AdaPT: An interactive procedure for multiple testing with side information

Lihua Lei and William Fithian

Setup

- Hypotheses $H_{0,i}, i \in [n]$ with $\mathcal{H}_0 = \{i : H_{0,i} \text{ is null}\}$.
- p_i: p-values, x_i: side information.
- Examples:

Ordered hypothesis testing	x_i: rank of H_i;
Spatio-temporal testing	x_i: geographic location;
Clinical meta-analysis	x_i: index of the experiments;
Genome-wide association study	x_i: indices of the gene and the disease
Differential expression analysis	x_i: number of reads
$\ldots$	$x_i : \ldots$

- False discovery proportion (FDP) and false discovery rate (FDR):

$$\text{FDP} = \frac{\#\text{ false rejections}}{\#\text{ rejections}}, \quad \text{FDR} = \mathbb{E}[\text{FDP}]$$

- Goal: incorporate side information to improve the power while controlling FDR at a pre-specified level.

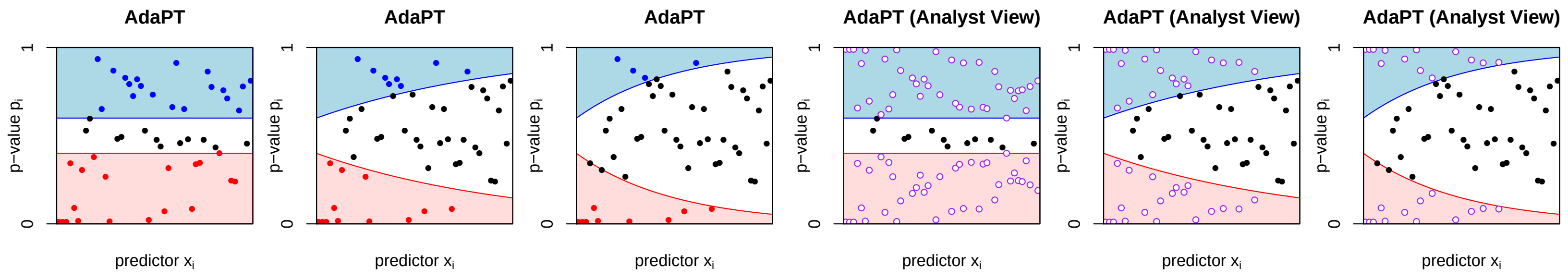
Adaptive P-Value Thresholding (AdaPT)

Define *partially masked p-values:*

$$\tilde{p}_{t,i} = \begin{cases} p_i & s_t(x_i) < p_i < 1 - s_t(x_i) \\ \{p_i,\ 1 - p_i\} & \text{otherwise.} \end{cases}$$

Visualization:

AdaPT (Oracle's view) AdaPT (Analysts' view)



FDP estimator of AdaPT:

$$\widehat{\text{FDP}}_t = \frac{\#\text{ blue points} + 1}{\#\text{ red points} \vee 1}.$$

Requirements on the update rule ($s_t(x_i) \to s_{t+1}(x_i)$):

- $s_{t+1}(x_i) \le s_t(x_i), \quad \forall i$;
- $s_{t+1}(x_i)$ only depends on $(x_i, \tilde{p}_{t,i})_{i=1}^n$, # blue points and # red points.

Theorem 1. *Assume that the null p-values are independent of each other and of the non-null p-values, and the null p-values are $U([0,1])$ or mirror-conservative. Then the AdaPT procedure controls the FDR at level α,* ***regardless of the update rule****.*

Guiding Principle for Updating Thresholds

Theorem 2. *Under mild assumptions, the optimal threshold $s(x)$ is* ***a level curve of local FDR****, defined as*

$$\text{fdr}(p \mid x) \;=\; \mathbb{P}(H_i \text{ is null } \mid x_i = x, p_i = p)$$

Guiding Principle

Step 1. Propose a **working model** (e.g. conditional two-group model);

Step 2. Use **your favorite method** to fit the model, based on $(x_i, \tilde{p}_{t,i})_{i=1}^n$;

Step 3. Estimate **level curves of local FDR**;

Step 4. Move the threshold towards a "near" level curve;

Consider the *conditional two-group model* as a **working model**:

$$H_i \mid x_i \sim \text{Bernoulli}(\pi_1(x_i))$$

$$p_i \mid H_i, x_i \sim \begin{cases} f_0(p \mid x_i) & \text{if } H_i = 0 \\ f_1(p \mid x_i) & \text{if } H_i = 1 \end{cases}$$

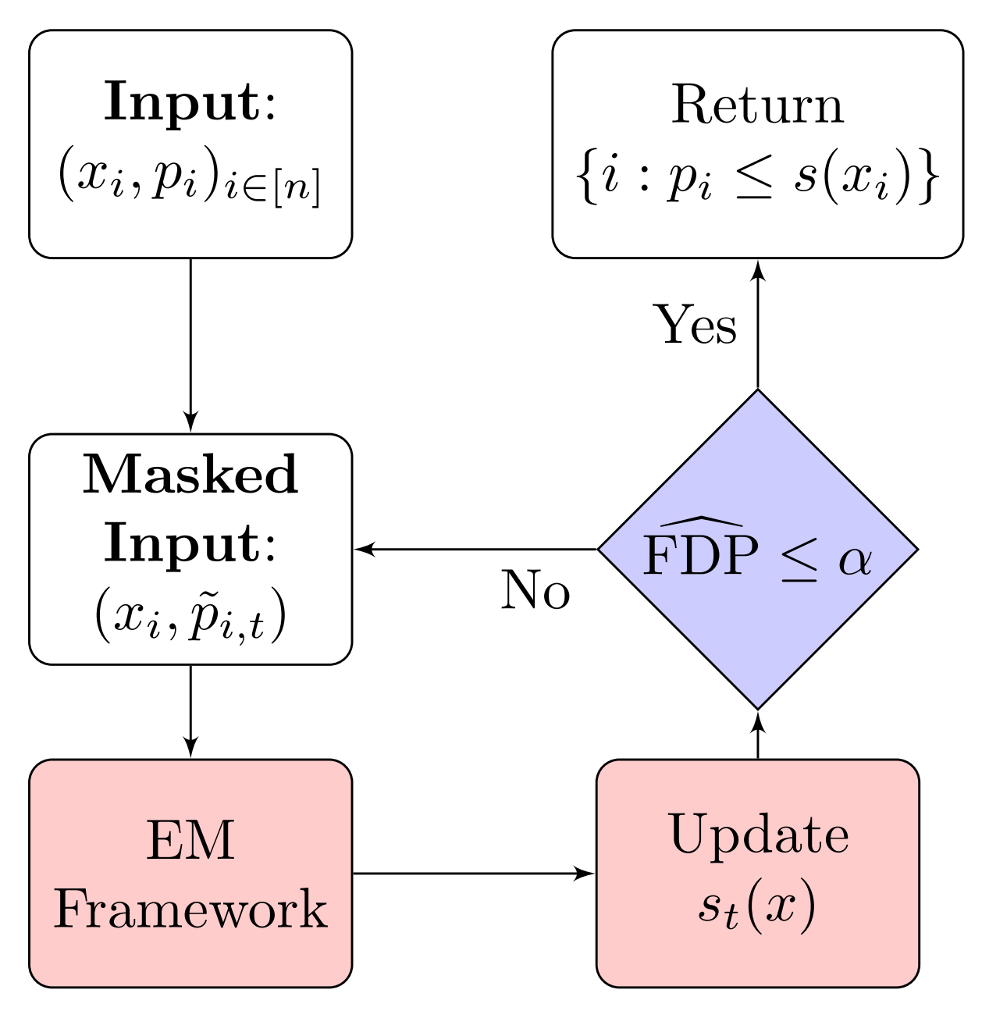
An example (conditional Gamma GLM):

$$\text{logit}(\pi_1(x)) = \beta^T \phi(x), f_0(p \mid x) = 1, f_1(p \mid x) \sim \text{Beta}(\gamma^T \phi(x), 1)$$

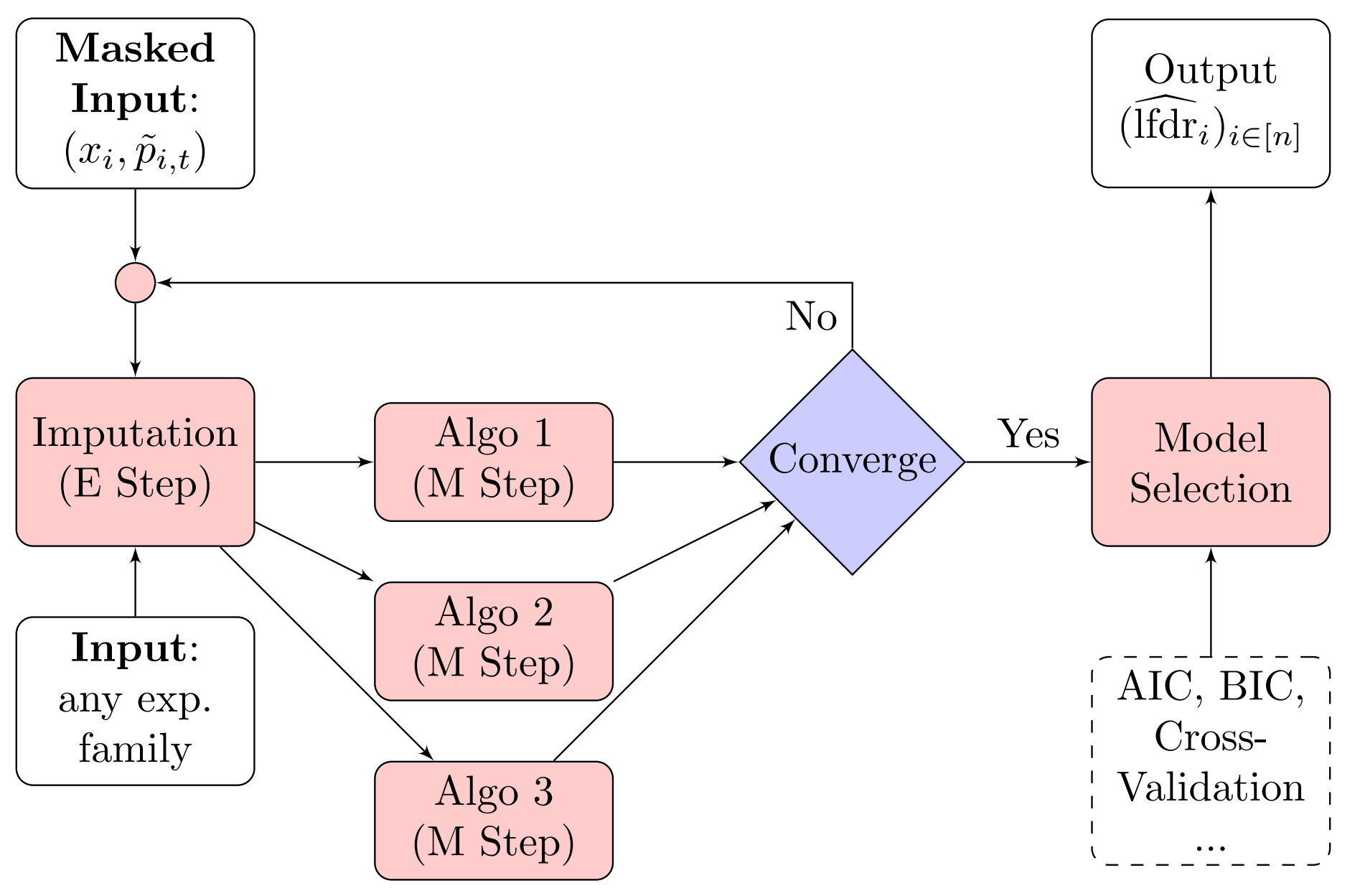
$$\text{in which case} \quad \text{lfdr}(x) = \frac{(1 - \pi_1(x)) f_0(p \mid x)}{f(p \mid x)} = \frac{1 - \pi_1(x)}{f(p \mid x)} = \frac{f(1 \mid x)}{f(p \mid x)}$$

Implementation

Pipeline of AdaPT


Input: $(x_i, p_i)_{i\in[n]}$
Masked Input: $(x_i, \tilde{p}_{i,t})$
EM Framework
Update $s_t(x)$
$\widehat{\text{FDP}} \le \alpha$
No
Yes
Return $\{i : p_i \le s(x_i)\}$

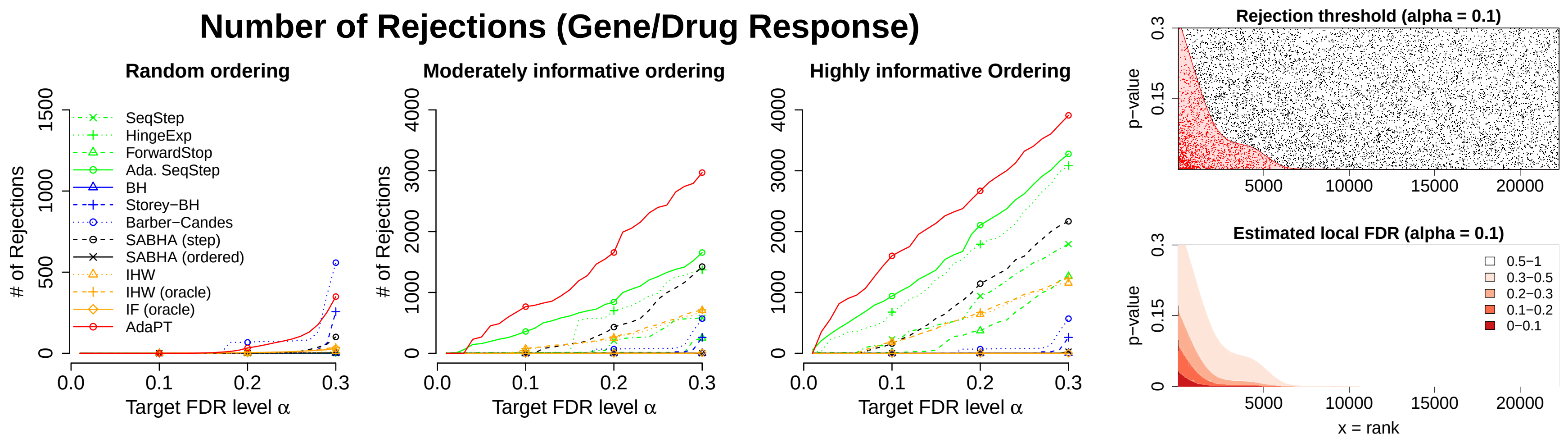
Sub-pipeline of EM framework


Masked Input: $(x_i, \tilde{p}_{i,t})$
Imputation (E Step)
Input: any exp. family
Algo 1 (M Step)
Algo 2 (M Step)
Algo 3 (M Step)
Converge
No
Yes
Model Selection
AIC, BIC, Cross-Validation ...
Output $(\widehat{\text{lfdr}}_i)_{i\in[n]}$

Applications

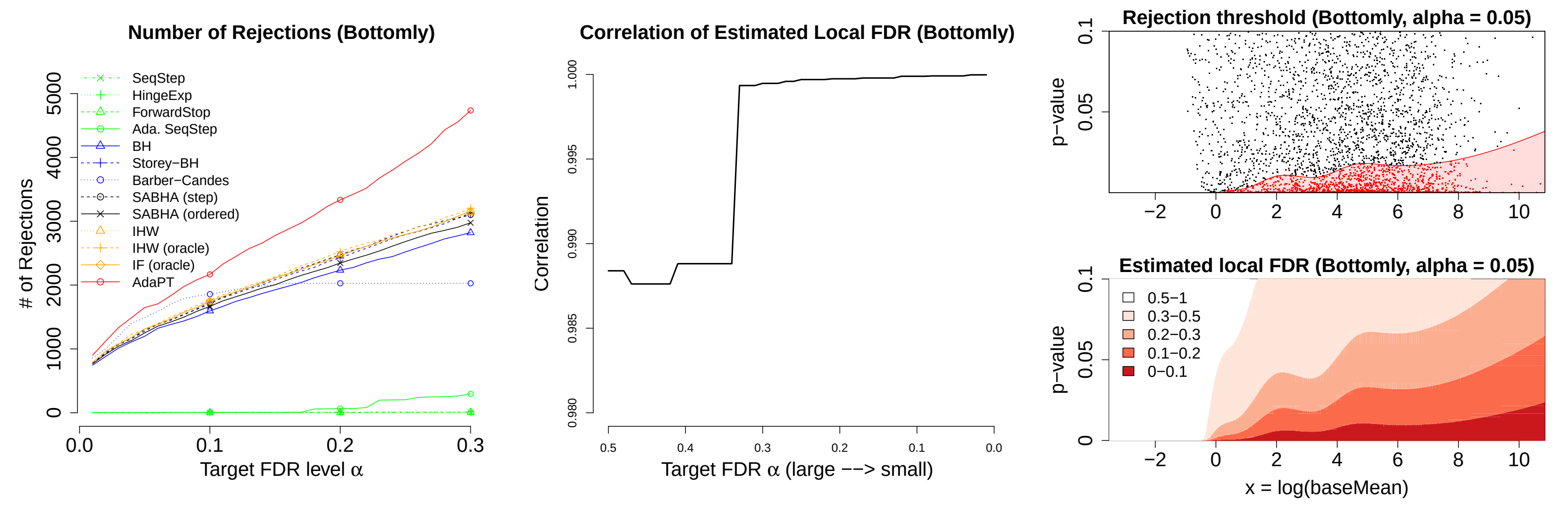
Example 1: Gene/drug response data (from GEO database):

- Gene expression in breast cancer cells in response to estrogen;
- $n = 22283$ genes, 25 trials at 5 doses including control;
- H_i: no differential response in low-dose vs. control;
- p_i: permutation t-test; x_i: rank of genes using other dosage groups;
- Working model: conditional Gamma GLM with $\phi(x)$ being the spline bases



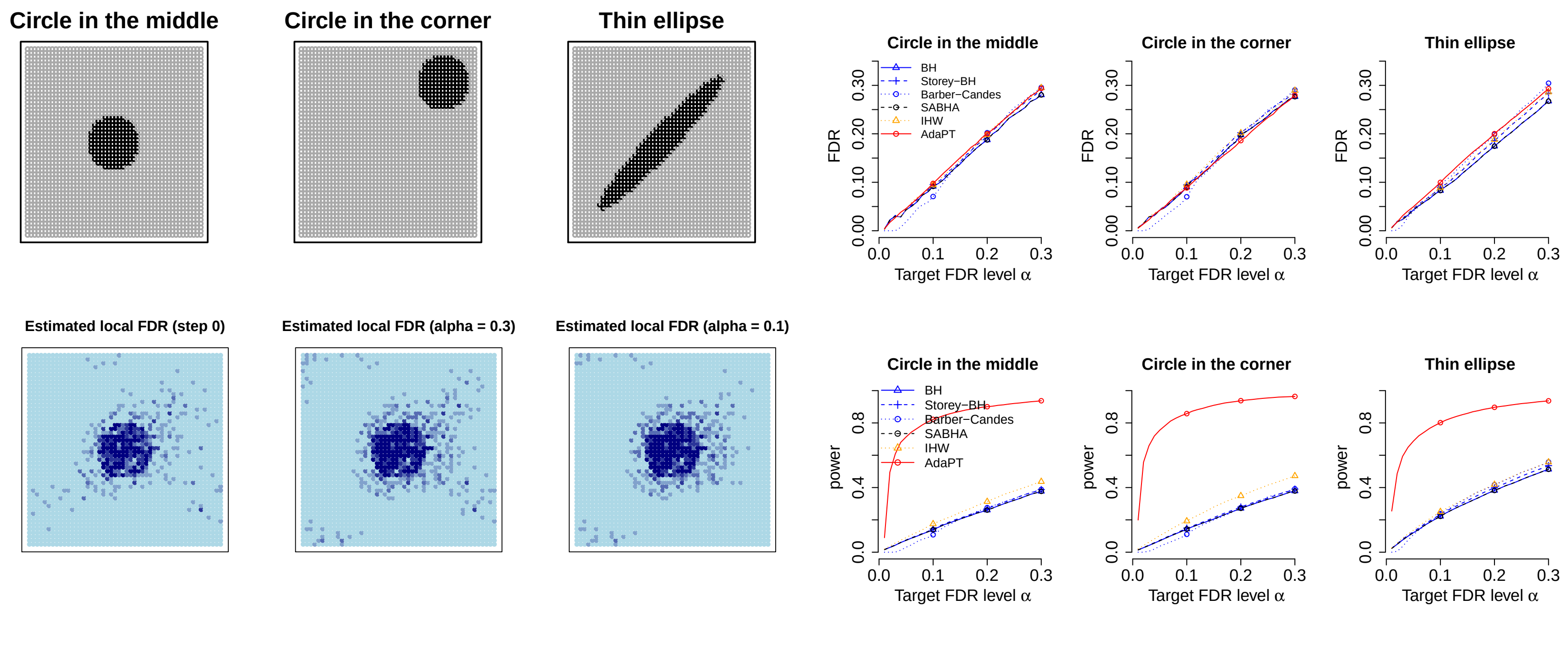
Example 2: RNA-seq data (Bottomly)

- Gene expression in two mouse strains C57BL/6J (B6) and DBA/2J (D2);
- $n = 13932$ genes, 21 samples (10 B6 and 11 D2);
- H_i : no differential response in gene i;
- p_i computed via `DEseq2` package; x_i : logarithmic normalized count;
- Working model: conditional Gamma GLM with $\phi(x)$ being the spline bases



Example 3: simulation study with two-dimensional covariates

- $x_i \overset{i.i.d.}{\sim} U([-100, 100] \times [-100, 100])$;
- $p_i = 1 - \Phi(z_i)$ where $z_i \sim N(0, 1)$ if $i \in \mathcal{H}_0$ and $z_i \sim N(2, 1)$ otherwise;
- Working model: conditional Gamma GAM with $\phi(x)$ being the spline bases


Circle in the middle
Circle in the corner
Thin ellipse
Estimated local FDR (step 0)
Estimated local FDR (alpha = 0.3)
Estimated local FDR (alpha = 0.1)
FDR
power
Target FDR level α
BH
Storey−BH
Barber−Candes
SABHA
IHW
AdaPT