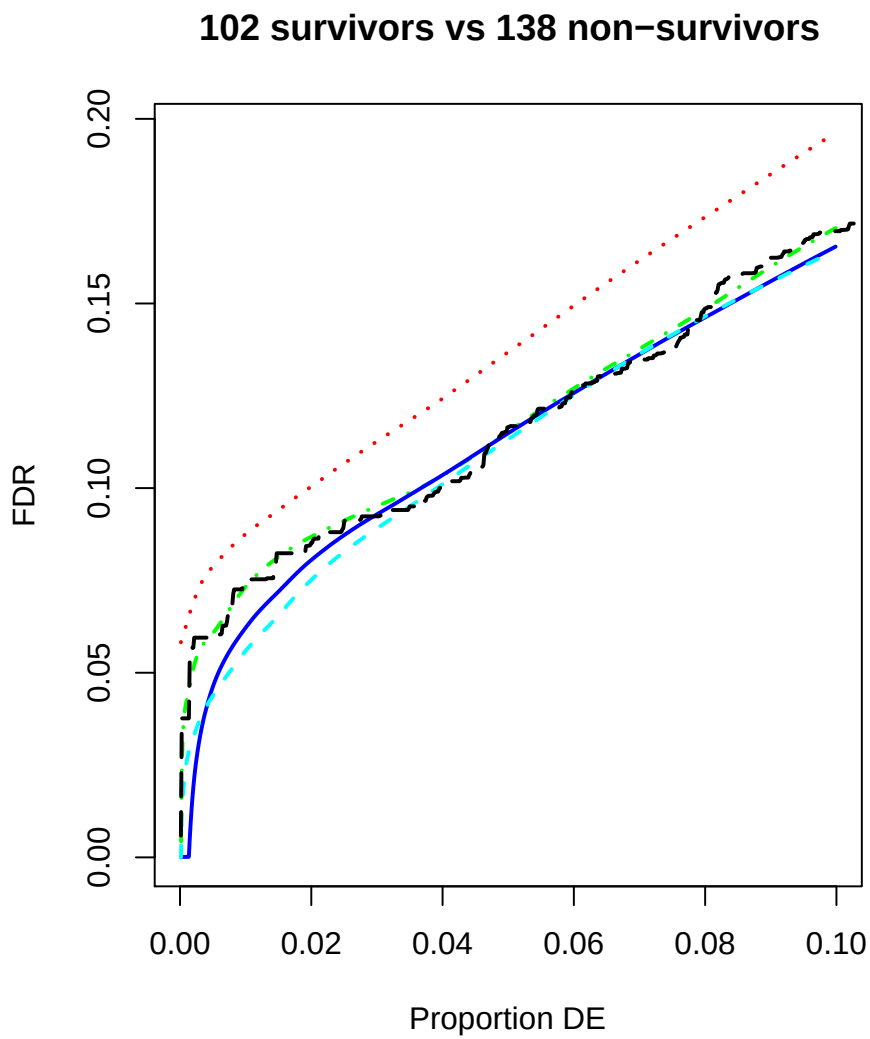
$p = 10000$, $n_1 = n_2 = 7$, $\pi_0 = 0.80$, expr. $\sim N(0, \sigma_i^2)$ and $N(\pm d \times \sigma_i, \sigma_i^2)$ (Smyth 2004)



Comparing S2d and t2d for real data

Hedenfalk et al. 2001 (breast cancer); Rosenwald et al. 2002 (lymphoma)





Summary

1. Translating theoretical optimality into practice is hard
2. The ODP test statistic is
   - powerful even out of context,
   - slow-ish to compute.
3. In a normal distribution setting, fdr2d with standard errors uses only genes with similar densities as reference set.
4. Pooling information across genes works best if the gene distributions are similar.
5. Availability:
   - t2d/t1d in Bioconductor R package `OCplus`
   - S2d/S1d as R code at `http://www.meb.ki.se/~aleplo`