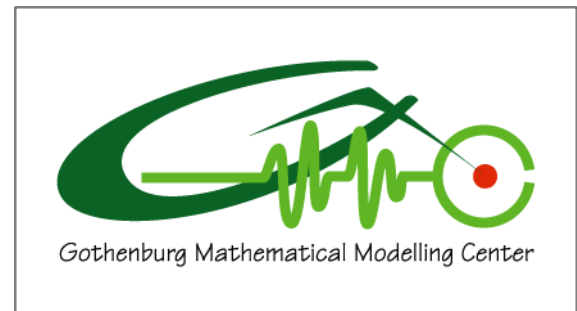
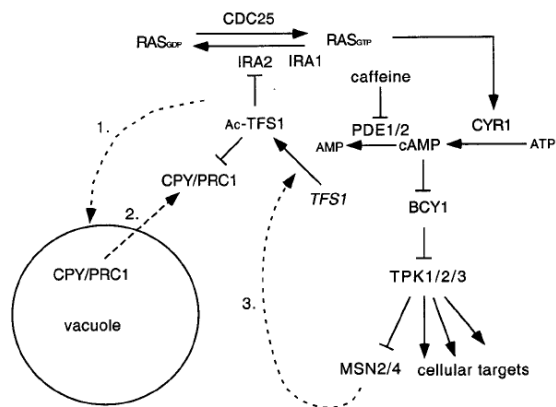


Tail estimation for false positives in high-throughput testing

Holger Rootzén & Dmitrii Zholud
GMMC & Stochastic Centre
Chalmers & Gothenburg University

www.math.chalmers.se/~rootzen/





Yeast
understanding basic
life functions

11,904 p-values

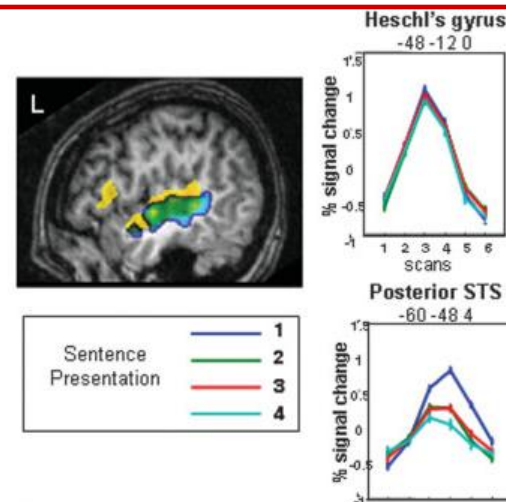
*Blomberg et al. 2003,
2010*



Arabidopsis Thaliana
association mapping

3,745 p-values

Zhao et al. 2007



fMRI brain scans
function of brain
language network

appr. 3 mill. p-values

Taylor et al. 2006



27 million p-values from study of interaction between pairs of genes in yeast, coming from 27 million experiments in 4 replications



... and this is the future




***how many of
the positive
test results are
false?***

***which pre-
processing
method is
best? *)***

*) typically true null
distribution is different from
the theoretical one

answers

- conditional distribution of $\#\{\text{false positives}\}$, given that there are r positives, is approximately binomial
- methods to estimate the success probability parameter (=pFDR) of this binomial distribution 
- estimates of the true null distribution of p-values resulting from a pre-processing method, and techniques to compare it with the theoretical uniform distribution

conditional binomial distribution of the number of false positives

$m_0 = \#\{\text{true null hypotheses}\}$, $m_1 = \#\{\text{false null hypotheses}\}$

$\#\{\text{false positives}\}$ asymptotically Poisson with parameter $m_0 F_0(\alpha)$

$\#\{\text{true positives}\}$ asymptotically Poisson with parameter $m_1 F_1(\alpha)$



$(\#\{\text{false positives}\} \mid \#\{\text{positives}\} = r) \approx \text{Bin}(r, \text{pFDR})$

$$\text{pFDR} = \frac{m_0 F_0(\alpha)}{m_0 F_0(\alpha) + m_1 F_1(\alpha)} \approx \frac{\pi_0 F_0(\alpha)}{F(\alpha)}$$



power model for distribution of p-values

$$F_0(x) = \Pr(P \leq x | H_0) = c_0 x^{1/\gamma_0} \quad 0 \leq x \leq u$$

$$F_1(x) = \Pr(P \leq x | H_1) = c_1 x^{1/\gamma_1} \quad 0 \leq x \leq u$$

$$F(x) = \Pr(P \leq x) = \pi_0 c_0 x^{1/\gamma_0} + \pi_1 c_1 x^{1/\gamma_1} \approx c x^{1/\gamma} \quad 0 \leq x \leq u$$



α *critical value*: null hypothesis "rejected" if p -value $\leq \alpha$

u *threshold*: model assumed to hold for $x \leq u \ll 1$
only p-values less than u are used for estimation

- only trust model for $x \leq u$, where u is chosen from data
- however u can be often be chosen much bigger than α
→ model-based estimates of $F_0(\alpha)$ and $F(\alpha)$ more accurate than empirical estimates

estimation of F (the TuT-method)

p_1, \dots, p_n observed p-values, $N = \#\{p_i \leq u\}$

- $\Pr(-\log(P/u) \leq x \mid P \leq u) = \frac{c(ue^{-x})^{1/\gamma}}{cu^{1/\gamma}} = e^{-x/\gamma}$
- $\Pr(P \leq \alpha) = \Pr(P \leq u)\Pr(P \leq \alpha \mid P \leq u)$



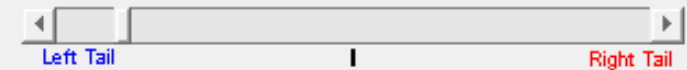
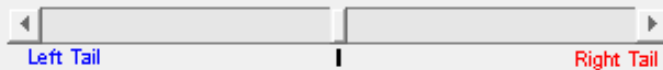
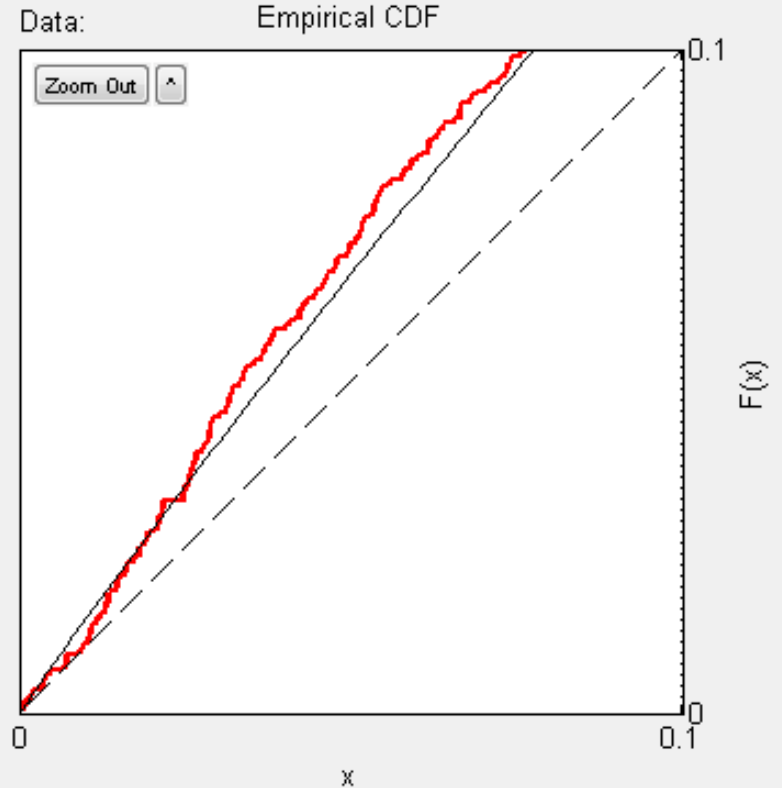
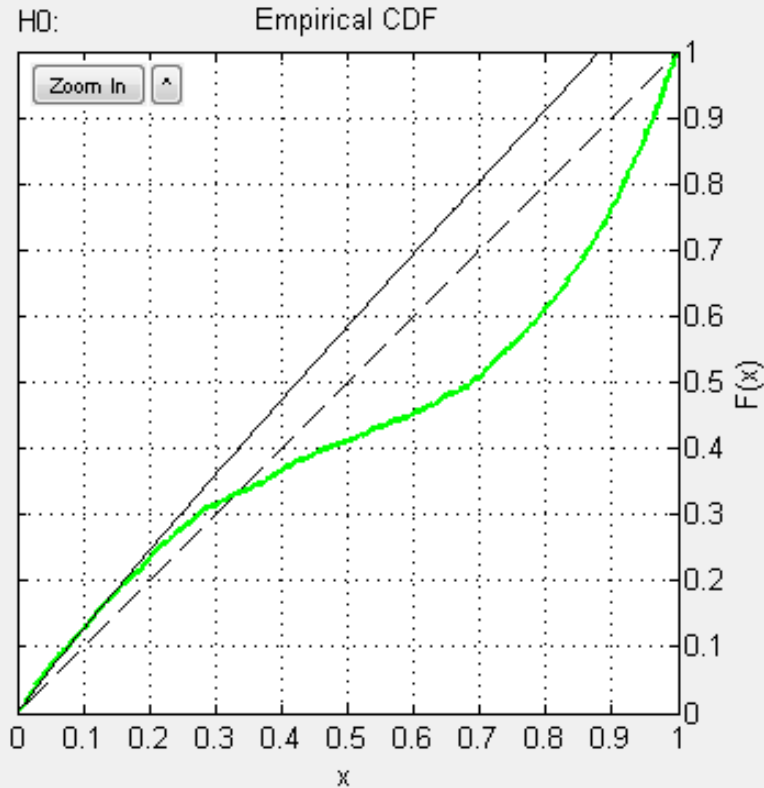
$$\hat{\gamma} = \frac{1}{N} \sum_{p_i \leq u} -\log(p_i/u), \quad \hat{F}(\alpha) = \frac{N}{n} \left(\frac{\alpha}{u}\right)^{1/\hat{\gamma}}$$

this together with “the same” estimator for F_0 , and “guesses” of π_0 directly gives estimates of

- FDR
- pFDR
- $\text{fdr}(x)$
- FWER
- k-FWER



SmartTail



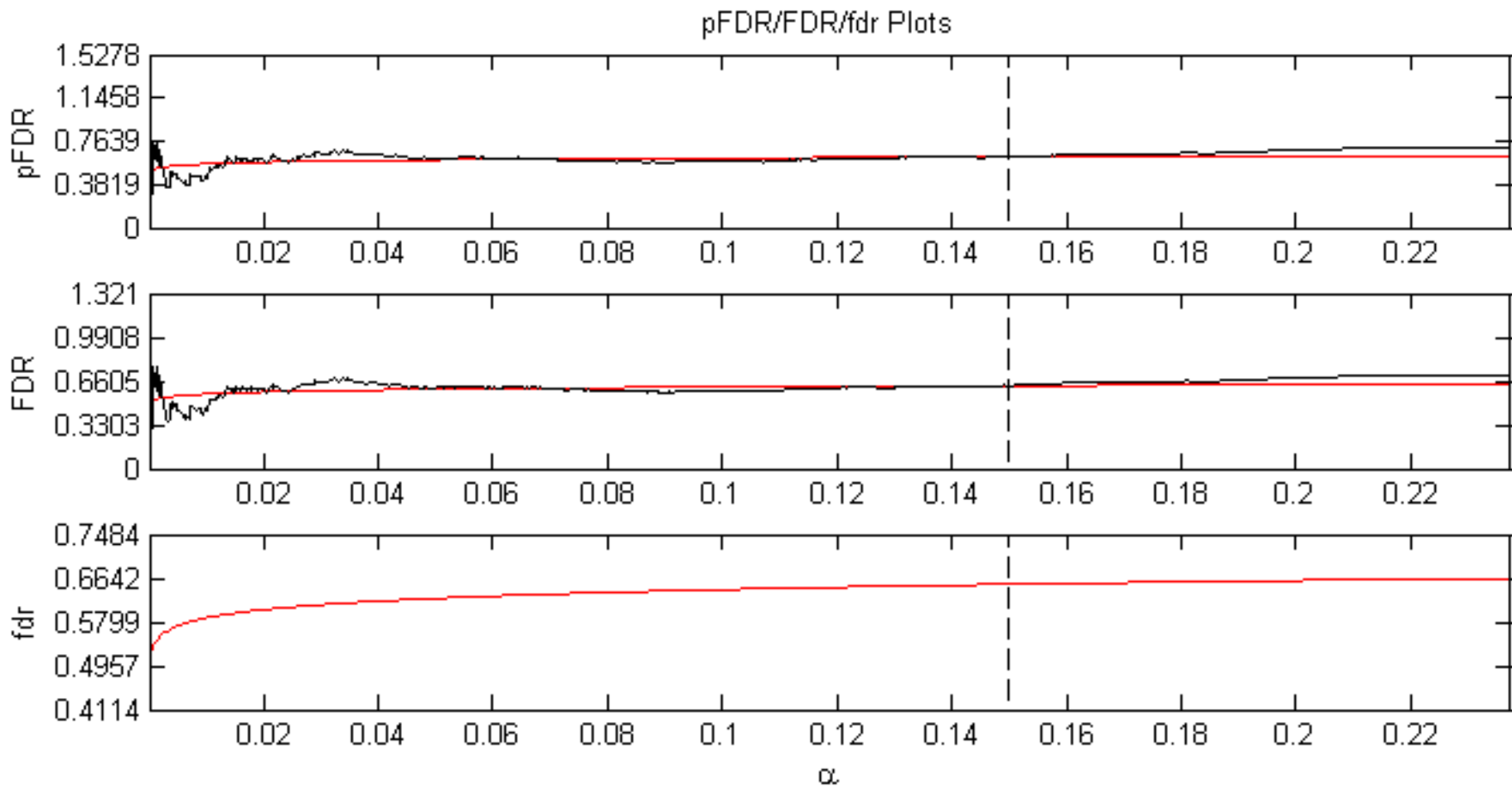
Statistics:

	Model	u	C	a	$\#\{x \text{ less/above } u\}$	y	$\hat{F}(y)$	CI
H0:	$F(x)=cx^a, x < u.$	0.100000	1.1259	0.9466	220 / 1508	<input type="text" value="0.01"/>	-	-
Data:	$F(x)=cx^a, x < u.$	0.100000	1.1259	0.9466	220 / 1508	<input type="text" value="0.01"/>	-	-

Bioscreen testing for gene expression in yeast: wildtype data

Binomial distribution of $\#\{\text{false positives}\}$, given $\#\{\text{positives}\}$,
 prob. parameter = pFDR

$$\text{pFDR} = \frac{\pi_0 F_0(\alpha)}{F(\alpha)}, \quad \text{fdr}(x) = \frac{\pi_0 \frac{d}{dx} F_0(x)}{\frac{d}{dx} F(x)}$$



motivation for power model

G_0 and G_h true and hypothetical null cdf-s of test statistic, respectively

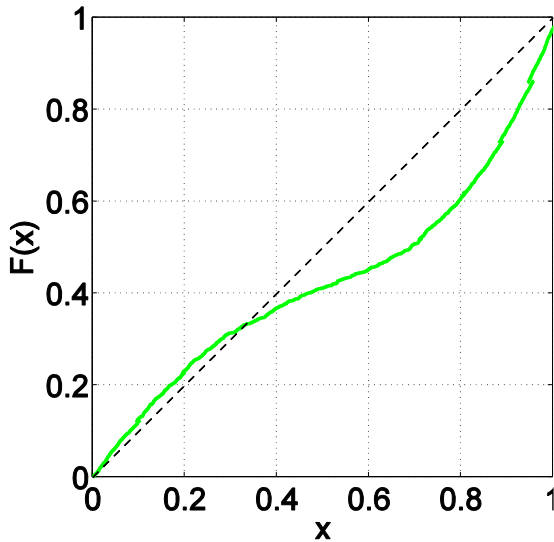
$$\bar{G}_0(x) = 1 - G_0(x) \approx C_0 \frac{1}{x^{1/\bar{\gamma}_0}} \quad \text{and} \quad \bar{G}_h(x) \approx C_h \frac{1}{x^{1/\bar{\gamma}_h}}, \quad x \downarrow 0$$



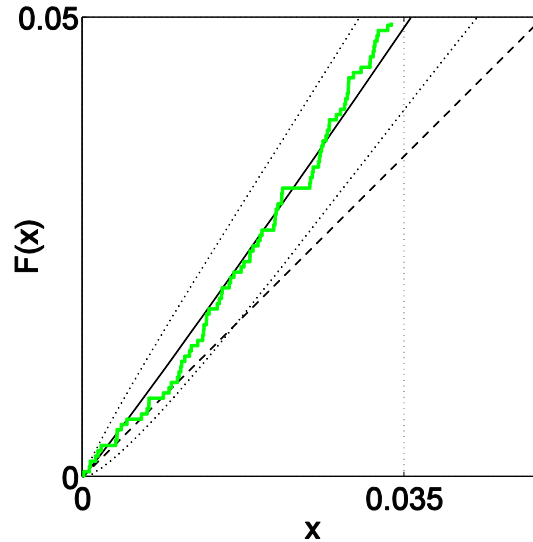
$$F_0(x) = G_0(G_h^{\leftarrow}(x)) \approx C_0 / ((x/C_h)^{-\bar{\gamma}_h})^{1/\bar{\gamma}_0} =: c_0 x^{\gamma_0}, \quad x \downarrow 0$$

- true for t- and F-statistics (talk by Zholud), more general motivation via extreme value theory
- same for F_1

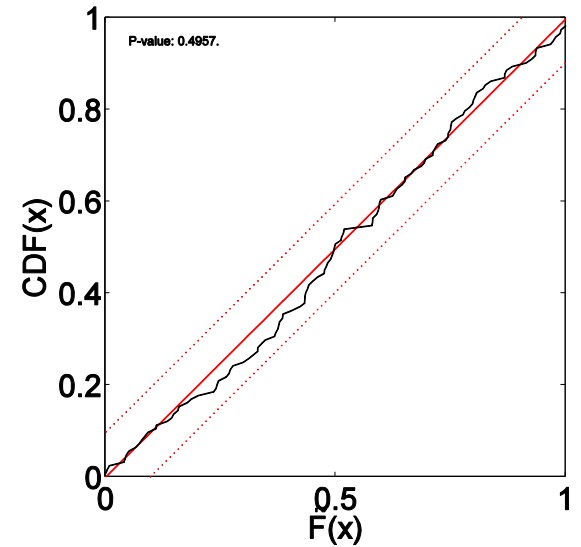
yeast: p-values for wildtype data set



empirical cdf of p-values (1728 p-values)

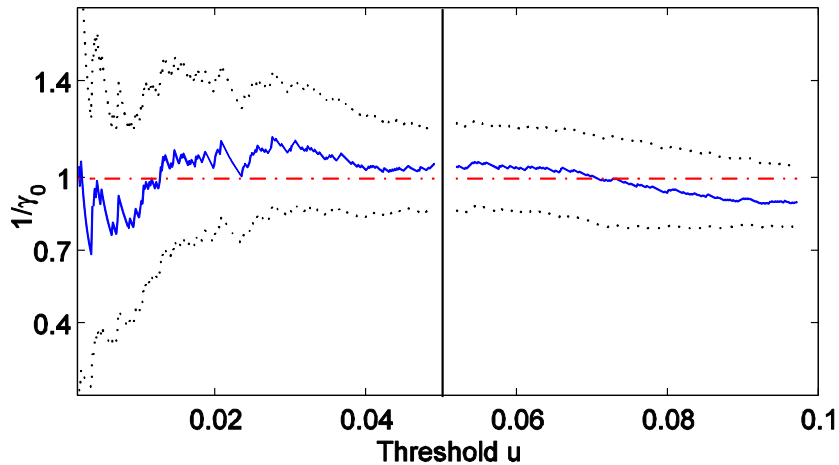


empirical cdf of p-values for $p < 0.05$, with fitted model

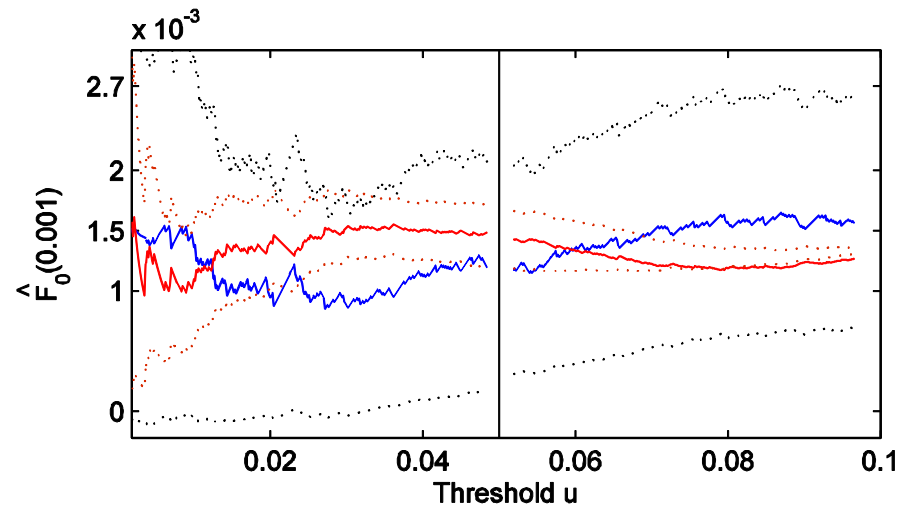


Kolmogorov - Smirnov 95% goodness of fit test ($p=0.49$)

yeast: choice of threshold u for wildtype data ("compromise between bias and varaince")

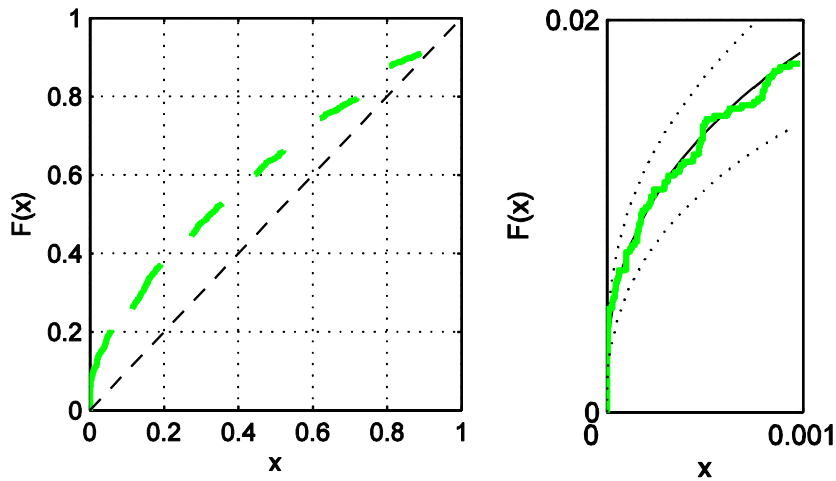


estimate of $1/\gamma_0$ as function
of threshold u

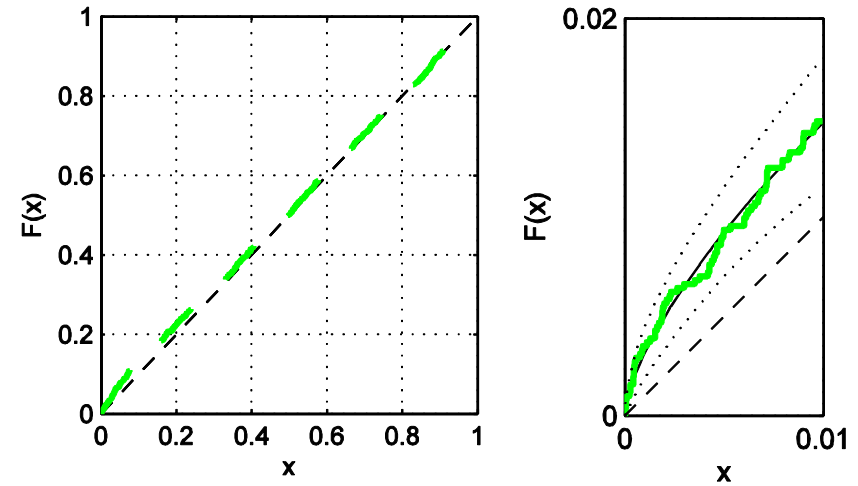


estimate of $F_0(\alpha)$ as function
of threshold u , for $\alpha = 0.001$
red line: estimate for γ_0 equal
to 1

Arabidopsis: comparing preprocessing methods

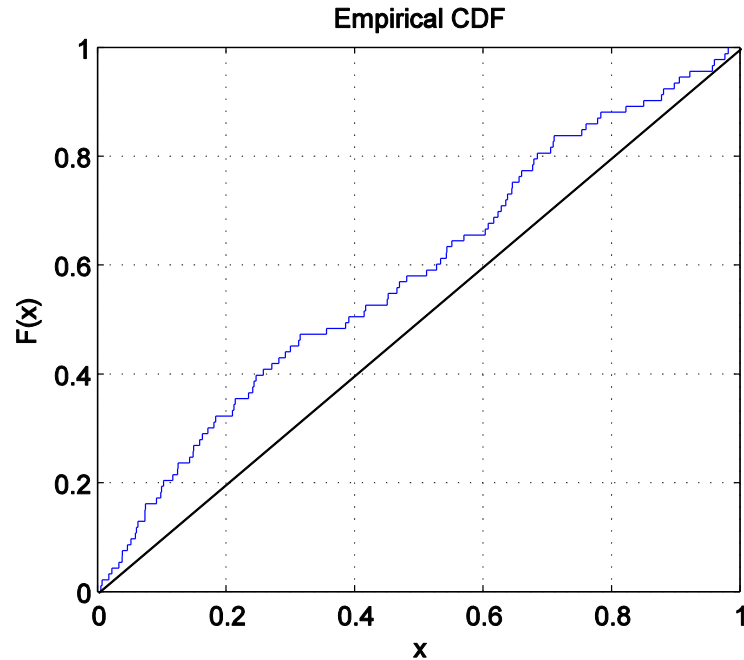


empirical cdf of p-values for
Kruskal-Wallis test (3745 p-
values)



empirical cdf of p-values for
Q+K test which corrects for
population structure

fMRI brain scans: accuracy of power model



Empirical cdf of 93 p-values from Kolmogorov-Smirnov goodness of fit tests of polynomial model, for 93 brain scans with each leading to appr. 35,000 p-values. The threshold was $u = 0.01$.

**False Positives in high-throughput testing →
it is tails (and not centers) of distributions
of p-values which matter!**

**Extreme Value Statistics is *the* instrument
for looking at tails.**



Clarke, S. and Hall, P. (2009). Robustness of multiple testing procedures against dependence. *Ann. Statist.* **37**, 332–358.

Knijnenburg, T. A., Wessels, L. F. A., Reinders, J. T. M., and Shmulevich, I. (2009). Fewer permutations, more accurate P-values. *Bioinformatics* **25**, 161-168.

Rootzén, H. and Zholud, D. (2010). Tail estimation methods for the number of false positives in high-throughput testing. *Submitted*.

Storey, J.D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B* **64**, 479-498.

Taylor, J. E. and Worsley, K. J. (2006). Inference for magnitudes and delays of response in the FIAC data using BRAINSTAT/FMRISTAT. *Human Brain Mapping* **27**, 434-41.

Zhao K., Aranzana M. J., Kim S, Lister C., Shindo C., et al. (2007). An Arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**, 71-82.

Zholud, D. (2010). Extremes of Student's one- and two-sample t-test, Welch's test, and F-test for non-normal and not necessarily i.i.d. random variables, *To be submitted*.

**Out of 10 mathematicians there is always 1 who knows the
1 binary number system and 1 who doesn't**