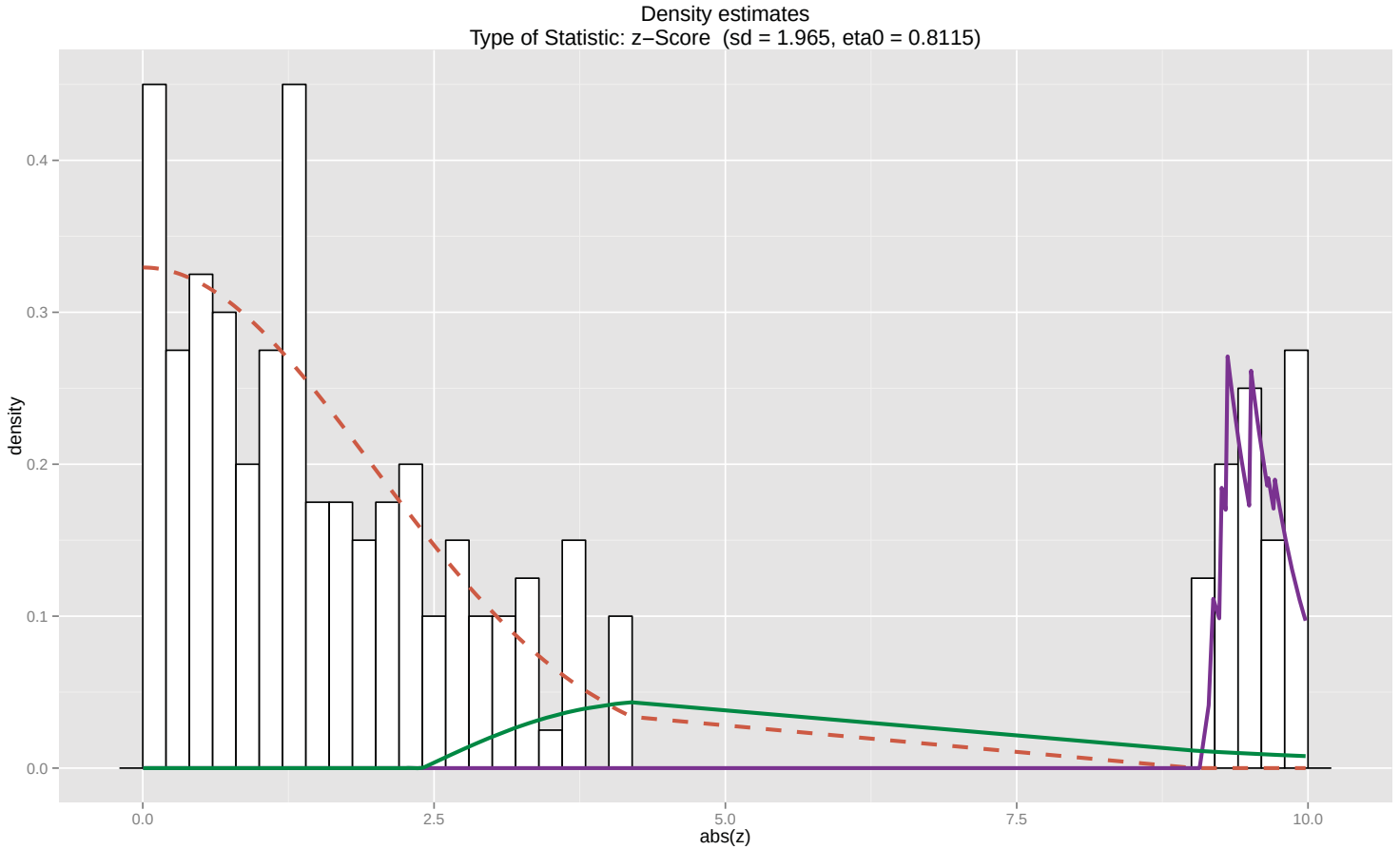
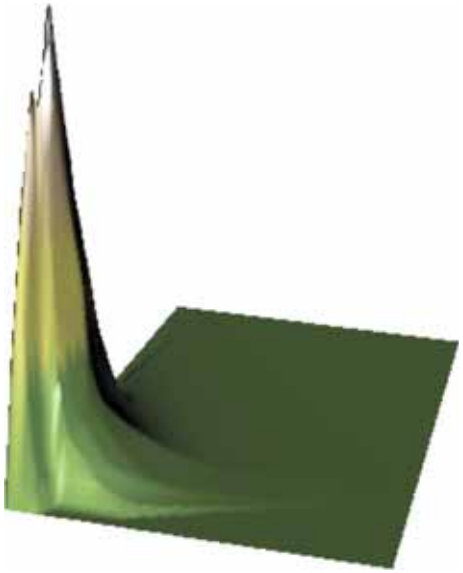
FDR estimation using log concave densities





Bernd Klaus
EMBL

27 August 2014 - MAS - Toulouse

European Molecular Biology Laboratory (EMBL)

European Intergovernmental Research Organisation

- **21 Member States**
- **Founded in 1974**
- **Sites in Heidelberg (D), Cambridge (GB), Roma (I), Grenoble (F), Hamburg (D)**
- **ca. 1400 staff (⊃1100 scientists) representing more than 60 nationalities**

EMBL’s five missions

- **Basic research**
- **Development of new technologies and instruments**
- **Technology transfer**
- **Services to the member states**
- **Advanced training**

What can *you* do at EMBL?

Biology

Chemistry

Physics

Mathematics

Informatics

Engineering

www.embl.org/phdprogramme
www.embl.org/postdocs
www.embl.org/jobs

Empirical Bayes FDR Analysis

- **Focus of the BH procedure:** Control of the false discovery rate
- High dimensional data allows to estimate rather than only to control the false discovery rate!

False Discovery Rate (FDR) Estimation

- **Ultimate goal of multiple testing**: Separation of signal (e.g. differentially expressed genes) from noise (e.g. inactive genes)
- $\Rightarrow$ **decision making is very difficult overlapping signals**

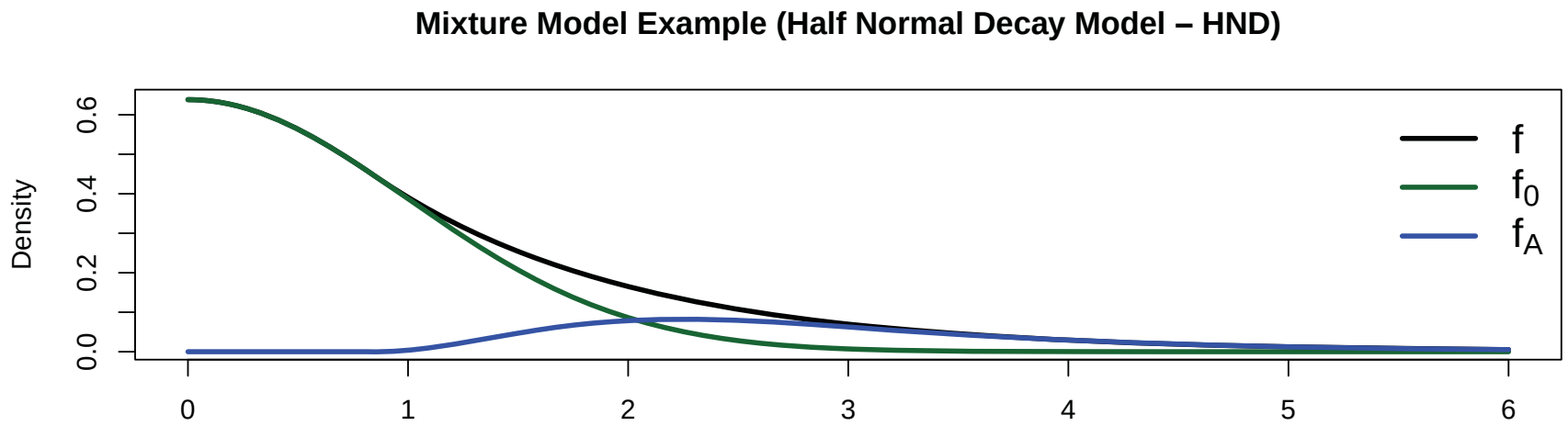
Rather than controlling the FDR we can estimate the FDR

- After the fit of a mixture model with a **"null" component** (noise) and an **"alternative" component** (signal), false discovery rates allow intuitive and simple signal identification
- $\Rightarrow$ High Dimension = Blessing
- Classical mixture model methods such as EM algorithms don't work reliably here

A Mixture Model for FDR Estimation

Mixture Model for Densities / Distributions

$$\{f(y); F(y)\} = \eta_0\{f_0(y); F_0(y)\} + (1 - \eta_0)\{f_A(y); F_A(y)\}$$



- y: **generalized test statistic**, large values indicate "interesting" case: absolute z-scores $|z|$, or $1 - p$, i.e the complement of p-values ...
- **Null model** f_0 / F_0 and **alternative component** f_A / F_A $\Rightarrow$ "interesting" cases
- η_0: **proportion of null features**

FDR Definitions

Via Bayes' formula:

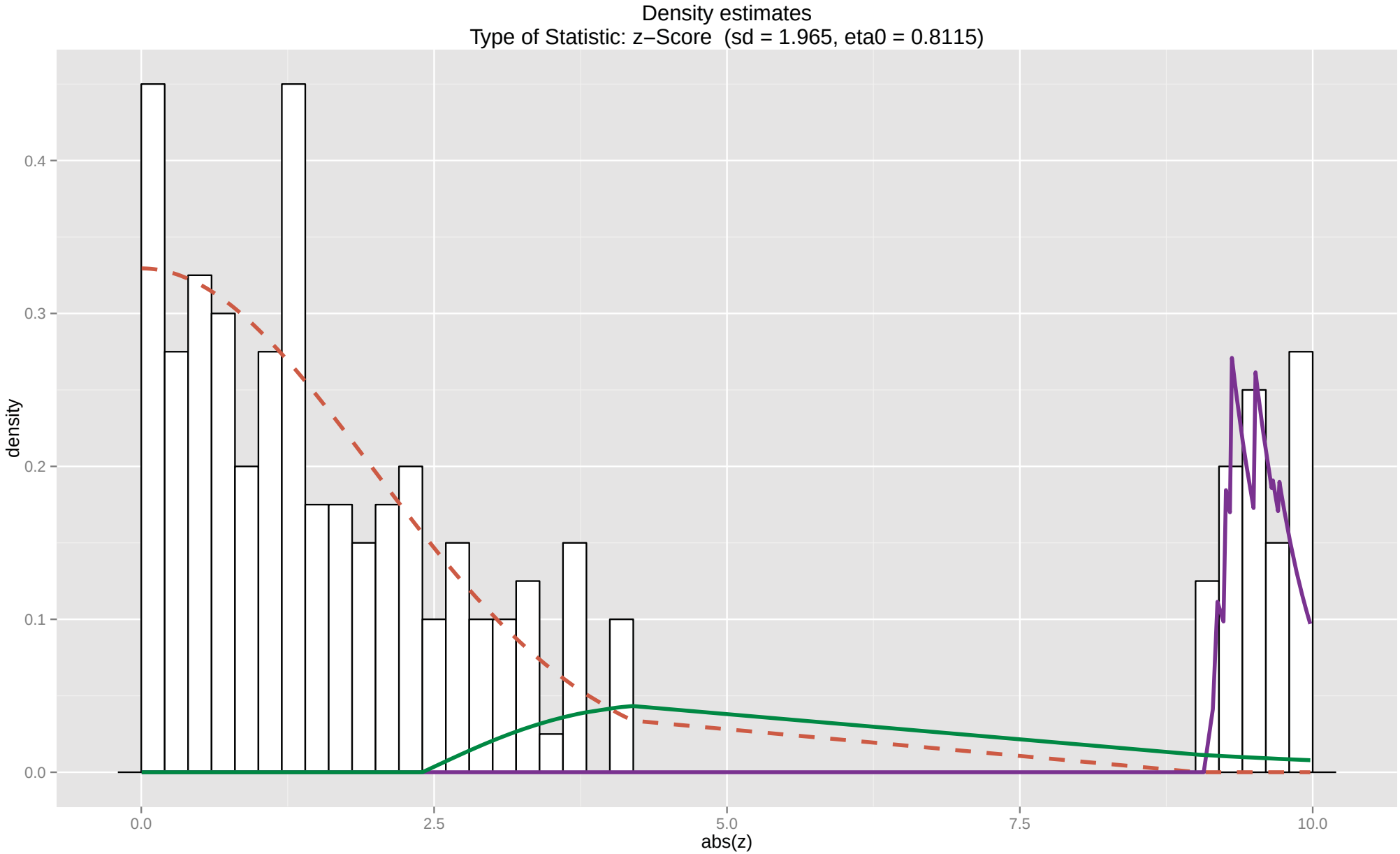
- **local FDR = fdr**:
  $\text{fdr}(y) = \text{Prob}(\text{“not interesting”} \mid Y = y) = \eta_0 \frac{f_0(y)}{f(y)}$
- **tail-area-based FDR = Fdr** (q-value):
  $\text{Fdr}(y) = \text{Prob}(\text{“not interesting”} \mid Y \geq y) = \eta_0 \frac{1-F_0(y)}{1-F(y)}$

FDR = local or tail–area based false discovery rates

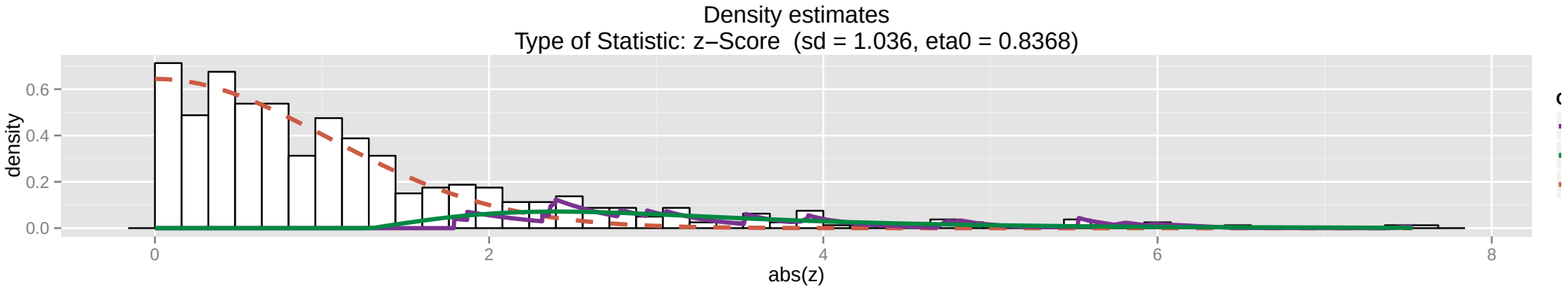
Common cutoff values for identifying differentially expressed genes) are $\text{Fdr}(y) < 0.05$ or $\text{fdr}(y) < 0.2$

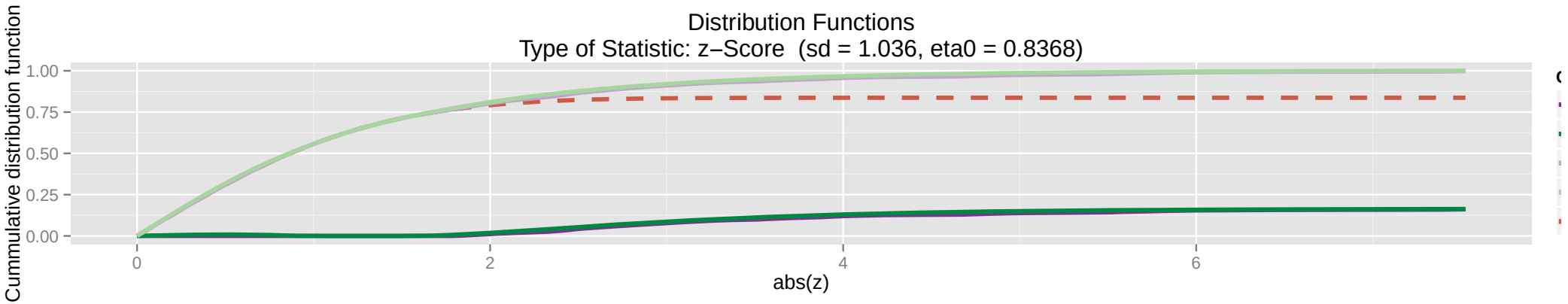
The BH rule corresponds to choosing $\eta_0 = 1$, the **ecdf as estimator of** F and the **theoretical null model** as estimator for F_0

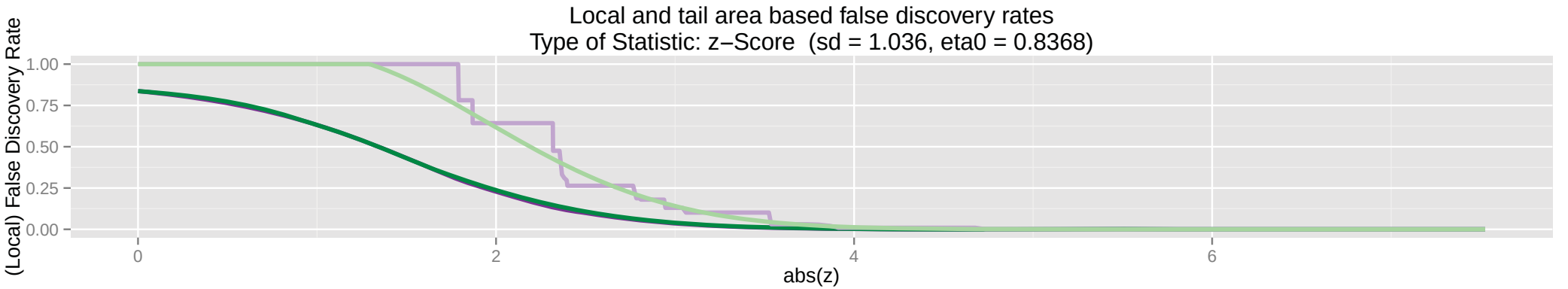
An easy two groups case



A hard two groups case

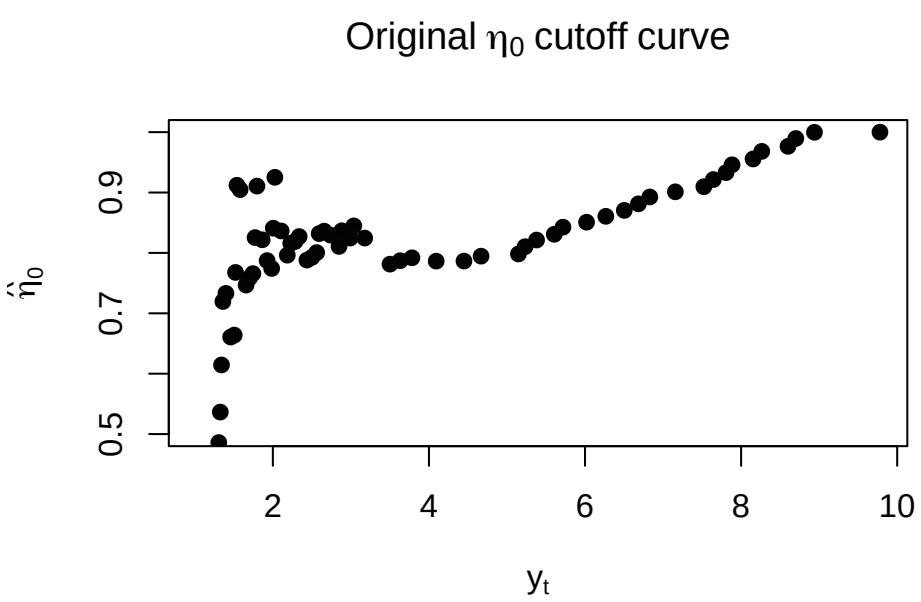


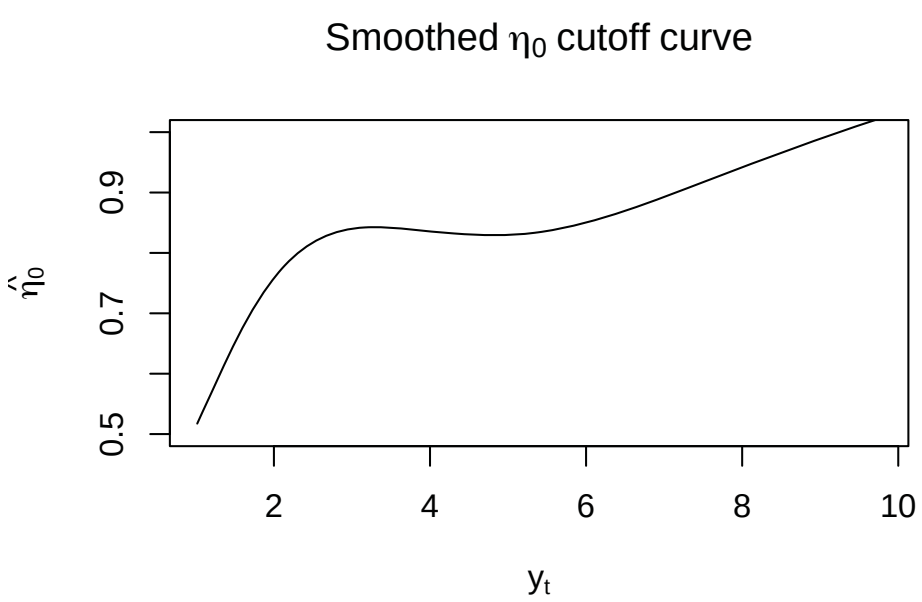




ConcaveFDR

1. **A novel empirical null model estimation technique**
   - truncated ML, **all statistics** y with $y < y_t$ are used to estimate the null model
   - y_t is found by an automated smoothing method
2. **Constrained maximum likelihood estimation** for the alternative model using **log–concave density estimation**





A Typical fdr Curve, the HND model

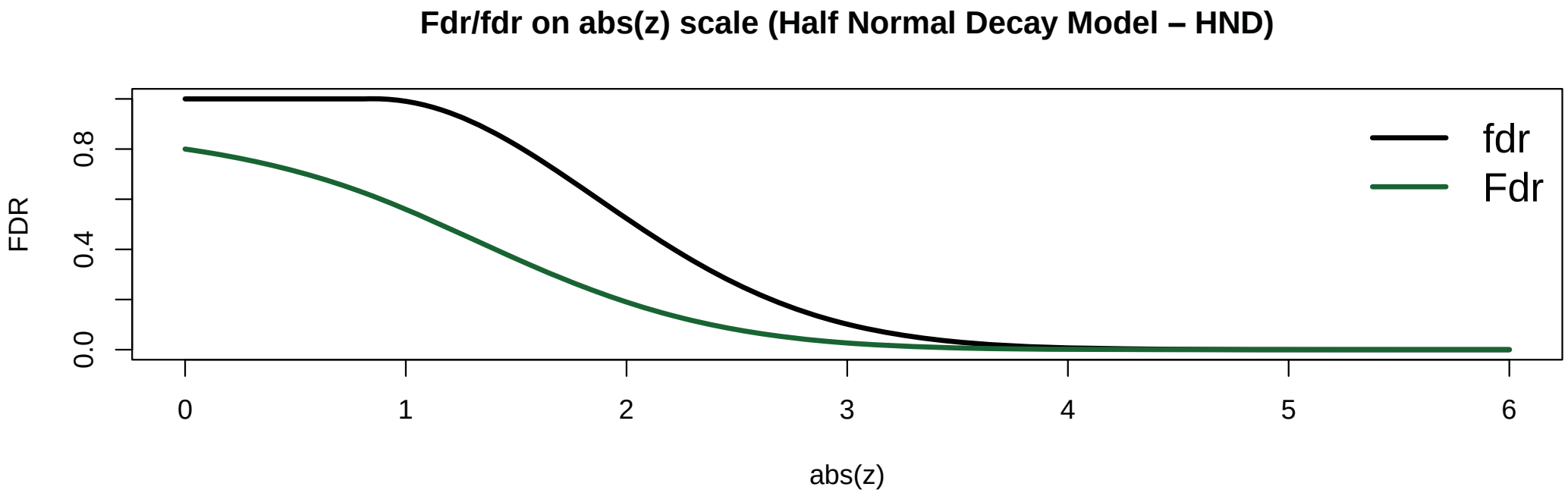
Half Normal Decay (HND) FDR-Model

- by Rice and Spiegelhalter (2008) for a **half normally distributed** Y:

$$\text{fdr}^{\text{HND}}(y|s) = \begin{cases} 1 & \text{for } y \leq s \\ e^{-(y-s)^2/2} & \text{for } y > s \end{cases}$$

- HND can be extended to include a scale parameter σ (Empirical null modeling)
- $\Rightarrow$ Null model for HND is then $N(0, \sigma)$.
- The parameter s can be mapped to the proportion of the null distribution η_0.

Graphical Representation of Fdr/fdr Using the HND Model



Noisy Splicing Drives mRNA Isoform Diversity in Human Cells

Joseph K. Pickrell[1]*, Athma A. Pai[1]*, Yoav Gilad[1]*, Jonathan K. Pritchard[1,2]*

1 Department of Human Genetics, The University of Chicago, Chicago, Illinois, United States of America, **2** Howard Hughes Medical Institute, The University of Chicago, Chicago, Illinois, United States of America

Abstract

While the majority of multiexonic human genes show some evidence of alternative splicing, it is unclear what fraction of observed splice forms is functionally relevant. In this study, we examine the extent of alternative splicing in human cells using deep RNA sequencing and *de novo* identification of splice junctions. We demonstrate the existence of a large class of low abundance isoforms, encompassing approximately 150,000 previously unannotated splice junctions in our data. Newly-identified splice sites show little evidence of evolutionary conservation, suggesting that the majority are due to erroneous splice site choice. We show that sequence motifs involved in the recognition of exons are enriched in the vicinity of unconserved splice sites. We estimate that the average intron has a splicing error rate of approximately 0.7% and show that introns in highly expressed genes are spliced more accurately, likely due to their shorter length. These results implicate noisy splicing as an important property of genome evolution.

PLoS Genetics 2010

"... we extrapolate that the majority of different mRNA isoforms present in a cell are not functionally relevant, though most copies of a pre-mRNA produce truly functional isoforms."

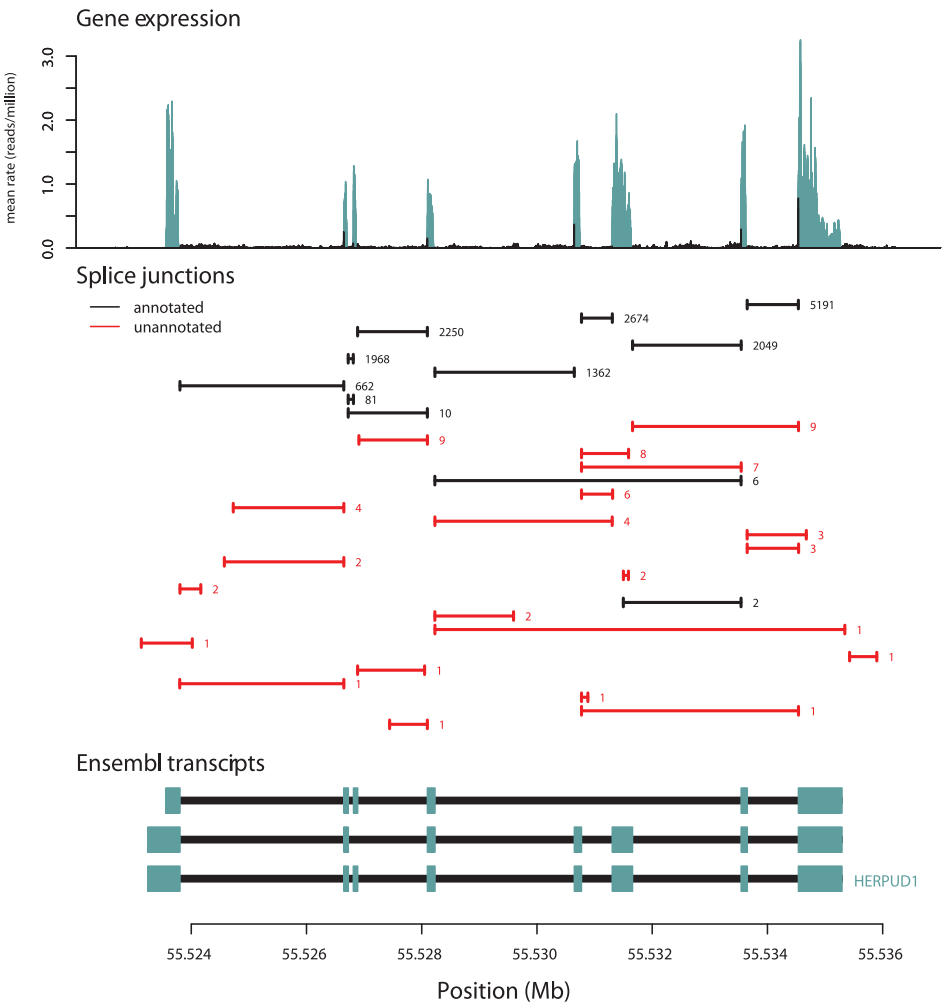
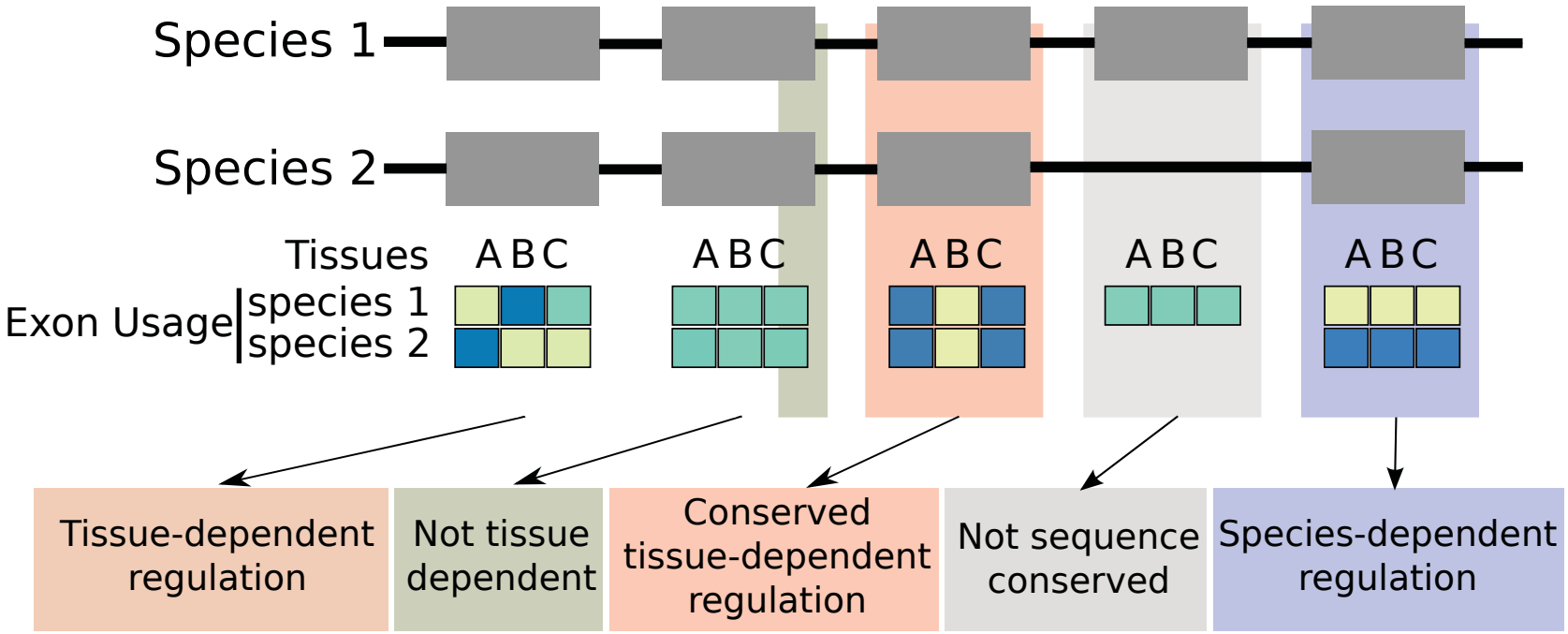


Figure 2. An example of splice junctions identified in a gene. In the top panel, we plot the average expression level at each base in a regior surrounding *HERPUD1*. In blue are bases annotated as exonic, and in black are those annotated as not exonic. In the middle panel, we plot the positions of all splice junctions in the region identified in our data. In black are splice junctions that are present in gene databases; in red are those that are not. The number of sequencing reads supporting each junction is written to the right of each junction, and junctions are ordered from top to bottom of the plot according to their coverage. In the bottom panel, we show the gene models in the region from Ensembl. The blue boxes show the positions of exons, and the black lines the positions of introns.
doi:10.1371/journal.pgen.1001236.g002

Regulation of (alternative) exon usage

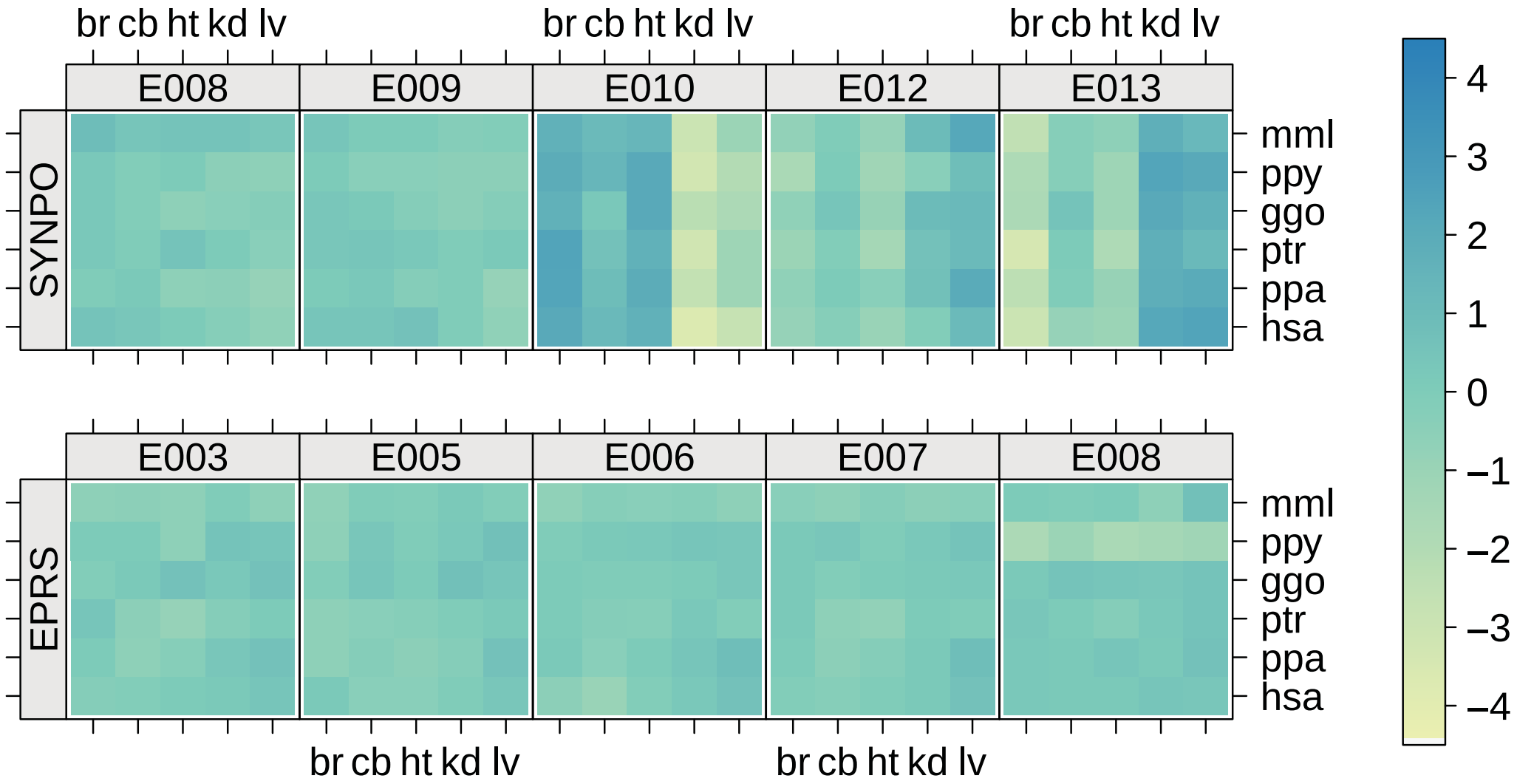


Data: multiple replicate samples each from:

- **6 primate species (hsa, ppa, ptr, ggo, ppy, mml)** X
- **5 tissues (heart, kidney, liver, brain, cerebellum)**

Brawand et al. Nature 2011 (Kaessmann Lab, Lausanne, CH)

Tissue and species dependence of relative exon usage



Drift and conservation of differential exon usage across tissues in primate species

Alejandro Reyes[a,1], Simon Anders[a,1], Robert J. Weatheritt[b,2], Toby J. Gibson[b], Lars M. Steinmetz[a,c], and Wolfgang Huber[a,3]

PNAS 2013

Application to exon usage conservation

- Exons = parts of the DNA sequence that are transcribed to mRNA and code for proteins

- (FDR) analysis of the correlation of exon usage across tissues between species pairs

Summary

- Publishing software implementing the methods developed is essential

- Use RMarkdown / knitr to enable reproducible research

http://yihui.name/knitr/

http://rmarkdown.rstudio.com/

- ConcaveFDR works reasonably well and will be available on Bioconductor

- FDR methods are useful in a lot of contexts

Huber Group @ EMBL

Simon Anders
Joseph Barry
Julian Gehring
Felix Klein
Malgorzata Oles
Andrzej Oles
Aleksandra Pekowska
Paul-Theodor Pyl
Sophie Rabe
Sascha Dietrich
Alejandro Reyes

Korbinian Strimmer

Verena Zuber
Sebastian Gibb

EMBL

Imperial College London