# Local Statistics for Genome-Wide Association Studies

Mickaël Guedj

Journées BIL, 28 Janvier 2010, Nantes

PHARNEXT

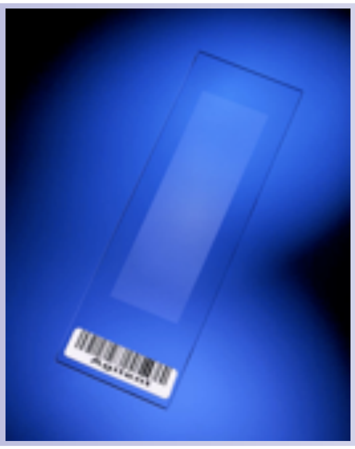
# I. Multiple-testing / local fdr

with S Robin (INAPG), A Celisse (INAPG), G Nuel (Univ Paris V)

# Multiple-testing

Advances in Molecular Biology and improvment of microarray technologies:

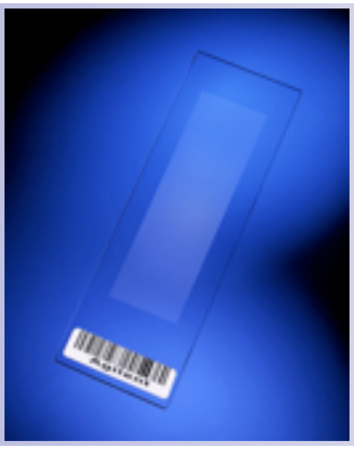
- □ Gene expression
- □ Genomic alterations (CGH)
- □ Genome-Wide Association Studies



# Multiple-testing

Advances in Molecular Biology and improvment of microarray technologies:

- □ Gene expression
- □ Genomic alterations (CGH)
- □ Genome-Wide Association Studies



The use of large-scale data requires the simultaneous evaluation of a huge number of statistical hypotheses.

30,000 genes / 1,000,000 genetic markers (SNPs) ...

▸ multiple-testing

# Multiple-testing

- □ $n$ tests at the $\alpha$ level:

|  | $H_0$ not rejected | $H_0$ rejected |  |
|---|---|---|---|
| $H_0$ true | $vn$ | $fp$ | $V$ |
| $H_0$ false | $fn$ | $vp$ | $F$ |
| total | $n-R$ | $R$ | $n$ |

# Multiple-testing

- $n$ tests at the $\alpha$ level:

true-negative

| | $H_0$ not rejected | $H_0$ rejected | |
|---|---|---|---|
| $H_0$ true | $vn$ | $fp$ | $V$ |
| $H_0$ false | $fn$ | $vp$ | $F$ |
| total | $n - R$ | $R$ | $n$ |

# Multiple-testing

- $n$ tests at the $\alpha$ level:

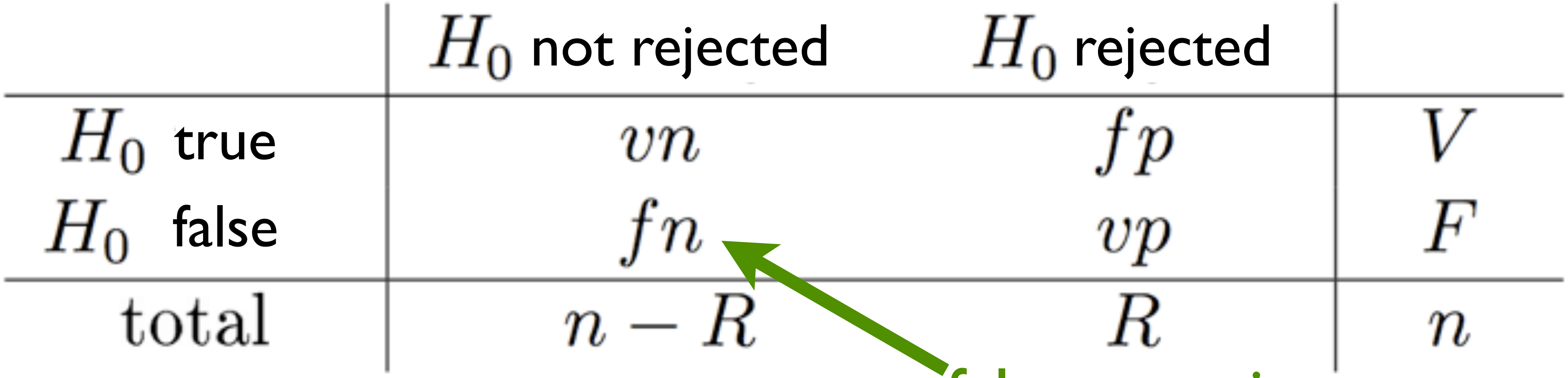
true-negative

| | $H_0$ not rejected | $H_0$ rejected | |
|---|---|---|---|
| $H_0$ true | $vn$ | $fp$ | $V$ |
| $H_0$ false | $fn$ | $vp$ | $F$ |
| total | $n - R$ | $R$ | $n$ |

true-positive

# Multiple-testing

□ $n$ tests at the $\alpha$ level:

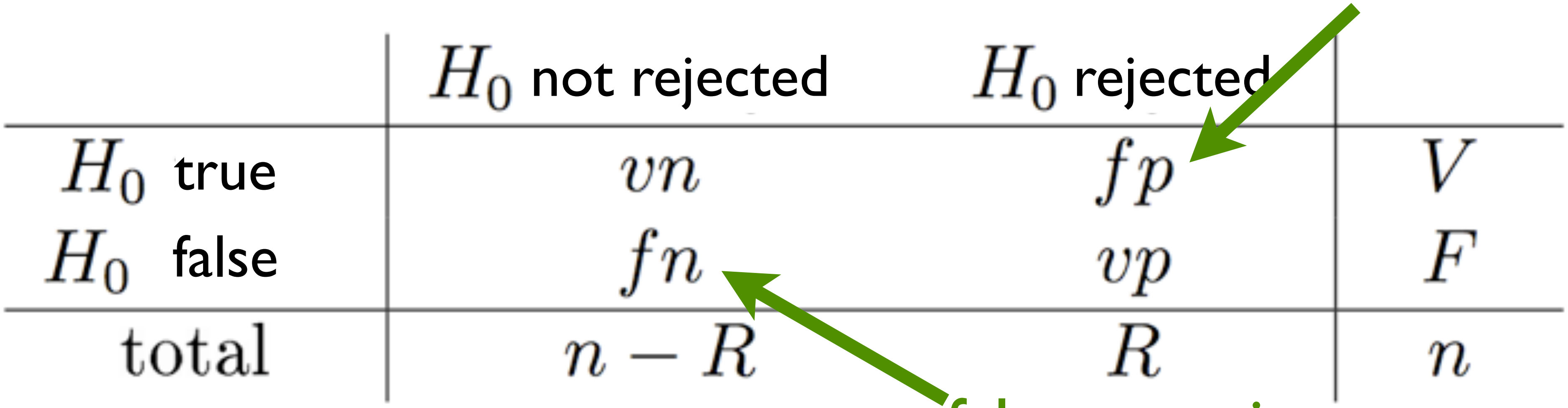


| | $H_0$ not rejected | $H_0$ rejected | |
|---|---|---|---|
| $H_0$ true | $vn$ | $fp$ | $V$ |
| $H_0$ false | $fn$ | $vp$ | $F$ |
| total | $n-R$ | $R$ | $n$ |

false-negative

# Multiple-testing

- □ $n$ tests at the $\alpha$ level:



|  | $H_0$ not rejected | $H_0$ rejected |  |
|---|---|---|---|
| $H_0$ true | $vn$ | $fp$ | $V$ |
| $H_0$ false | $fn$ | $vp$ | $F$ |
| total | $n-R$ | $R$ | $n$ |

false-positive

false-negative

# Multiple-testing

- $n$ tests at the $\alpha$ level:

| | $H_0$ not rejected | $H_0$ rejected | |
|---|---|---|---|
| $H_0$ true | $vn$ | $fp$ | $V$ |
| $H_0$ false | $fn$ | $vp$ | $F$ |
| total | $n - R$ | $R$ | $n$ |

- $n = 100,000 \qquad \alpha = 5\%$
    - 5,000 false-positives >> # true-positives

# Multiple-testing

- □ $n$ tests at the $\alpha$ level:

|  | $H_0$ not rejected | $H_0$ rejected |  |
|---|---|---|---|
| $H_0$ true | $vn$ | $fp$ | $V$ |
| $H_0$ false | $fn$ | $vp$ | $F$ |
| total | $n - R$ | $R$ | $n$ |

- □ $n = 100,000 \qquad \alpha = 5\%$
  - 5,000 false-positives >> # true-positives
  - the control of false-positives is a crucial issue.
  - type-I error-rate not adapted anymore

# Multiple-testing

- □ $n$ tests at the $\alpha$ level:

| | | |
|---|---|---|
| $H_0$ true | | $V$ |
| $H_0$ false | | $F$ |
| total | | $n$ |

➽ **error rates that consider the whole family of tests**

- □ $n = 100,000$    $\alpha = 5\%$
  - 5,000 false-positives >> # true-positives
  - the control of the *fp* is a crucial issue.
  - type-I error-rate not adapted anymore

# FWER

- Family-Wise Error-Rate: prob to falsely reject at least one hypothesis

$$\mathrm{FWER} = \mathbb{P}_{H_0}(fp > 0).$$

  - Bonferroni's majoration:

$$\mathrm{FWER} = 1 - \mathbb{P}_{H_0}(fp = 0) = 1 - (1-\alpha)^n \leqslant \max(n\alpha; 1).$$

→ $\alpha' = \frac{\alpha}{n}$

  - Estimation with Monte-Carlo simulations.
  - Conservative, loss of power.

(Benjamini et Hochberg 95)

# FDR

- less conservative than the FWER
- more intuitive interpretation

□ **False Discovery Rate:** prop of *fp* over the rejected hypotheses

$$\mathrm{FDR} = \mathbb{E}(Q),$$

with $Q = \frac{fp}{R}$ if $R > 0$ or $Q = 0$ otherwise.

(Benjamini et Hochberg 95)
(Forner, Guedj et al 07)

# FDR

- less conservative than the FWER
- more intuitive interpretation

□ False Discovery Rate: prop of *fp* over the rejected hypotheses

$$\mathrm{FDR} = \mathbb{E}(Q),$$

with $Q = \frac{fp}{R}$ if $R > 0$ or $Q = 0$ otherwise.

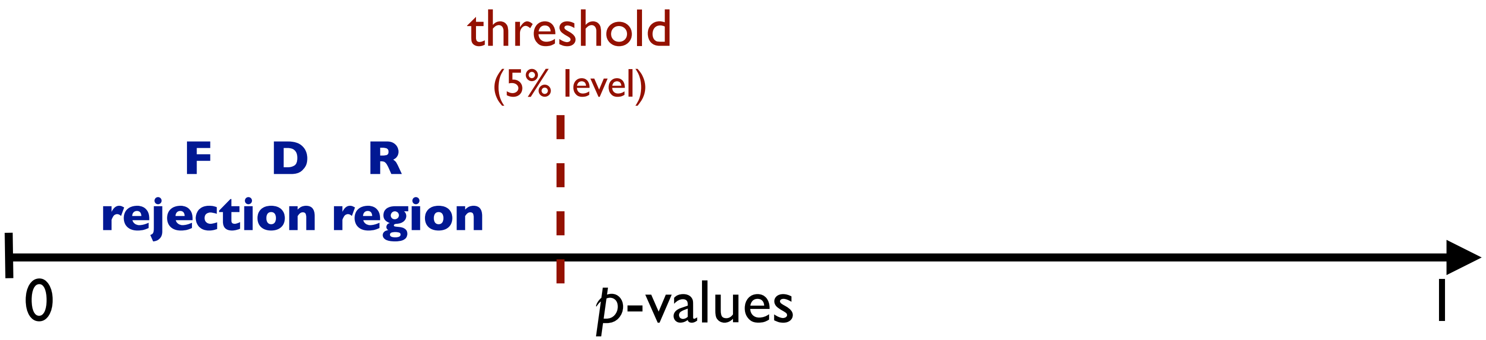
□ Benjamini-Hochberg's majoration:

$$\mathrm{FDR} \leqslant \min\left(\frac{n\alpha}{R(\alpha)};1\right)$$

□ Estimation with Monte-Carlo simulations.

# FDR

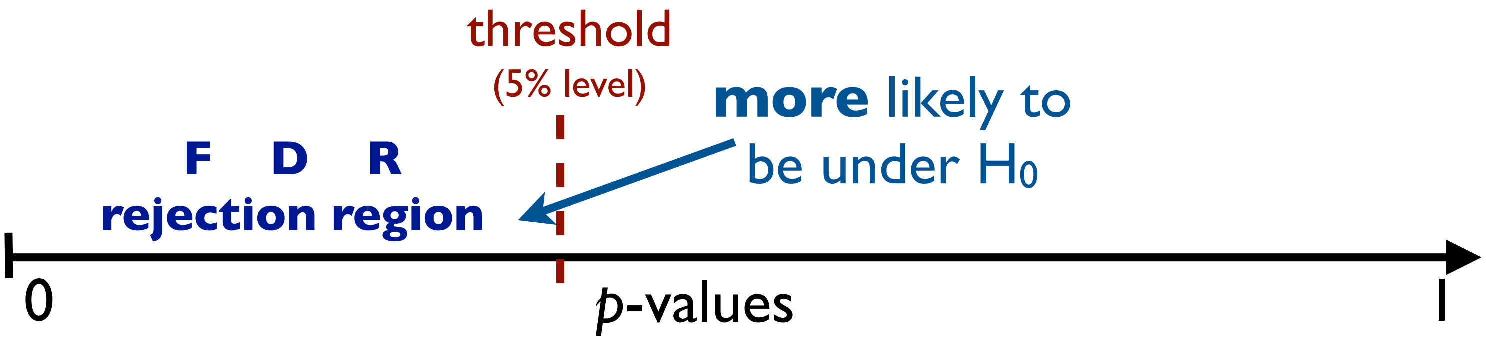
- False Discovery Rate:



- **Global criterion,** can not be used to assess the reliability of a specific hypothesis.
- **Associated to a given rejection region** without distinguishing statistics/*p*-values that are close to the threshold and those that are not.

# FDR

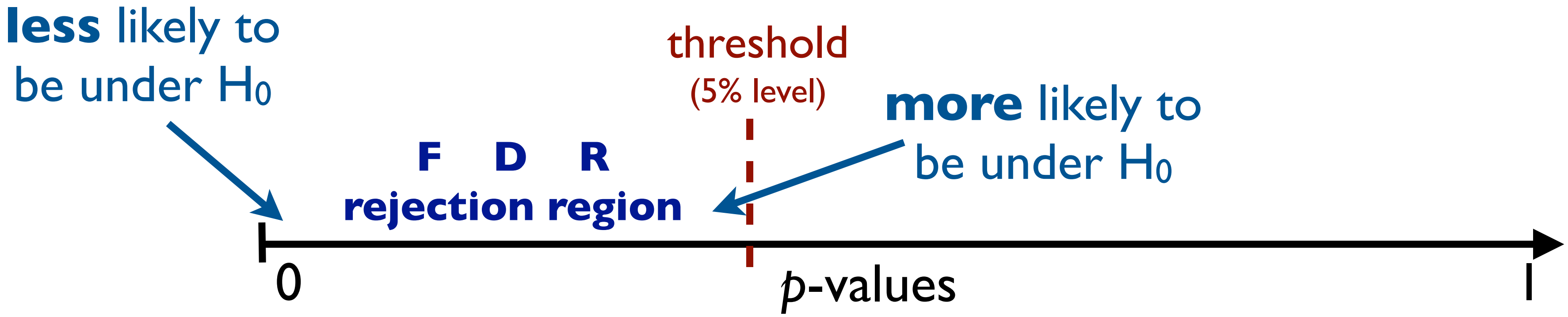
- False Discovery Rate:



- **Global criterion,** can not be used to assess the reliability of a specific hypothesis.
- **Associated to a given rejection region** without distinguishing statistics/$p$-values that are close to the threshold and those that are not.

# FDR

- False Discovery Rate:



**less** likely to be under $H_0$

threshold (5% level)

**more** likely to be under $H_0$

**F D R rejection region**

0

$p$-values

1

- **Global criterion,** can not be used to assess the reliability of a specific hypothesis.
- **Associated to a given rejection region** without distinguishing statistics/$p$-values that are close to the threshold and those that are not.

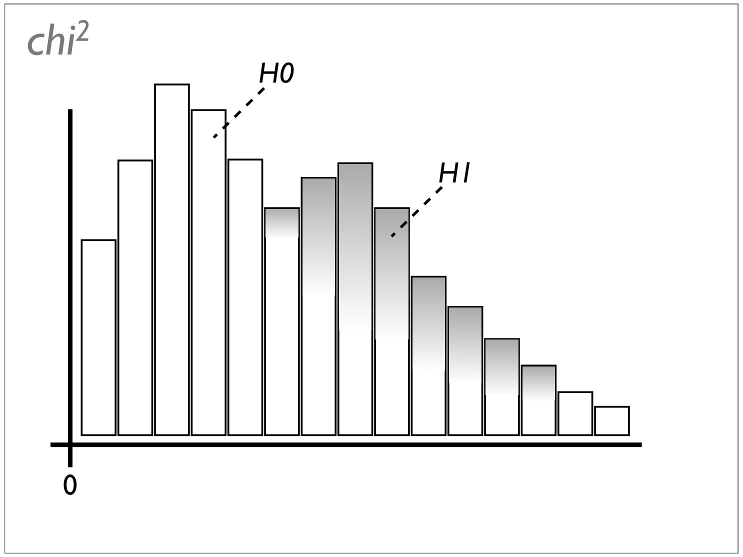
(Efron 04)

# Local FDR

- □ Local False Discovery Rate: prob of a given null hypothesis to be true

$$\mathrm{fdr}_i = \mathbb{P}\left(H = H0 | \mathcal{S} = \mathcal{S}_i\right)$$

- □ Mixture model: general and statistically convenient framework



$$f = \pi_0 f_0 + \pi_1 f_1,$$

$$\mathrm{fdr}_i \equiv \frac{\pi_0 f_0(\mathcal{S}_i)}{f(\mathcal{S}_i)}$$

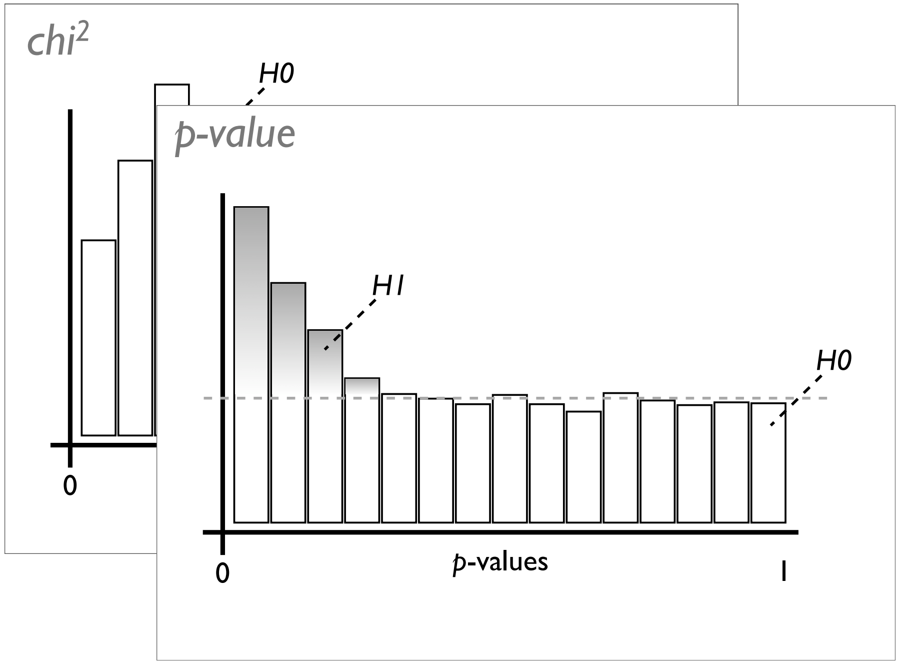
(Efron 04)

# Local FDR

□ Local False Discovery Rate: prob of a given null hypothesis to be true

$$\mathrm{fdr}_i = \mathbb{P}\left(H = H0 | \mathcal{S} = \mathcal{S}_i\right)$$

□ Mixture model: general and statistically convenient framework



$$f = \pi_0 f_0 + \pi_1 f_1,$$

$$\mathrm{fdr}_i \equiv \frac{\pi_0 f_0 (p\mathrm{v}_i)}{f(p\mathrm{v}_i)}$$

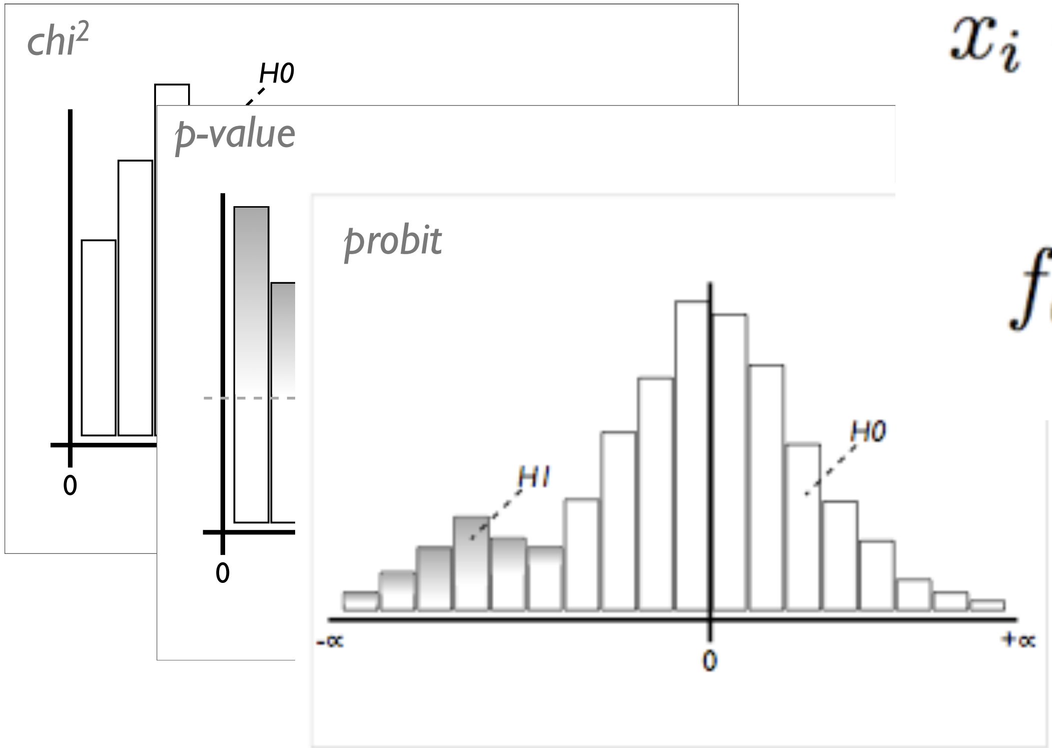
(Efron 04)

# Local FDR

- □ Local False Discovery Rate: prob of a given null hypothesis to be true

$$\mathrm{fdr}_i = \mathbb{P}\left(H = H0 | \mathcal{S} = \mathcal{S}_i\right)$$

- □ Mixture model: general and statistically convenient framework



$$x_i = \mathrm{probit}(pv_i) = \Phi^{-1}(pv_i),$$

$$f_{\theta_j}(x_i) = \frac{1}{\sigma_j\sqrt{2\pi}} e^{\frac{-(x_i - \widehat{\mu}_j)^2}{2(\sigma_j)^2}},$$

$$f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

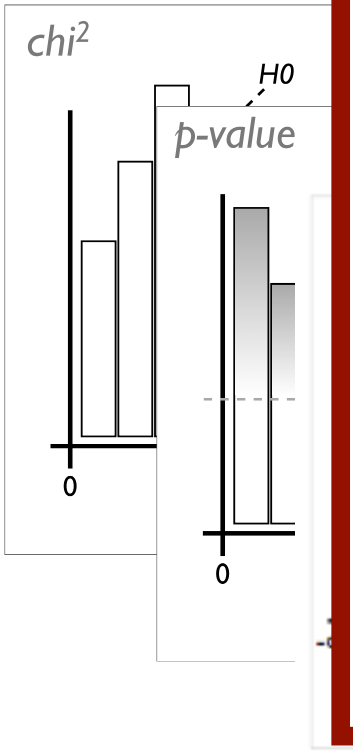
$$f_1 = \mathcal{N}(\mu_1, \sigma_1)$$

(Efron 04)

# Local FDR

□ Local ... to be true

□ Mixt ... work



$(pv_i),$

$-\hat{\mu}_j)^2$ ... $_j)^2$ ,

- **EM algorithm**
- **fully parametric**

(unknown parameters are estimated at the same time)

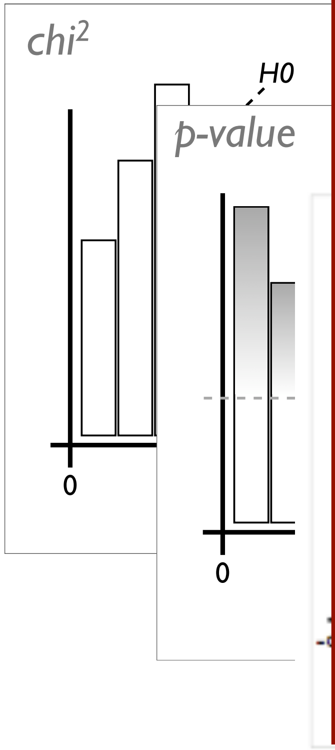
- **easy to implement** (in R)
- **fast**

(Efron 04)

# Local FDR

□ Local … to be true

□ Mixt… ork



$(pv_i),$

$-\hat{\mu}_j)^2$ $j)^2$ ,

- **EM algorithm**
- **fully parametric**

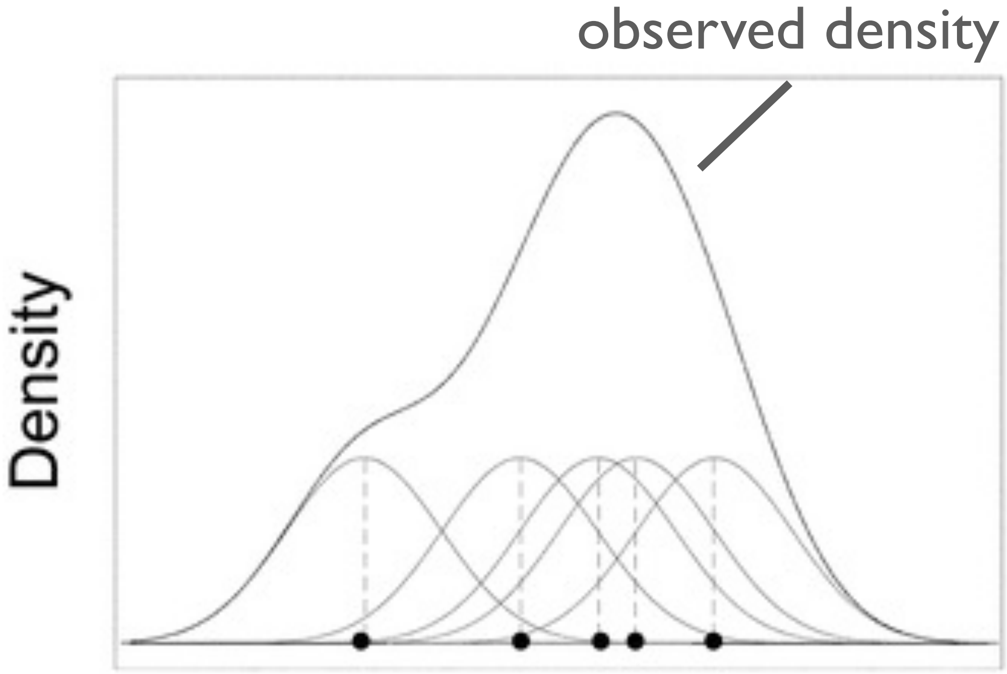
(unknown parameters are estimated at the same time)

- **easy to implement** (in R)
- **fast**
- **tests must be independent**
- **Gaussian assumption reasonable for $H_0$ but not for $H_1$**

# kerfdr

- □ Kernel-based alternative: non-parametric estimation of $f_1$ by convolving the data with a kernel

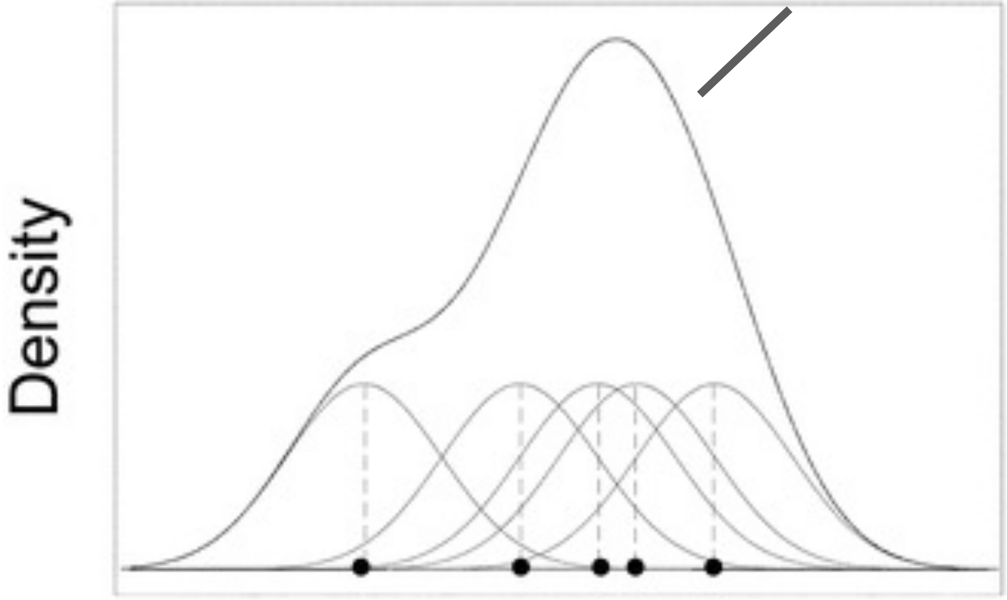
2 parameters



# kerfdr

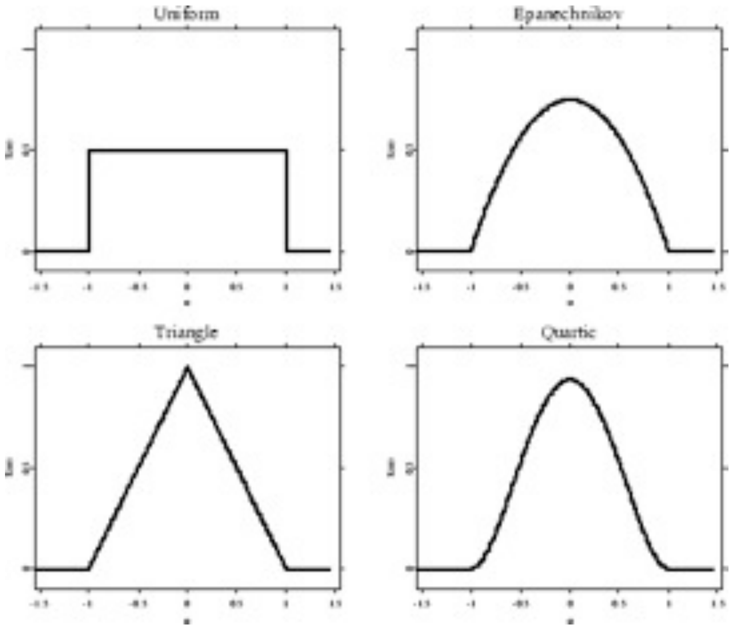
- Kernel-based alternative: non-parametric estimation of $f_1$ by convolving the data with a kernel

2 parameters



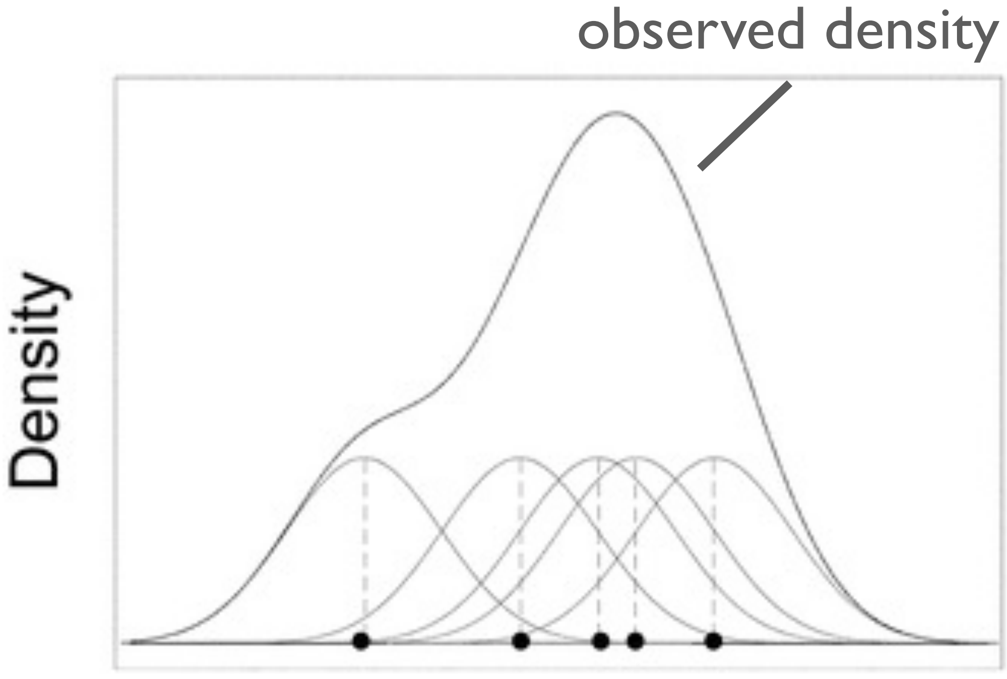
observed density

- kernel function (shape)



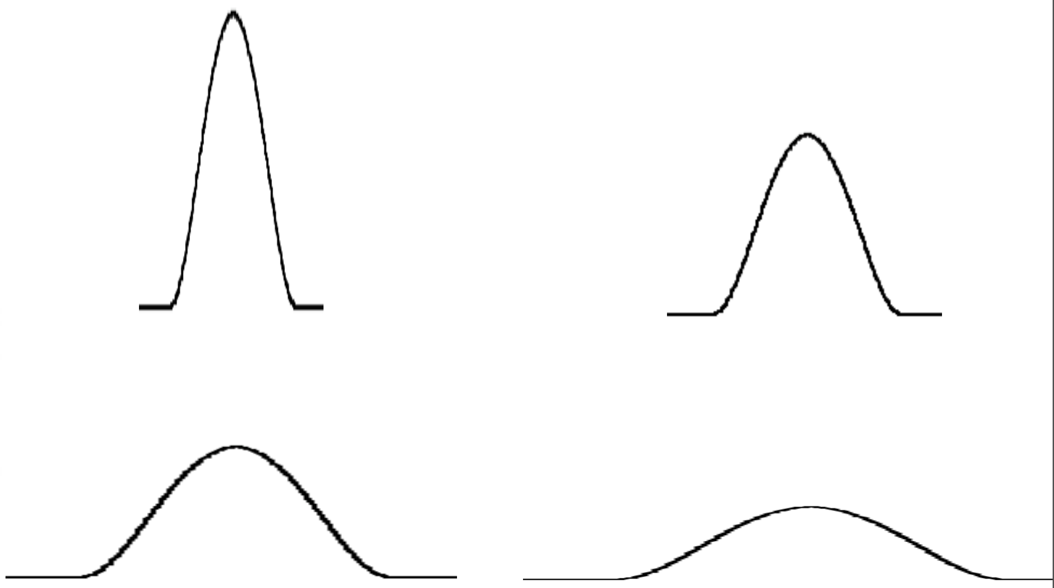
# kerfdr

- □ Kernel-based alternative: non-parametric estimation of $f_1$ by convolving the data with a kernel



2 parameters

- kernel function (shape)
- bandwidth (smoothing)

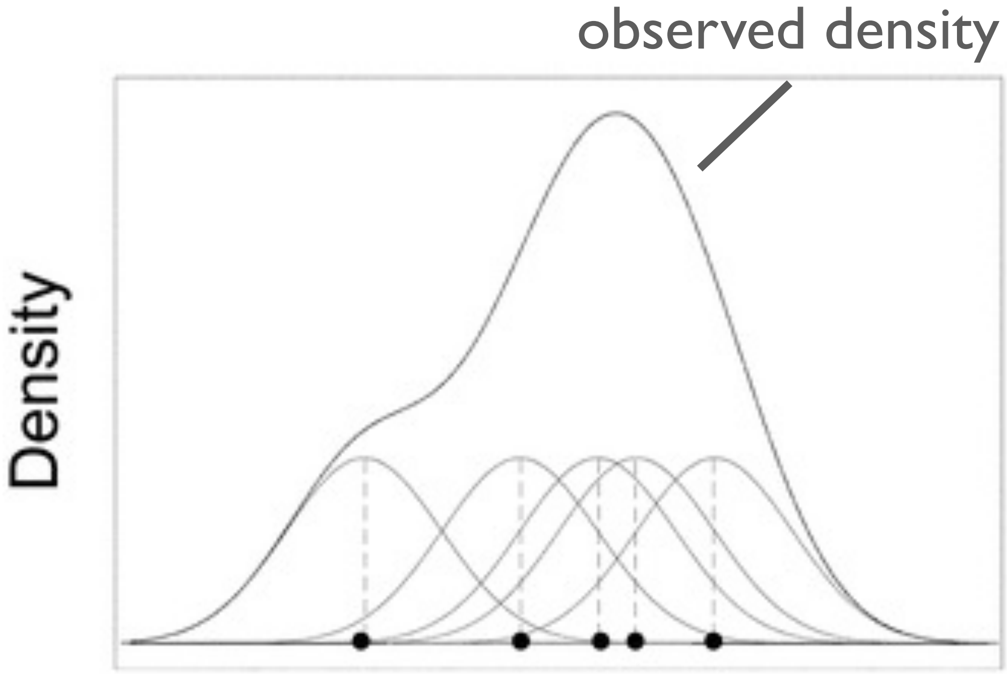


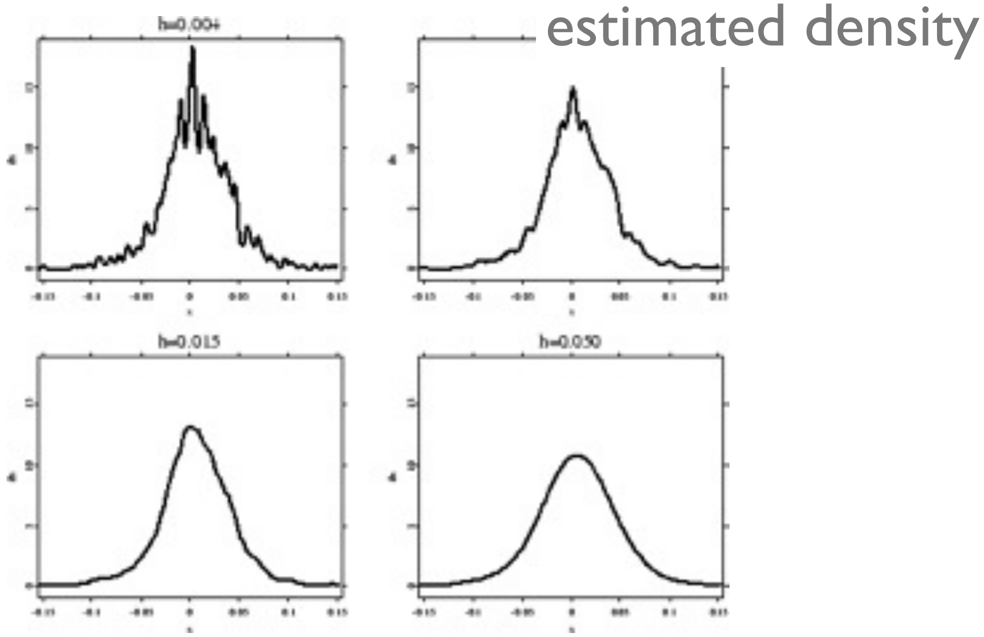
# kerfdr

- **Kernel-based alternative:** non-parametric estimation of $f_I$ by convolving the data with a kernel

2 parameters

- kernel function (shape)
- bandwidth (smoothing)





(Robin et al 07)

# kerfdr

- Local fdr kernel-based estimation:

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

(Robin et al 07)

# kerfdr

- Local fdr kernel-based estimation:

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

local FDR

$$\widehat{\tau}_{i0} = \widehat{\pi}_0 f_0(x_i) / \widehat{f}(x_i),$$

(Robin et al 07)

# kerfdr

- Local fdr kernel-based estimation:

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

local FDR

kernel function

$$\widehat{\tau}_{i0} = \widehat{\pi}_0 f_0(x_i)/\widehat{f}(x_i),$$

bandwidth

$$\widehat{f}_1(x) = \left[\sum_{i=1}^{n} \frac{1-\widehat{\tau}_{i0}}{h} k\left(\frac{x-x_i}{h}\right)\right] \Big/ \left(n - \sum_{j=1}^{n} \widehat{\tau}_{j0}\right)$$

(Robin et al 07)

# kerfdr

- Local fdr kernel-based estimation:

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

$$\widehat{\tau}_{i0} = \widehat{\pi}_0 f_0(x_i)/\widehat{f}(x_i),$$

$$\widehat{f}_1(x) = \left[\sum_{i=1}^{n} \frac{1-\widehat{\tau}_{i0}}{h} k\left(\frac{x-x_i}{h}\right)\right] / \left(n - \sum_{j=1}^{n} \widehat{\tau}_{j0}\right)$$

(Robin et al 07)

# kerfdr

- Local fdr kernel-based estimation:

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

$$\widehat{\tau}_{i0} = \widehat{\pi}_0 f_0(x_i) / \widehat{f}(x_i),$$

$$\widehat{f}_1(x) = \left[\sum_{i=1}^{n} \frac{1-\widehat{\tau}_{i0}}{h} k\left(\frac{x - x_i}{h}\right)\right] / \left(n - \sum_{j=1}^{n} \widehat{\tau}_{j0}\right)$$

(Robin et al 07)

# kerfdr

- Local fdr kernel-based estimation:

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

iterative algorithm
(EM-like)

$$\widehat{\tau}_{i0} = \widehat{\pi}_0 f_0(x_i) / \widehat{f}(x_i),$$

$$\widehat{f}_1(x) = \left[ \sum_{i=1}^{n} \frac{1 - \widehat{\tau}_{i0}}{h} k\left( \frac{x - x_i}{h} \right) \right] \Big/ \left( n - \sum_{j=1}^{n} \widehat{\tau}_{j0} \right)$$

# kerfdr

- Local fdr kernel-based estimation:
  - Semi-parametric.
  - Does not require any assumption on the H1 distribution ($f_1$).
  - Provides more realistic estimates.
  - $\pi_0$, $h$ and $k$ must be pre-determined.
  - Tests must be independent.

# kerfdr

□ Implementation

- Estimation of $\pi_0$
- Determination of the bandwitdh
- Computation of $f_l$
- Semi-supervised situations
- Truncated distributions

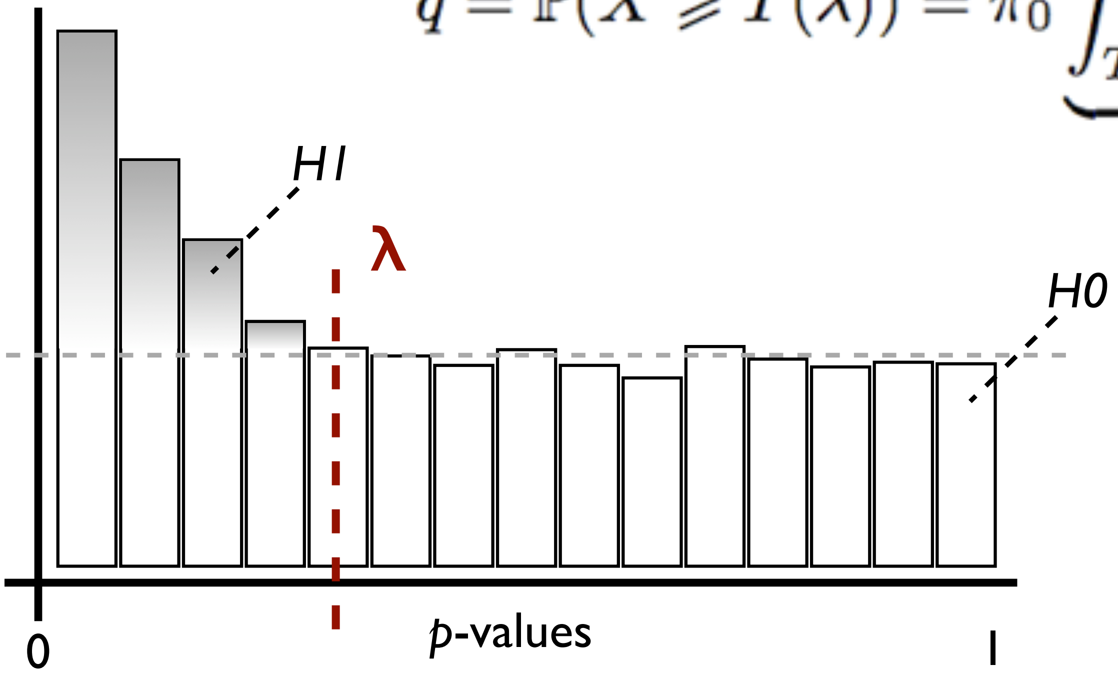
practical generalizations

(Storey 01)

# kerfdr

- □ Implementation
- ▸ Estimation of $\pi_0$
- □ Many methods already implemented

$$q = \mathbb{P}(X \geqslant T(\lambda)) = \pi_0 \underbrace{\int_{T(\lambda)}^{+\infty} f_0(x)dx}_{q_0} + \pi_1 \underbrace{\int_{T(\lambda)}^{+\infty} f_1(x)dx}_{q_1}$$



$$\pi_0 = \frac{q - q_1}{q_0 - q_1}$$

→ $$\widehat{\pi_0} = \frac{\widehat{q}}{1 - \lambda}$$

(Sheather and Jones 91)
(Silverman 86)
(Scott 92)

# kerfdr

- □ Implementation
- ▸ Determination of the bandwidth
- □ Many methods already implemented:
  - □ Biased and unbiased cross-validation estimations.
  - □ Derivative-based methods.

(Silverman 82)

# kerfdr

- Implementation
- Computation of $\widehat{f}_1(x)$
  - Naive computation requires a quadratic complexity.
  - Discrete convolution through Fast-Fourier-Transforms allows a far more efficient linear complexity.

$$\widehat{f}_1(x) = \left[\sum_{i=1}^{n} \frac{1-\widehat{\tau}_{i0}}{h} k\left(\frac{x - x_i}{h}\right)\right] / \left(n - \sum_{j=1}^{n} \widehat{\tau}_{j0}\right) .$$

# kerfdr

- Implementation
- Semi-supervised situations
  - Among the null hypotheses ➽ some are known to be true while other are known to be false (control-genes).
  - Prior information is taken into account in the estimation procedure.
  - Known local FDR $\tau_{i0}$ are kept fixed: contribute to the estimation for the other observations / not updated at each step of the algorithm.

# kerfdr

- Implementation

nb of MC simulations

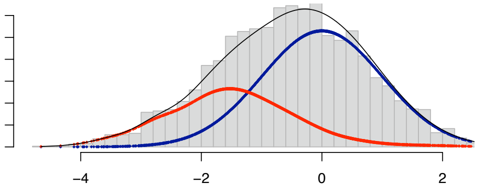
- Truncated distributions within an interval $I$
  - e.g. : $p$-values computed by Monte-Carlo → $p$-values > $1/S$
  - the restrictions of $f_1$, $f_0$ and $f$ to $I$ need to be normalized.

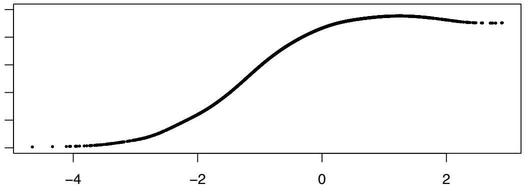
$$q = \int_I f(x)dx = \pi_0 \underbrace{\int_I f_0(x)dx}_{q_0} + \pi_1 \underbrace{\int_I f_1(x)dx}_{q_1}$$

# kerfdr

- Implementation

- R package ‘kerfdr’
  - Simple and straightforward to use
  - Many options for more advanced users
  - Fast thanks to Fast-Fourier-Transforms
  - Includes the estimation of $\pi_0$ and of the bandwidth
  - Handles semi-supervised situations and truncated distributions
  - Produces graphics





# kerfdr

- Application 1: simulations
  - $p$-values simulated according to the mixture model
  - $f_0$ is the uniform distribution over [0,1]
  - 4 proportions of null hypotheses: $\pi_0 = 0.99 / 0.95 / 0.90 / 0.70$
  - $f_1$ is either an exponential $\varepsilon(\mu_1)$ or a uniform distribution over $[0,2\mu_1]$
  - 2 different means for $f_1$: $\mu_1 = 0.01 / 0.001$
  - Number of observations: $n$ = 1,000
  - Number of simulations: $S$ = 500

# kerfdr

## □ Application 1: simulations

- $p$-values simulated according to the mixture model
- $f_0$ is the uniform distribution over [0,1]
- 4 proportions of null hypotheses: $\pi_0$ = 0.99 / 0.95 / 0.90 / 0.70
- $f_1$ is either an exponential $\varepsilon(\mu_1)$ or a uniform distribution over $[0,2\mu_1]$
- 2 different means for $f_1$: $\mu_1$ = 0.01 / 0.001
- Number of observations: $n$ = 1,000
- Number of simulations: $S$ = 500
- Performances are assessed by means of the Root-Mean-Square Error :

$$RMSE(\pi_0, f) = \frac{1}{S}\sum_s \sqrt{\frac{1}{n}\sum_i (\widehat{\tau}_i^s - \tau_i)^2}.$$

expected value

estimated value

# kerfdr

## Application 1: simulations

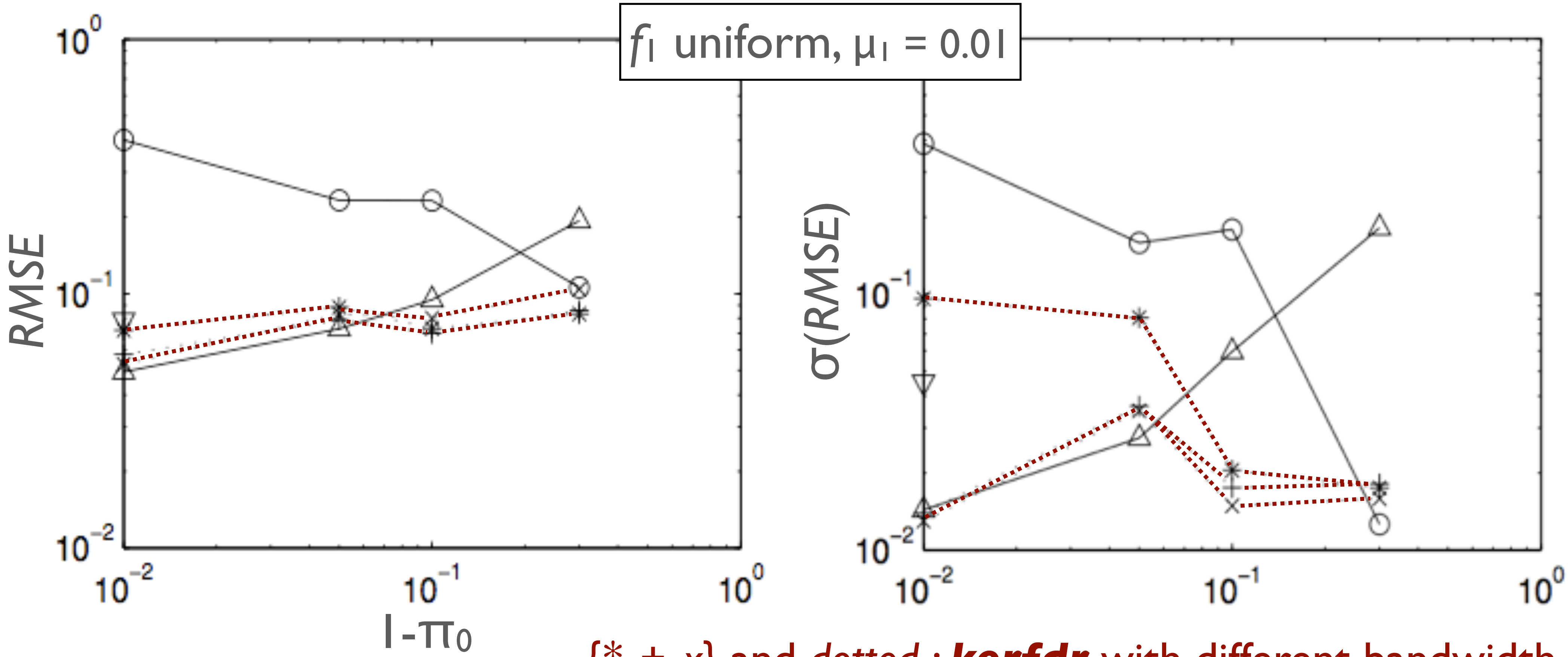
- *p*-values simulated according to the mixture model
- $f_0$ is the uniform distribution over [0,1]
- 4 proportions of null hypotheses: $\pi_0$ = 0.99 / 0.95 / 0.90 / 0.70
- $f_1$ is either an exponential $\varepsilon(\mu_1)$ or a uniform distribution over $[0,2\mu_1]$
- 2 different means for $f_1$: $\mu_1$ = 0.01 / 0.001
- Number of observations: $n$ = 1,000
- Number of simulations: $S$ = 500
- Performances are assessed by means of the Root-Mean-Square Error :

$$RMSE(\pi_0, f) = \frac{1}{S}\sum_s \sqrt{\frac{1}{n}\sum_i (\widehat{\tau}_i^s - \tau_i)^2}.$$

- **The smaller the *RMSE*, the better the performances.**

# kerfdr

- **Application 1:** comparison with existing methods



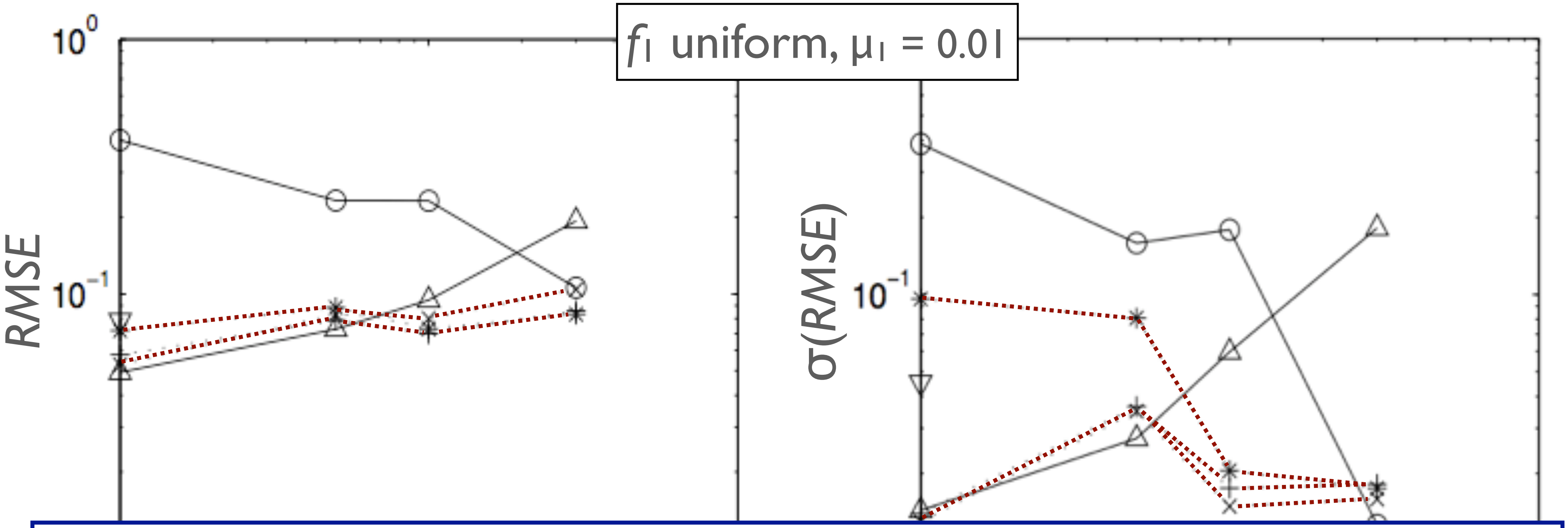
{*, +, *x*} and *dotted* : ***kerfdr*** with different bandwidth

-Δ- : Splines-based density estimation (Efron 04)

-O- : EM 2-components Gaussian mixture model (McLachlan et al 06)

# kerfdr

- Application 1: comparison with existing methods

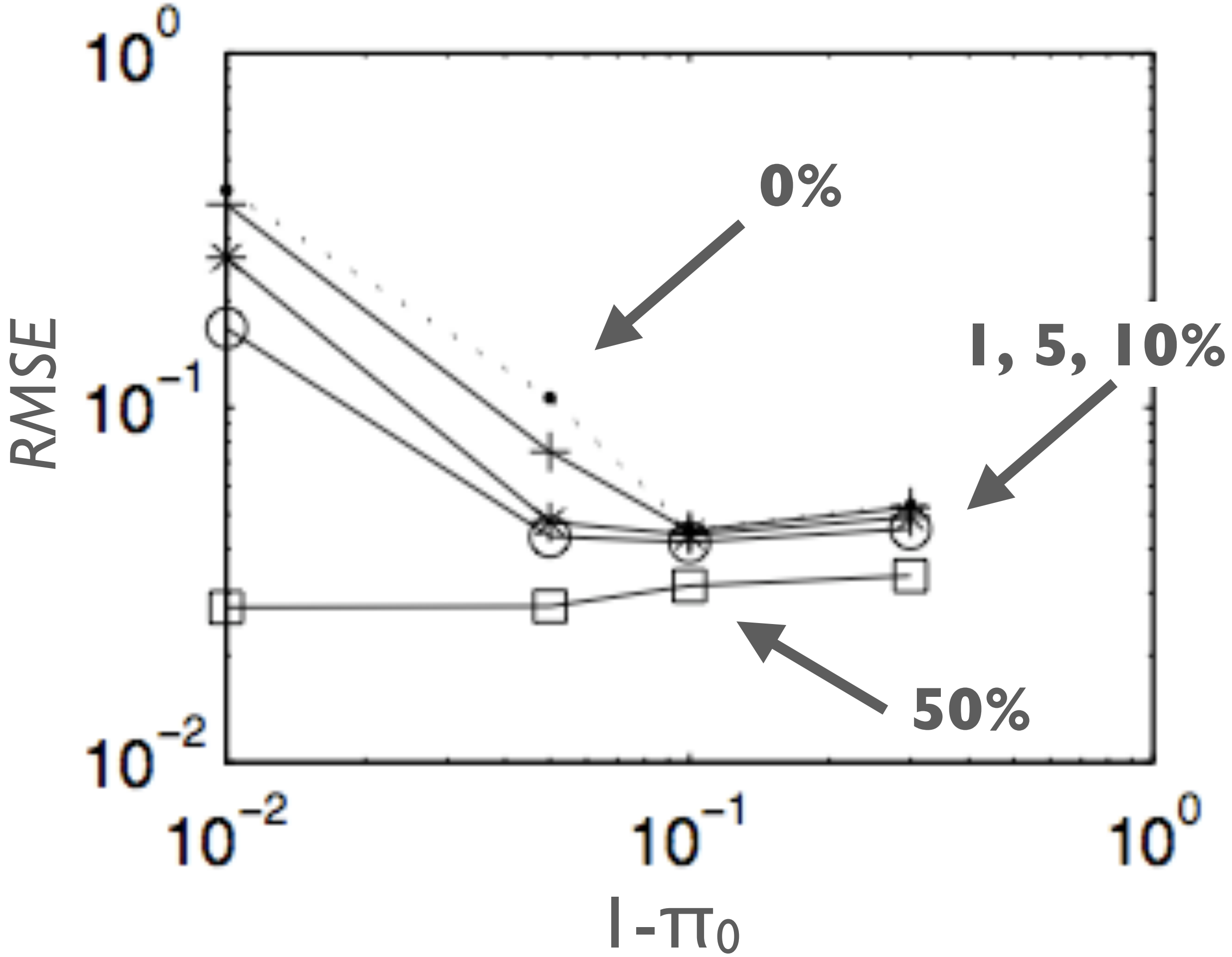


$f_1$ uniform, $\mu_1 = 0.01$

- Estimates of *kerfdr* not very sensitive to the bandwidth
- *kerfdr* performs as well the other methods when $f_0$ and $f_1$ are well separated ($\mu_1 = 0.001$, data not shown)
- It outperforms them in more difficult situations ($\mu_1 = 0.01$) especially in terms of stability.

# kerfdr

- □ Application 1: semi-supervised : from 0% to 50% of known hypotheses



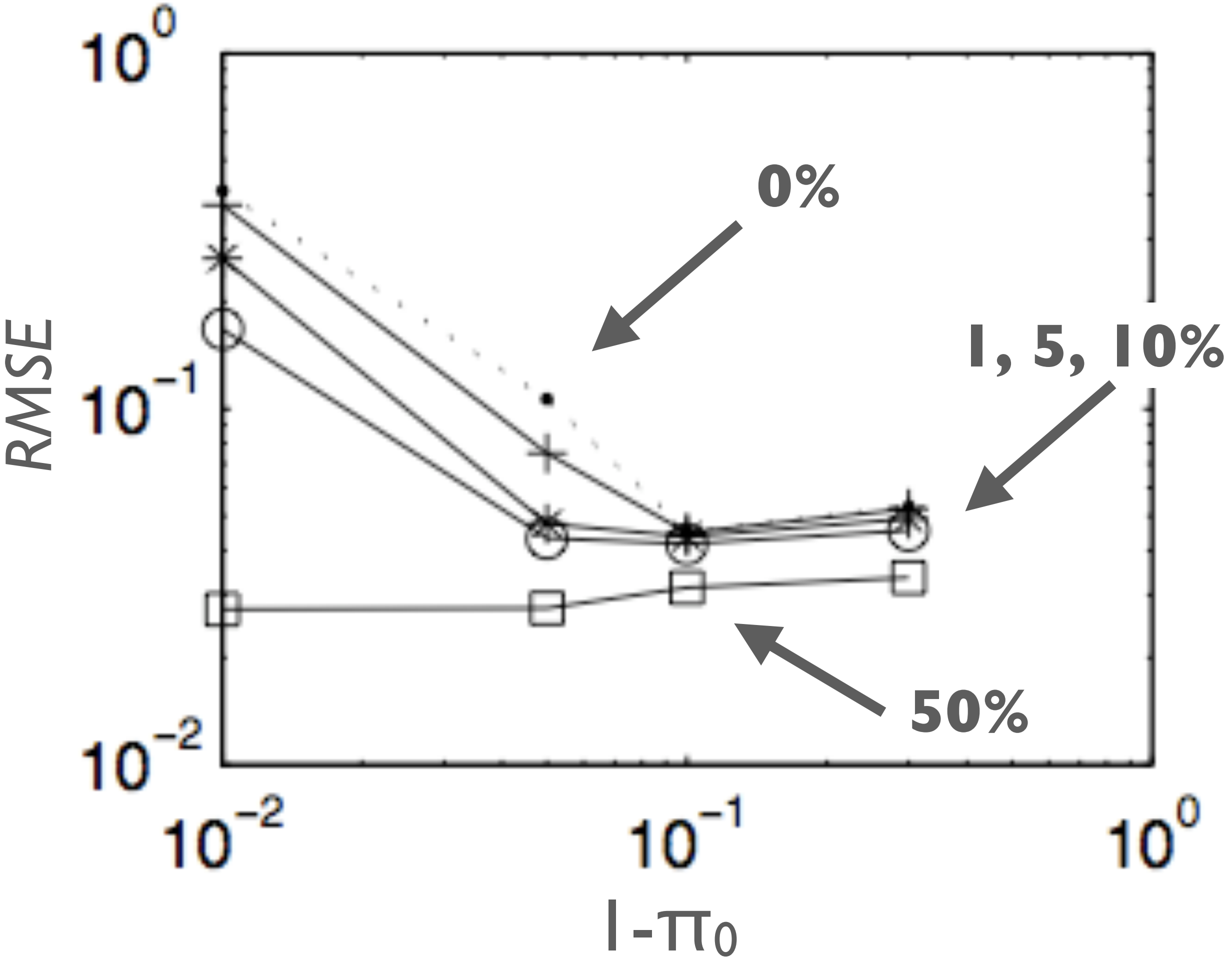
0%

1, 5, 10%

50%

# kerfdr

□ Application 1: semi-supervised : from 0% to 50% of known hypotheses



0%

1, 5, 10%

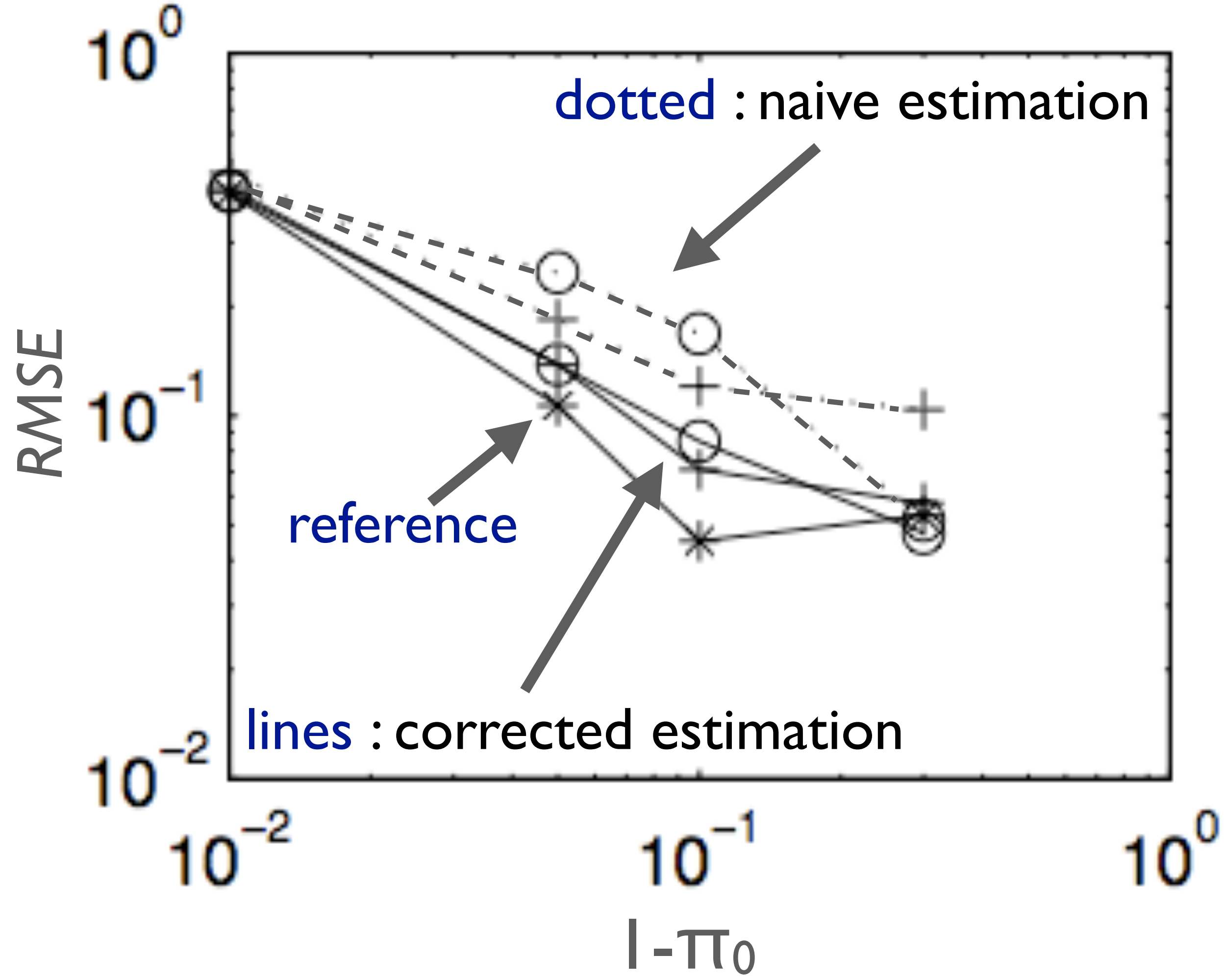
50%

The proportion of known hypotheses improves the estimates.

Even a small proportion of 1 or 5 % !!!

# kerfdr

- □ Application 1: truncated distributions : $p$-value are truncated to a given threshold $p^*$



dotted : naive estimation

reference

lines : corrected estimation

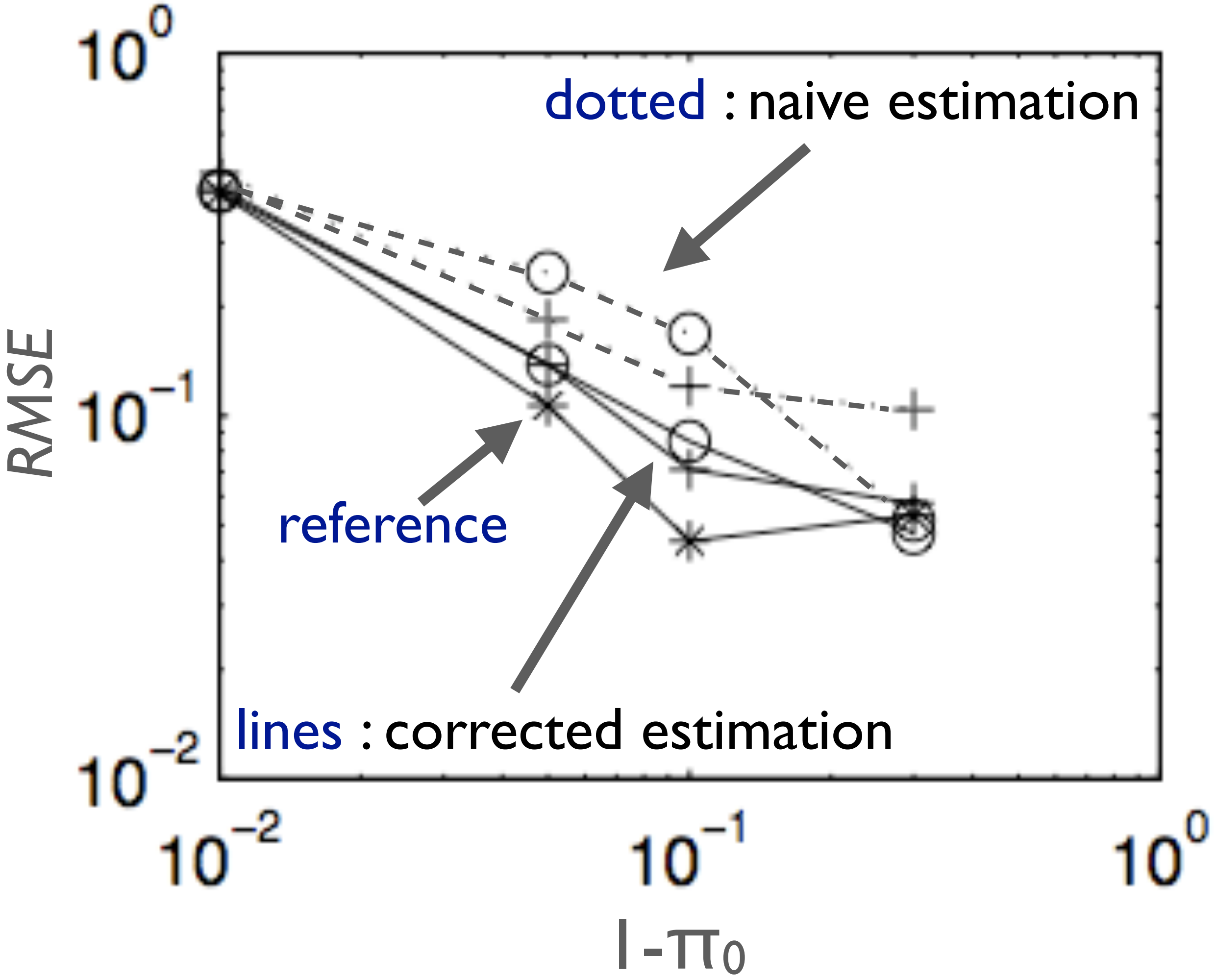
* : $p^* = 0$ (reference)

o : $p^* = 10^{-3}$

+ : $p^* = 10^{-2}$

# kerfdr

- □ Application 1: truncated distributions : *p*-value are truncated to a given threshold *p**



dotted : naive estimation

reference

lines : corrected estimation

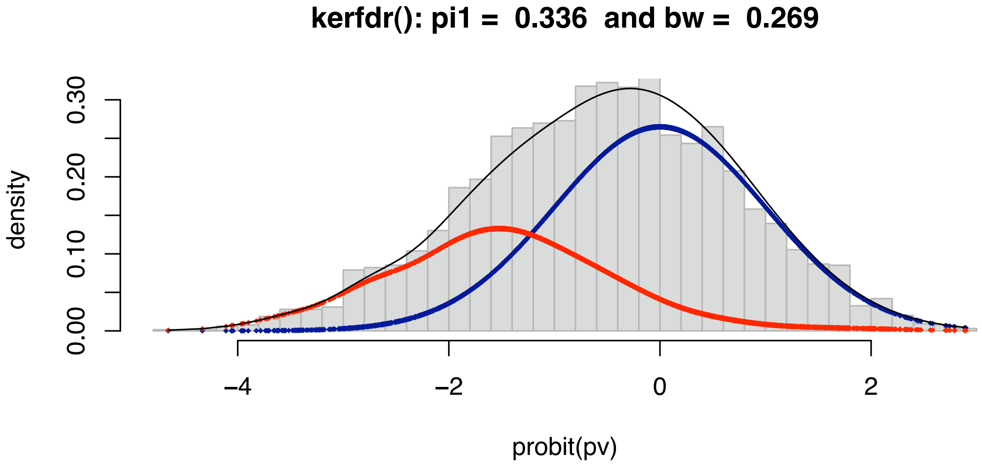
Correction improves the quality of the estimates.

Corrected estimates almost as good as the untruncated reference !!!

(Hedenfalk et al 01)

# kerfdr

- **Application 2:** differential gene-expressions
  - 3,226 genes studied among two groups of BRCA1 (7 patients) and BRCA2 (8 patients).
  - Test: t test-like statistic (Delmar et al 05).

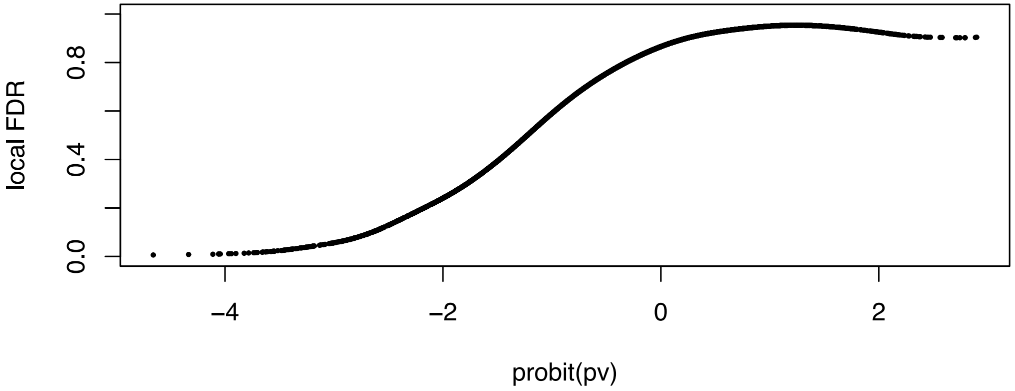


🟦 $f_0(x)$

🟥 $f_1(x)$

⬛ $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$

- 1 - $\pi_0$ = 0.336
- # of genes < 1% = 5
- running time < 1 sec



(data provided by Merck-Serono)

# kerfdr

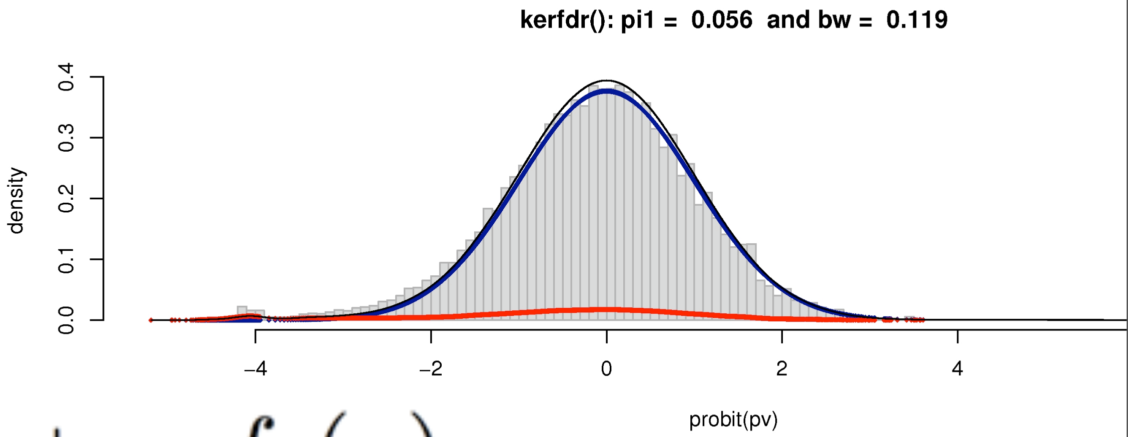
## Application 3: genome-wide association

- 203 controls from Rennes genotyped using a 100K Affy (100,000 SNPs covering the genome).
- Test: Hardy-Weinberg equilibrium test.

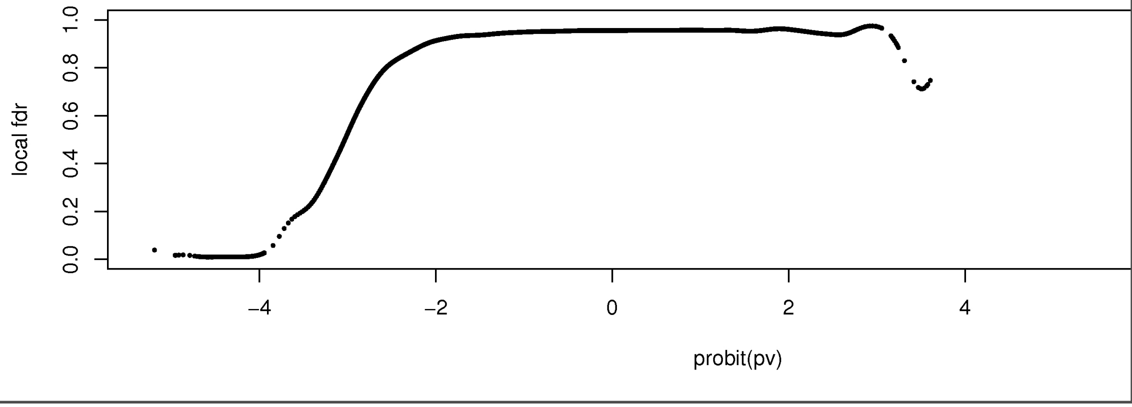
🟦 $f_0(x)$

🟥 $f_1(x)$

⬛ $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$



- $1 - \pi_0 = 0.056$
- # of SNPs < 1% = 29
- running time < 3 sec



# kerfdr

- Algorithm available *via* the CRAN or at

http://stat.genopole.cnrs.fr/software/kerfdr

- Guedj et al. *kerfdr: a semi-parametric kernel-based approach to local FDR estimations*. BMC Bioinfo. 2009
- Strimmer. *A unified approach to FDR estimation.* BMC Bioinfo. 2008

# II. Local replications / local score

with G Nuel (Univ Paris V), B Prum (Univ Evry), J Wojcik (Merck-Serono)

# Introduction

- Replication in independent populations as the gold standard for results validation.
- Performed at the marker or haplotypic level.

(Ioannidis 03)

# Introduction

- Replication in independent populations as the gold standard for results validation.
- Performed at the marker or haplotypic level.
- However replications are still difficult to obtain:

(Ioannidis 03)

# Introduction

- Replication in independent populations as the gold standard for results validation.
- Performed at the marker or haplotypic level.
- However replications are still difficult to obtain:
  - Lack of Power
  - Multiple-Testing
  - Genotyping Error, Missing Values
  - Population Stratifications

# Introduction

- Beside these study-design and data-analysis related factors ...
- ... inconsistent findings might also result from real biological differences between populations:

(Lohmueller et al 05)

# Introduction

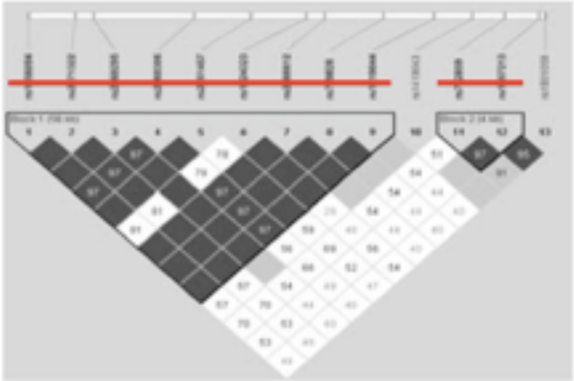
- □ Beside these study-design and data-analysis related factors ...
- □ ... inconsistent findings might also result from real biological differences between populations:

  Differences in allele frequencies.

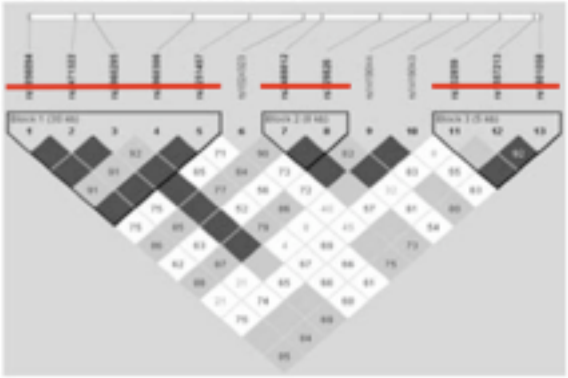
  Allele and locus heterogeneity.

  Variation in the strength of LD:

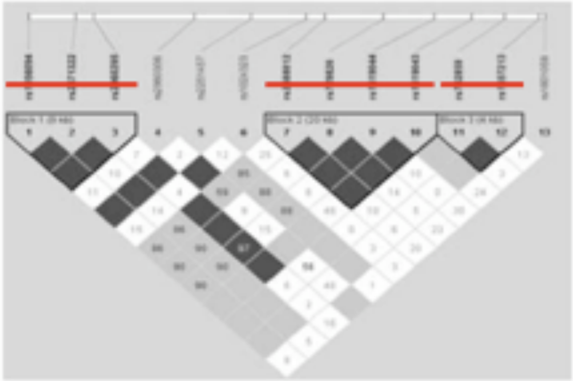
Caucasian



African-American



Asian



# Introduction

- Local Replication:

# Introduction

- □ Local Replication:
- □ We expect to observe an accumulation of high statistics of association around a disease susceptibility locus (DSL):

# Introduction

- Local Replication:
- We expect to observe an accumulation of high statistics of association around a disease susceptibility locus (DSL):

    Linkage Disequilibrium with surrounding markers.

    Aggregation of several DSL in a same genomic location.

# Introduction

- Local Replication:
- We expect to observe an accumulation of high statistics of association around a disease susceptibility locus (DSL):

  Linkage Disequilibrium with surrounding markers.

  Aggregation of several DSL in a same genomic location.

- Such accumulations may be locally replicated across populations ...

# Introduction

- Local Replication:
- We expect to observe an accumulation of high statistics of association around a disease susceptibility locus (DSL):

  Linkage Disequilibrium with surrounding markers.

  Aggregation of several DSL in a same genomic location.

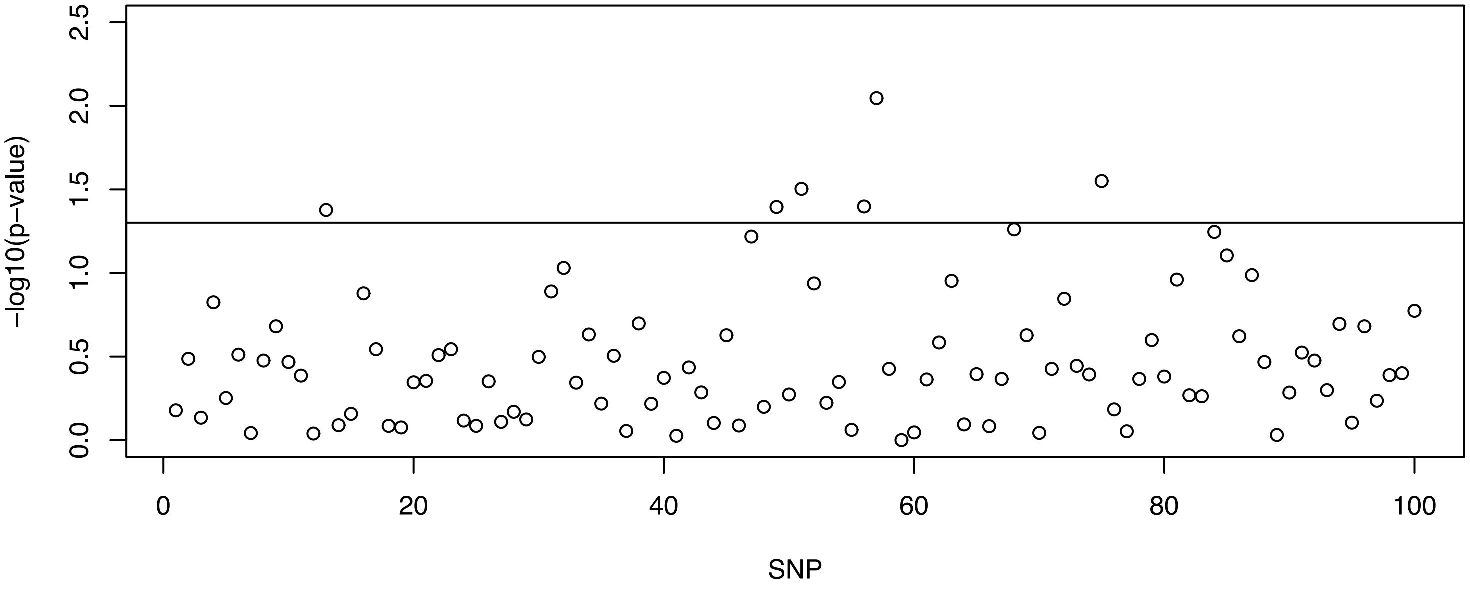
- Such accumulations may be locally replicated across populations ...
- ... without restraint about the specific allele or pattern of alleles to be replicated.

# Introduction

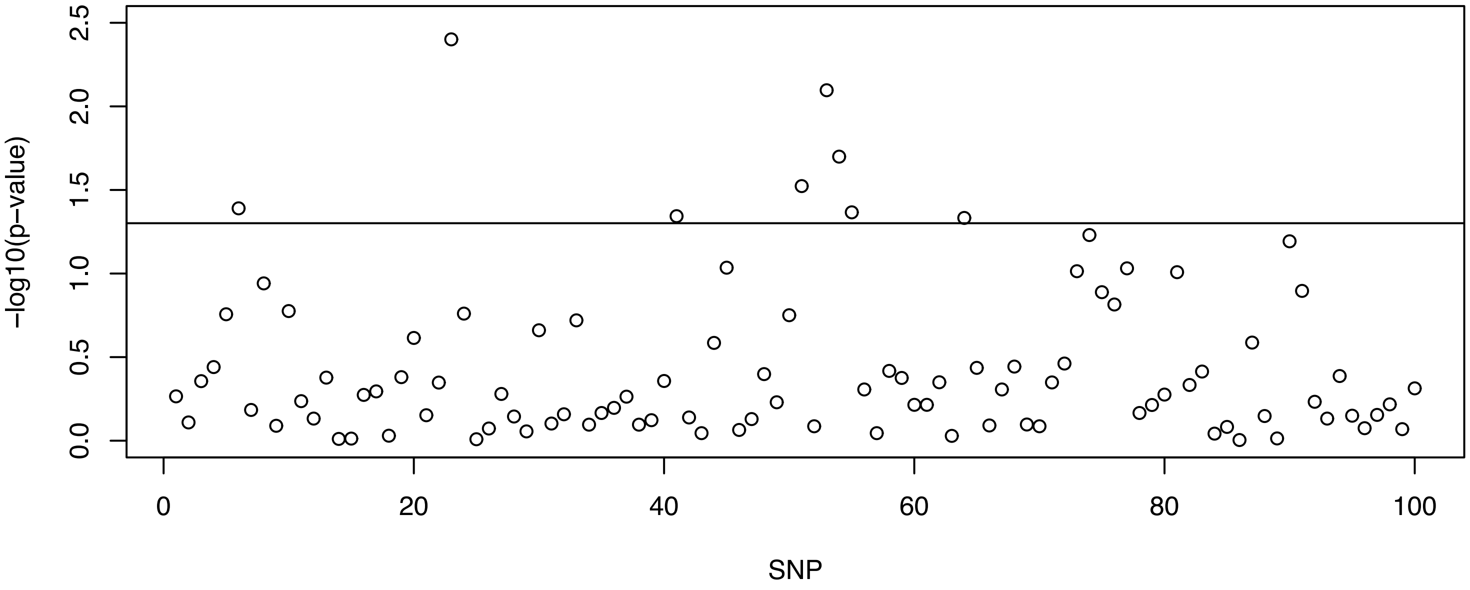
- Local Replication: **definition**

A local accumulation of high statistics of association in a given genomic region...

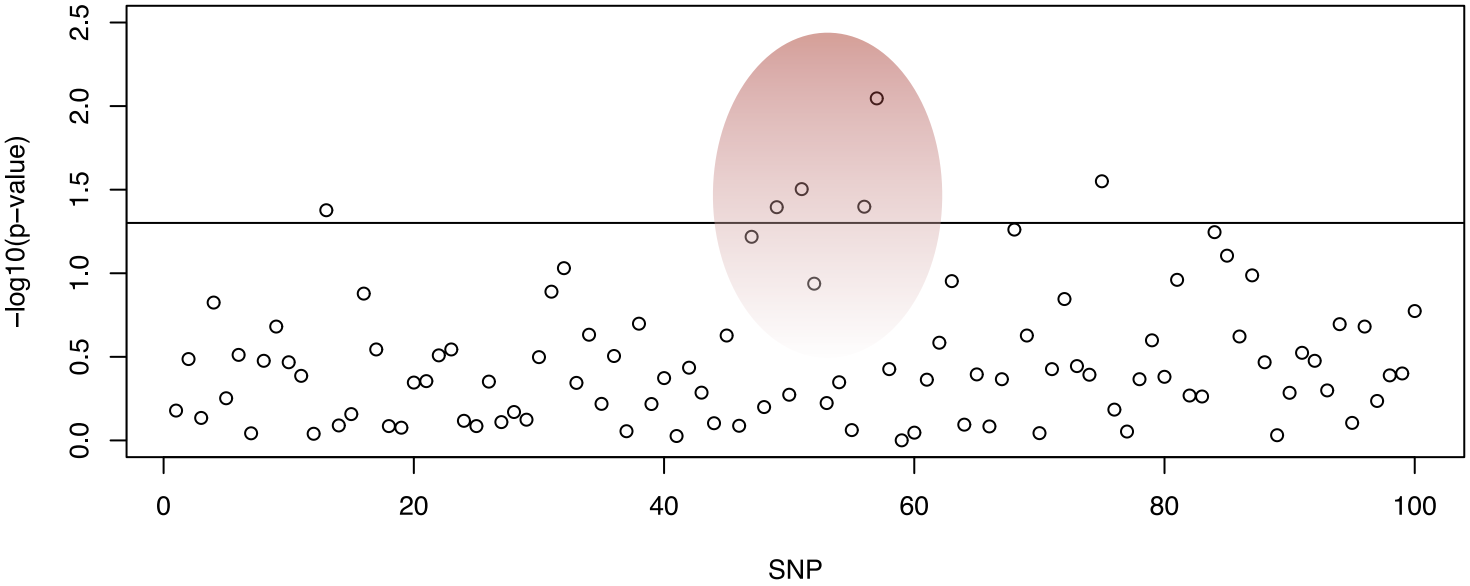
... replicated among the different populations.



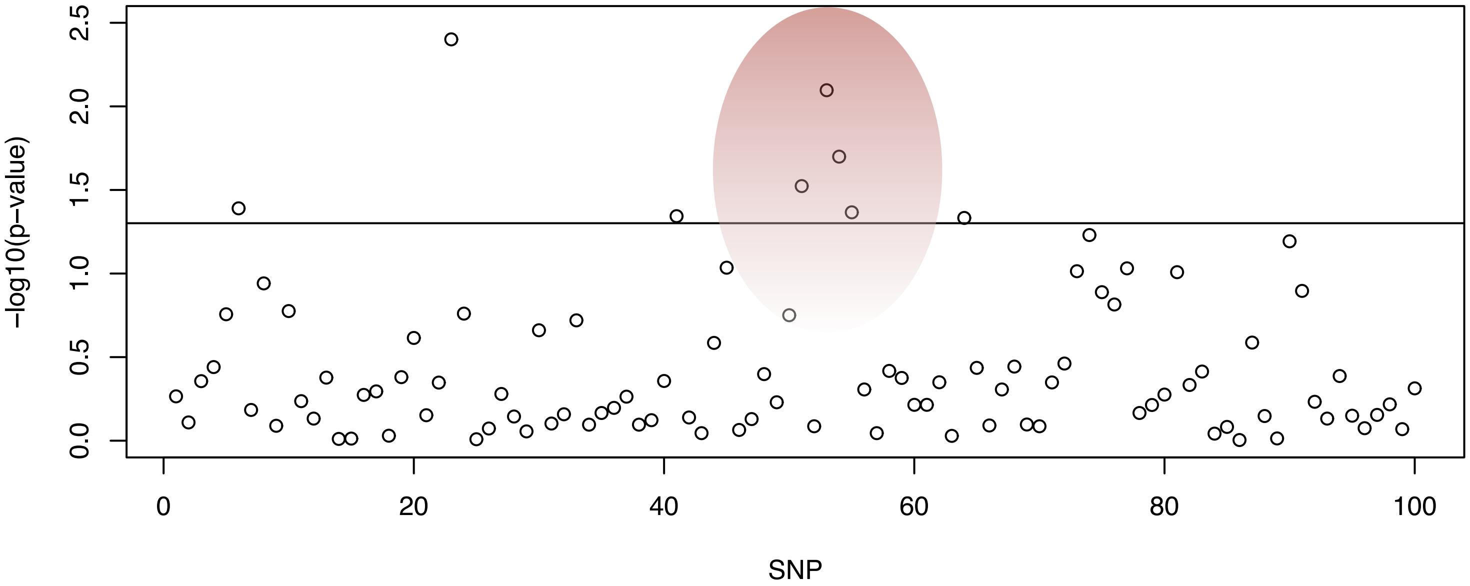
Population 1



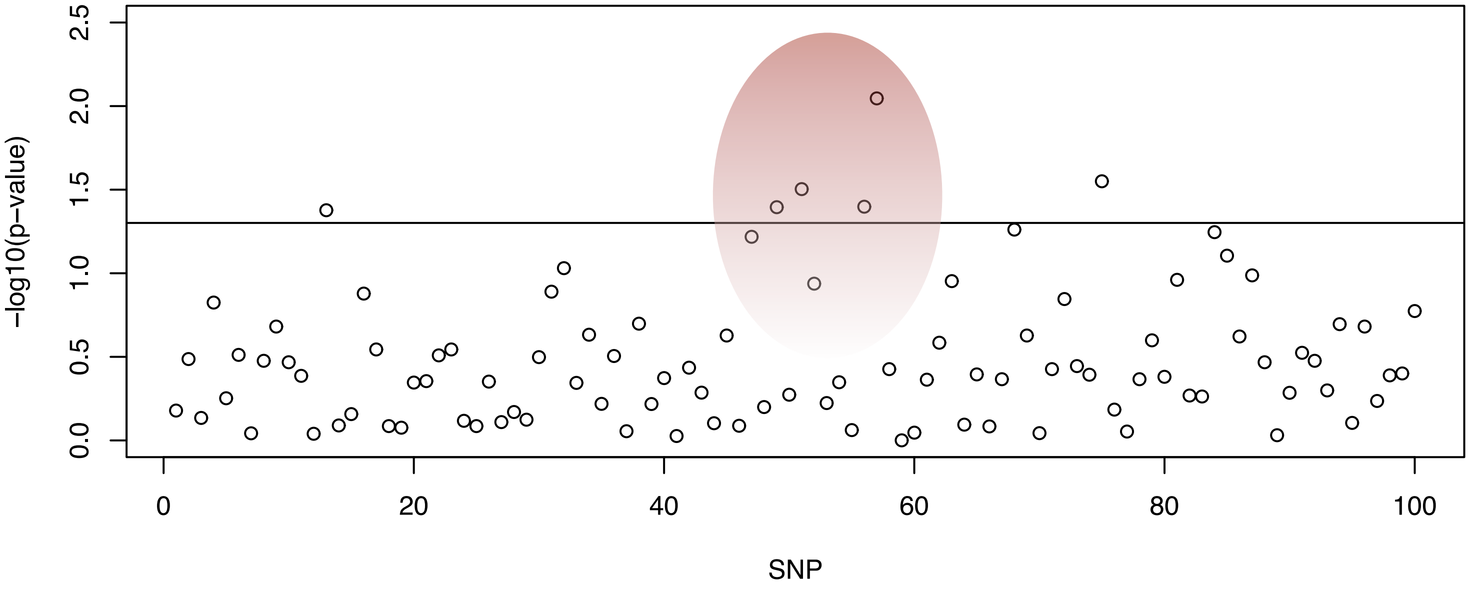
Population 2



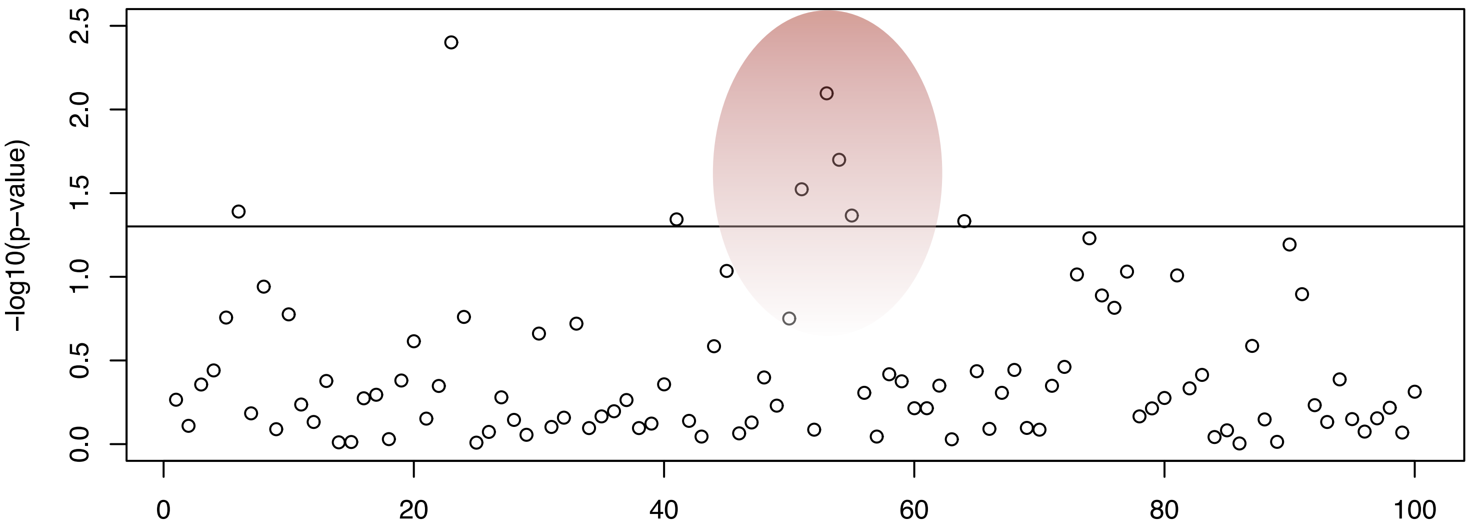
Population 1



Population 2

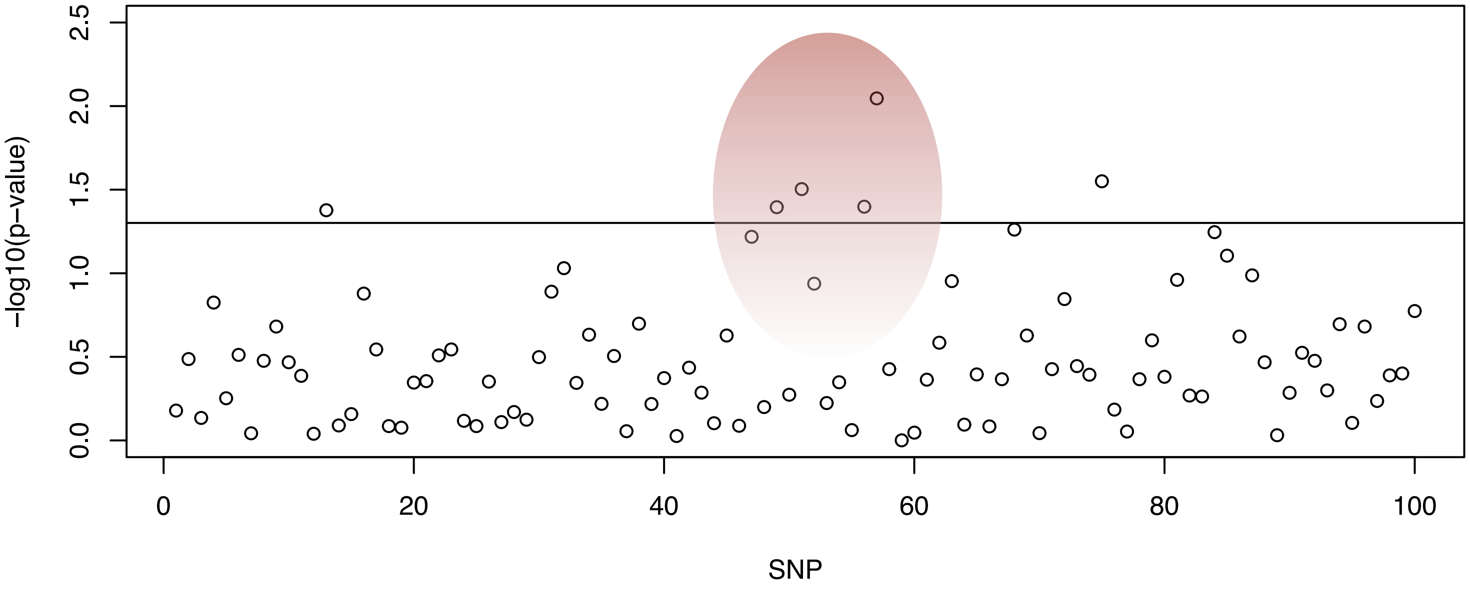


Population 1

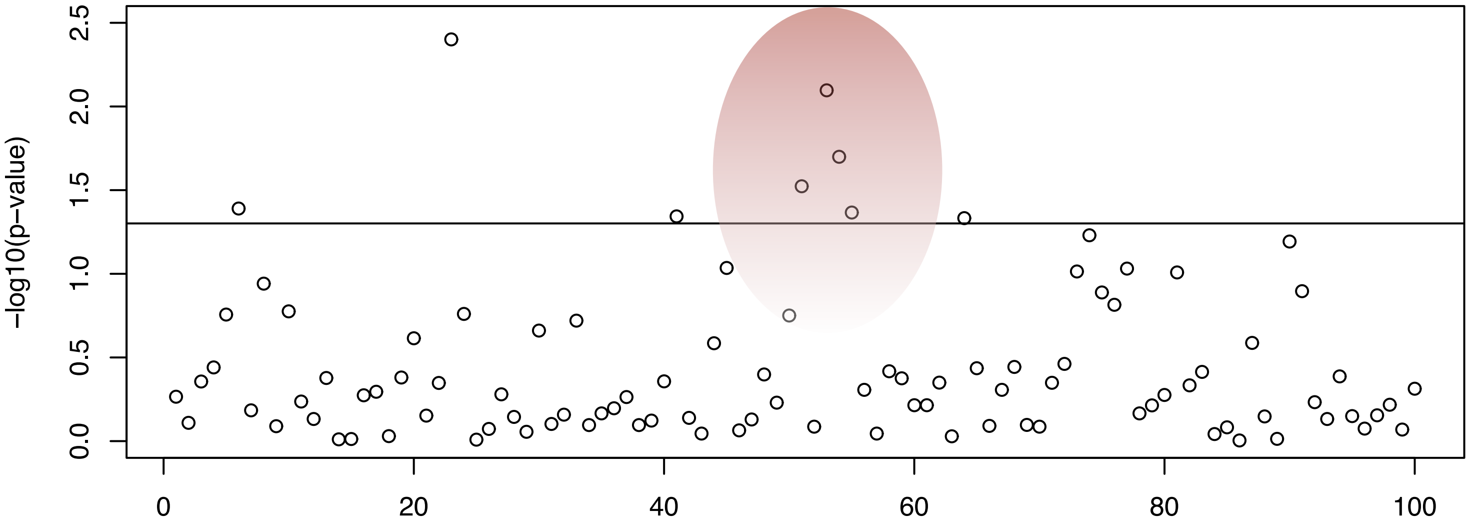


Population 2

Sliding-Frames ?! >> the frame size has to be specified



Population 1

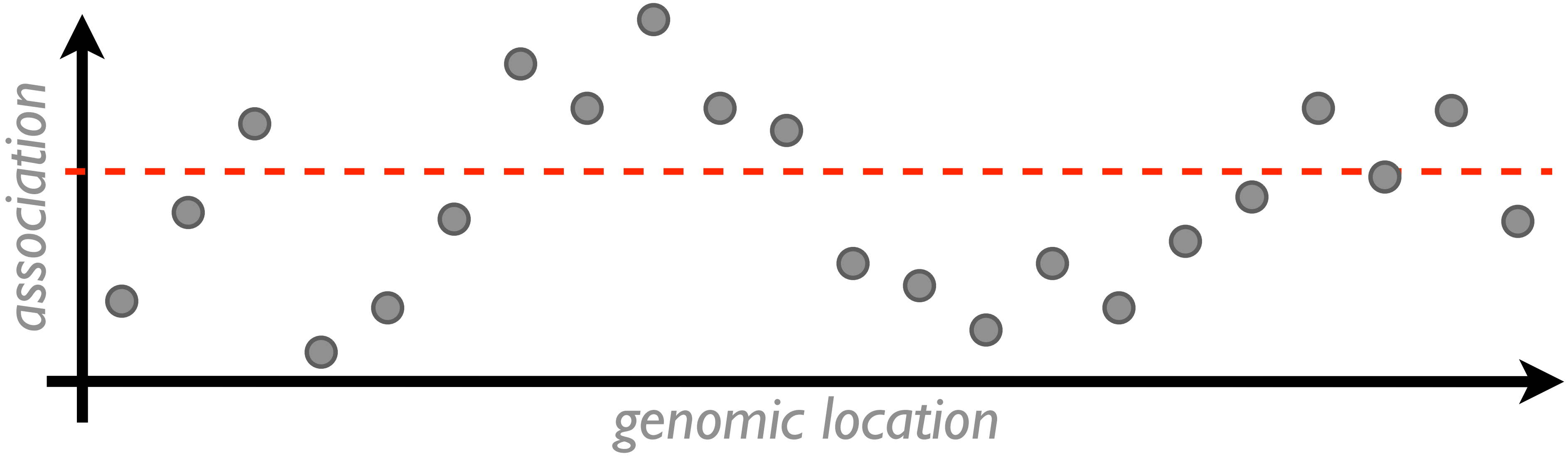


Population 2

## Sliding-Frames ?! >> Local Score

# Local Score

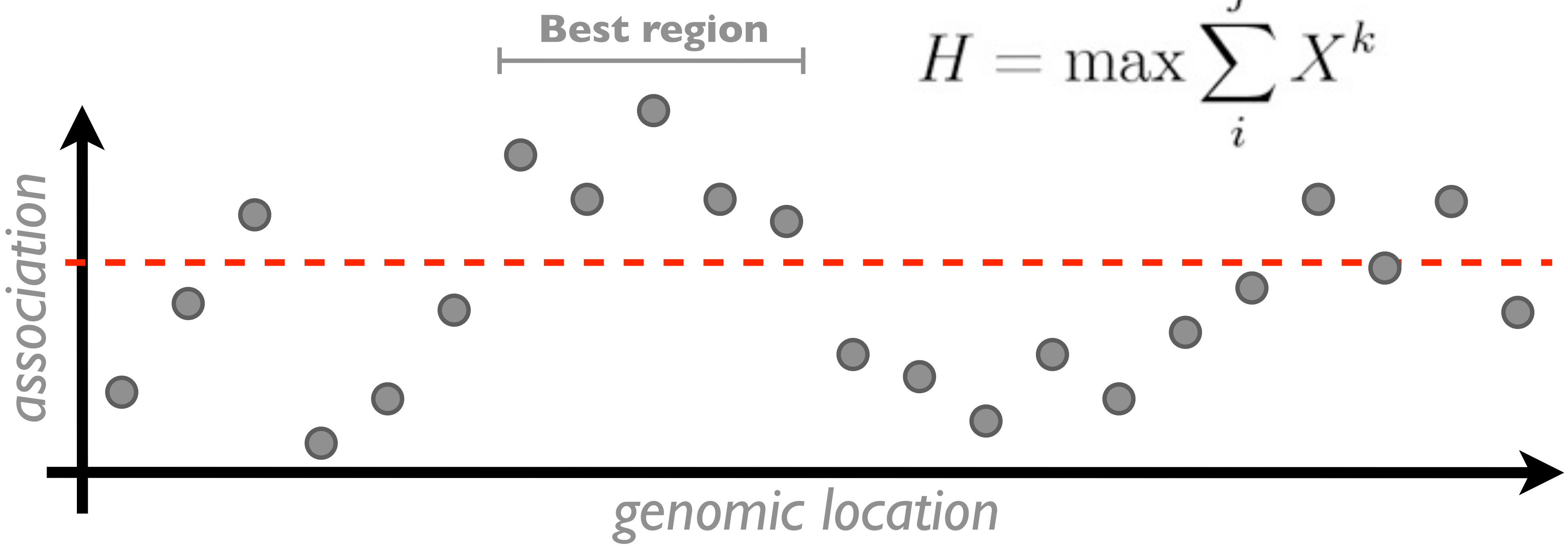
- □ Definition: Let $\mathbf{X} = (X_i)_{i = 1 \ldots n}$ be a sequence of random variables → association statistics:
  e.g. Pearson $\chi^2$ on case/control genotype frequencies.



# Local Score

- □ Definition: Let $\mathbf{X} = (X_i)_{i=1...n}$ be a sequence of random variables → association statistics:
  e.g. Pearson $\chi^2$ on case/control genotype frequencies.

$$H = \max \sum_{i}^{j} X^k$$



# Local Score

**1 -2 -4 2 1 1 -3 1 -2**

# Local Score

$$\mathbf{1} \quad \mathbf{-2} \quad \mathbf{-4} \quad \boxed{\mathbf{2} \quad \mathbf{1} \quad \mathbf{1}} \quad \mathbf{-3} \quad \mathbf{1} \quad \mathbf{-2}$$

$H = 4$

# Local Score

$$\mathbf{1} \quad \mathbf{-2} \quad \mathbf{-4} \quad \boxed{\mathbf{2} \quad \mathbf{1} \quad \mathbf{1}} \quad \mathbf{-3} \quad \mathbf{1} \quad \mathbf{-2}$$

$H = 4$

$$\mathbf{-1} \quad \mathbf{2} \quad \mathbf{1} \quad \mathbf{-4} \quad \mathbf{-2} \quad \mathbf{-2} \quad \mathbf{2} \quad \mathbf{1} \quad \mathbf{-1} \quad \mathbf{3} \quad \mathbf{1} \quad \mathbf{-2}$$

# Local Score

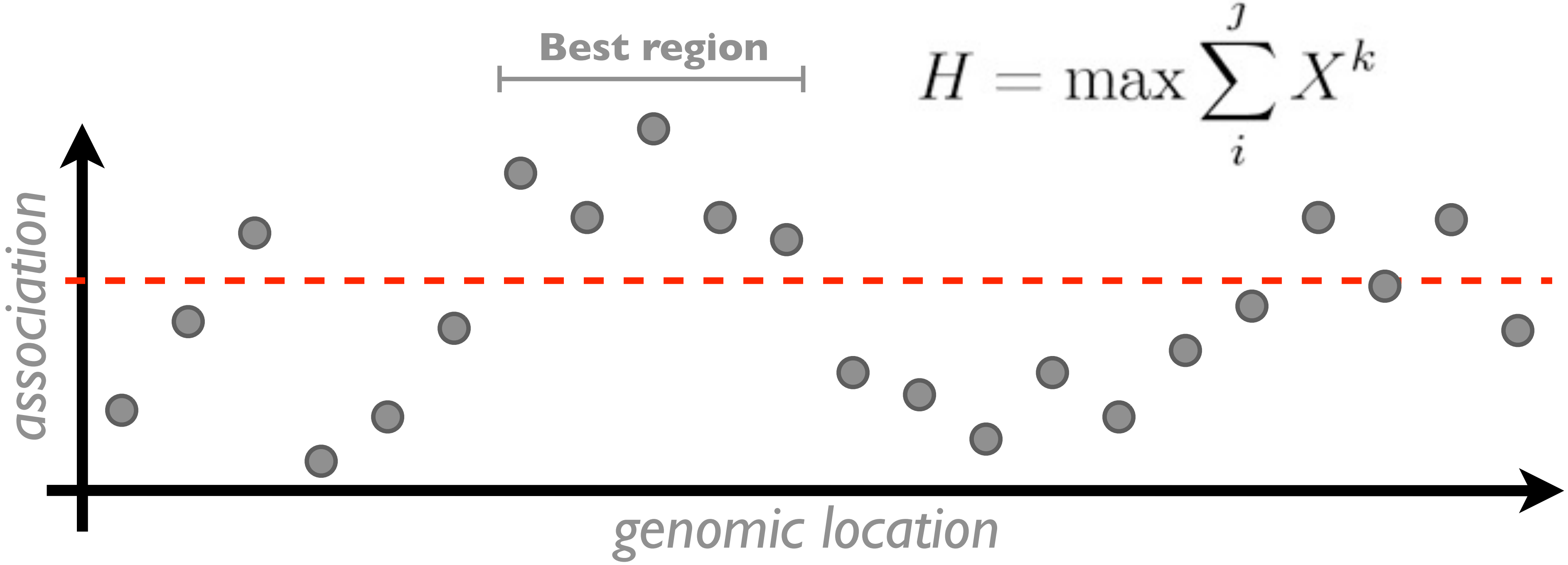
$$\mathbf{1} \quad \mathbf{-2} \quad \mathbf{-4} \quad \underset{H = 4}{\boxed{\mathbf{2} \quad \mathbf{1} \quad \mathbf{1}}} \quad \mathbf{-3} \quad \mathbf{1} \quad \mathbf{-2}$$

$$\mathbf{-1} \quad \mathbf{2} \quad \mathbf{1} \quad \mathbf{-4} \quad \mathbf{-2} \quad \mathbf{-2} \quad \underset{H = 6}{\boxed{\mathbf{2} \quad \mathbf{1} \quad \mathbf{-1} \quad \mathbf{3} \quad \mathbf{1}}} \quad \mathbf{-2}$$

# Local Score

- □ Definition: Let $\boldsymbol{X} = (X_{i})_{i = 1 \ldots n}$ be a sequence of random variables → association statistics:

  e.g. Pearson $\chi^2$ on case/control genotype frequencies.

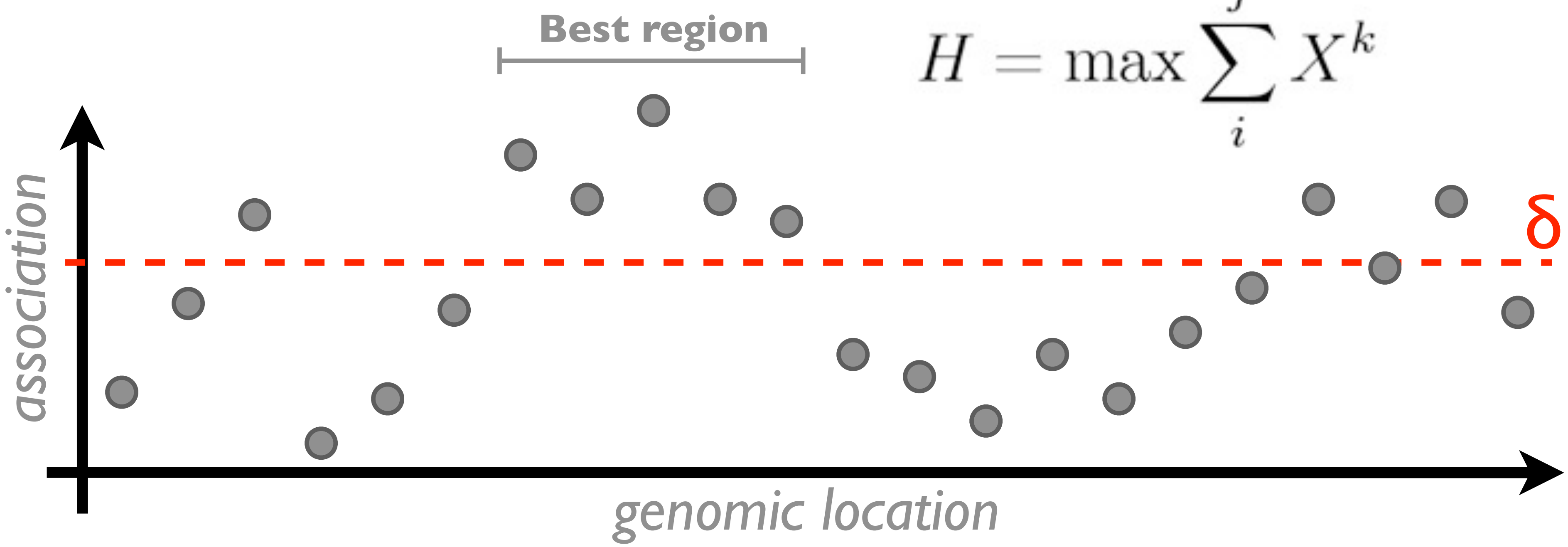


$$H = \max \sum_{i}^{j} X^{k}$$

- □ On average, the sequence ***X*** must be negative otherwise the best region would easily span the entire sequence.

# Local Score

- Definition: Let $\boldsymbol{X} = (X_i)_{i=1\ldots n}$ be a sequence of random variables → association statistics:
  e.g. Pearson $\chi^2$ on case/control genotype frequencies.



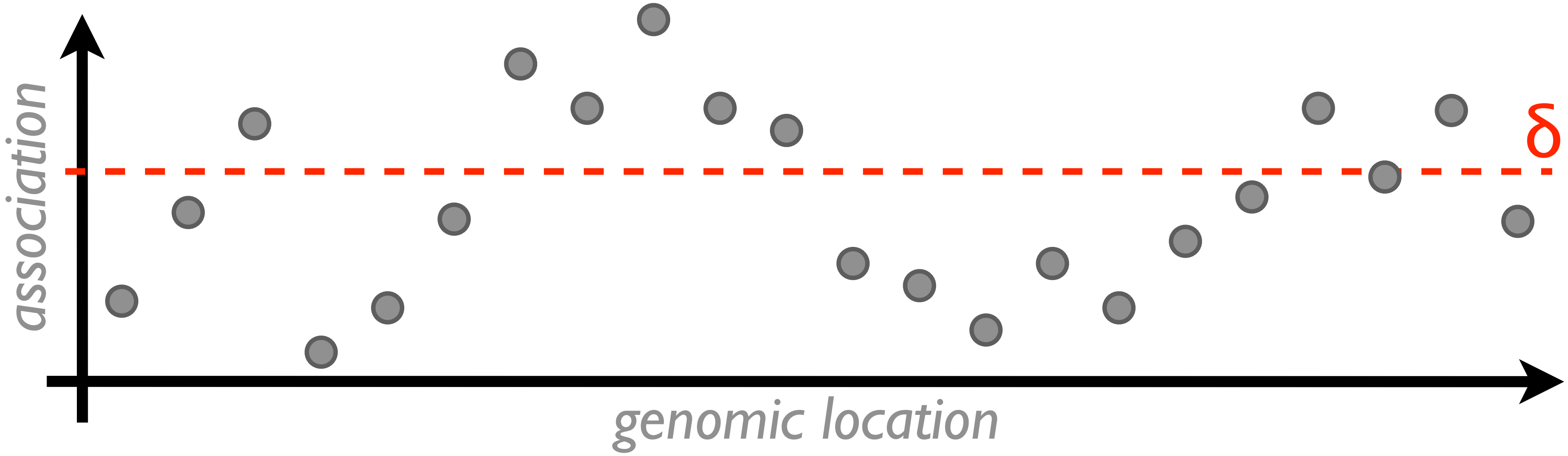
$$H = \max \sum_{i}^{j} X^k$$

- On average, the sequence $\boldsymbol{X}$ must be negative otherwise the best region would easily span the entire sequence → $\boldsymbol{X}' = \boldsymbol{X} - \delta$ ($\delta$ = 5% level)

# Local Score

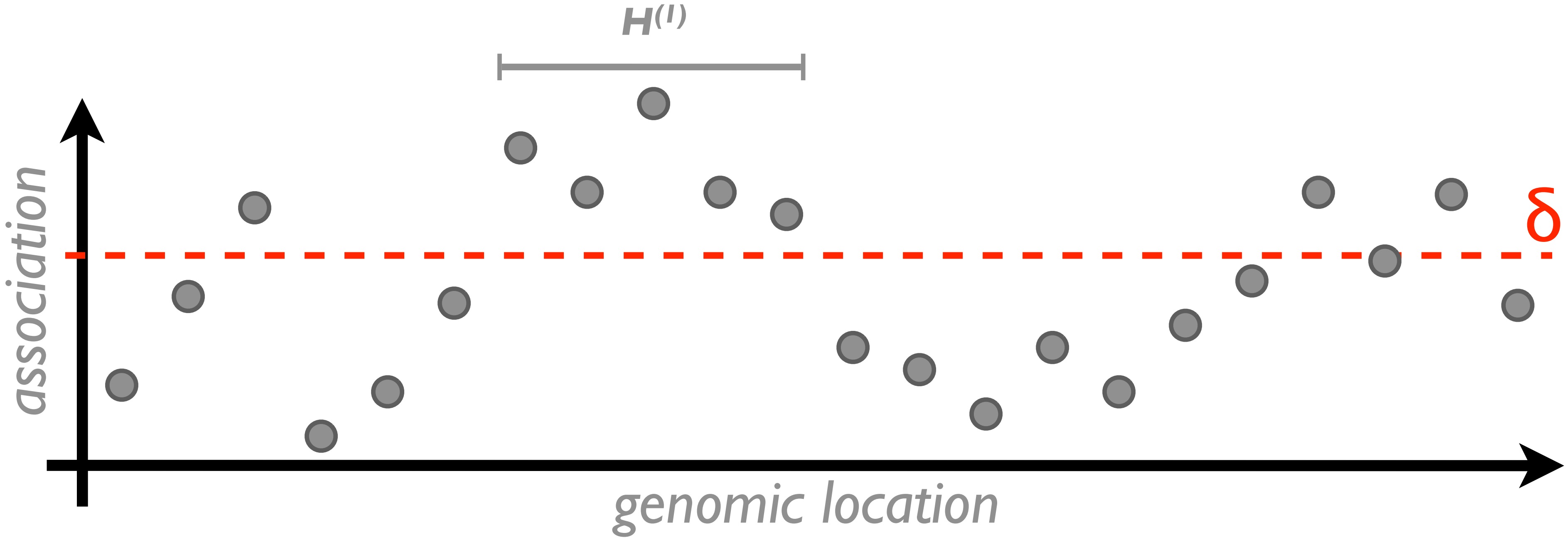
□ The *k* first best regions: $H^{(1)}, \ldots, H^{(k)}$.

□ $H^{(k)}$ is defined as the Local Score of the initial sequence disjoint from the preceding *k*-1 best regions.



# Local Score

- The *k* first best regions: $H^{(1)}, \ldots, H^{(k)}$.
- $H^{(k)}$ is defined as the Local Score of the initial sequence disjoint from the preceding *k*-1 best regions.

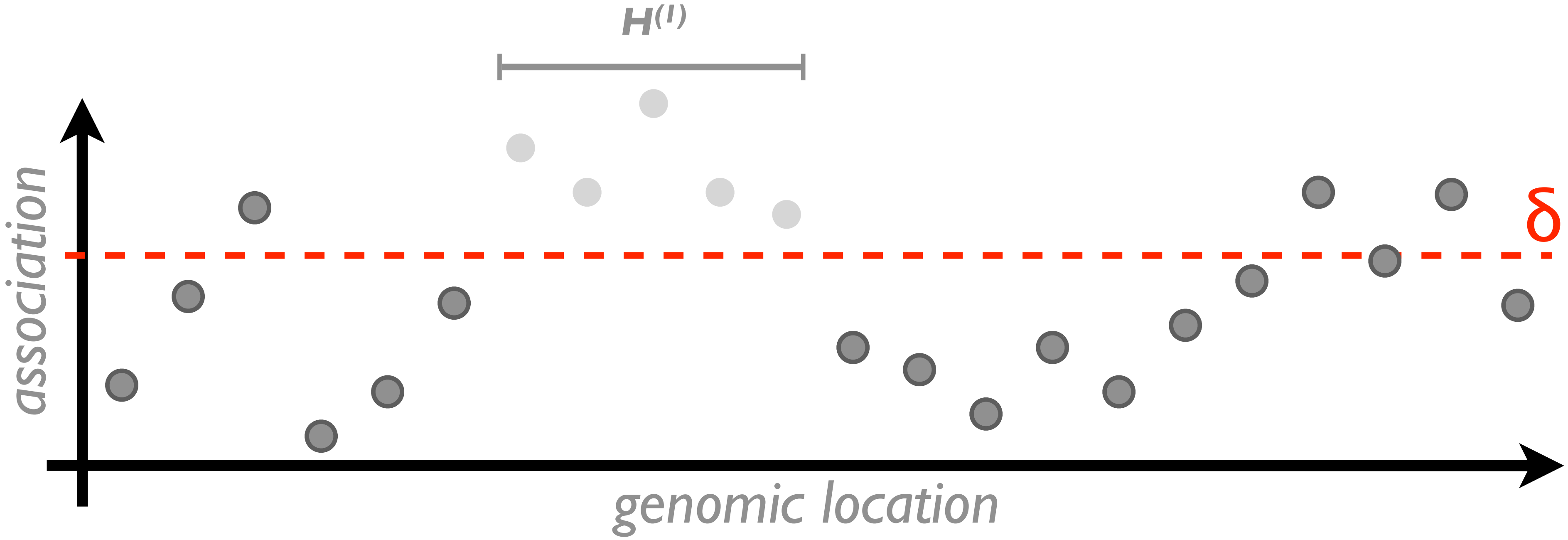


- Find the first best region.

# Local Score

□ The *k* first best regions: $H^{(1)}, \ldots, H^{(k)}$.

□ $H^{(k)}$ is defined as the Local Score of the initial sequence disjoint from the preceding *k*-1 best regions.

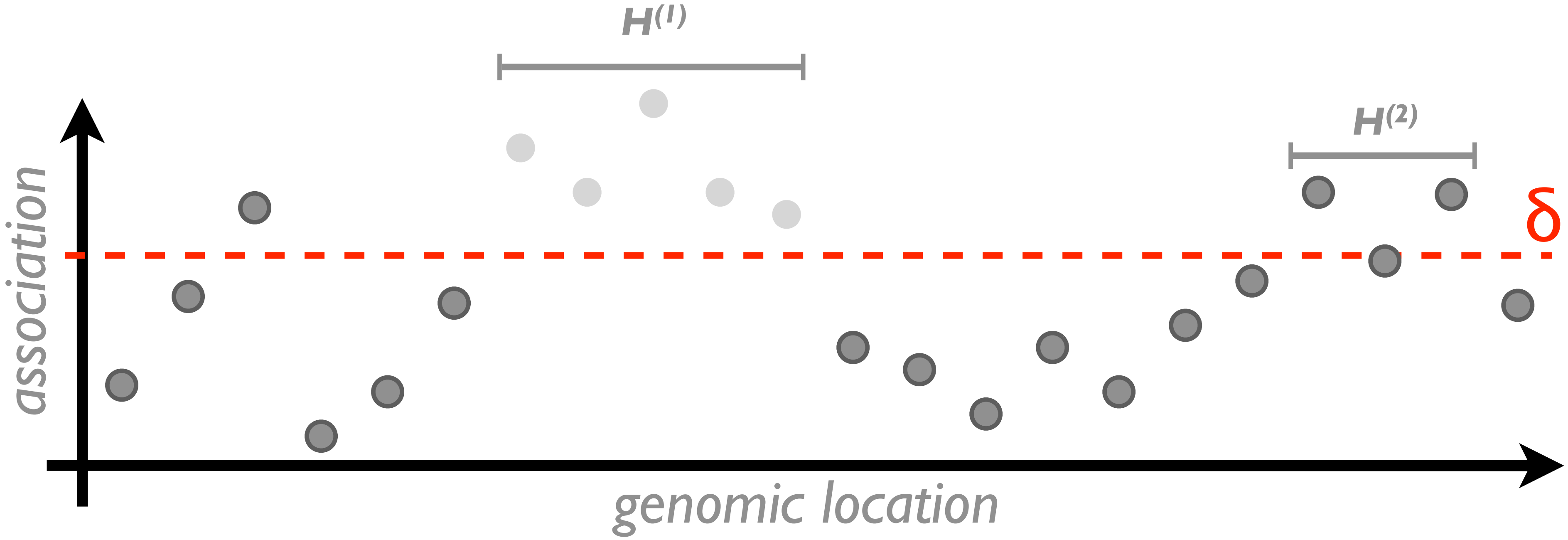


☑ Find the first best region.

□ Remove it from the sequence.

# Local Score

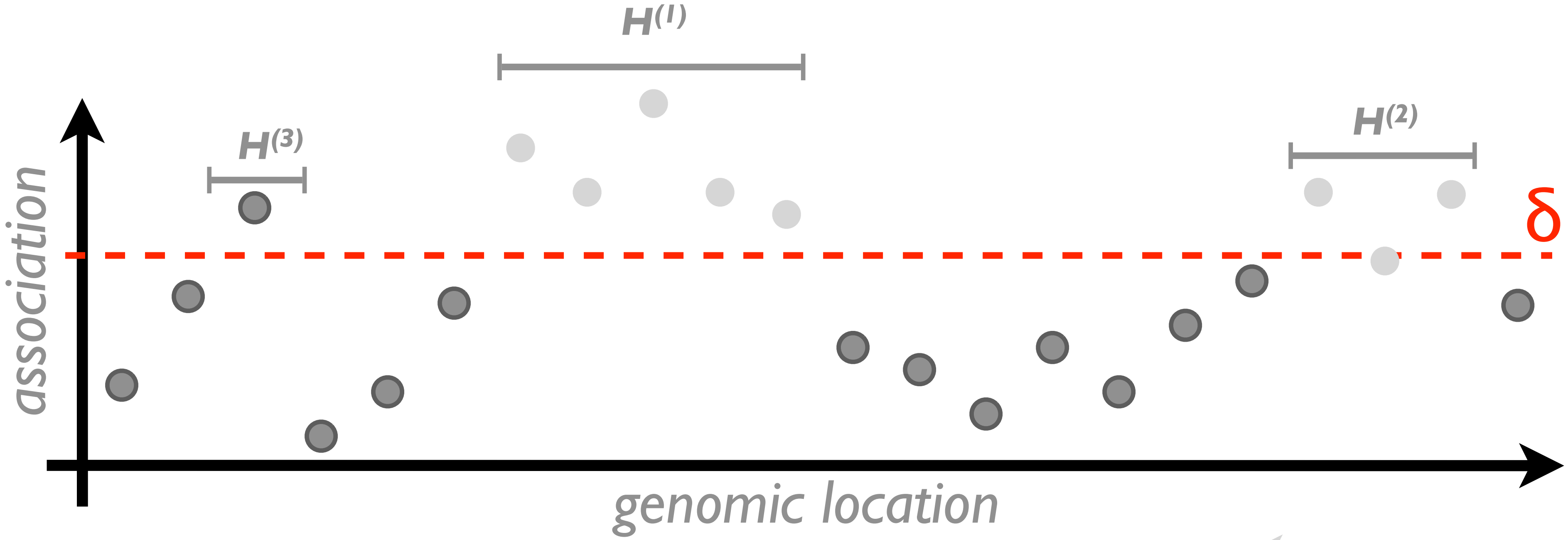
- □ The *k* first best regions: $H^{(1)}, \ldots, H^{(k)}$.
- □ $H^{(k)}$ is defined as the Local Score of the initial sequence disjoint from the preceding *k*-1 best regions.



- ☑ Find the first best region.
- ☑ Remove it from the sequence.
- □ Then find the second best region.

# Local Score

- □ The *k* first best regions: $H^{(1)}, ..., H^{(k)}$.
- □ $H^{(k)}$ is defined as the Local Score of the initial sequence disjoint from the preceding *k*-1 best regions.



- ☑ Find the first best region.
- ☑ Remove it from the sequence.
- ☑ Then find the second best region.

until $H^{(k+1)} < 0$

# Local Score

- Statistical significance of the regions:

| | |
|---|---|
| Region 1 | $H^{(1)}$ |
| Region 2 | $H^{(2)}$ |
| Region 3 | $H^{(3)}$ |
| Region 4 | $H^{(4)}$ |
| Region 5 | $H^{(5)}$ |
| $\vdots$ | $\vdots$ |
| Region *k* | $H^{(k)}$ |

# Local Score

- Statistical significance of the regions:

Region 1 $H^{(1)} \longrightarrow pv^{(1)}$

Region 2 $H^{(2)} \longrightarrow pv^{(2)}$

Region 3 $H^{(3)} \longrightarrow pv^{(3)}$

Region 4 $H^{(4)} \longrightarrow pv^{(4)}$

Region 5 $H^{(5)} \longrightarrow pv^{(5)}$

$\vdots \qquad \vdots \qquad \vdots$

Region $k$ $H^{(k)} \longrightarrow pv^{(k)}$

# Local Score

- □ Statistical significance of the regions:
- □ Extreme-Value theory but requires restrictive assumptions (e.g. independence of markers):

$$Pr\left(H \geq \frac{\ln n}{\lambda} + x\right) \simeq 1 - \exp(-Ke^{-\lambda x})$$ Gumbel distribution

# Local Score

- Statistical significance of the regions:
- Extreme-Value theory but requires restrictive assumptions (e.g. independence of markers):

$$Pr\left(H \geq \frac{\ln n}{\lambda} + x\right) \simeq 1 - \exp(-Ke^{-\lambda x})$$ Gumbel distribution

- Monte-Carlo simulations permuting case-control labels but a more important time of execution.

# Local Score

- In Statistics: **asymtoptic and exact distributions**

  e.g. Iglehart (1972)
  Extreme values in the in the gi/g/1 queues. *Annals of Mathematical Statistics.*

- In Computer Science: **clever detection of Local Scores**

  e.g. Ruzzo and Tompa (1999)
  A linear time algorithm for finding all maximal scoring subsequences. *Proceedings from ISMB.*

- In Genomics: **biological sequences analysis/alignment**

  e.g. Karlin (2005)
  Statistical signals in Bioinformatics. *PNAS.*

# Local Score

- In Genetic Epidemiology:

  **Fast and simple tool to detect associated genomic regions at the first-stage of GWAS:**

  Guedj, Robelin et al (2006)
  Detecting local high-scoring segments: a first-stage approach to genome-wide association studies. *Stat. App. Genet. Mol. Bio.*

  **Application in a two-stage design:**

  Aschard, Guedj and Demenais (2007)
  A two-step multiple-marker strategy for genome-wide association studies. *Proceedings of GAW15.*

# Local Score

□ Application to Local Replications:

# Local Score

□ Application to Local Replications:

□ Let pop$_A$ and pop$_B$ denote the two populations and

$$\boldsymbol{X_A} = (X_{Ai})_{i\,=\,1\ldots n} \text{ and } \boldsymbol{X_B} = (X_{Bi})_{i\,=\,1\ldots n}$$

their respective sequences of test statistics for the same set of markers.

# Local Score

- Application to Local Replications:
- Let pop$_A$ and pop$_B$ denote the two populations and

$$\boldsymbol{X_A} = (X_{Ai})_{i\,=\,1 \ldots n} \text{ and } \boldsymbol{X_B} = (X_{Bi})_{i\,=\,1 \ldots n}$$

their respective sequences of test statistics for the same set of markers.

- Let $\boldsymbol{X'_A} = \boldsymbol{X_A} - \delta$ and $\boldsymbol{X'_B} = \boldsymbol{X_B} - \delta$.

# Local Score

- □ Application to Local Replications:
- □ Let pop$_A$ and pop$_B$ denote the two populations and

$$\boldsymbol{X_A} = (X_{Ai})_{i\,=\,1\ldots n} \text{ and } \boldsymbol{X_B} = (X_{Bi})_{i\,=\,1\ldots n}$$

their respective sequences of test statistics for the same set of markers.

- □ Let $\boldsymbol{X'_A} = \boldsymbol{X_A} - \delta$ and $\boldsymbol{X'_B} = \boldsymbol{X_B} - \delta$.
- □ $\boldsymbol{X'_{AB}} = \boldsymbol{X'_A} + \boldsymbol{X'_B}$ : on which we apply the Local Score.

# Local Score

- Application to Local Replications:
- Let $pop_A$ and $pop_B$ denote the two populations and

  $$\boldsymbol{X_A} = (X_{Ai})_{i = 1 \ldots n} \text{ and } \boldsymbol{X_B} = (X_{Bi})_{i = 1 \ldots n}$$

  their respective sequences of test statistics for the same set of markers.
- Let $\boldsymbol{X'_A} = \boldsymbol{X_A} - \delta$ and $\boldsymbol{X'_B} = \boldsymbol{X_B} - \delta$.
- $\boldsymbol{X'_{AB}} = \boldsymbol{X'_A} + \boldsymbol{X'_B}$ : on which we apply the Local Score.
- Easily extended to more than two populations and different sets of markers.

# Power study

# Power study

- Based on Monte-Carlo simulations.

# Power study

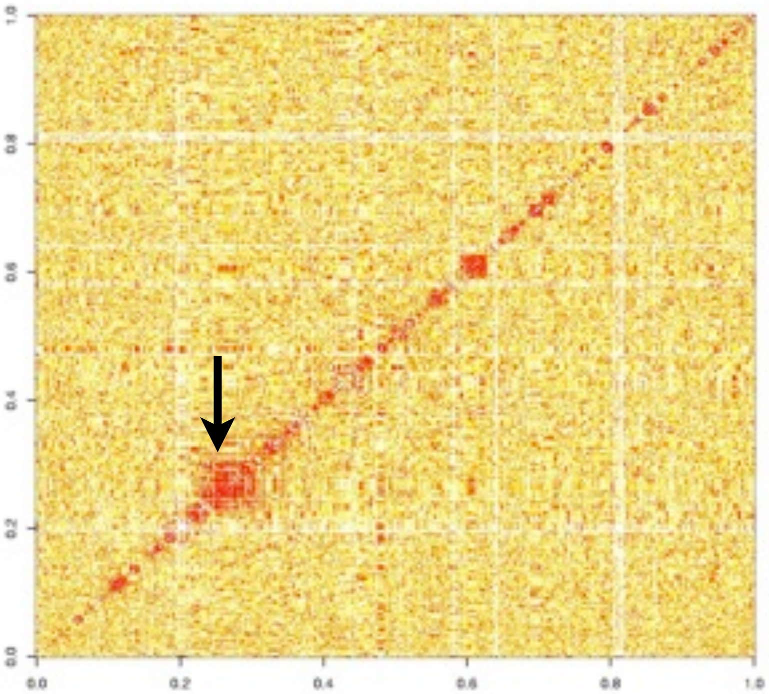
- Based on Monte-Carlo simulations.
- Based on Real Data (to preserve a realistic pattern of LD).
- 301 and 289 chr19 from *French* ($pop_A$) and *Swedish* ($pop_B$) controls (as an empirical distribution of possible diplotypes).
- chr 19 = 674 SNPs genotyped using a 100K Affymetrix chip.
- This data set is used as the basis to simulate new cases and controls.

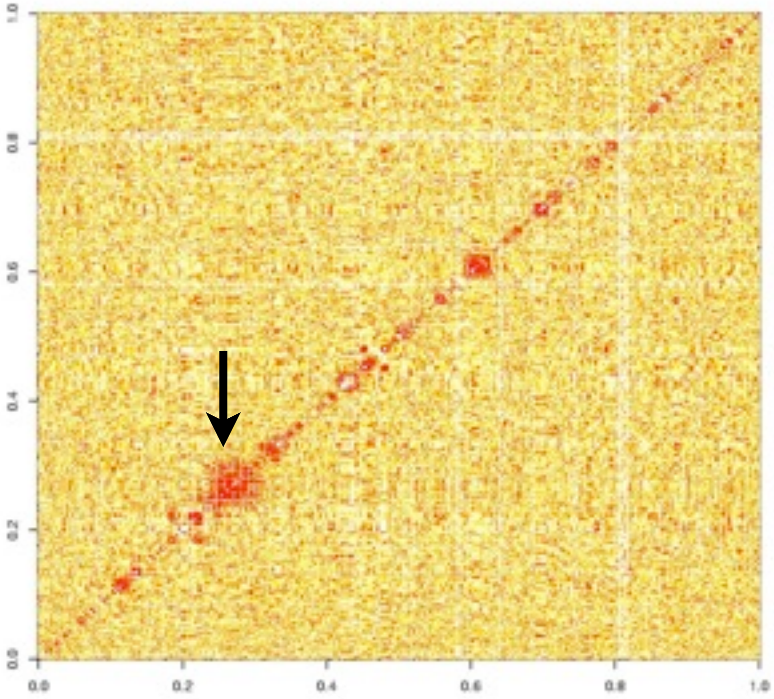
# Power study

- Genetic and Disease Model:
- One bi-allelic DSL (*aa*, *aA* and *AA*)
- Susceptibility allele frequency: $p_A = 0.3$
- Coef. of consanguinity in the general population: $F = 0$
- Relative Risk of the homozygous susceptibility genotype: $RR_{AA}$ from 1 to 2.5
- Additive Mode of Transmission → $RR_{aA} = (RR_{AA}+1)/2$
- The DSL is hidden after the sampling of cases and controls

# Power study

- Situation 1/4:
- The two populations have similar patterns of LD.
- The DSL is localised in a block of LD.



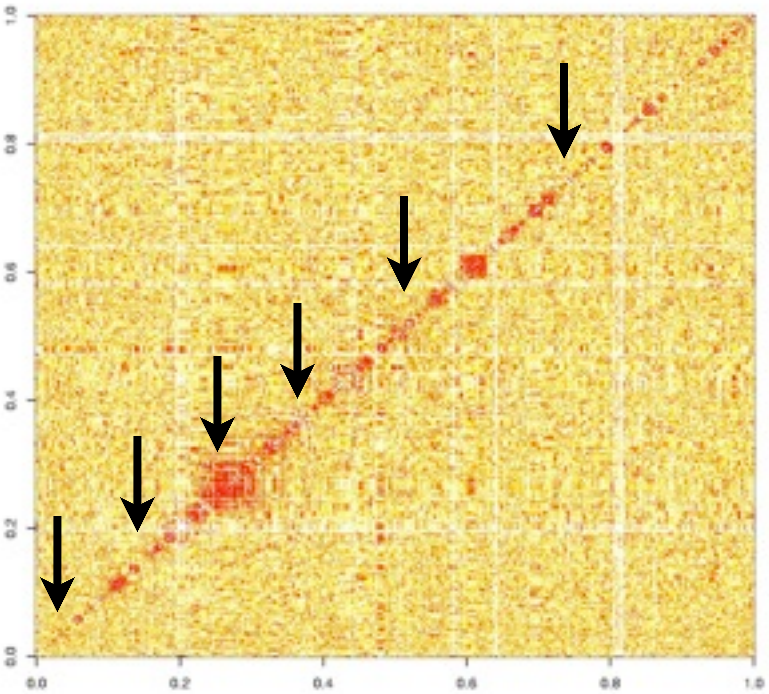


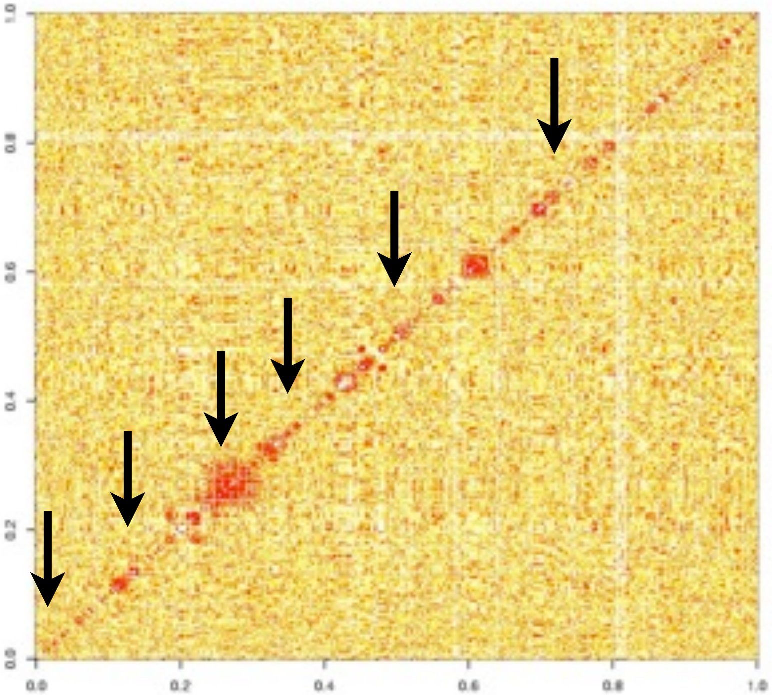
# Power study

□ Situation 2/4:

□ The two populations have similar patterns of LD.

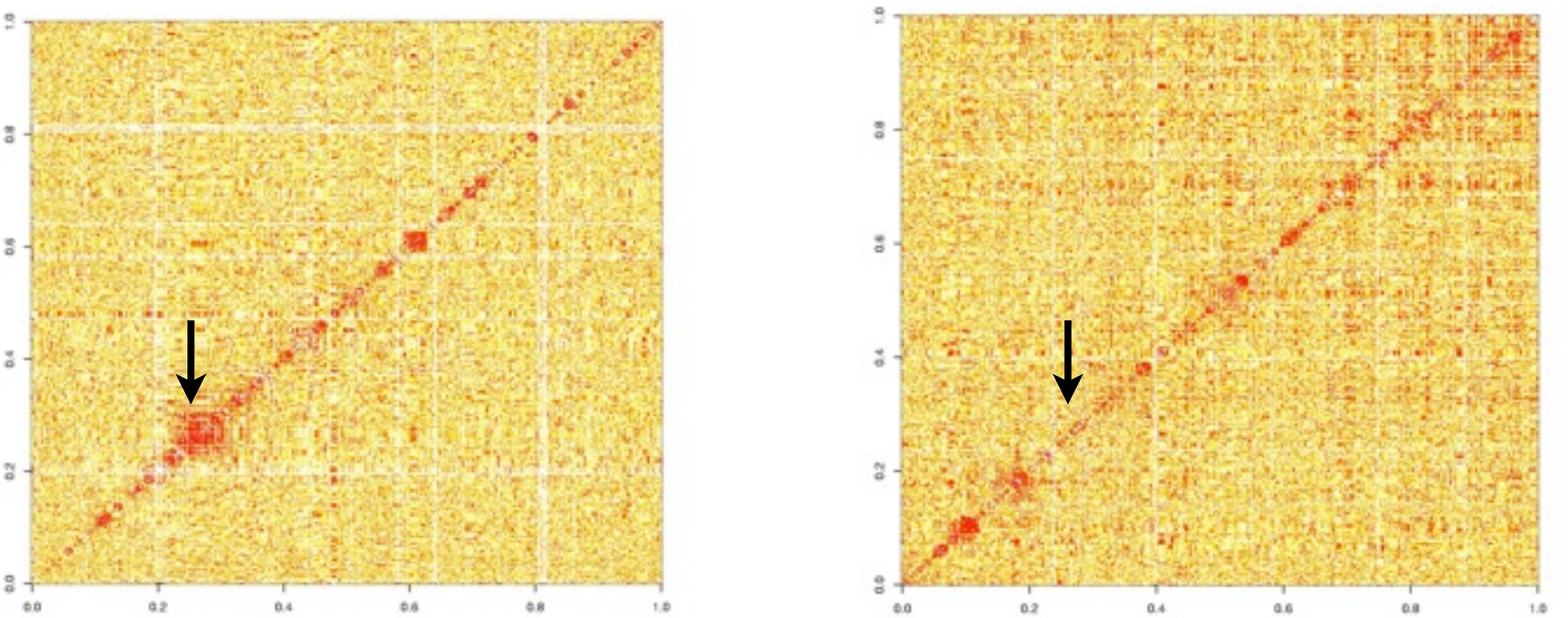
□ The DSL is randomly chosen (among SNPs that present a Minor-Genotype-Frequency of at least 1%).





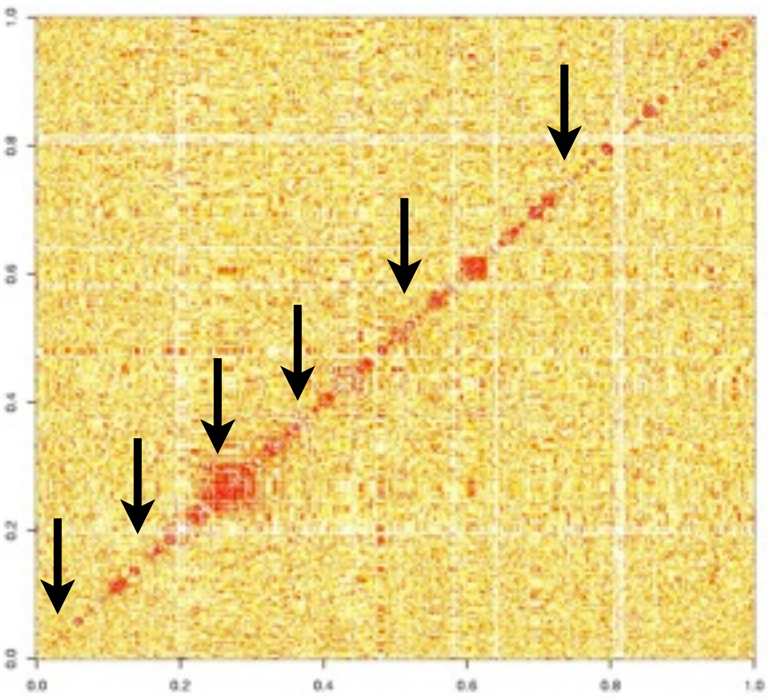
# Power study

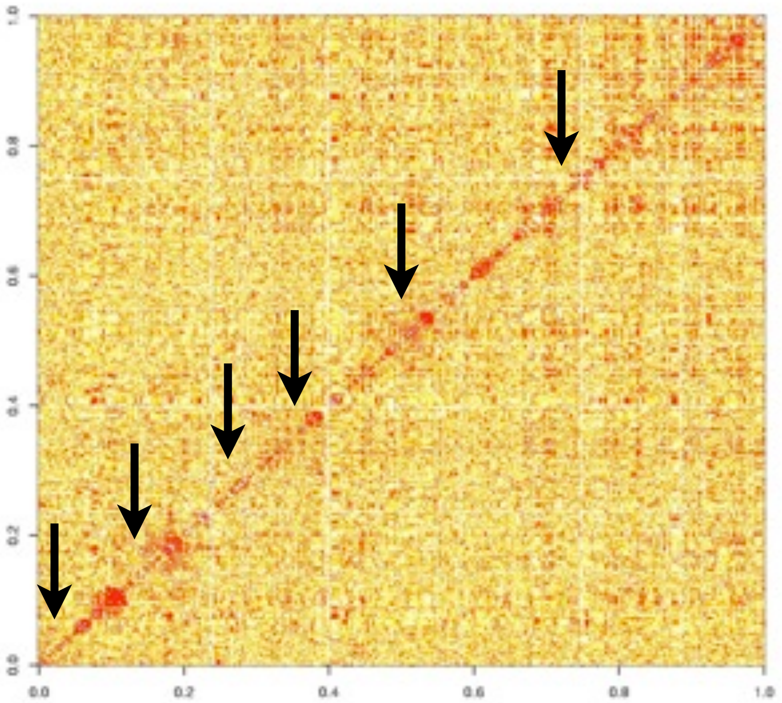
- □ Situation 3/4:
- □ The two populations have different patterns of LD.
- □ The DSL is localised in a block of LD.



# Power study

- Situation 4/4:
- The two populations have different patterns of LD.
- The DSL is randomly chosen (among SNPs that present a Minor-Genotype-Frequency of at least 1%).





(Guedj, Wojcik et al 06)

# Power study

- Test statistic: exact and unbiased allelic test
- $\boldsymbol{X_A} = [\ -\log_{10}(pv_{Ai})\ ]_{i\,=\,1 \ldots n}$ and $\boldsymbol{X_B} = [\ -\log_{10}(pv_{Bi})\ ]_{i\,=\,1 \ldots n}$

# Power study

- □ Test statistic: exact and unbiased allelic test
- □ $\boldsymbol{X_A} = [\ -\log_{10}(pv_{Ai})\ ]_{i\,=\,1\ldots n}$ and $\boldsymbol{X_B} = [\ -\log_{10}(pv_{Bi})\ ]_{i\,=\,1\ldots n}$
- □ Local Score: *H0* is rejected if the Local Score of at least the best region is significant at the 5% level.

# Power study

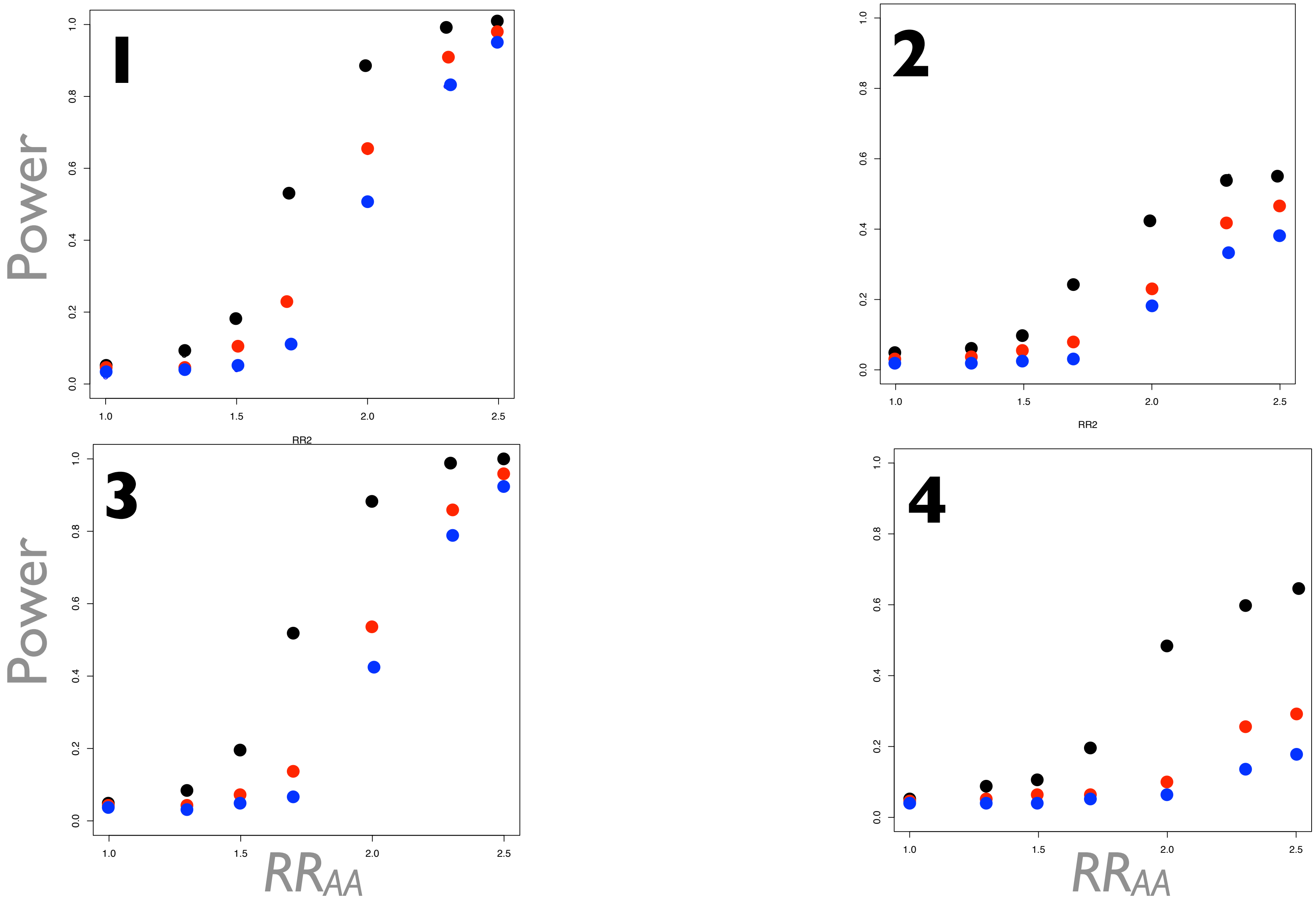
- □ Test statistic: exact and unbiased allelic test
- □ $\boldsymbol{X}_{\boldsymbol{A}} = [\ -\log_{10}(pv_{Ai})\ ]_{i\,=\,1\ldots n}$ and $\boldsymbol{X}_{\boldsymbol{B}} = [\ -\log_{10}(pv_{Bi})\ ]_{i\,=\,1\ldots n}$
- □ Local Score: *H0* is rejected if the Local Score of at least the best region is significant at the 5% level.
- □ Single-marker analysis: *H0* is rejected if at least one SNP is replicated in the two populations.
- □ $pv_{Ai} \leq \alpha$ AND $pv_{Bi} \leq \alpha$

Corrected for multiple-testing by Bonferroni (FWER) and Benjamini-Hochberg (FDR).

# Power study

- ● Local Score
- ● FWER
- ● FDR

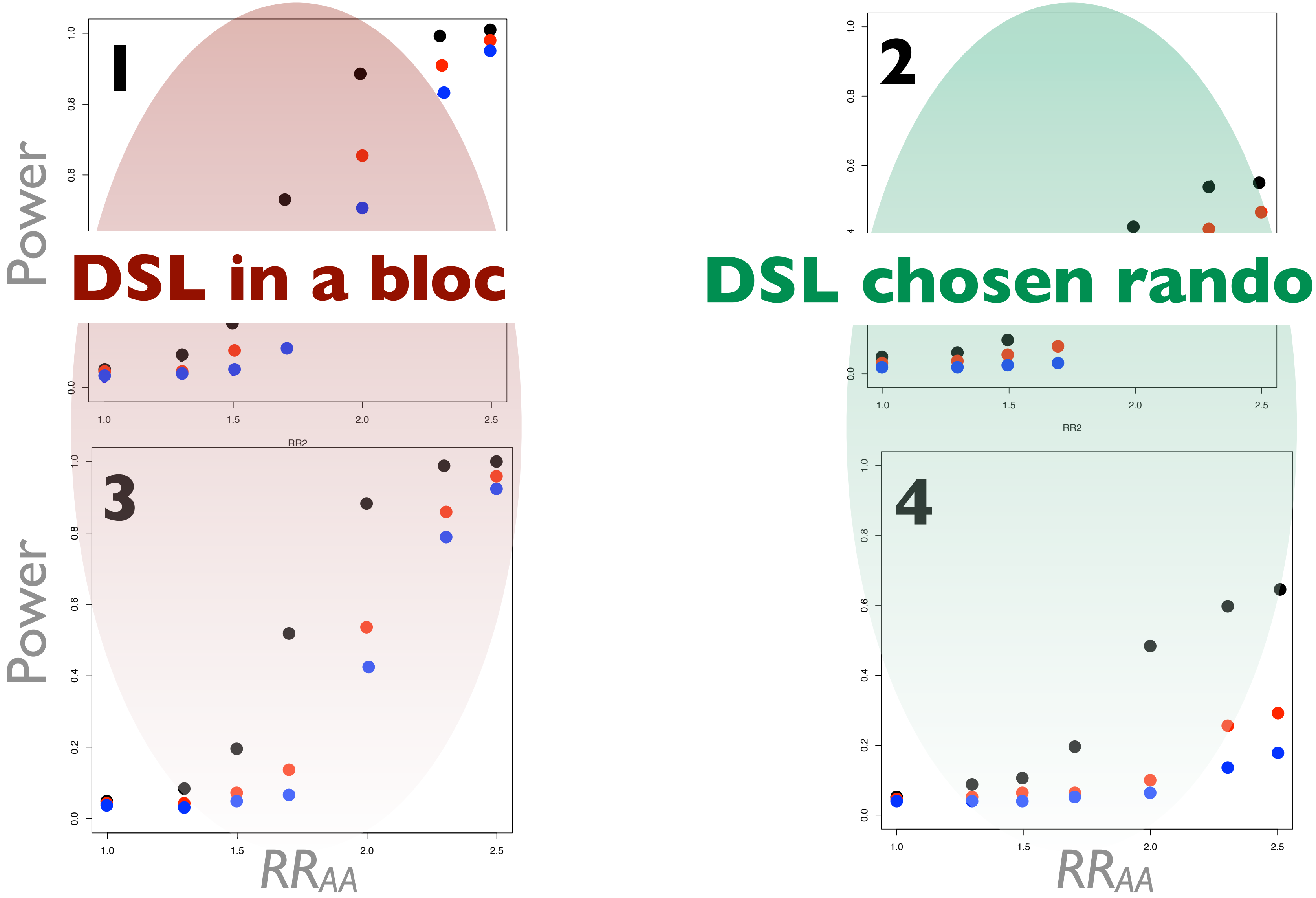
- Results:



# Power study

● Local Score
● FWER
● FDR

□ Results:



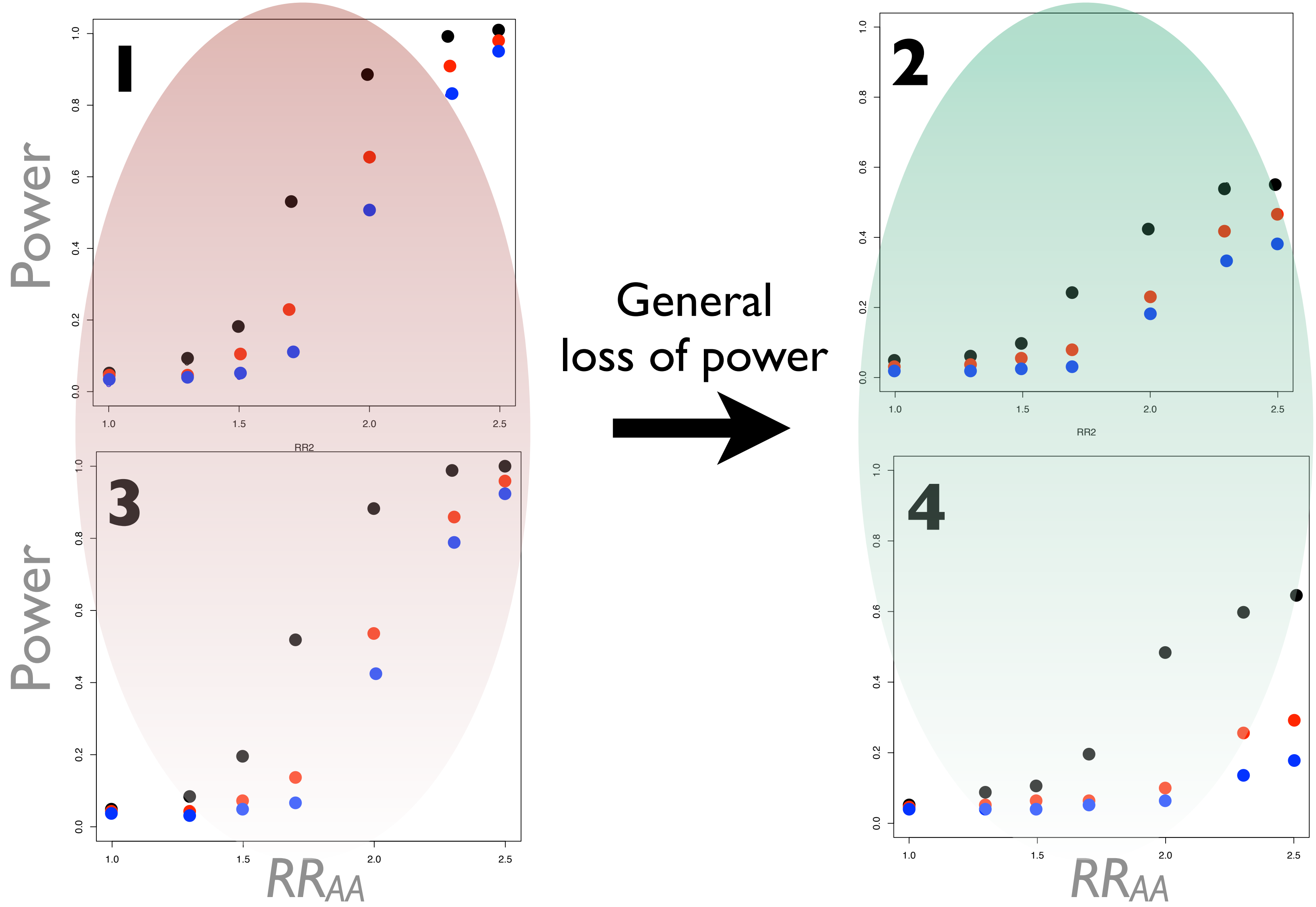
**DSL in a bloc**

**DSL chosen randomly**

# Power study

● Local Score
● FWER
● FDR

□ Results:

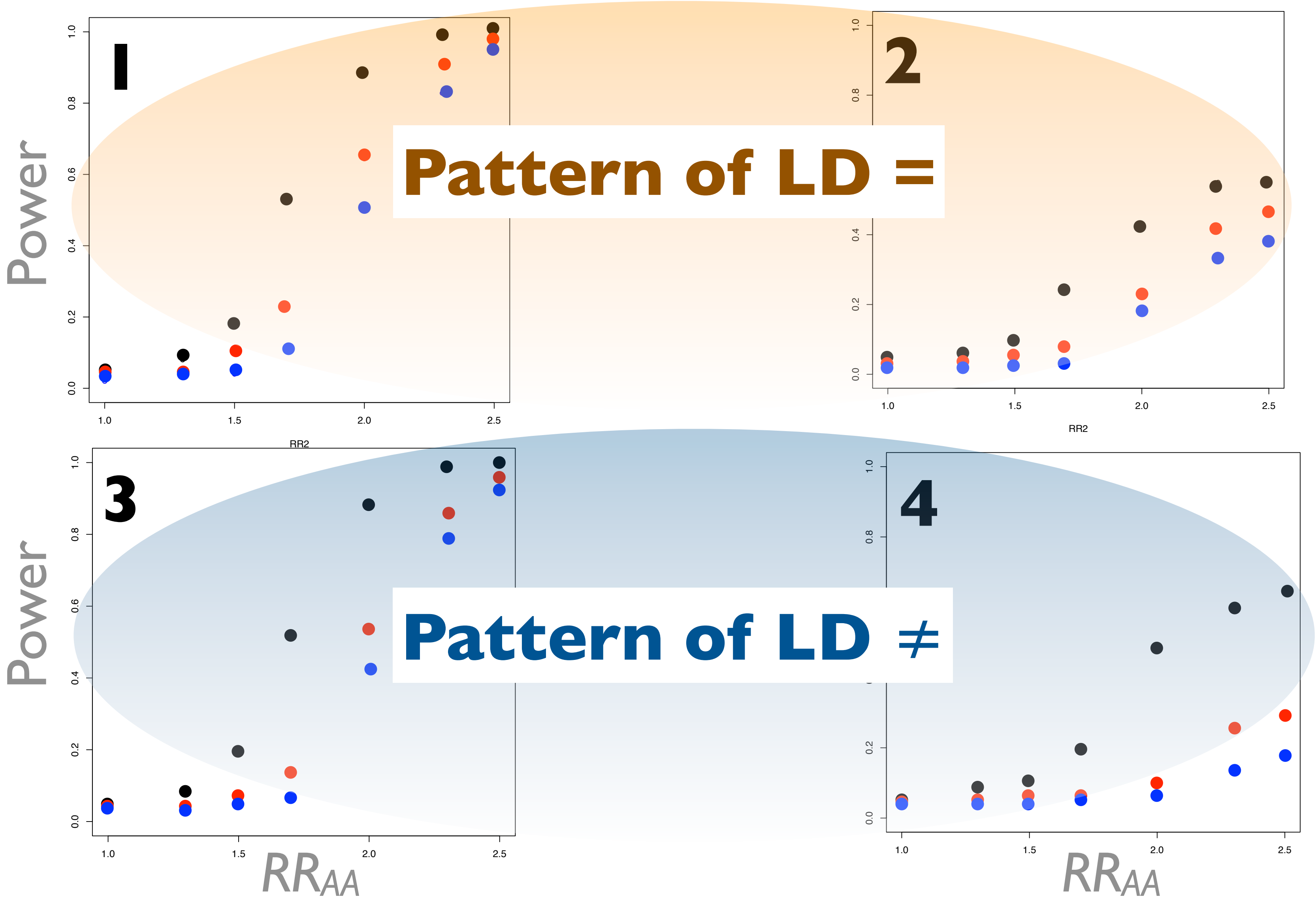


General loss of power

# Power study

● Local Score
● FWER
● FDR

□ Results:



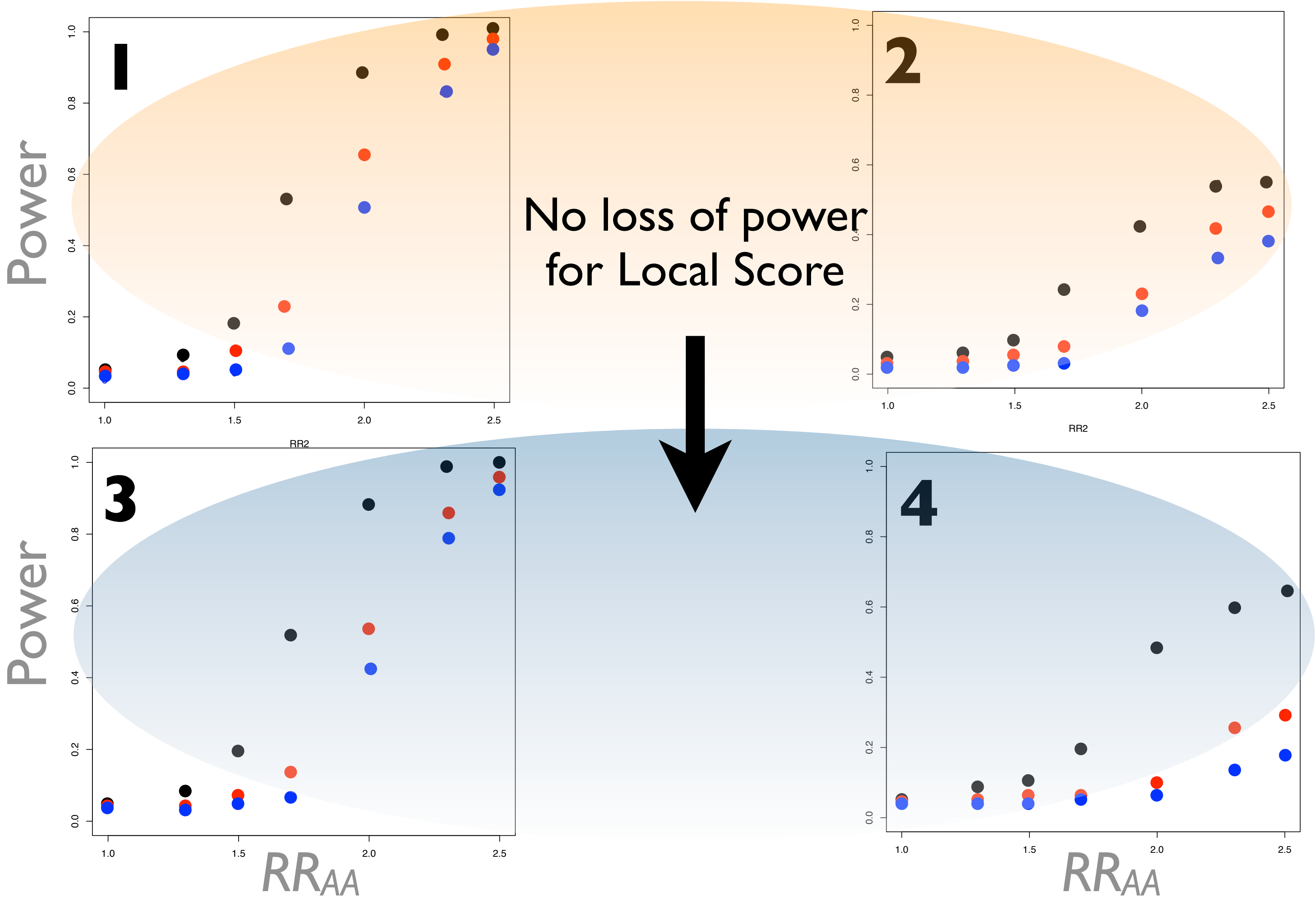
Pattern of LD =

Pattern of LD ≠

# Power study

● Local Score
● FWER
● FDR

□ Results:



No loss of power for Local Score

# Application

- Data: Systemic Lupus Erythematosus.
- 2 populations:

  *Argentina:* 255 cases and 256 controls.

  *Sweden:* 279 cases and 515 controls.

- 100K Affymetrix chip.
- Results: 3 regions are 'locally replicated' (significant at the 5% level) with the Local Score approach.
- 2 of them do not share any marker with the results of marker-based replications.

# Conclusions

- Looking at Local Replications appears more robust to biological differences between populations.
- Local Score as a simple and natural framework.

# Conclusions

- ☐ Looking at Local Replications appears more robust to biological differences between populations.
- ☐ Local Score as a simple and natural framework.
- ☐ Strict Replications show a stronger evidence for true replication.

# Conclusions

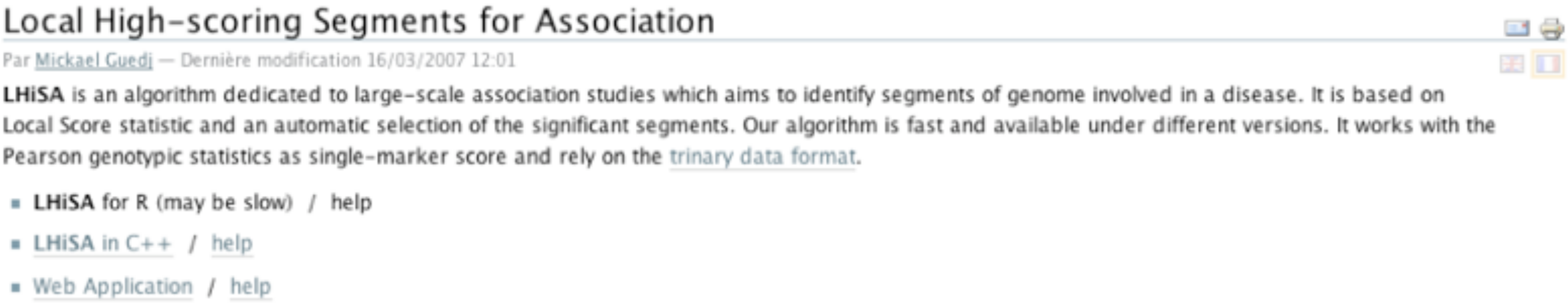
- Looking at Local Replications appears more robust to biological differences between populations.
- Local Score as a simple and natural framework.
- Strict Replications show a stronger evidence for true replication.
- Considering Local Replications can help to identify DSL shared across populations ...
- ... but also across diseases: auto-immune diseases (e.g. pop$_A$ : lupus / pop$_B$ : psoriasis).

# Software : LHiSA

- C++ (not maintained anymore)

- R can work for any study design (case-control, families), with any test statistic (if specified by the user) and handles more than one population (for Local Replications).

http://stat.genopole.cnrs.fr/software/lhisa



Local High-scoring Segments for Association

Par Mickael Guedi — Dernière modification 16/03/2007 12:01

**LHiSA** is an algorithm dedicated to large-scale association studies which aims to identify segments of genome involved in a disease. It is based on Local Score statistic and an automatic selection of the significant segments. Our algorithm is fast and available under different versions. It works with the Pearson genotypic statistics as single-marker score and rely on the trinary data format.

- **LHiSA** for R (may be slow) / help
- LHiSA in C++ / help
- Web Application / help

# Acknowledgements

G Nuel, J Wojcik, B Prum, S Robin, A Célisse.

Merck-Serono for the data.

F Demenais for useful discussions.

Email: mickael.guedj@gmail.com

# Any questions ??

« That's what I want to say. See if you can find some statistics to prove it! »

# Annexe 1:

| | | |
|---|---|---|
| Region 1 | $H^{(1)}$ | $pv^{(1)}$ |
| Region 2 | $H^{(2)}$ | $pv^{(2)}$ |
| Region 3 | $H^{(3)}$ | $pv^{(3)}$ |
| Region 4 | $H^{(4)}$ | $pv^{(4)}$ |
| Region 5 | $H^{(5)}$ | $pv^{(5)}$ |
| ⋮ | ⋮ | ⋮ |

**Sequential testing procedure on ordered statistics.**

**Control the resulting type-I error rate.**

# Annexe 2:

## Same Marker Set

| | | | | | |
|---|---|---|---|---|---|
| $\mathbf{X'_{A}} =$ | $X'_{A1}$ | $X'_{A2}$ | $X'_{A3}$ | $X'_{A4}$ | $X'_{A5}$ |
| $\mathbf{X'_{B}} =$ | $X'_{B1}$ | $X'_{B2}$ | $X'_{B3}$ | $X'_{B4}$ | $X'_{B5}$ |
| $\mathbf{X'_{AB}} =$ | $X'_{A1}+X'_{B1}$ | $X'_{A2}+X'_{B2}$ | $X'_{A3}+X'_{B3}$ | $X'_{A4}+X'_{B4}$ | $X'_{A5}+X'_{B5}$ |

## Different Marker Sets

| | | | | | |
|---|---|---|---|---|---|
| $\mathbf{X'_{A}} =$ | $X'_{A1}$ | $X'_{A2}$ | $X'_{A3}$ | _ | $X'_{A5}$ |
| $\mathbf{X'_{B}} =$ | $X'_{B1}$ | _ | $X'_{B3}$ | $X'_{B4}$ | $X'_{B5}$ |
| $\mathbf{X'_{AB}} =$ | $X'_{A1}+X'_{B1}$ | $X'_{A2}$ | $X'_{A3}+X'_{B3}$ | $X'_{B4}$ | $X'_{A5}+X'_{B5}$ |