Continous Testing for Poisson Processes Intensities

F. Picard[*], E. Roquain[‡], A.-L. Fougères [†], P. Reynaud-Bouret[°]

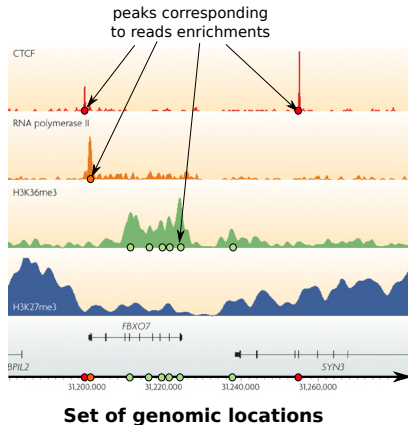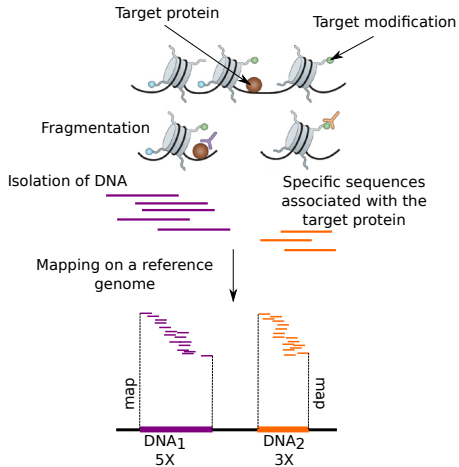[*] *Laboratoire de Biométrie et Biologie Évolutive*, Univ. Lyon 1,
[‡] *Laboratoire de Probabilités et Modèles Aléatoires*, Univ. Paris 6,
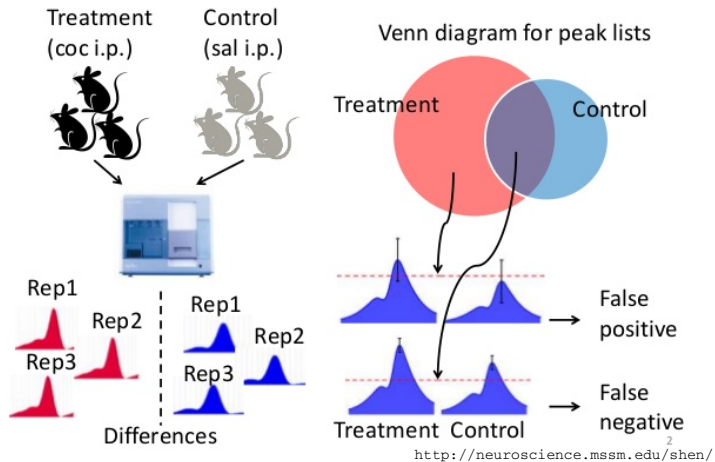[†] *Institut Camille Jordan*, Univ. Lyon 1,
[°] *Laboratoire J.A. Dieudonné*, Univ. Nice.

August 2016

Mapping Protein-DNA interactions

Set of genomic locations

Differential Analysis of Peak-like data

http://neuroscience.mssm.edu/shen/

Aims of the project

- Develop a test procedure to **compare the location of the peaks**
- Assess the **confidence of the test** in terms of family wise error rate and false discovery rate
- Perform simulations & analyze experimental data
- Provide a computationally sounded procedure & package

Outline

1 Point Process modeling of Genomic features

2 Test statistics and associated *p*-value process

3 Two error rates in continuous time

4 Simulations

5 Application

6 Conclusions

Observations are random sets of points

- We observe two independent sets of peaks location:

$$N_A = \{T_1, \ldots, T_{n_A}\} \quad \text{and} \quad N_B = \{T_1, \ldots, T_{n_B}\}.$$

- We model those sets by two heterogeneous Poisson processes with intensity λ_A, λ_B in $L^2[0,1]$.

- For any interval $I \subseteq [0,1]$,

$$N_A(I) \sim \mathcal{P}\left(\int_I \lambda_A(t)dt\right) \quad \text{and} \quad N_B(I) \sim \mathcal{P}\left(\int_I \lambda_B(t)dt\right)$$
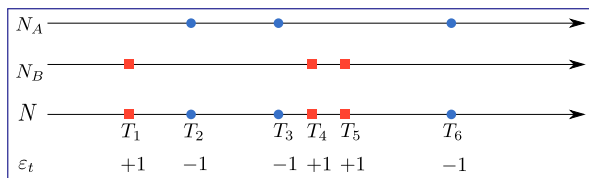
Aim

Testing $\lambda_A = \lambda_B$, or more precisely, detecting zones where $\lambda_A \neq \lambda_B$

Definition of the joint process

- From (N_A, N_B) we define the couple (N, ε)
- $N = N_A \cup N_B$ is the **joint process** of intensity $\lambda = \lambda_A + \lambda_B$,
- and where $\varepsilon = (\varepsilon_T)_{T \in N}$ is a **set of marks**:

$$\varepsilon_T = \begin{cases} +1, & \text{if } T \in N_A, \\ -1, & \text{if } T \in N_B. \end{cases}$$

Conditional distribution of the marks

- Conditionally to N, the distribution of the marks is:

$$\mathbb{P}(\varepsilon_T = +1 | N) = \frac{\lambda_A(T)}{\lambda_A(T) + \lambda_B(T)}$$

- We introduce notation:

$$\forall t \in [0, 1], \ \theta(t) = \frac{\lambda_A(t) - \lambda_B(t)}{\lambda_A(t) + \lambda_B(t)}.$$

- Conditionally to N, the distribution of the marks becomes:

$$\varepsilon_T | N \sim 2\mathcal{B}\left(\frac{\theta(T) + 1}{2}\right) - 1,$$

Nuisance parameters and conditional testing

- The distribution of the joint process (N, ε) can be re-parametrized:

$$(N, \varepsilon) \sim \mathbb{P}_{\theta, \lambda}$$

- λ and θ are unknown under the null, but are not "really" of interest
- We propose procedures that are **conditional to the observed joint process** N.

Reparametrization of the test

Conditional to N, the new hypothesis focuses on ε and becomes $\theta = 0$.

An infinite set of local null hypothesis

- We propose a functional testing framework : $\lambda_A = \lambda_B$ or $\theta = 0$.
- The global strategy corresponds to the **global null** hypothesis.
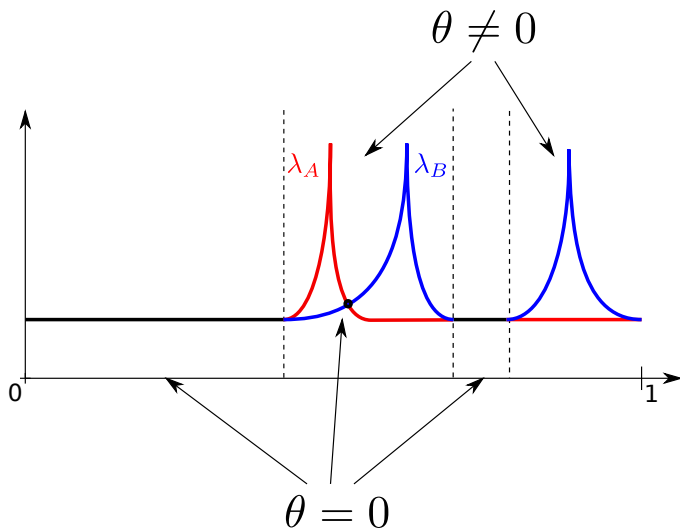- We consider **local hypothesis**:

$$H_{0,t} : \big\{\theta(t) = 0\big\} \quad \text{against } H_{1,t} : \quad \big\{\theta(t) \neq 0\big\}.$$

- The null hypothesis corresponding to

$$\Big\{\forall t \in J, \, \theta(t) = 0\Big\} \Leftrightarrow \Big\{\mathcal{H}_0\left\{J\right\} = \bigcap_{t \in J} H_{0,t}.\Big\}$$

- The *global null* hypothesis corresponds to $\mathcal{H}_0\left\{[0,1]\right\}$.
- The null function on $[0, 1]$ is denoted by θ_0 in the sequel.

Local testing with a cartoon

Definition of scanning windows

- We introduce a resolution parameter η that is fixed
- Using the **continuous testing** framework, we perform a whole *continuum* of tests for each interval of length η contained in $[0, 1]$.
- We will distinguish sets of points (denoted by t) from sets of windows center (denoted by x)

$$\forall x \in \mathcal{X}_\eta = [\eta/2, 1 - \eta/2], \ I_\eta(x) = [x - \eta/2, x + \eta/2]$$

Our multiple testing procedures are based on single tests on $\mathcal{H}_0 \{I_\eta(x)\}$ for all possible window centers
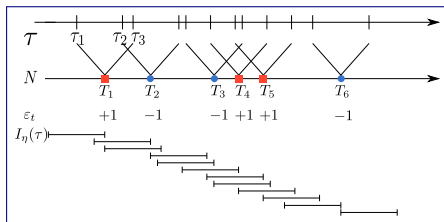
Is continuous testing computationally tractable ?

- Each observation T_i has a span η and will be used by the testing procedure on $[T_i - \eta/2, T_i + \eta/2]$
- There exists a partition τ of $\mathcal{X}_\eta$ consisting in M intervals and with inner breaks given by

$$\boldsymbol{\tau} = \left(\bigcup_{T \in N} \{T - \eta/2\} \cup \{T + \eta/2\} \right) \bigcap \mathcal{X}_\eta,$$

- The set τ is chosen as the center of the observed windows

$]\tau_{m-1}, \tau_m]$ are homogeneous intervals in terms of *composition* $N \cap I_\eta(x)$

Outline

1 Point Process modeling of Genomic features

2 Test statistics and associated p-value process

3 Two error rates in continuous time

4 Simulations

5 Application

6 Conclusions

Introducing kernel-based statistics

- We would like to define a statistics based an estimator of:

$$\|\lambda_A - \lambda_B\|^2_{I_\eta(x)} = \int_{I_\eta(x)} \left(\lambda_A(s) - \lambda_B(s)\right)^2 ds$$

- $H_0\{I_\eta(x)\}$ is rejected if this norm is high w.r.t. its expectation
- Our test statistics is ($n = N([0, 1])$:

$$S_\eta(x) = \frac{1}{n(n-1)} \sum_{T \neq T' \in N \cap I_\eta(x)} K_h \left(T - T'\right) \varepsilon_T \varepsilon_{T'}$$

- $S_\eta(x + dx)$ can be computed from $S_\eta(x)$ which results in an **online** algorithm

Conditional testing and the *p*-value process

- We are interested in the distribution of $S_\eta(x)$ under $H_0\{I_\eta(x)\}$:
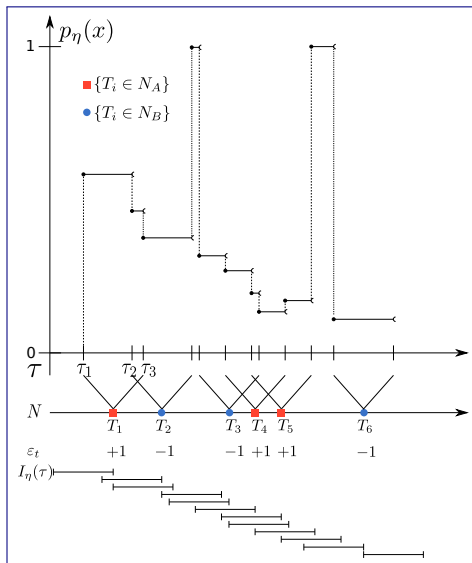
$$\forall x \in \mathcal{X}_\eta, \ F_{\theta_0,N}(s;x) = \mathbb{P}_{\theta_0}\Big(S_\eta(x) \geq s|N\Big)$$

- Since the intensities are heterogeneous, we rather consider *p*-values (normalize between $[0,1]$):

$$\forall x \in \mathcal{X}_\eta, \ p_\eta(x) = F_{\theta_0,N}\Big(S_\eta(x);x\Big)$$

- Since $S_\eta(x)$ is piece-wise constant, $\Big(p_\eta(x)\Big)_x$ is a piece-wise constant process on $[0,1]$.

The *p*-value process with a cartoon

Conditional Monte-Carlo approximation of the _p_-values

- Sample B independent draws of i.i.d. Rademacher sets of marks:

$$\varepsilon^b := (\varepsilon_T^b)_{T \in N}, \text{ for } b = 1, ..., B$$

- Label the observed marks such that $\varepsilon^0 := (\varepsilon_T)_{T \in N}$, (first term of a $B + 1$-sample of marks)

- The conditional distribution given N of the Rademacher process is:

$$\varepsilon_T^b | N \sim 2\mathcal{B}(1/2) - 1,$$

- We obtain the estimated _p_-value process

$$\widehat{p}_\eta(x) = \frac{1}{B + 1} \left(1 + \sum_{b=1}^{B} 1_{\left\{ S_\eta^b(x) \geq S_\eta^0(x) \right\}} \right).$$

> This parametrization guarantees that under $H_0\{I_\eta(x)\}$:
> $$\forall \alpha \in [0, 1], \ \mathbb{P}_{\theta, \lambda}\left(\widehat{p}_\eta(x) \leq \alpha \right) \leq \alpha.$$

Outline

1 Point Process modeling of Genomic features

2 Test statistics and associated *p*-value process

3 Two error rates in continuous time

4 Simulations

5 Application

6 Conclusions

Acceptation and Rejection Sets

- A multiple testing procedure is defined by a **rejection set**:

$$\mathcal{R}_\eta(u) := \{x \in \mathcal{X}_\eta \ : \ p_\eta(x) < u\},$$

- u is a threshold potentially depending on the data.
- The set of **accepted windows** is denoted by

$$\mathcal{A}_\eta(u) := \{x \in \mathcal{X}_\eta \ : \ p_\eta(x) \geq u\}.$$

- $\mathcal{A}_\eta(u)$ is an approximation of

$$J_0^\eta := \left\{x \in \mathcal{X}_\eta \ : \ \forall t \in I_\eta(x), \theta(t) = 0\right\}$$

Challenge

How to evaluate the quality of threshold u ?

False Positive Windows and the continuous FWER

- The target is the set of **false positive windows**

$$J_0^\eta \cap \mathcal{R}_\eta(u)$$

- Its size can be measured by its Lebesgue measure:

$$\Lambda\big(J_0^\eta \cap \mathcal{R}_\eta(u)\big)$$

- The **Family-Wise Error Rate** in continuous time can be defined by

$$\text{FWER}_{\theta,\lambda}^\eta(u) = \mathbb{P}_{\theta,\lambda}\Big(\Lambda\big(J_0^\eta \cap \mathcal{R}_\eta(u)\big) > 0\Big).$$

Aim

Calibrate $u^\alpha \in [0,1]$ such that $\text{FWER}_{\theta,\lambda}^\eta(u^\alpha)$ is controlled at level α

False Positive Windows and the continuous FDR

- The target is the set of **false positive windows**

$$J_0^\eta \cap \mathcal{R}_\eta(u)$$

- Its size can be measure by its Lebesgue measure:

$$\Lambda\big(J_0^\eta \cap \mathcal{R}_\eta(u)\big)$$

- The **False Discovery Rate** in continuous time can be defined by

$$\mathsf{FDR}_{\theta,\lambda}^\eta(v) = \mathbb{E}_{\theta,\lambda}\left(\frac{\Lambda\left(J_0^\eta \cap \mathcal{R}_\eta(v)\right)}{\Lambda\left(\mathcal{R}_\eta(v)\right)}\right)$$

Aim

Calibrate $v^\alpha \in [0,1]$ such that $\mathsf{FDR}_{\theta,\lambda}^\eta(v^\alpha)$ is controlled at level α

Controlling the FWER in continuous time - 1

- The starting point is that we have for all u,

$$\left\{ J_0^\eta \cap \mathcal{R}_\eta(u) \neq \emptyset \right\} = \left\{ \exists x \in J_0^\eta \,:\, p_\eta(x) < u \right\}$$
$$= \left\{ \inf_{x \in J_0^\eta} \{p_\eta(x)\} < u \right\}.$$

- Control the FWER by learning the distribution of the min. *p*-values under the null.

- Consider the conditional α-quantile of the min. *p*-value process on $[0,1]$:

$$U_{J_0^\eta}^\alpha = \min \left\{ u \in [0,1] \,:\, \mathbb{P}_{\theta_0} \left(\inf_{x \in J_0^\eta} \{p_\eta(x)\} \leq u \,\big|\, N \right) \right\}.$$

Controlling the FWER in continuous time - 2

- But the set of windows J_0^η is unknown: **Choose the worst-case scenario**

- We compute the quantile of the min. of the *p*-value process on $\mathcal{X}_\eta$:

$$U_{\mathcal{X}_\eta}^\alpha = \min \left\{ u \in [0,1] \: : \: \mathbb{P}_{\theta_0} \left(\inf_{x \in \mathcal{X}_\eta} \{p_\eta(x)\} \le u \, \big| N \right) \right\}.$$

- This **ensures the control of the FWER at level** α

- This procedure can be extended to step-down approaches.

FWER-Adjusted *p*-value process: the min-*p* procedure

- In practice we would like to use the *adjusted p*-value process:

$$\forall x \in \mathcal{X}_\eta, \ q_\eta(x) = F_{\theta_0, N}^{\min}\Big(p_\eta(x)\Big)$$

- This requires to compute the distribution of the min-*p* under the null

$$\forall z \in [0, 1], \ F_{\theta_0, N}^{\min}(z) = \mathbb{P}_{\theta_0, N}\left(\inf_{x \in \mathcal{X}_\eta}\{p_\eta(x)\} \leq z \mid N\right)$$
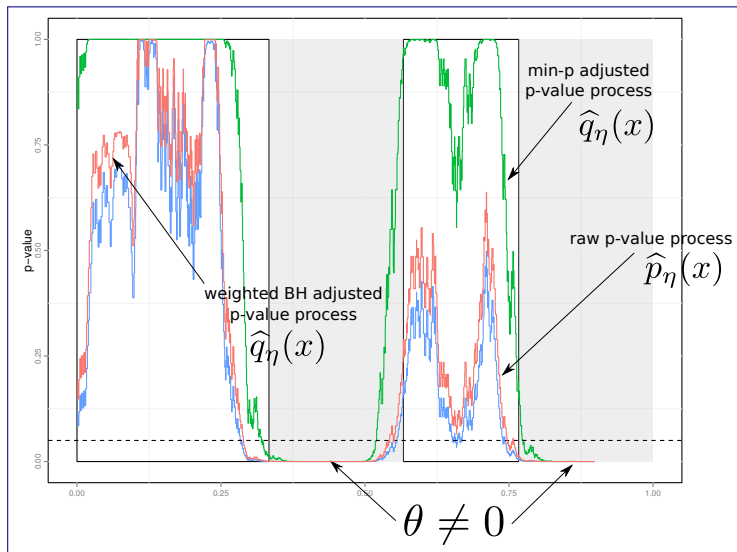
- In practice we control the FWER using:

$$\forall x \in \mathcal{X}_\eta, \ \widehat{q}_\eta(x) = \widehat{F}_{\theta_0, N}^{\min}\Big(\widehat{p}_\eta(x)\Big)$$

The min-p procedure with a cartoon
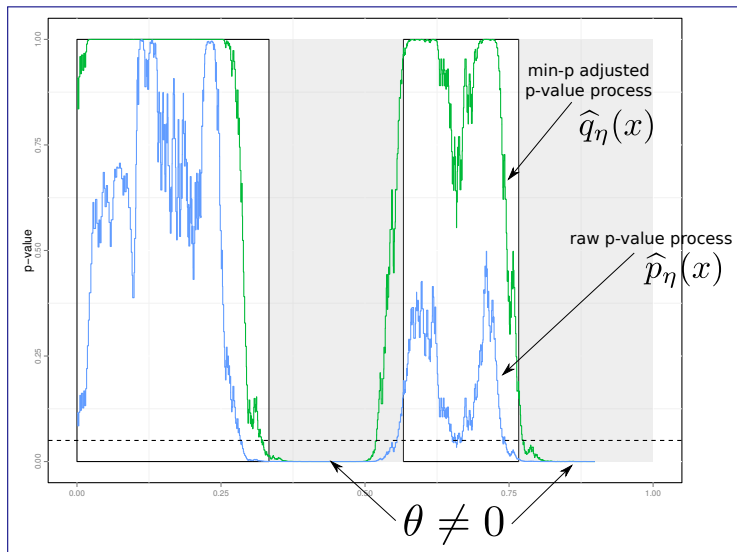
The weighted BH procedure with a cartoon

Outline

1 Point Process modeling of Genomic features

2 Test statistics and associated *p*-value process

3 Two error rates in continuous time

4 Simulations

5 Application

6 Conclusions

Simulation design in a cartoon
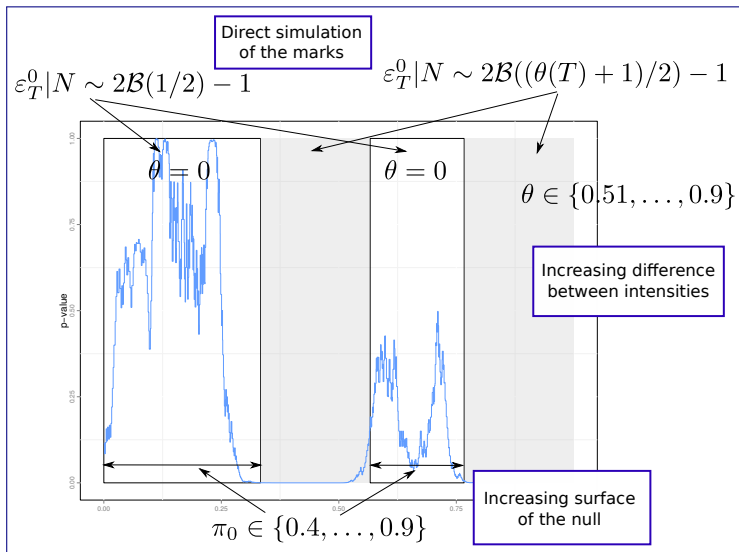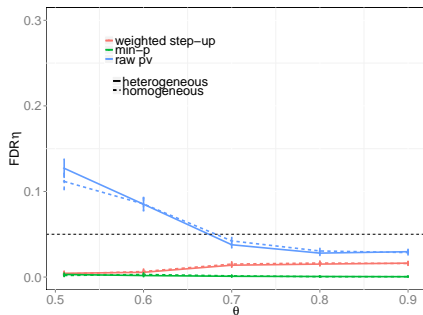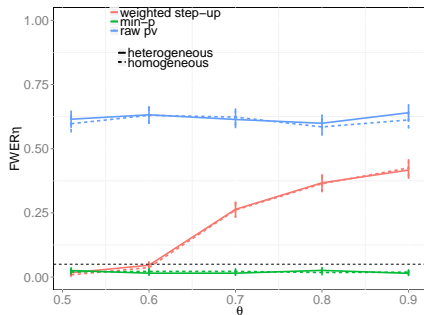
Error Rates control for different signal strengths

Figure: Influence of θ on FWER_η & FDR_η, $\eta = 0.1$.

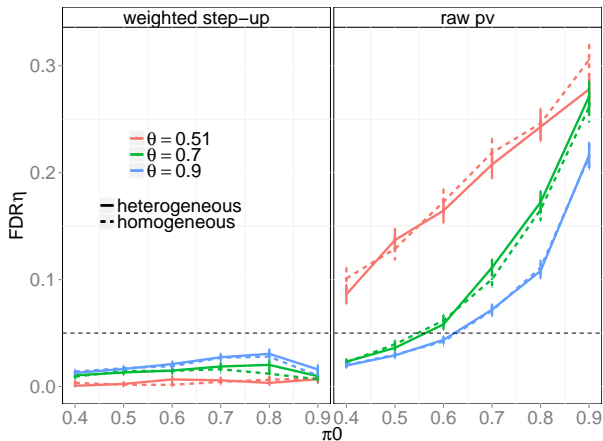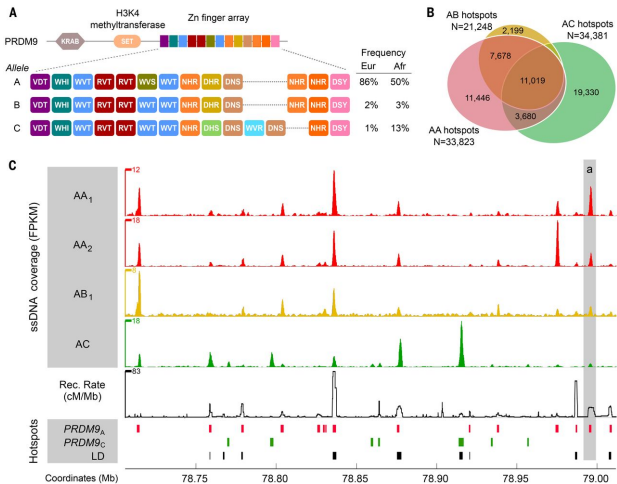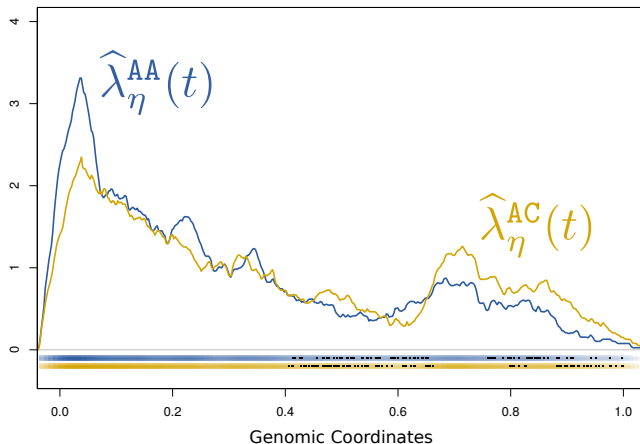# FDR control for different π_0

Figure: FDR_η, $\alpha = 5\%$ and $\eta = 0.1$. π_0: length of intervals on which $\theta = 0$.

Outline

1 Point Process modeling of Genomic features

2 Test statistics and associated *p*-value process

3 Two error rates in continuous time

4 Simulations

❺ Application

6 Conclusions

Recombination and PRDM9 binding sites

Comparison of PRDM9 binding between genotypes

Comparison of PRDM9 binding between genotypes

$$S_\eta^0(x) \propto \sum_{T,T'} K_h(T,T') \varepsilon_T^0 \varepsilon_T'^0$$

Comparison of PRDM9 binding between genotypes

Outline

1 Point Process modeling of Genomic features

2 Test statistics and associated *p*-value process

3 Two error rates in continuous time
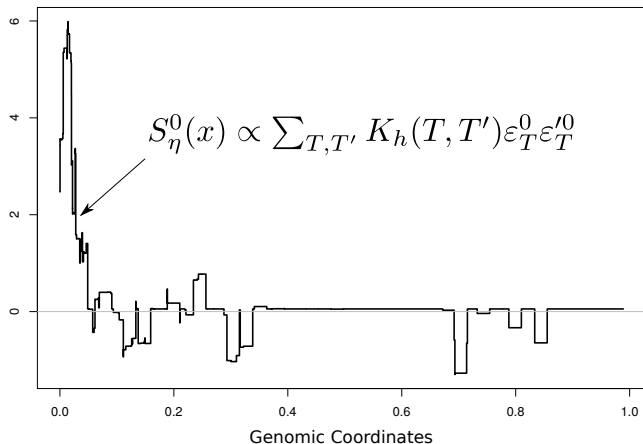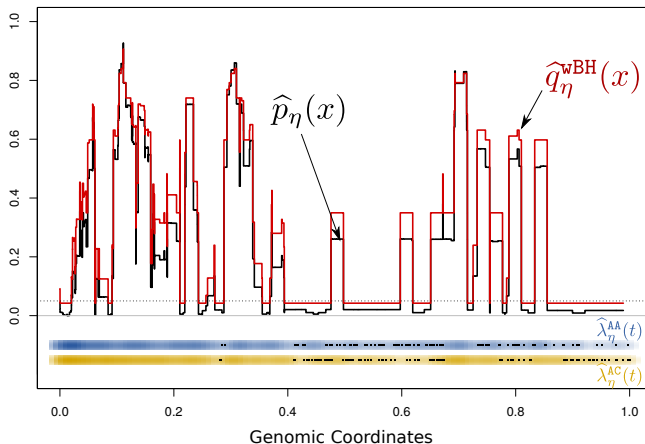
4 Simulations

5 Application

6 Conclusions

Perspectives of our work

- We provide a framework to locally compare Poisson processes intensities
- How procedures control the FWER and the FDR in continuous time
- This framework can be extended to one-sided hypothesis, and one-sample testing (homogeneity)
- Provides a new look on scanning statistics (lack of proper definition for FDR)
- Extension to 2D / 3D scans ?

A weighted step-up BH procedure

- Let τ be the partition that defines the windows:

$$\Lambda\left(\mathcal{R}_\eta(v)\right) = \sum_{m=0}^{M-1} (\tau_{m+1} - \tau_m) 1_{\{p_\eta(\tau_m) \leq v\}}.$$

- Compute the weights $w_m = (\tau_{m+1} - \tau_m)/(1 - \eta)$

- Denote $\{p_m, 1 \leq m \leq M\} = \{p_\eta(\tau_m), 0 \leq m \leq M-1\}$ and order this *p*-values in increasing order $p_{\sigma(1)} \leq \cdots \leq p_{\sigma(M)}$ for an appropriate permutation σ of $\{1, \ldots, M\}$;

- Consider $\widehat{k} = \max\{k \in \{1, \ldots, M\} \ : \ p_{\sigma(k)} \leq \alpha \sum_{l=1}^{k} w_{\sigma(l)}\}$

- Compute V^α as $\alpha \sum_{l=1}^{\widehat{k}} w_{\sigma(l)}$.

BH-adjusted *p*-value process

- Let us denote by $(q_\eta(x))_{x \in \mathcal{X}_\eta}$ the adjusted *p*-values of the step-up procedure:

$$q_\eta(x) = \min_{k : p_{\sigma(k)} \geq p_\eta(x)} \left\{ \frac{p_{\sigma(k)}}{\sum_{l=1}^{k} w_{\sigma(l)}} \right\}.$$

- The decision at level α is simply to reject the nulls corresponding to windows $I_\eta(x)$ with *adjusted p*-values lower than α.

- We can check that

$$\mathcal{R}_\eta(V^\alpha) = \{x \in \mathcal{X}_\eta \; : \; q_\eta(x) \leq \alpha\}.$$

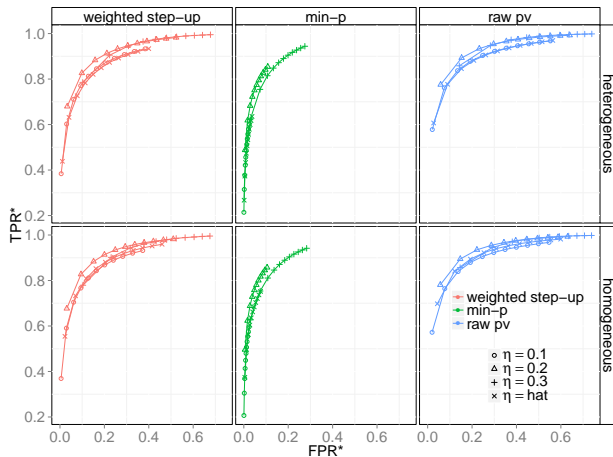◄ FDR control

Calibration of the window size η

Figure: FPR* & TPR* for different values of η ($\pi_0 = 0.5$). On point corresponds to one value for α (from 0.001 to 0.5).