Catching Local Replications:

a Local Score-based approach to replicated association studies.

Mickaël Guedj[1,2], Jérôme Wojcik[2] and Grégory Nuel[1].

IGES 2007, York UK

[1] Statistics and Genome laboratory, CNRS, INRA, University of Evry, FRANCE
[2] Serono Pharmaceutical Research, Geneva, SWITZERLAND

Introduction

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

(Ioannidis 03)

Introduction

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

  Successful replication rate of 16-30%.

(Ioannidis 03)

Introduction

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

  Successful replication rate of 16-30%.

  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)

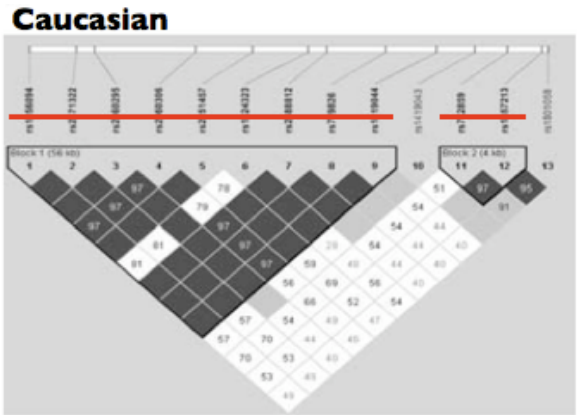
Introduction

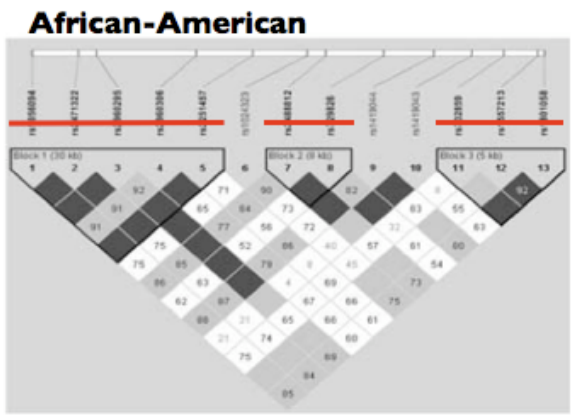
- Beside these study-design and data-analysis related factors ...
- ... inconsistent findings might also result from real biological differences between populations:

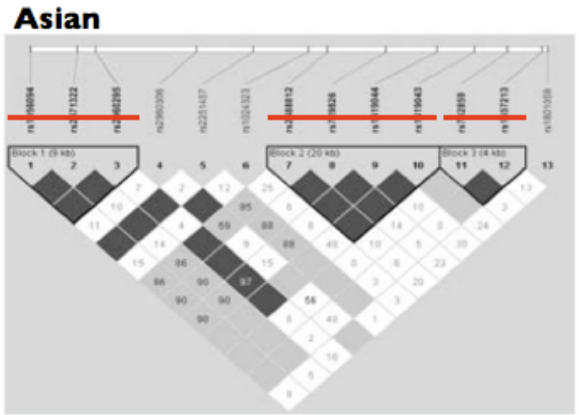
  Differences in allele frequencies.

  Allele and locus heterogeneity.

  Variation in the strength of LD:







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.

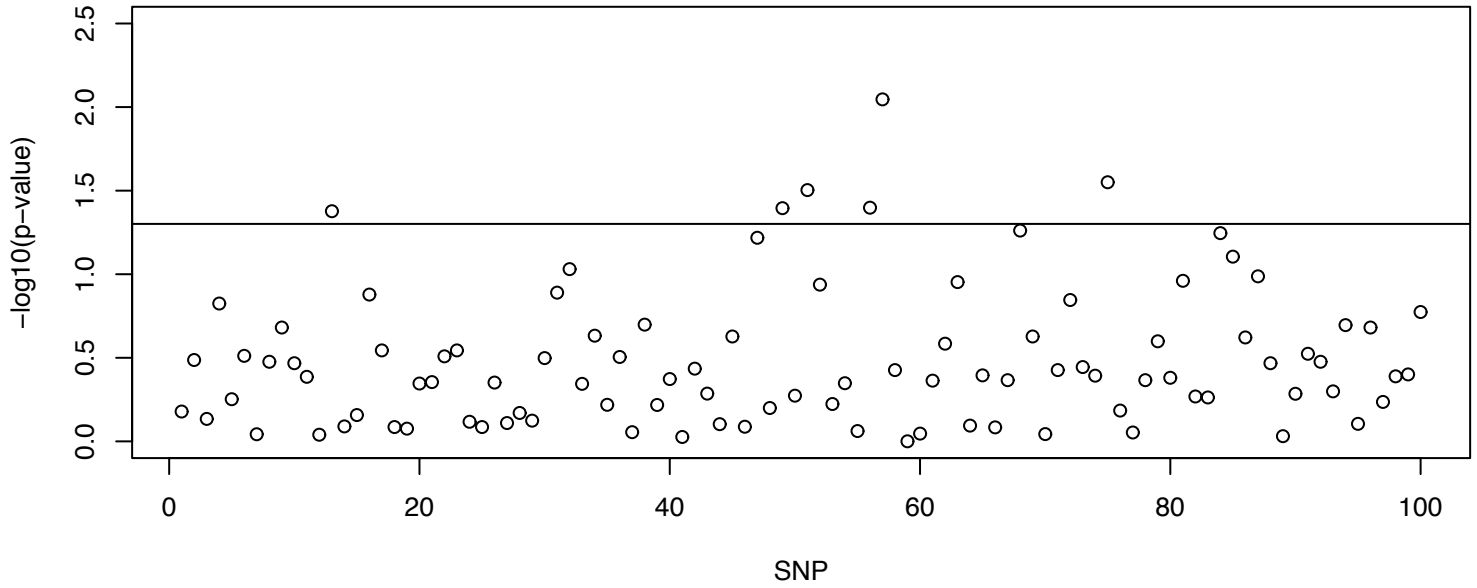
  Aggregation of several DSL in a same genomic location.
- Such accumulations may be locally replicated across populations ...
- ... without restraint about the specific allele or pattern of alleles to be replicated.

Introduction

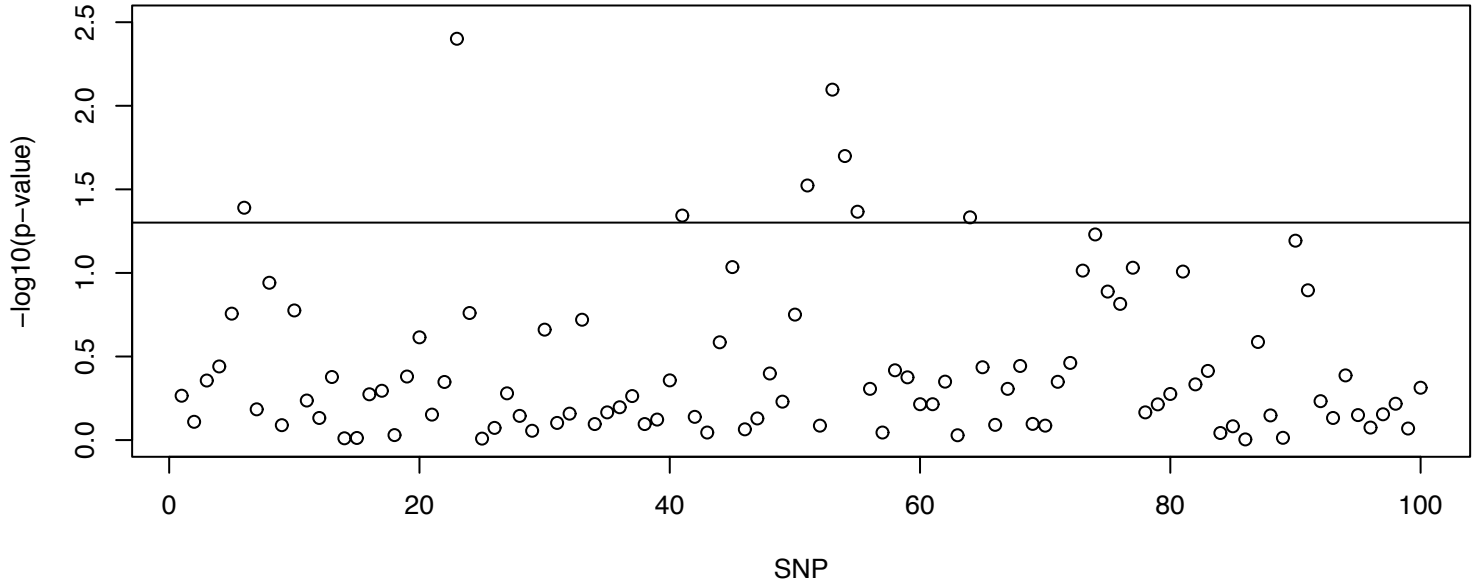
- Local Replication: **definition**

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

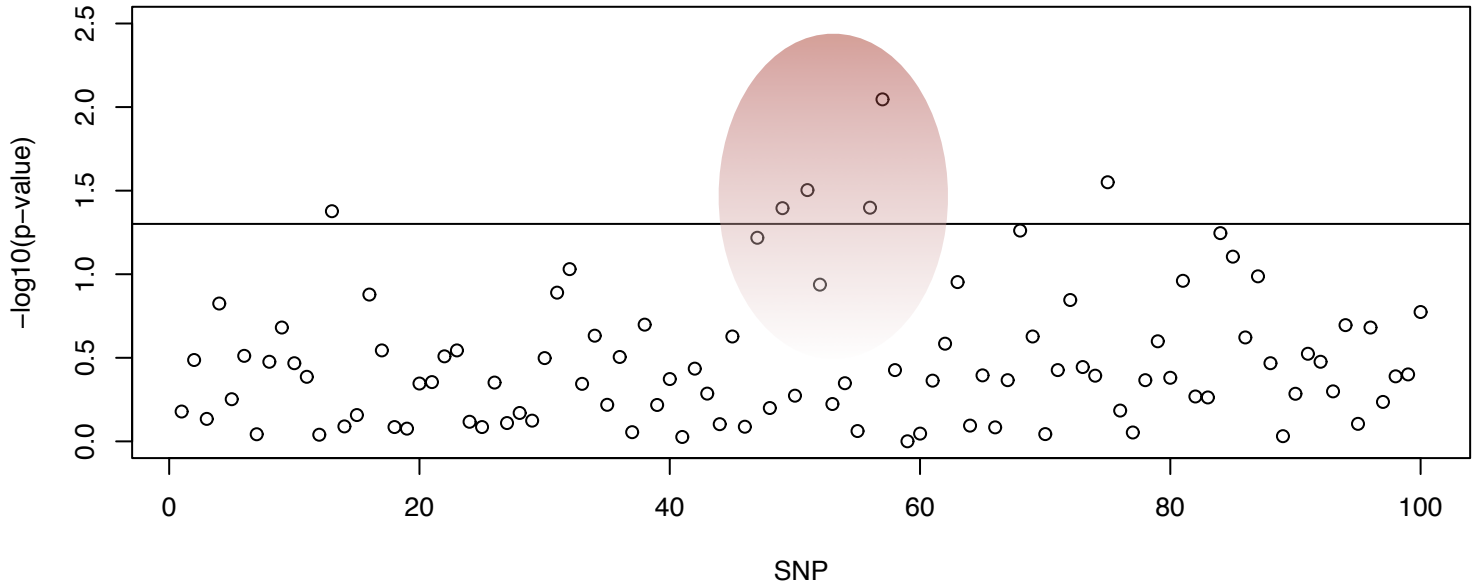
... replicated among the different populations.



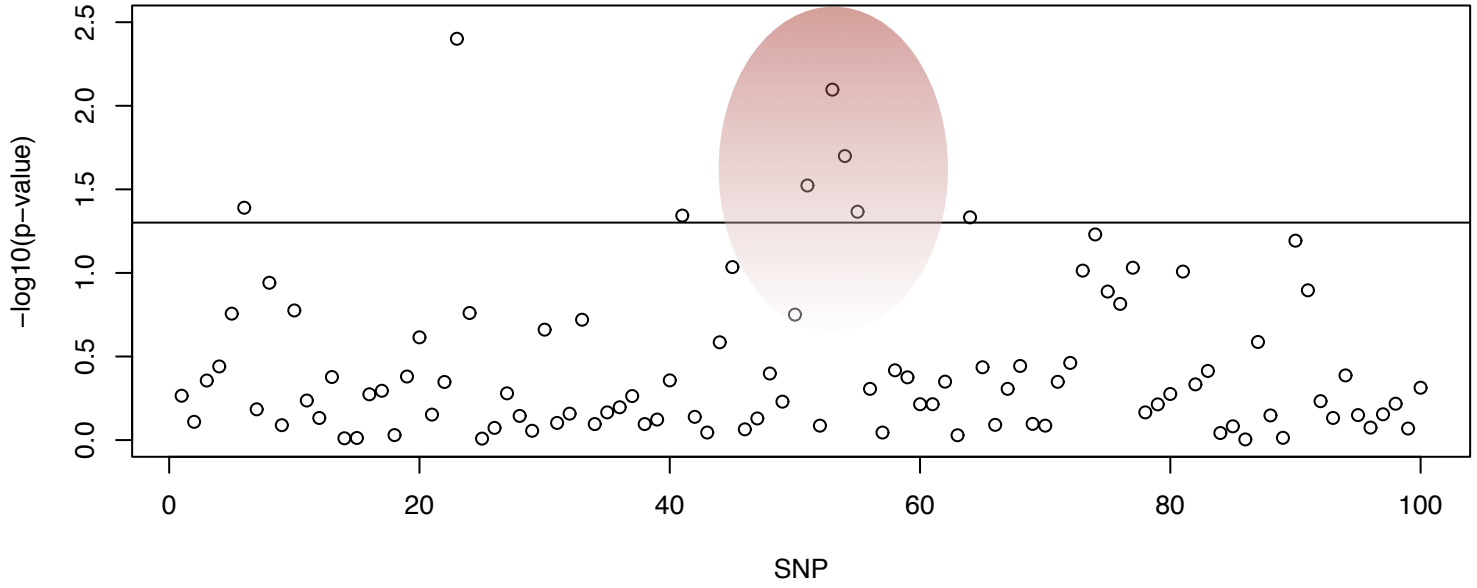
Population 1



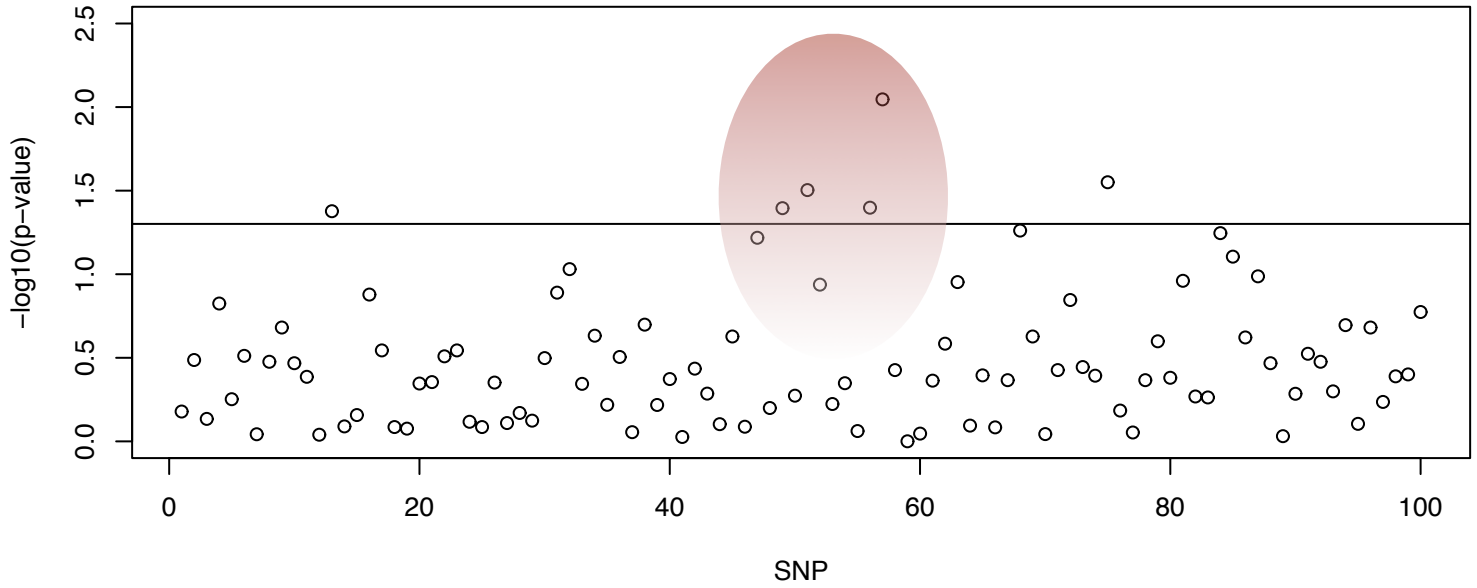
Population 2



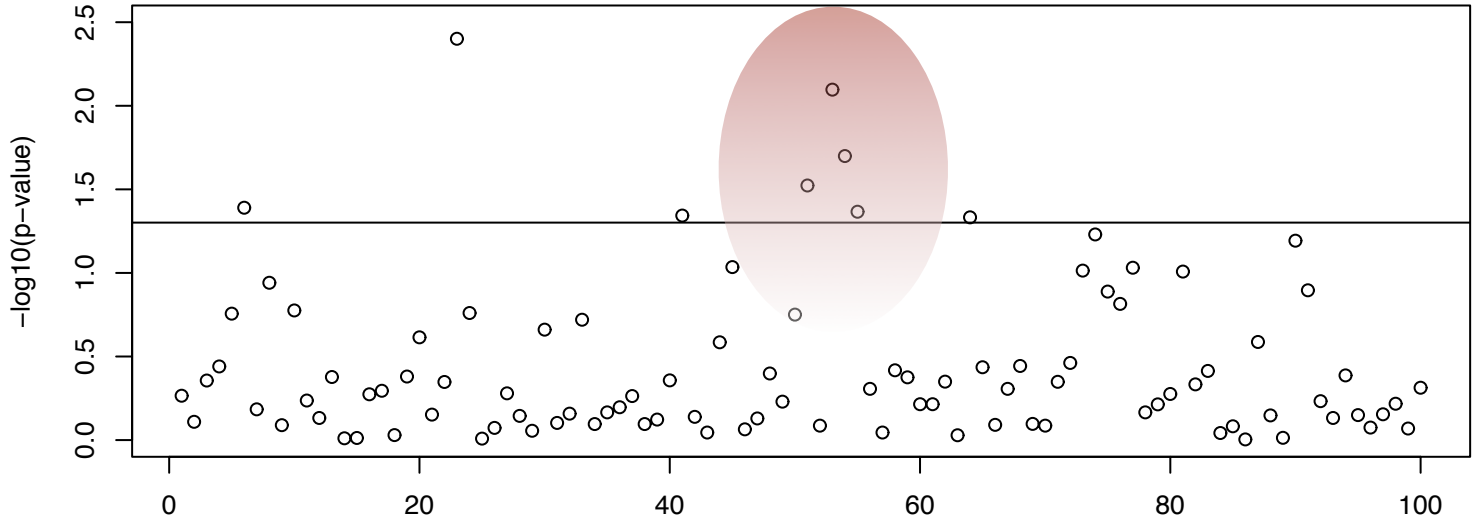
Population 1



Population 2

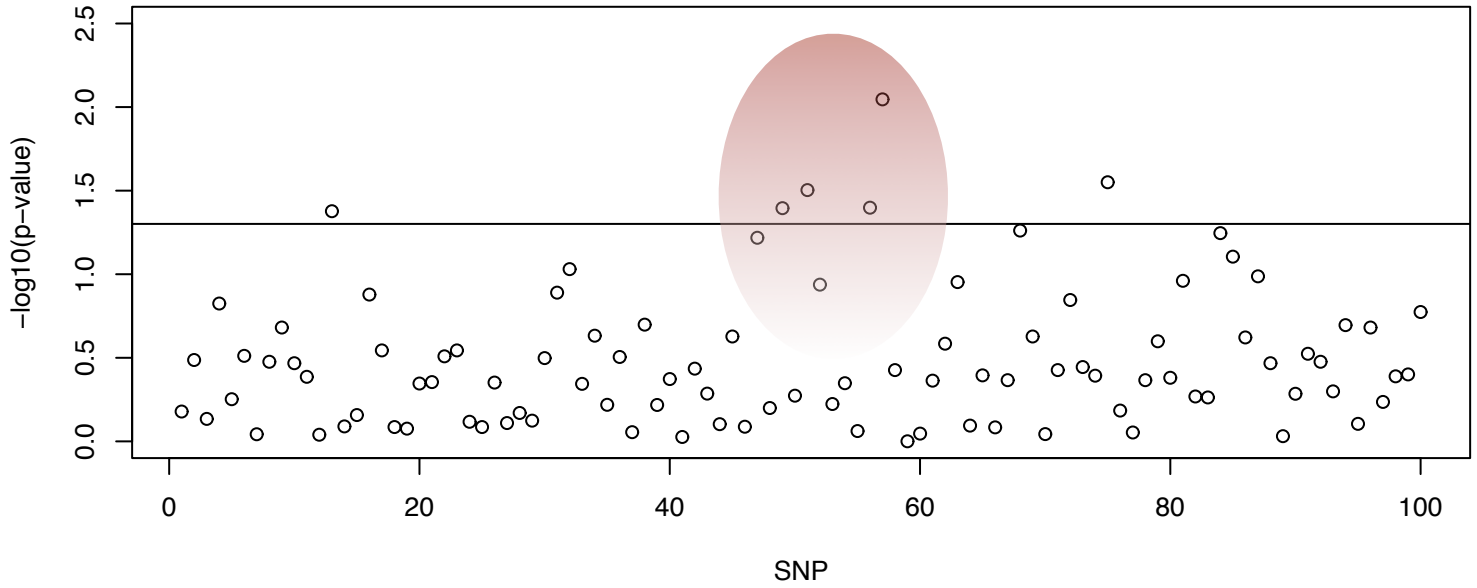


Population 1

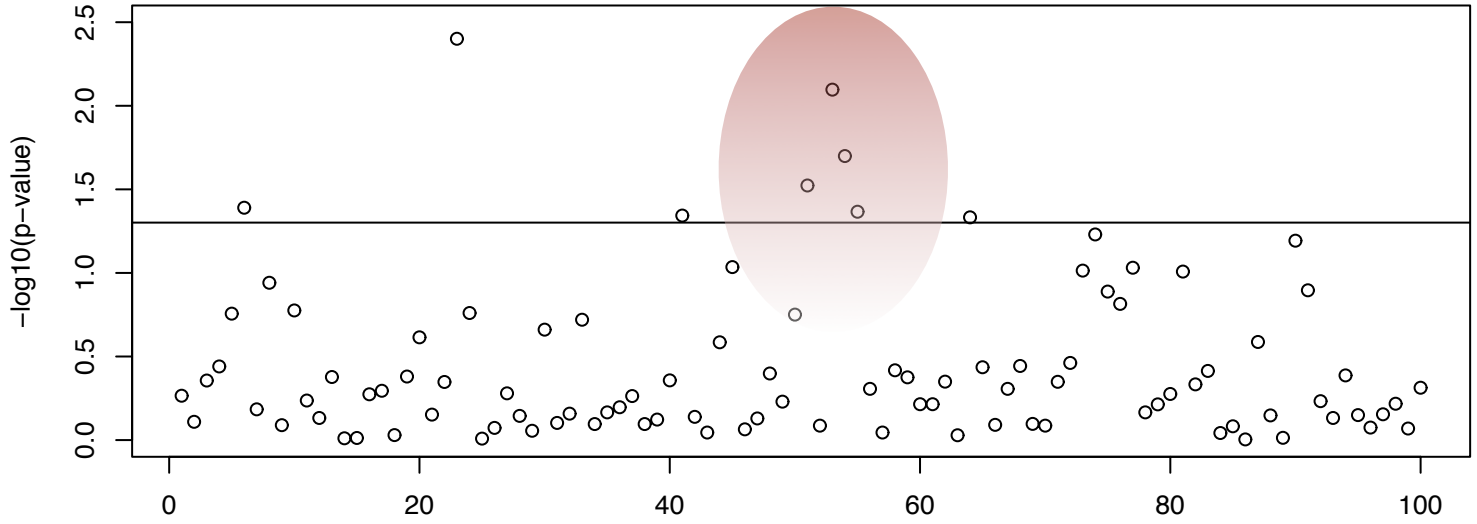


Population 2

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



Population 1



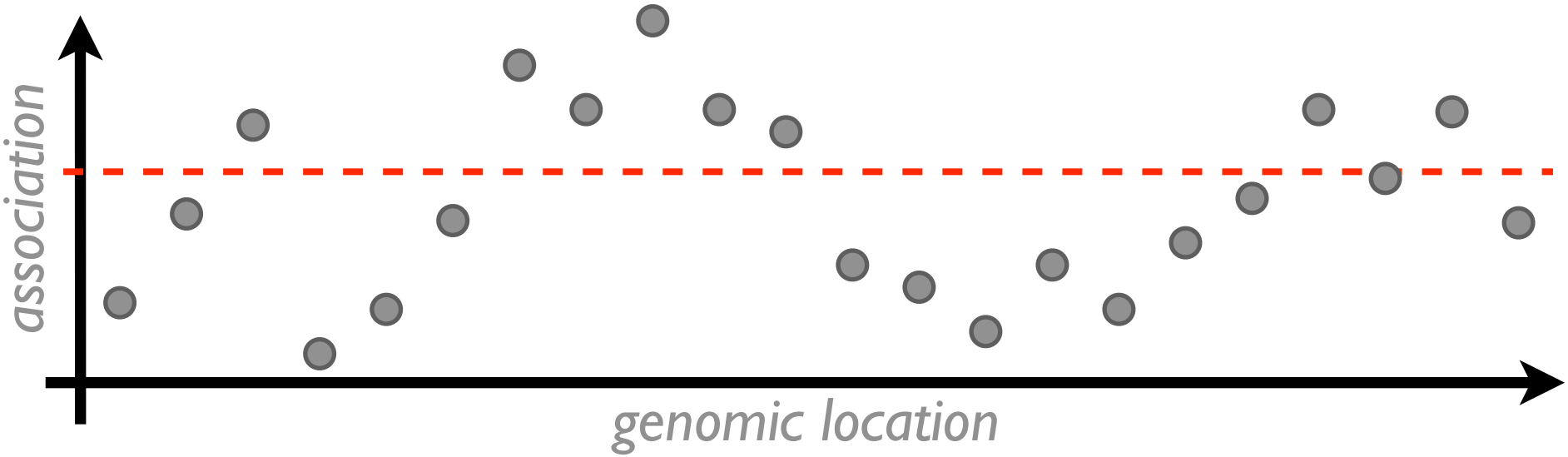
Population 2

Sliding-Frames ?! >> Local Score

Local Score

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

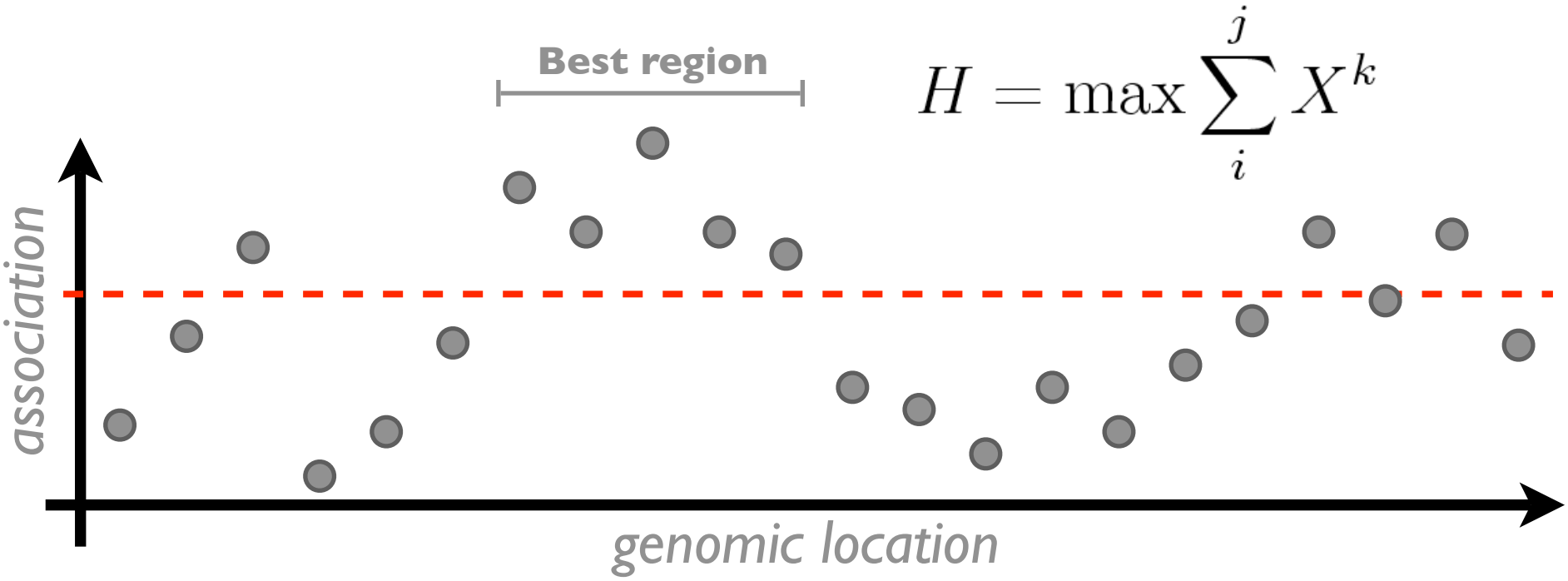
  e.g. Pearson χ^2 on case/control genotype frequencies.



Local Score

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

  e.g. Pearson χ^2 on case/control genotype frequencies.



Local Score

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

Local Score

$H = 4$

Local Score

$H = 4$

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

Local Score

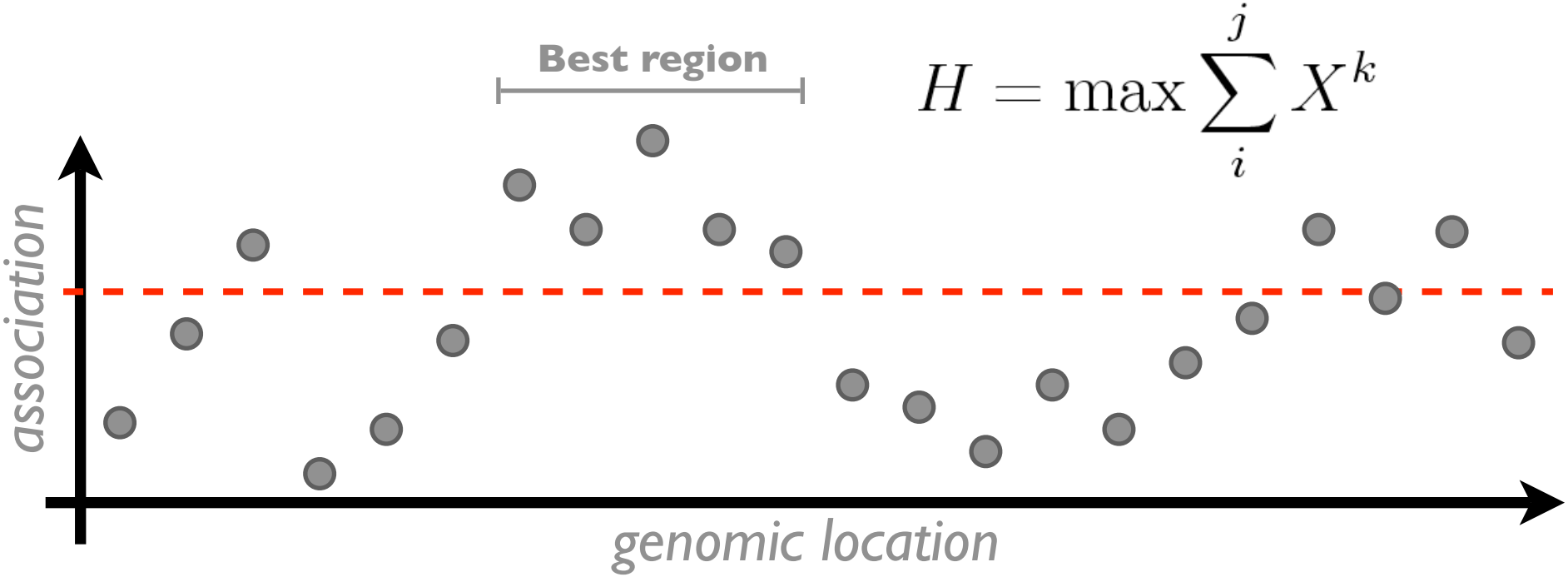
$H = 4$

$H = 6$

Local Score

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

  e.g. Pearson χ^2 on case/control genotype frequencies.

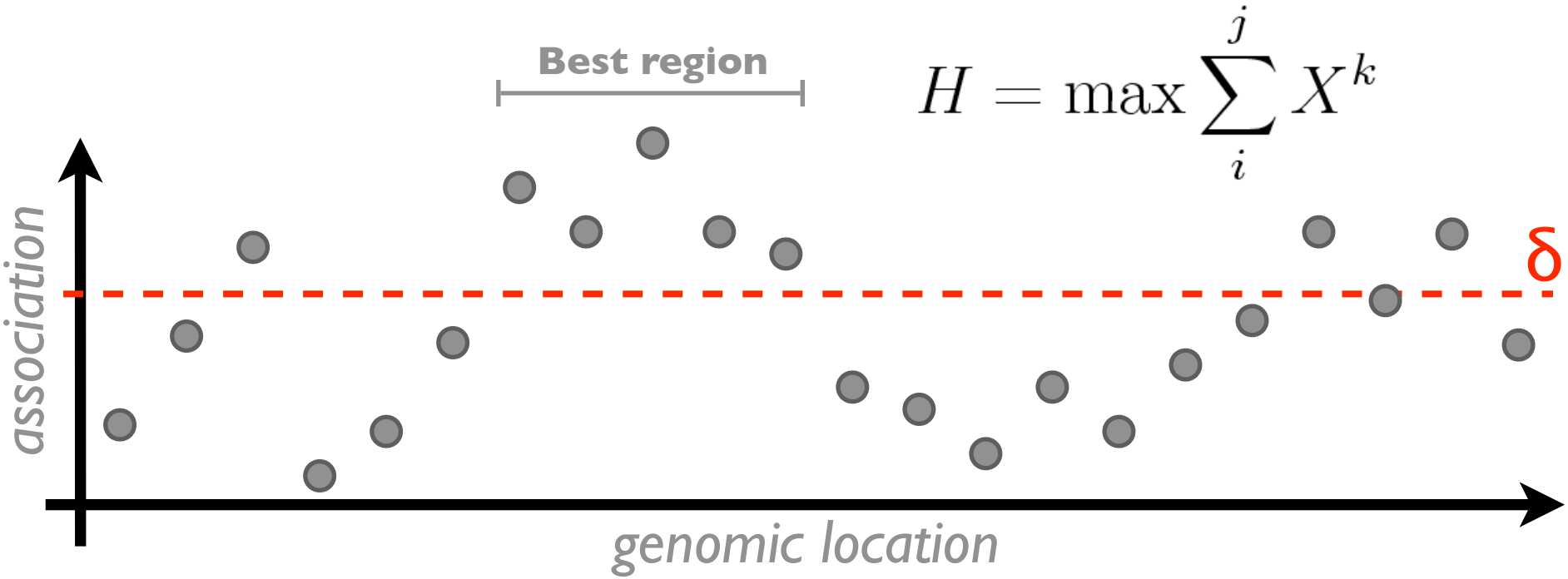


- On average, the sequence $\boldsymbol{X}$ must be negative otherwise the best region would easily span the entire sequence.

Local Score

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

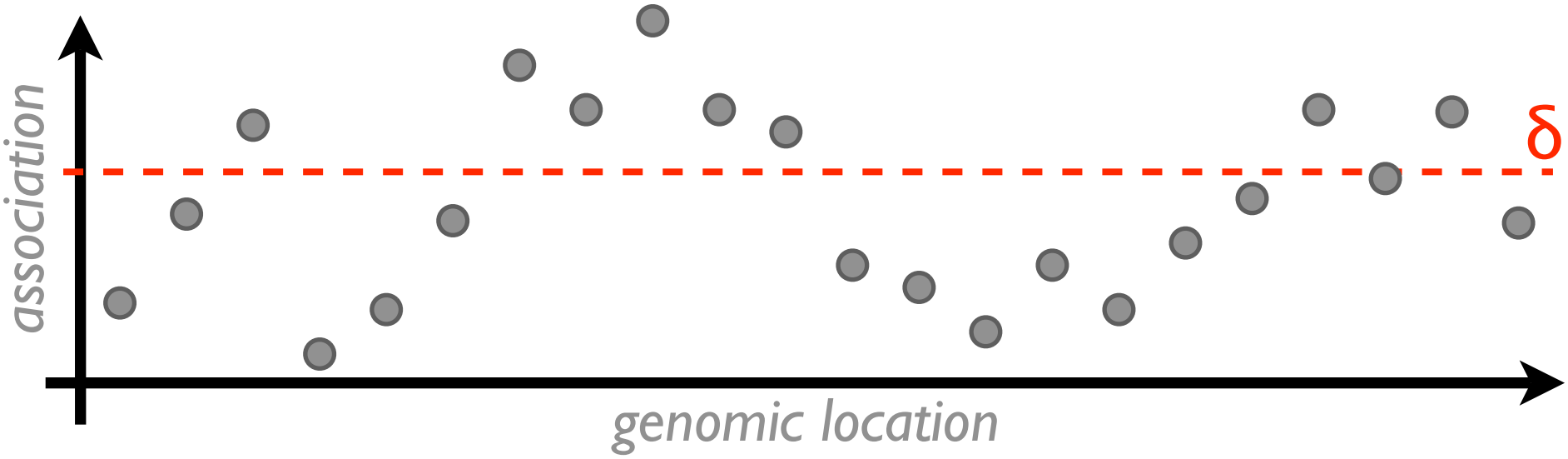
  e.g. Pearson χ^2 on case/control genotype frequencies.



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

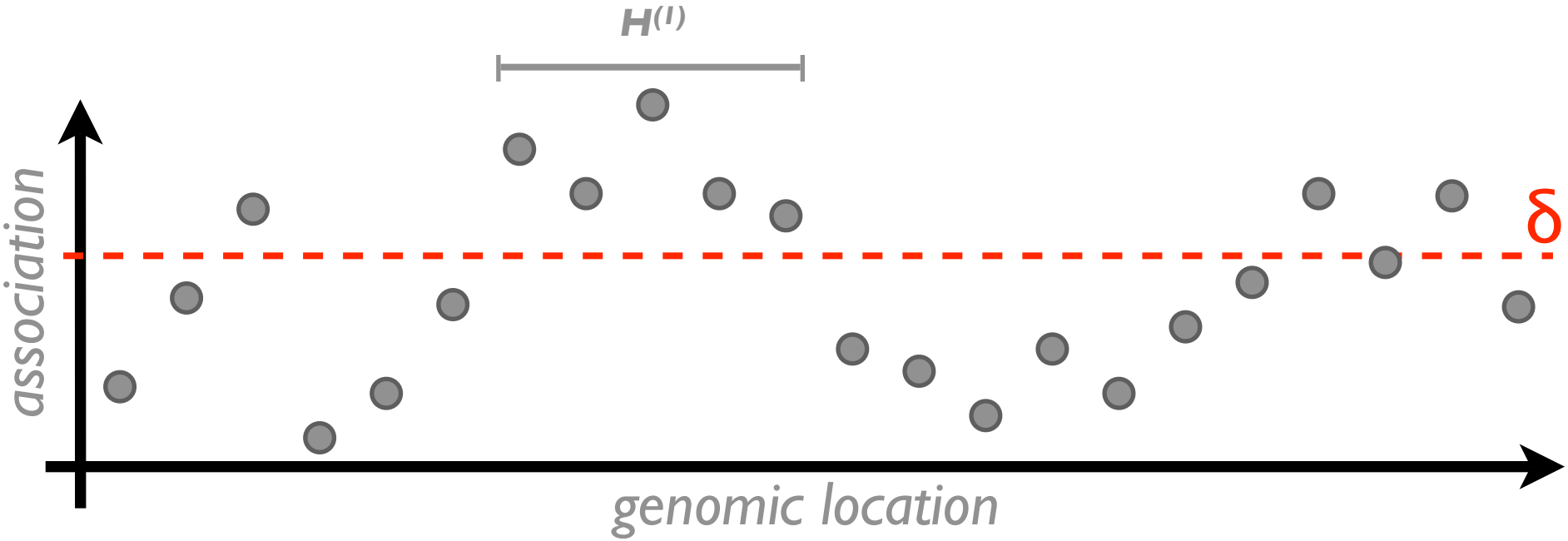
Local Score

- 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.



Local Score

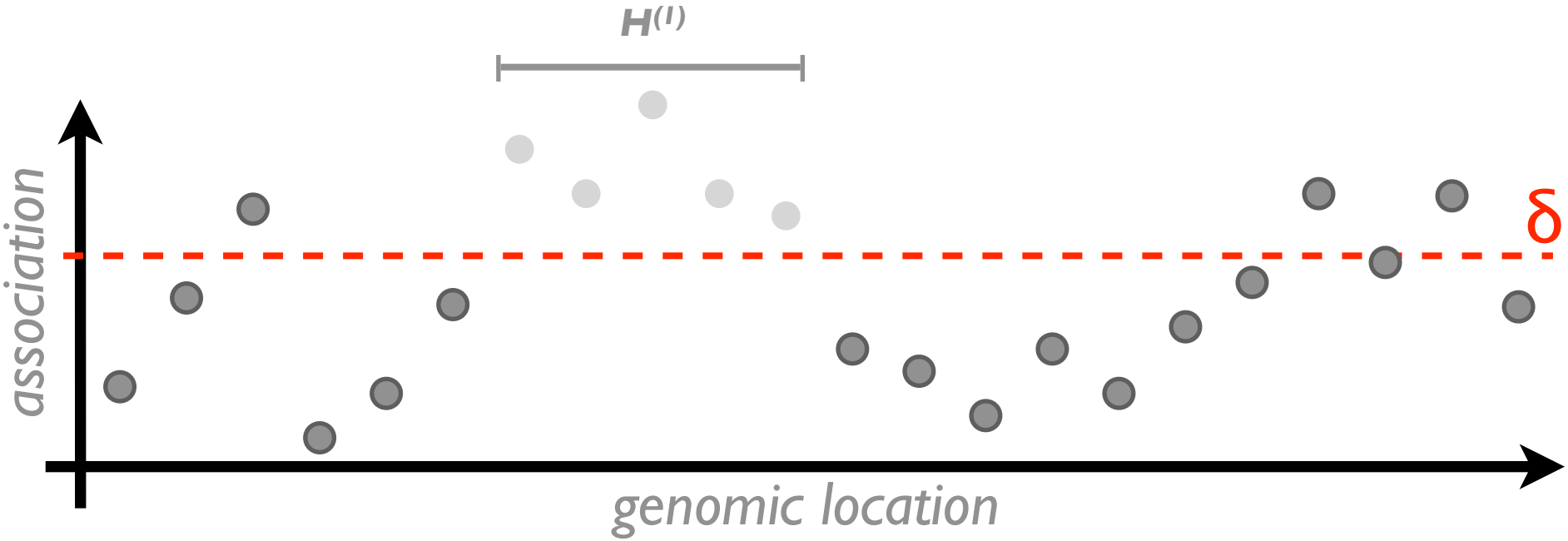
- The *k* first best regions: $H^{(1)}$, ..., $H^{(k)}$.
- $H^{(k)}$ is defined as the Local Score of the initial sequence disjoint from the preceding *k*-1 best regions.



- Find the first best region.

Local Score

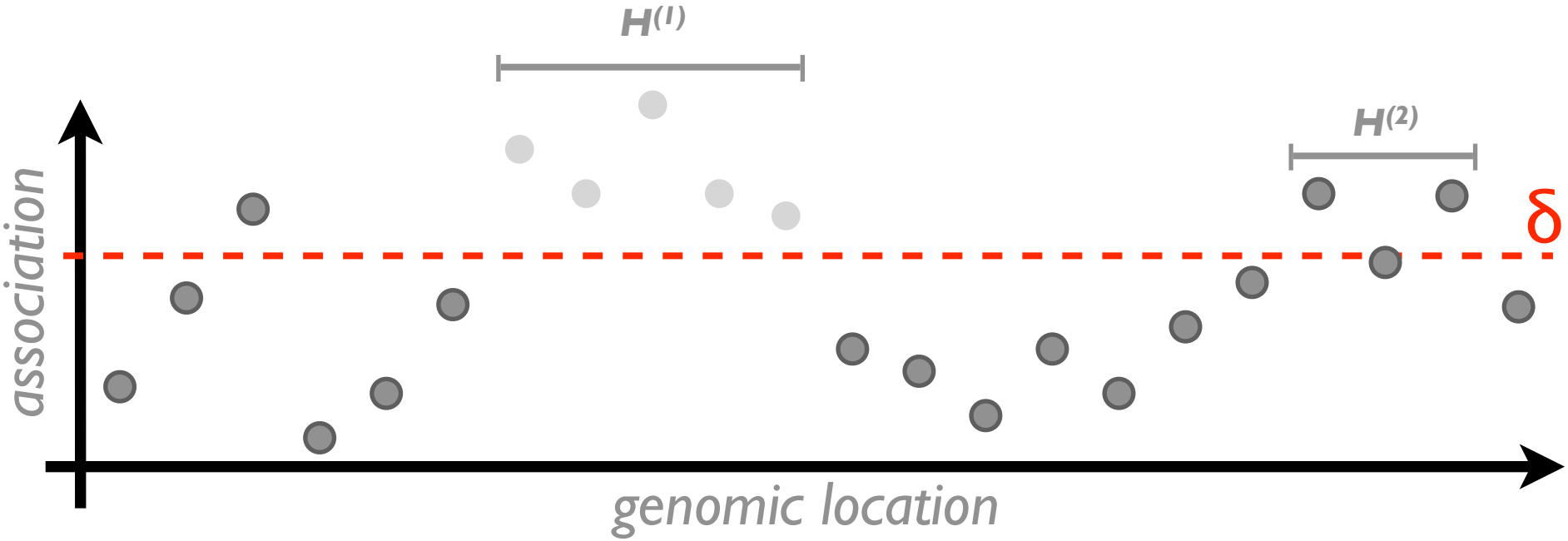
- The *k* first best regions: $H^{(1)}$, ..., $H^{(k)}$.
- $H^{(k)}$ is defined as the Local Score of the initial sequence disjoint from the preceding *k*-1 best regions.



- [x] Find the first best region.
- [ ] Remove it from the sequence.

Local Score

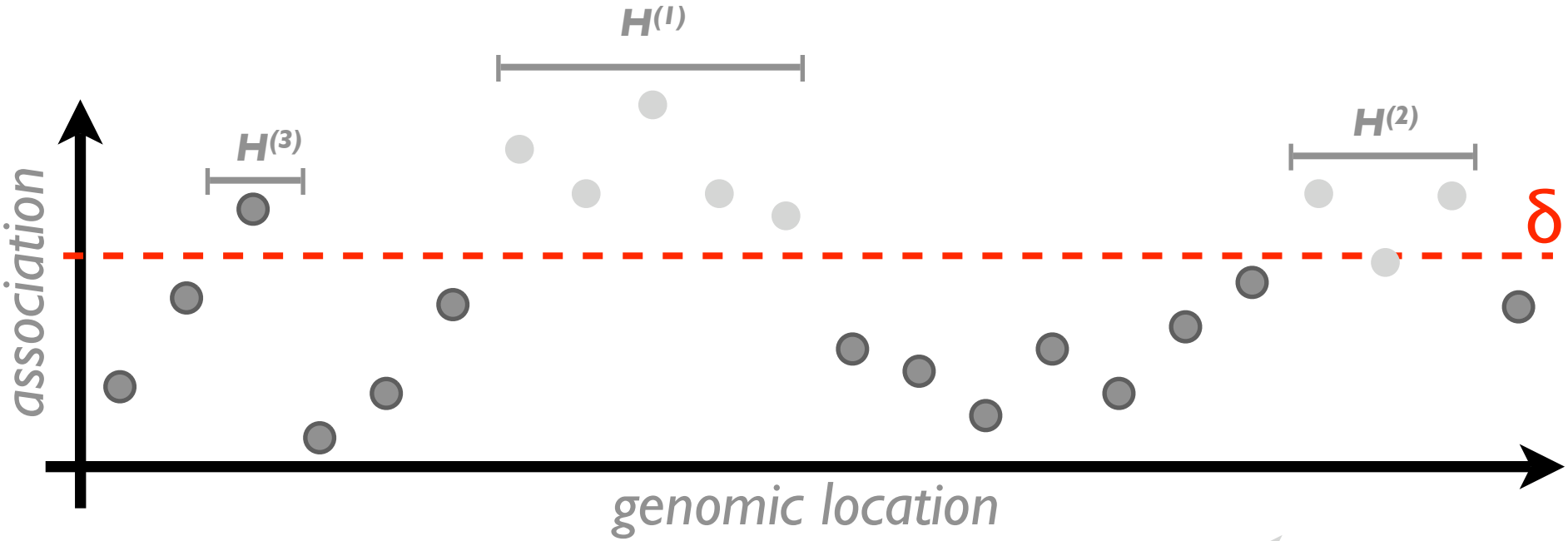
- The *k* first best regions: $H^{(1)}$, ..., $H^{(k)}$.
- $H^{(k)}$ is defined as the Local Score of the initial sequence disjoint from the preceding *k*-1 best regions.



- [x] Find the first best region.
- [x] 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.



- [x] Find the first best region.
- [x] Remove it from the sequence.
- [x] 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	$H^{(5)}$
⋮	⋮
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	$H^{(5)}$	$\longrightarrow$	$pv^{(5)}$
$\vdots$	$\vdots$		$\vdots$
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 (in press)
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

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

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

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

  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

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

  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

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

  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

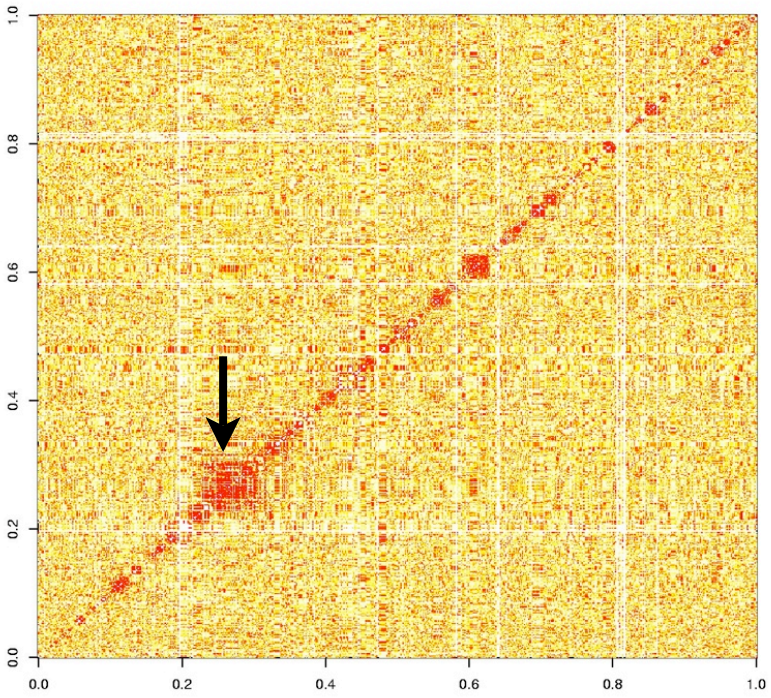
- Based on Monte-Carlo simulations.
- Based on Real Data (to preserve a realistic pattern of LD).
- 301 and 289 chr19 from *French* (pop_A) and *Swedish* (pop_B) controls as an empirical distribution of possible diplotypes.
- chr 19 = 674 SNPs genotyped using a 100K Affymetrix chip.
- This data set is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

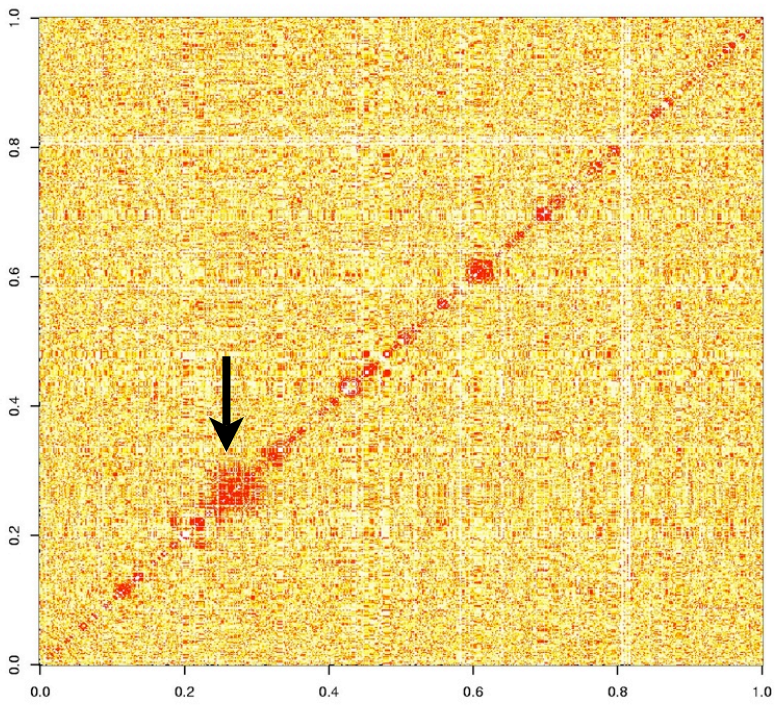
Power study

- 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

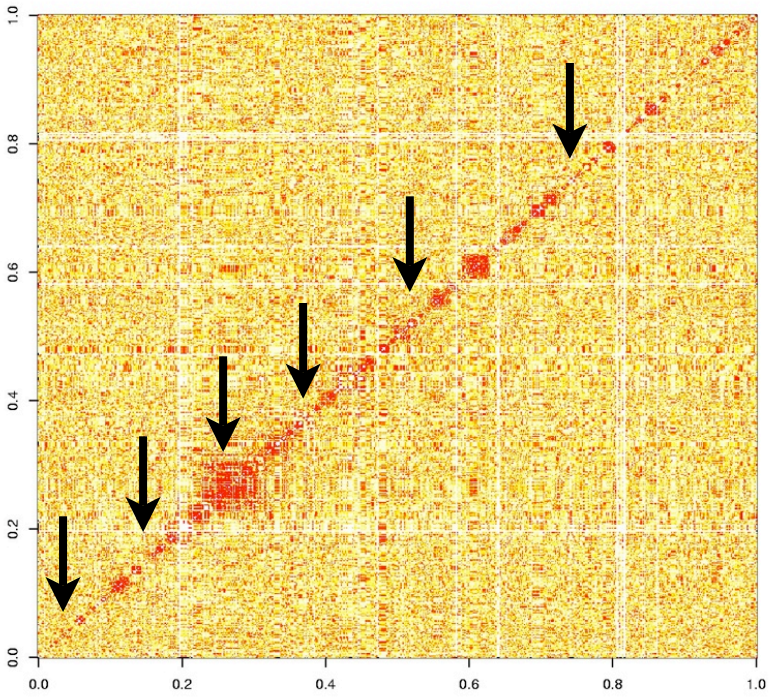
- Situation 1/4:
- The two populations have similar patterns of LD.
- The DSL is localised in a block of LD.

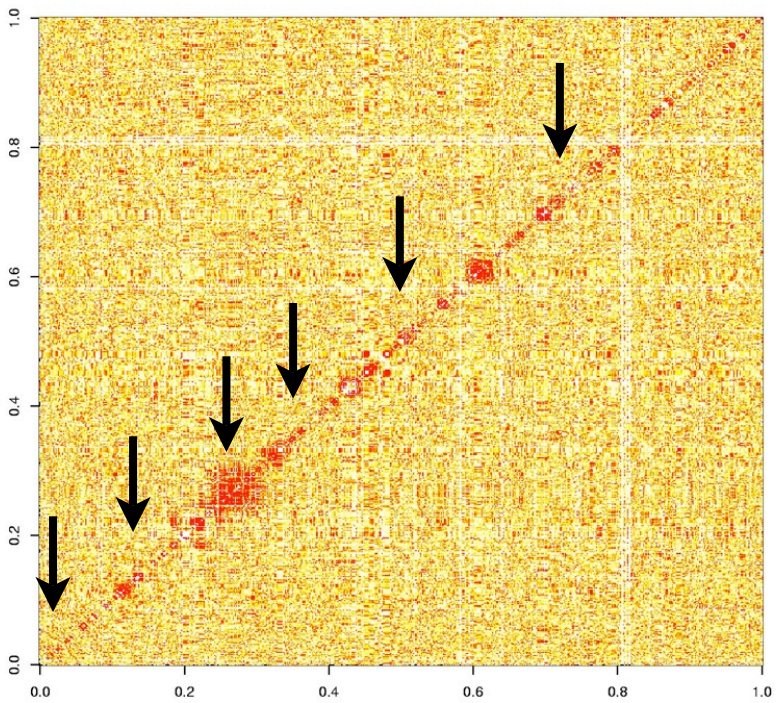




Power study

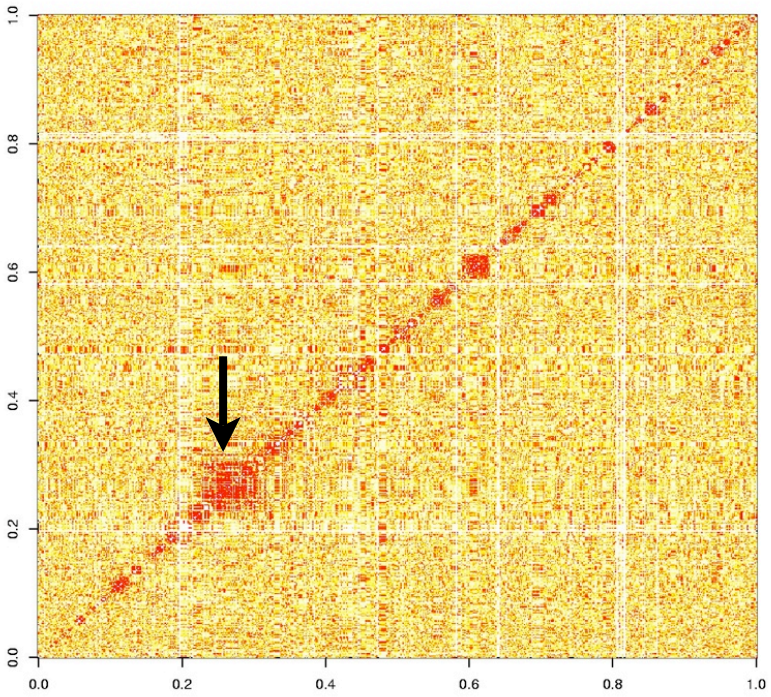
- Situation 2/4:
- The two populations have similar patterns of LD.
- The DSL is randomly chosen among SNPs that present a Minor Genotype Frequency of at least 1%.

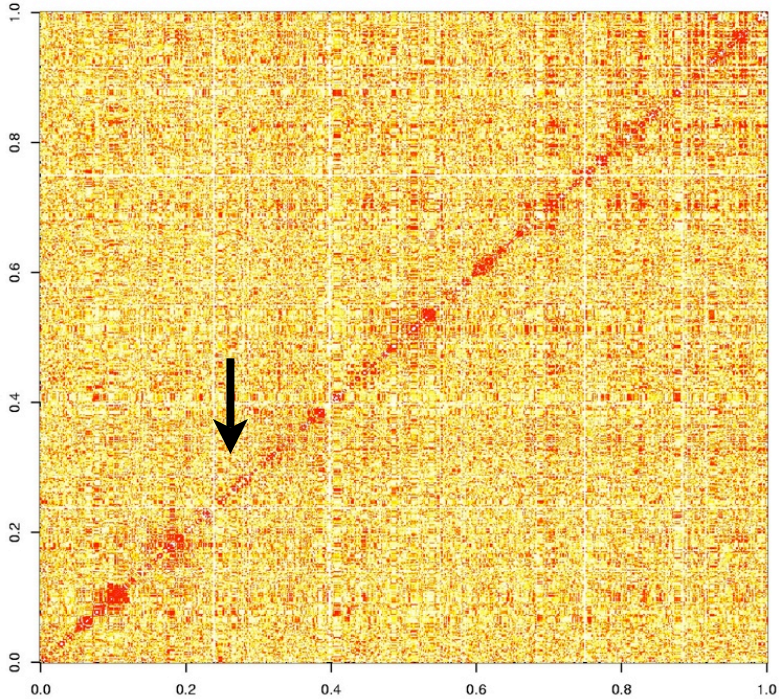




Power study

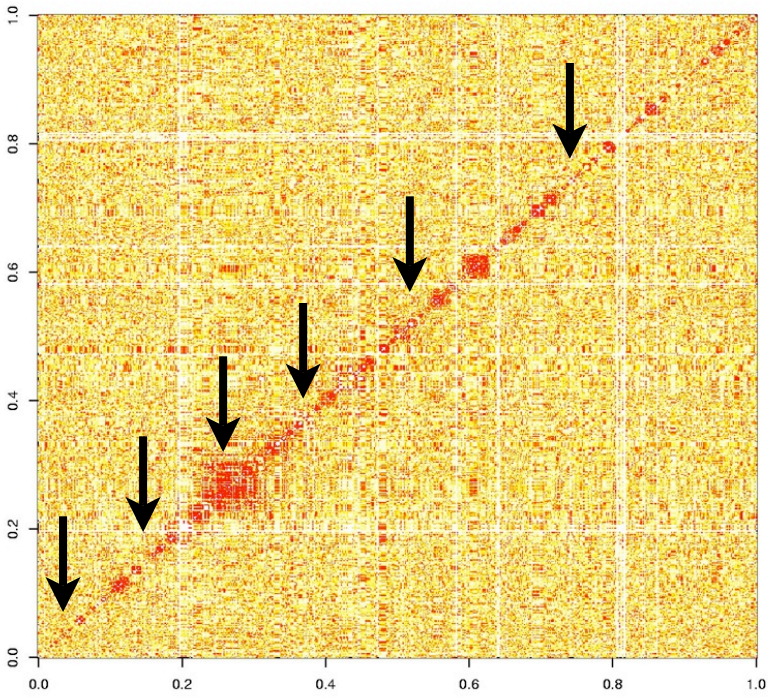
- Situation 3/4:
- The two populations have different patterns of LD.
- The DSL is localised in a block of LD.

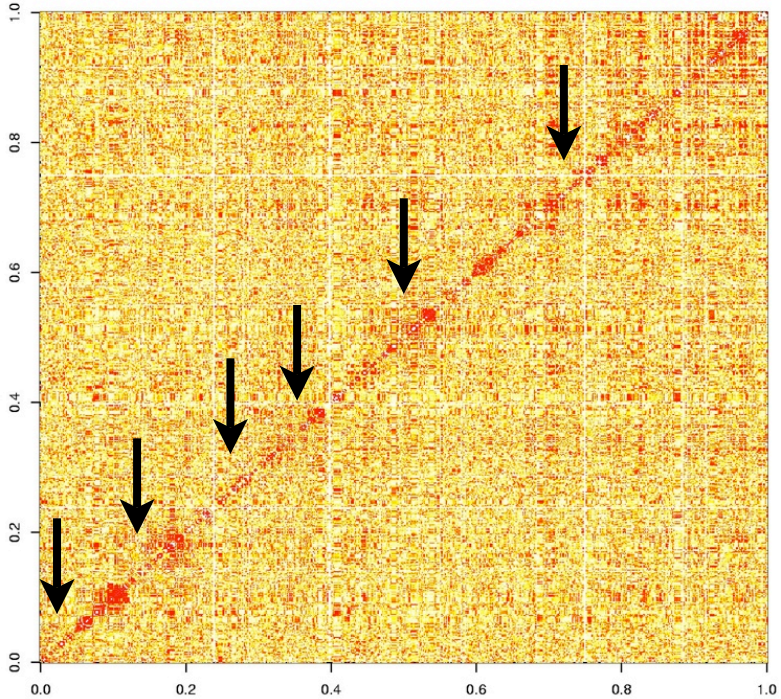




Power study

- 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: $-\log_{10}(pv)$

  → (unbiased) exact allelic test.

Power study

- Test statistic: $-\log_{10}(pv)$
- Local Score: *H0* is rejected if the Local Score of at least the best region is significant at the 5% level.
- $\boldsymbol{X_A} = [\ -\log_{10}(pv_{Ai})\ ]_{i\,=\,1\ldots n}$ and $\boldsymbol{X_B} = [\ -\log_{10}(pv_{Bi})\ ]_{i\,=\,1\ldots n}$
- $\delta = -\log_{10}(0.05)$
- $\boldsymbol{X'_A} = [\ -\log_{10}(pv_{Ai}) - \delta]_{i\,=\,1\ldots n}$
- $\boldsymbol{X'_B} = [\ -\log_{10}(pv_{Bi}) - \delta]_{i\,=\,1\ldots n}$

$\Big]\ \boldsymbol{X'_{AB} = X'_A + X'_B}$

Power study

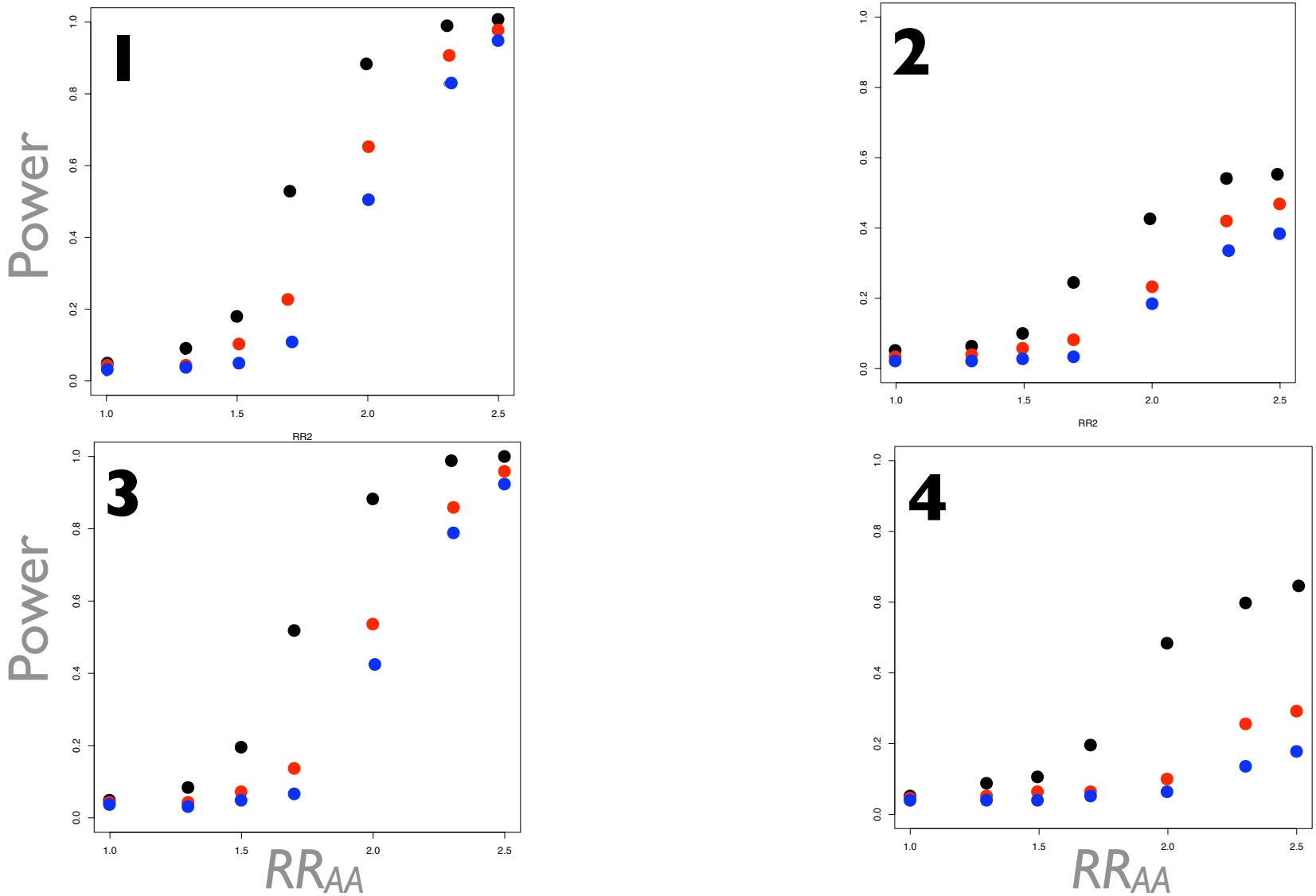
- Test statistic: $-\log_{10}(pv)$
- Local Score: *H0* is rejected if the Local Score of at least the best region is significant at the 5% level.
- $\boldsymbol{X_A} = [\ -\log_{10}(pv_{Ai})\ ]_{i = 1\ldots n}$ and $\boldsymbol{X_B} = [\ -\log_{10}(pv_{Bi})\ ]_{i = 1\ldots n}$
- Single-marker analysis: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

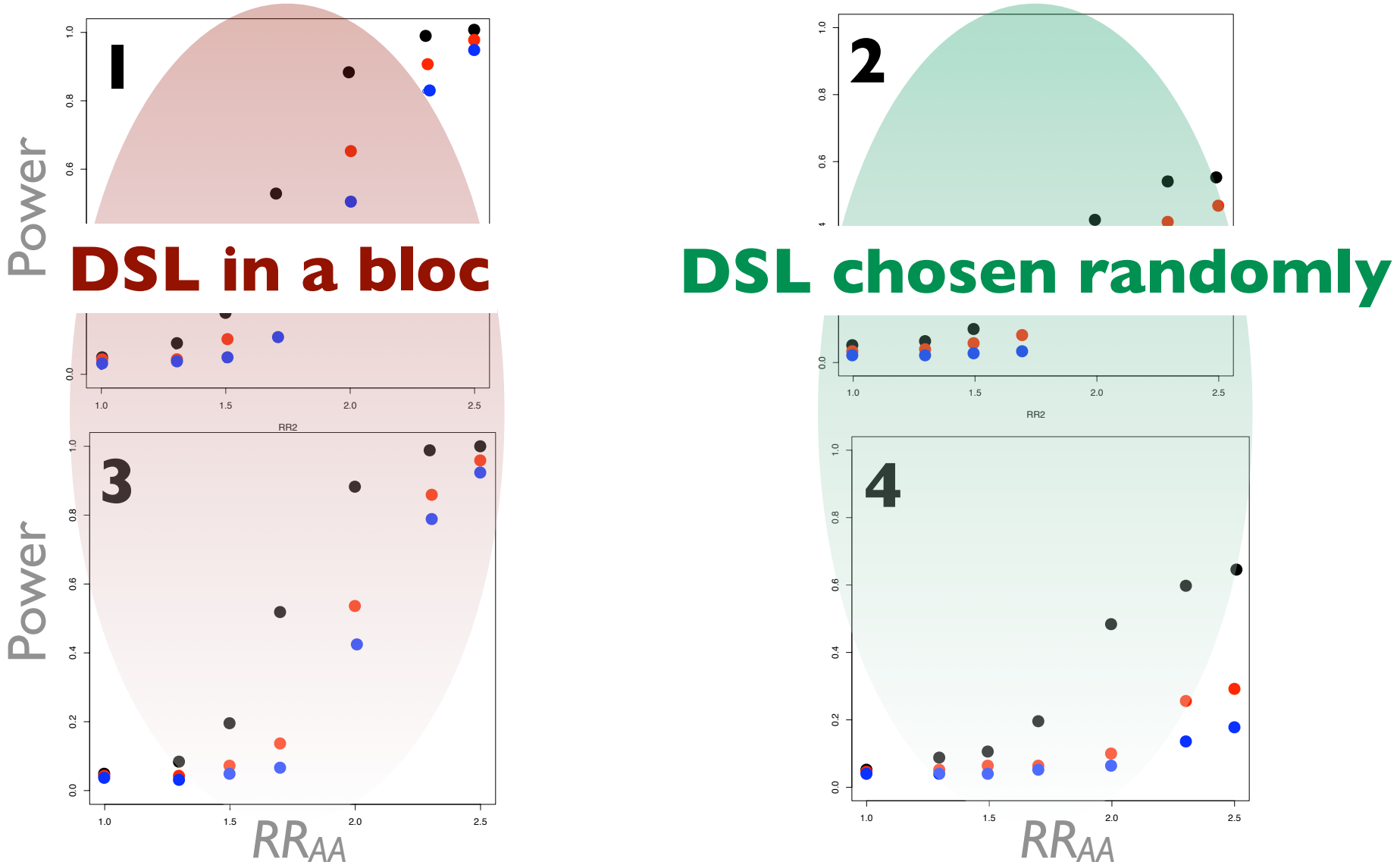
- Results:



Power study

- Local Score
- FWER
- FDR

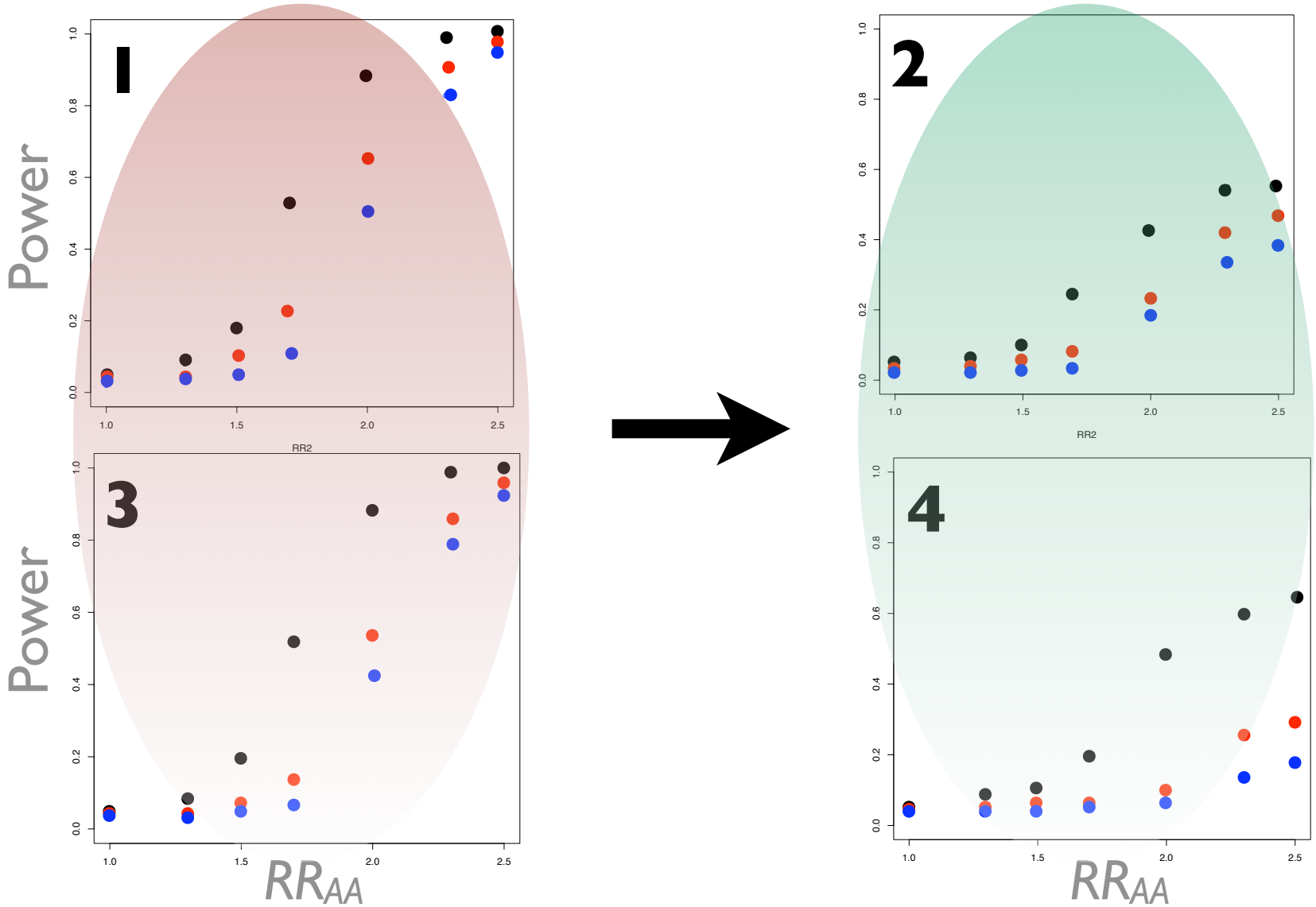
- Results:



Power study

Local Score
FWER
FDR

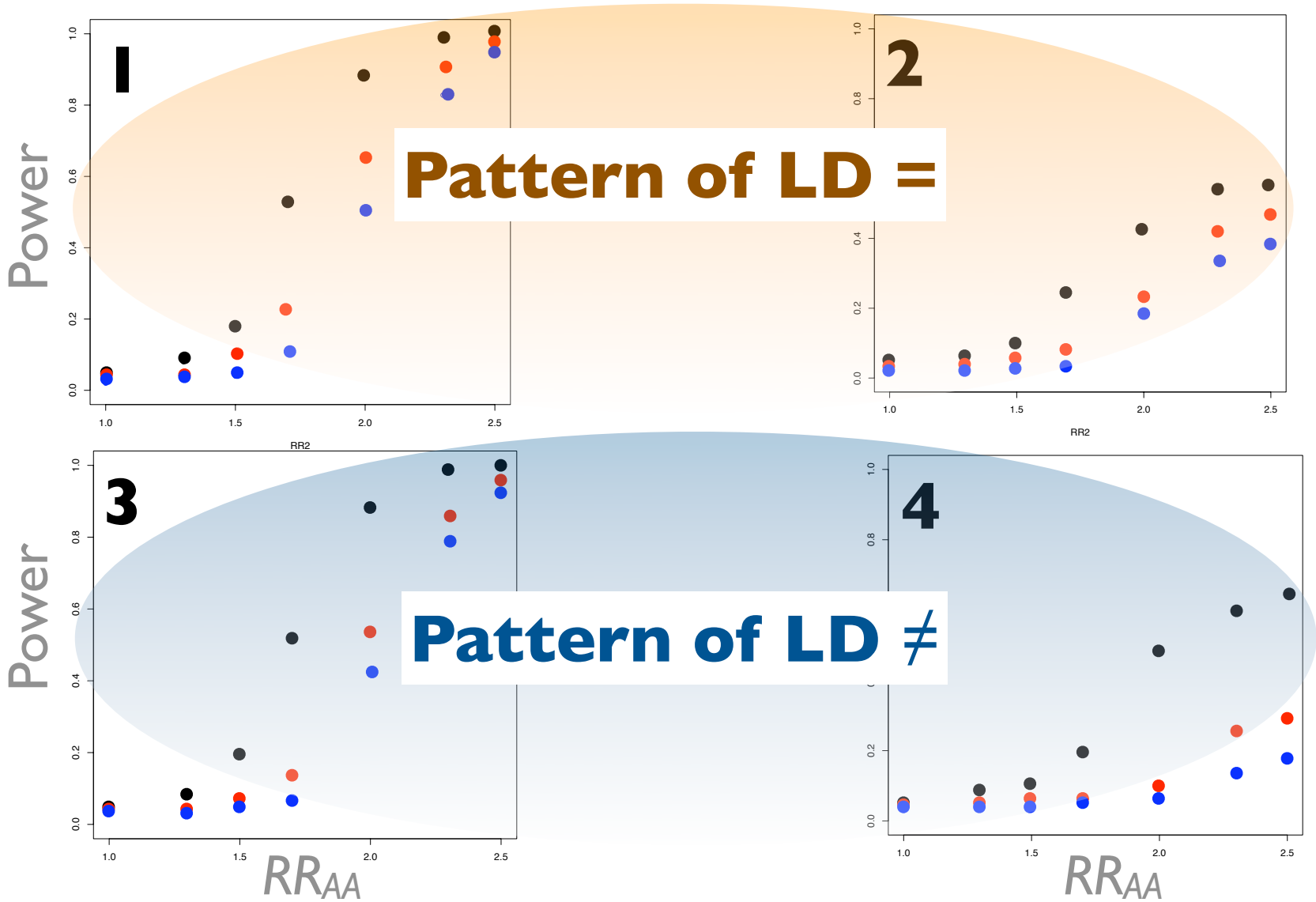
- Results:



Power study

- Local Score
- FWER
- FDR

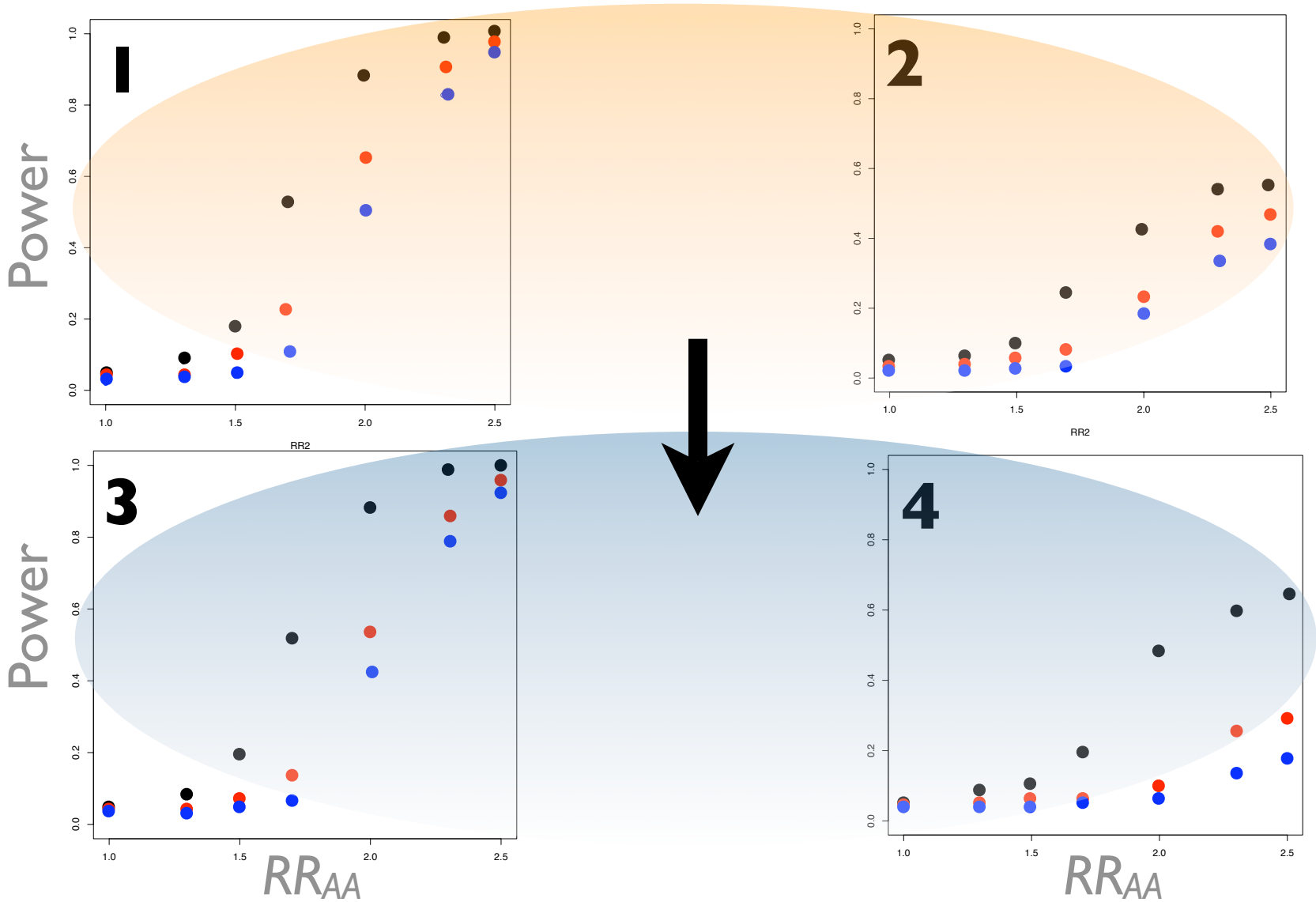
- Results:



Power study

- Local Score
- FWER
- FDR

Results:



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

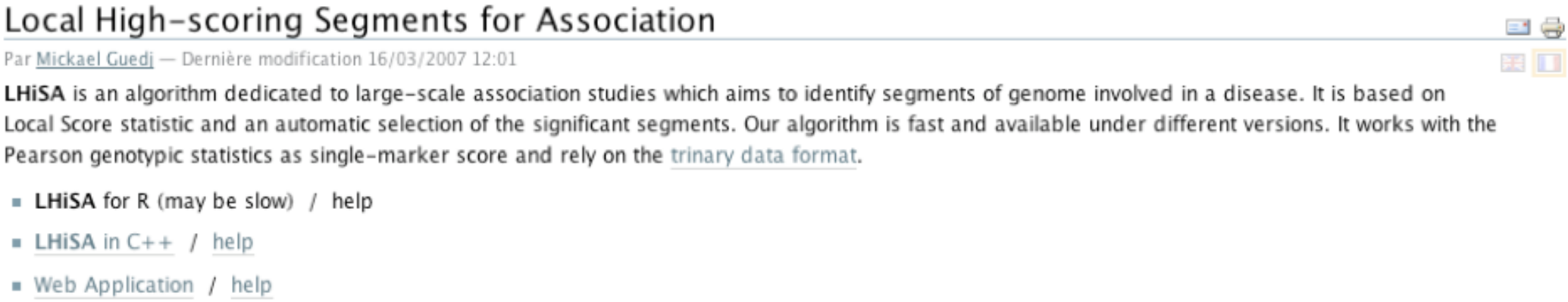
- Looking at Local Replications appears more robust to biological differences between populations.
- Local Score as a simple and natural framework.
- Strict Replications show a stronger evidence for true replication.
- Considering Local Replications can help to identify DSL shared across populations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

Software : LHiSA

- C++

- R (new) 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 Guedj — 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 and B Prum for supervision.

Merck-Serono for the data.

F Demenais for useful discussions.

IGES Scientific Program Comittee.

Email: mickael.guedj@genopole.cnrs.fr

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

$\mathbf{X'_A} = X'_{A1} \quad X'_{A2} \quad X'_{A3} \quad X'_{A4} \quad X'_{A5}$

$\mathbf{X'_B} = X'_{B1} \quad X'_{B2} \quad X'_{B3} \quad X'_{B4} \quad X'_{B5}$

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

Different Marker Sets

$\mathbf{X'_A} = X'_{A1} \quad X'_{A2} \quad X'_{A3} \quad \_ \quad X'_{A5}$

$\mathbf{X'_B} = X'_{B1} \quad \_ \quad X'_{B3} \quad X'_{B4} \quad X'_{B5}$

$\mathbf{X'_{AB}} = X'_{A1}+X'_{B1} \quad X'_{A2} \quad X'_{A3}+X'_{B3} \quad X'_{B4} \quad X'_{A5}+X'_{B5}$