Catching Local Replications: a Local Score-based approach to replicated association studies.

Mickaël Guedj\textsuperscript{1,2}, Jérôme Wojcik\textsuperscript{2} and Grégory Nuel\textsuperscript{1}.

IGES 2007, York UK

\textsuperscript{1} Statistics and Genome laboratory, CNRS, INRA, University of Evry, FRANCE
\textsuperscript{2} Serono Pharmaceutical Research, Geneva, SWITZERLAND
Introduction

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

![Images of population distributions](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
Sliding-Frames

Population 1

Population 2

the frame size has to be specified
Population 1

Population 2

Sliding-Frames ?! >> Local Score
Definition: Let $\mathbf{X} = (X_i)_{i=1}^n$ be a sequence of random variables $\rightarrow$ association statistics:
e.g. Pearson $\chi^2$ on case/control genotype frequencies.
Local Score

Definition: Let $X = (X_i)_{i=1}^n$ be a sequence of random variables associated with 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

\[
\begin{array}{cccccccc}
1 & -2 & -4 & 2 & 1 & 1 & -3 & 1 & -2 \\
\end{array}
\]

\[H = 4\]
Local Score

\[ \begin{array}{ccccccccccc}
1 & -2 & -4 & 2 & 1 & 1 & -3 & 1 & -2 \\
-1 & 2 & 1 & -4 & -2 & -2 & 2 & 1 & -1 & 3 & 1 & -2 \\
\end{array} \]

\[ H = 4 \]
Local Score

\[
\begin{array}{cccccccc}
1 & -2 & -4 & 2 & 1 & 1 & -3 & 1 & -2 \\
-1 & 2 & 1 & -4 & -2 & -2 & 2 & 1 & -1 & 3 & 1 & -2 \\
\end{array}
\]

\[H = 4\]

\[
\begin{array}{cccccccc}
1 & -2 & -4 & 2 & 1 & 1 & -3 & 1 & -2 \\
-1 & 2 & 1 & -4 & -2 & -2 & 2 & 1 & -1 & 3 & 1 & -2 \\
\end{array}
\]

\[H = 6\]
Local Score

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

On average, the sequence \( X \) must be negative otherwise the best region would easily span the entire sequence.

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

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

On average, the sequence $X$ must be negative otherwise the best region would easily span the entire sequence $\Rightarrow X' = X - \delta$ ($\delta = 5\%$ 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: $H^{(1)}, ..., H^{(k)}$.
- $H^{(k)}$ 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: $H^{(1)}, \ldots, H^{(k)}$.
- $H^{(k)}$ 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.

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

- Statistical significance of the regions:

<table>
<thead>
<tr>
<th>Region</th>
<th>$H^{(k)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 1</td>
<td>$H^{(1)}$</td>
</tr>
<tr>
<td>Region 2</td>
<td>$H^{(2)}$</td>
</tr>
<tr>
<td>Region 3</td>
<td>$H^{(3)}$</td>
</tr>
<tr>
<td>Region 4</td>
<td>$H^{(4)}$</td>
</tr>
<tr>
<td>Region 5</td>
<td>$H^{(5)}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Region $k$</td>
<td>$H^{(k)}$</td>
</tr>
</tbody>
</table>
Local Score

- Statistical significance of the regions:

Region 1 \( H^{(1)} \rightarrow pv^{(1)} \)
Region 2 \( H^{(2)} \rightarrow pv^{(2)} \)
Region 3 \( H^{(3)} \rightarrow pv^{(3)} \)
Region 4 \( H^{(4)} \rightarrow pv^{(4)} \)
Region 5 \( H^{(5)} \rightarrow pv^{(5)} \)

\[ \vdots \]
Region \( k \) \( H^{(k)} \rightarrow pv^{(k)} \)
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):

  \[ Pr \left( H \geq \frac{\ln n}{\lambda} + x \right) \approx 1 - \exp(-Ke^{-\lambda x}) \]  

  Gumbel distribution

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

- **In Statistics:** asymptotic and exact distributions
  
  e.g. Iglehart (1972)
  Extreme values 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:


Application in a two-stage design:

Aschard, Guedj and Demenais (in press)
Local Score

- Application to Local Replications:
Local Score

- Application to Local Replications:

  - Let $\text{pop}_A$ and $\text{pop}_B$ denote the two populations and

  $\mathbf{X}_A = (X_{Ai})_{i=1...n}$ and $\mathbf{X}_B = (X_{Bi})_{i=1...n}$

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

Application to Local Replications:

Let $\text{pop}_A$ and $\text{pop}_B$ denote the two populations and

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

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

Let $X'_A = X_A - \delta$ and $X'_B = X_B - \delta$. 
Local Score

- Application to Local Replications:
- Let $\text{pop}_A$ and $\text{pop}_B$ denote the two populations and 
  
  \[ X_A = (X_{Ai})_{i=1}^{n} \text{ and } X_B = (X_{Bi})_{i=1}^{n} \]

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

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

- $X'_{AB} = X'_A + X'_B$ : on which we apply the Local Score.
Local Score

- Application to Local Replications:
  - Let $\text{pop}_A$ and $\text{pop}_B$ denote the two populations and
    
    \[ X_A = (X_{Ai})_{i=1}^{n} \text{ and } X_B = (X_{Bi})_{i=1}^{n} \]
    
    their respective sequences of test statistics for the same set of markers.
  
  - Let $X'_A = X_A - \delta$ and $X'_B = X_B - \delta$.
  
  - $X'_{AB} = X'_A + X'_B$: 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\textsubscript{A}) and Swedish (pop\textsubscript{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 $\Rightarrow RR_{aA} = (RR_{AA}+1)/2$

- 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%.
Power study

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

  \[\Rightarrow\] (unbiased) exact allelic test.
Power study

- **Test statistic:** \(-\log_{10}(p_v)\)

- **Local Score:** \(H_0\) is rejected if the Local Score of at least the best region is significant at the 5% level.

- \(X_A = \left[ -\log_{10}(p_{v_{Ai}}) \right]_{i = 1 \ldots n}\) and \(X_B = \left[ -\log_{10}(p_{v_{Bi}}) \right]_{i = 1 \ldots n}\)

- \(\delta = -\log_{10}(0.05)\)

- \(X'_A = \left[ -\log_{10}(p_{v_{Ai}}) - \delta \right]_{i = 1 \ldots n}\)

- \(X'_B = \left[ -\log_{10}(p_{v_{Bi}}) - \delta \right]_{i = 1 \ldots n}\)

- \(X'_{AB} = X'_A + X'_B\)
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.

- $X_A = \left[ -\log_{10}(pv_{Ai}) \right]_{i=1}^{n}$ and $X_B = \left[ -\log_{10}(pv_{Bi}) \right]_{i=1}^{n}$

- 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

Results:

1. Local Score
2. FWER
3. FDR

Power

$RR_2$

$RR_{AA}$
Power study

Results:

1. DSL in a bloc

2. DSL chosen randomly

- Local Score
- FWER
- FDR
Power study

Results:

1. Local Score
2. FWER
3. FDR

Graphs showing the relationship between Power and RR_{AA} for different values of RR2.
Power study

Results:

Pattern of LD =

Pattern of LD ≠

Local Score
FWER
FDR
Power study

Results:

1. Local Score
2. FWER
3. FDR

Power study results shown in the graph with different markers representing Local Score, FWER, and FDR.
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
Acknowledgements

G Nuel, J Wojcik and B Prum for supervision.
Merck-Serono for the data.
F Demenais for useful discussions.
IGES Scientific Program Committee.

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

<table>
<thead>
<tr>
<th></th>
<th>$X'_A$</th>
<th>$X'_A$</th>
<th>$X'_A$</th>
<th>$X'_A$</th>
<th>$X'_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X'_A$</td>
<td>$X'_{A1}$</td>
<td>$X'_{A2}$</td>
<td>$X'_{A3}$</td>
<td>$X'_{A4}$</td>
<td>$X'_{A5}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$X'_B$</th>
<th>$X'_B$</th>
<th>$X'_B$</th>
<th>$X'_B$</th>
<th>$X'_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X'_B$</td>
<td>$X'_{B1}$</td>
<td>$X'_{B2}$</td>
<td>$X'_{B3}$</td>
<td>$X'_{B4}$</td>
<td>$X'_{B5}$</td>
</tr>
</tbody>
</table>

### Different Marker Sets

<table>
<thead>
<tr>
<th></th>
<th>$X'_A$</th>
<th>$X'_A$</th>
<th>$X'_A$</th>
<th>$X'_A$</th>
<th>$X'_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X'_A$</td>
<td>$X'_{A1}$</td>
<td>$X'_{A2}$</td>
<td>$X'_{A3}$</td>
<td>_</td>
<td>$X'_{A5}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$X'_B$</th>
<th>$X'_B$</th>
<th>$X'_B$</th>
<th>$X'_B$</th>
<th>$X'_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X'_B$</td>
<td>_</td>
<td>$X'_{B3}$</td>
<td>$X'_{B4}$</td>
<td>$X'_{B5}$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$X'_A$</th>
<th>$X'_A$</th>
<th>$X'_A$</th>
<th>$X'_A$</th>
<th>$X'_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X'_A$</td>
<td>$X'_{A1}$</td>
<td>$X'_{A2}$</td>
<td>_</td>
<td>$X'_{A3}$</td>
<td>$X'<em>{A4}$ + $X'</em>{B4}$</td>
</tr>
</tbody>
</table>