

Genome-wide association in the presence of high linkage disequilibrium

Quantifying The Influence of a Single SNP

Georgia Tsiliki, Maria De Iorio, David A. Stephens

Department of Mathematics, Imperial College London

Epidemiology, Public Health and Primary Care, Imperial College London

Department of Mathematics and Statistics, McGill University

Imperial College
London



- SNP Association Studies
- The Diabetes Case Control Study
- Quantifying The Causal Effect of a Single SNP
- Linkage Disequilibrium and Haplotype Blocks

Single Nucleotide Polymorphisms (SNPs)

Observed polymorphisms are of interest because

- (a) they contain information about the unobserved underlying genealogy of the population of interest,
- (b) they are believed (individually or collectively) to cause the diseases under study

Single Nucleotide Polymorphisms (SNPs) are the most frequent form of polymorphisms in the human genome (1 SNP every few hundred *bp*)

A	G	A	G	A	C	C	G	A	...	C	T	A	G
A	G	A	G	A	C	C	T	A	...	C	T	A	G
A	G	A	G	A	C	C	G	A	...	C	G	A	G
							X_1						

A genome-wide association study identifies novel risk loci for type 2 diabetes

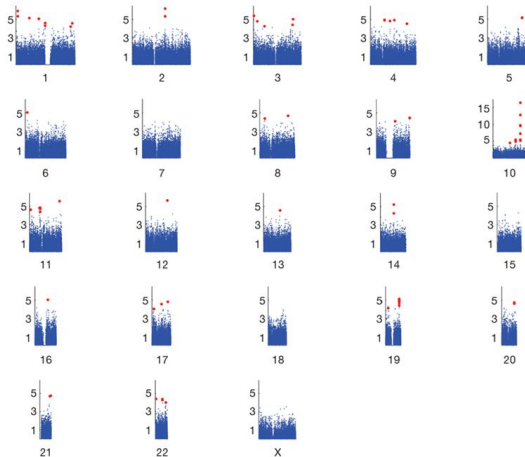
Robert Sladek^{1,2,4}, Ghislain Rocheleau^{1*}, Johan Rung^{1*}, Christian Dina^{5*}, Lishuang Shen¹, David Serre¹, Philippe Boutin⁵, Daniel Vincent⁴, Alexandre Belisle⁴, Samy Hadjadj⁶, Beverley Balkau⁷, Barbara Heude⁷, Guillaume Charpentier⁸, Thomas J. Hudson^{4,9}, Alexandre Montpetit⁴, Alexey V. Pshezhetsky¹⁰, Marc Prentki^{10,11}, Barry I. Posner^{2,12}, David J. Balding^{1,3}, David Meyre⁵, Constantin Polychronakos^{1,3} & Philippe Froguel^{5,14}

Nature 445, 881-885 (22 February 2007)

- French population case-control study
- 1,363 T2DM cases and controls
- 392,935 SNPs genome wide
- Two high-throughput platforms
 - ▶ Illumina Infinium Human1 BeadArrays
 - ▶ Human Hap300 BeadArrays

T2DM-association study

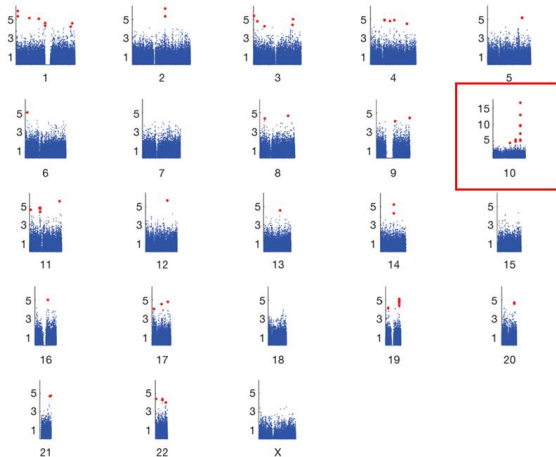
Plots of $-\log_{10}(p)$ for p-values from composite tests of association.



(Sladek *et al.*, Nature, 2007)

T2DM-association study

Plots of $-\log_{10}(p)$ for p-values from composite tests of association.



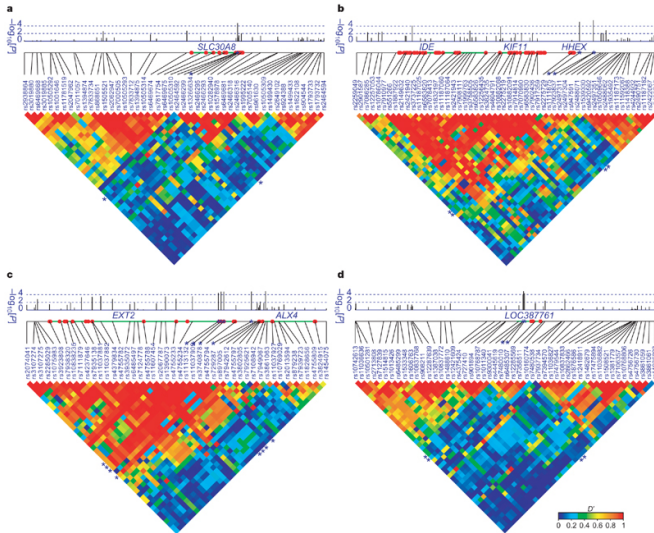
(Sladek *et al.*, Nature, 2007)

Second Phase:

- 57 SNPs identifies
- Typed in 2,617 T2DM cases and 2,894 controls
- Adjustments for potential population stratification, genotyping errors
- Tests for association
- Statistical significance via permutation procedures.

Result: Eight SNPs in five loci increase risk of T2DM
(heterozygous relative risks 1.15-1.65)

T2DM-association study



(Sladek *et al.*, Nature, 2007)

Background and Notation

- Study Design: Cohort or Case/Control, sample size n

$$n = n_0 + n_1.$$

- Binary outcome Y_i for $i = 1, \dots, n$.
- Genomic data $X_{i1}, \dots, X_{iK}$ for $i = 1, \dots, n$.
 - ▶ SNP Haplotype: $2n$ conditionally independent binary vectors

$$X_{ik} \in \{0, 1\}$$

- ▶ SNP Genotype: n conditionally independent vectors,

$$X_{ik} \in \{0, 1, 2\}$$

- Minor Allele Frequency p_k

$$p_k = \Pr[X_{ik} = 1]$$

The Standard Method

We seek the SNP loci that are associated with the condition.

Most commonly

- perform single point analysis
- implement significance tests
- adjust for multiplicity
- report highly ranked loci

Often the results feed forward into a second-phase analysis;

- First phase genome wide in cases and controls, returns candidate loci
- Second phase targeted to candidate regions

Problems with The Standard Method

- single point analysis does not acknowledge the **dependence** between the loci/traits
 - ▶ **structural** dependence due to physical linkage or evolutionary history (i.e. dependence between X_j and X_k)
 - ▶ **functional** dependence due to differential impact on trait (i.e. covariance between (Y, X_j, X_k)).
- adjustment for multiplicity does not acknowledge the dependence between the tests
- highly ranked polymorphisms may not actually be **causal**.

Plus: not clear whether p -values correctly quantify the evidence for association

Problems with The Standard Method

- single point analysis does not acknowledge the **dependence** between the loci/traits
 - ▶ **structural** dependence due to physical linkage or evolutionary history (i.e. dependence between X_j and X_k)
 - ▶ **functional** dependence due to differential impact on trait (i.e. covariance between (Y, X_j, X_k)).
- adjustment for multiplicity does not acknowledge the dependence between the tests
- highly ranked polymorphisms may not actually be **causal**.

Plus: not clear whether p -values correctly quantify the evidence for association

Problems with The Standard Method

- single point analysis does not acknowledge the **dependence** between the loci/traits
 - ▶ **structural** dependence due to physical linkage or evolutionary history (i.e. dependence between X_j and X_k)
 - ▶ **functional** dependence due to differential impact on trait (i.e. covariance between (Y, X_j, X_k)).
- adjustment for multiplicity does not acknowledge the dependence between the tests
- highly ranked polymorphisms may not actually be **causal**.

Plus: not clear whether p -values correctly quantify the evidence for association

Problems with The Standard Method

- single point analysis does not acknowledge the **dependence** between the loci/traits
 - ▶ **structural** dependence due to physical linkage or evolutionary history (i.e. dependence between X_j and X_k)
 - ▶ **functional** dependence due to differential impact on trait (i.e. covariance between (Y, X_j, X_k)).
- adjustment for multiplicity does not acknowledge the dependence between the tests
- highly ranked polymorphisms may not actually be **causal**.

Plus: not clear whether p -values correctly quantify the evidence for association

Problems with The Standard Method

- single point analysis does not acknowledge the **dependence** between the loci/traits
 - ▶ **structural** dependence due to physical linkage or evolutionary history (i.e. dependence between X_j and X_k)
 - ▶ **functional** dependence due to differential impact on trait (i.e. covariance between (Y, X_j, X_k)).
- adjustment for multiplicity does not acknowledge the dependence between the tests
- highly ranked polymorphisms may not actually be **causal**.

Plus: not clear whether p -values correctly quantify the evidence for association

Regression/Association in Observational Studies

These problems are manifestations of the classic problems with regression in observational studies;

effect modification and confounding

e.g. we cannot interpret the “true” effect of predictor X_j in an observational study if

- X_j is correlated with other predictors

$$X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_K$$

and/or

- some of the other predictors are correlated with Y .

Case I: Haplotype Data

Standard Epidemiology Measures:

- Risk Difference

$$\begin{aligned}\delta_k &= \Pr[Y = 1|X_k = 1] - \Pr[Y = 1|X_k = 0] \\ &= E[Y = 1|X_k = 1] - E[Y = 1|X_k = 0] \\ &= \pi_{k11} - \pi_{k01}\end{aligned}$$

- Relative Risk

$$\phi_k = \frac{\pi_{k11}}{\pi_{k01}}$$

- Odds Ratio

$$\psi_k = \frac{\pi_{k11}/(1 - \pi_{k11})}{\pi_{k01}/(1 - \pi_{k01})} = \frac{\pi_{k11}\pi_{k00}}{\pi_{k01}\pi_{k10}}$$

We estimate these from the 2×2 table of SNP/phenotype counts.

Other standard Statistical Measures of Independence in 2×2 tables:

- Pearson Chi-squared

$$\chi^2 = \sum_{j=1}^2 \sum_{k=1}^2 \frac{(n_{jk} - \hat{n}_{jk})^2}{\hat{n}_{jk}}$$

- Likelihood Ratio

Problem: we cannot estimate

$$\delta_k = E[Y = 1|X_k = 1] - E[Y = 1|X_k = 0]$$

by

$$\hat{\delta}_k = \bar{y}_{x_k=1} - \bar{y}_{x_k=0}$$

without bias from the available data if X_k is correlated with any of the other X s.

Simple Example

$K = 3$: 8 possible haplotypes

000	001	010	011	100	101	110	111
p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8

Three scenarios:

- NO DEPENDENCE : $p_8 = 0.125$
- MILD DEPENDENCE : $p_8 = 0.3$
- STRONG DEPENDENCE : $p_8 = 0.5$

equal probabilities elsewhere.

	NO			MILD			STRONG		
	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
X_1	1	0	0	1	$\frac{1}{6}$	$\frac{1}{6}$	1	$\frac{3}{10}$	$\frac{3}{10}$
X_2	0	1	0	$\frac{1}{6}$	1	$\frac{1}{6}$	$\frac{3}{10}$	1	$\frac{3}{10}$
X_3	0	0	1	1	$\frac{1}{6}$	1	$\frac{3}{10}$	$\frac{3}{10}$	1

Simple Example

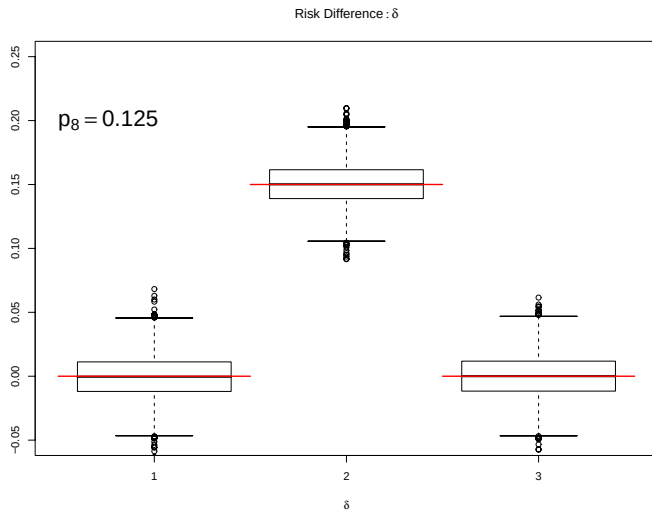
Causal SNP in locus 2:

	X_1	X_2	X_3
$\Pr[Y = 1 X = 0]$	0.10	0.10	0.10
$\Pr[Y = 1 X = 1]$	0.10	0.25	0.10
δ	0.00	0.15	0.00
$\log(\psi)$	0.00	1.10	0.00

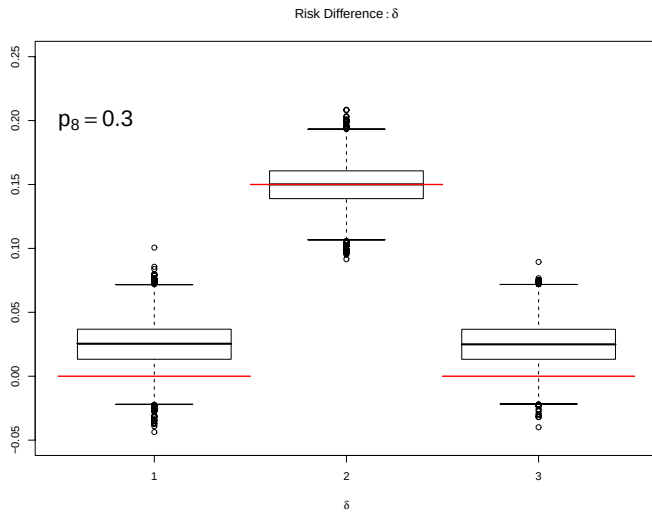
Sample size: $n=2000$ (cohort simulation)

Examine sampling distribution of the estimates; number of replicates is 5000

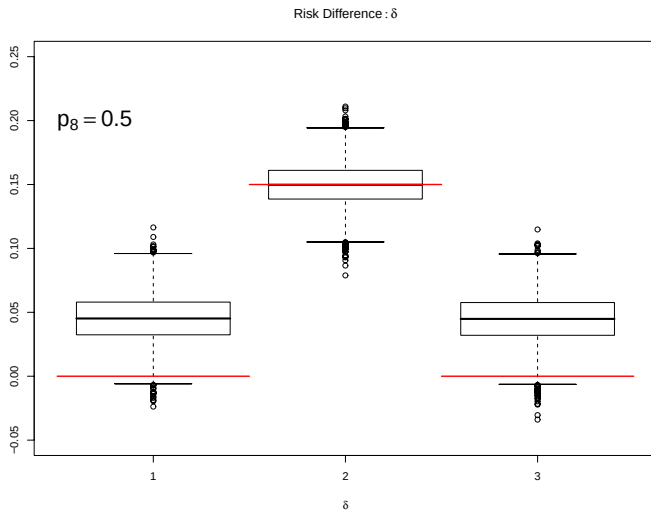
Results: No Dependence



Results: Mild Dependence



Results: Strong Dependence



Realistic Genomic Simulation

$K = 1000$, $n = 2000$.

- dependence structure simulated using latent Gaussian process
- could use population genomics simulation (e.g. CoaSim)

Cohort simulation, single causal SNP at locus $k = 501$:

$$\pi_{k01} = 0.05 \quad \pi_{k11} = 0.20 \quad \implies \quad \psi = 4.75, \log(\psi) = 1.558$$

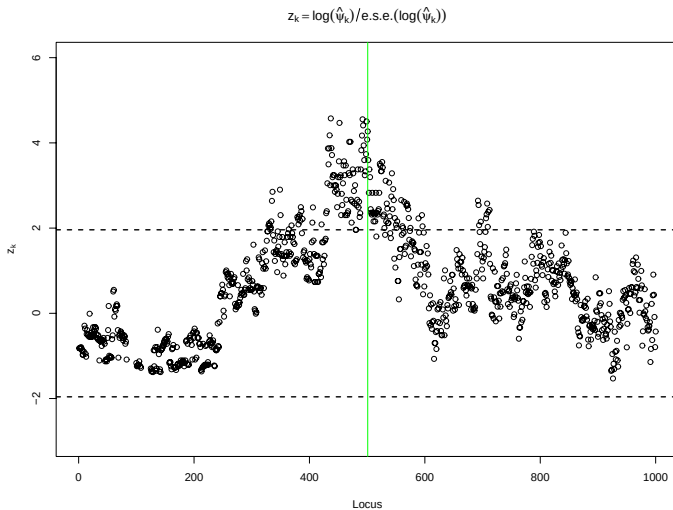
yields 114 cases.

Inspect standardized log-odds ratio statistic:

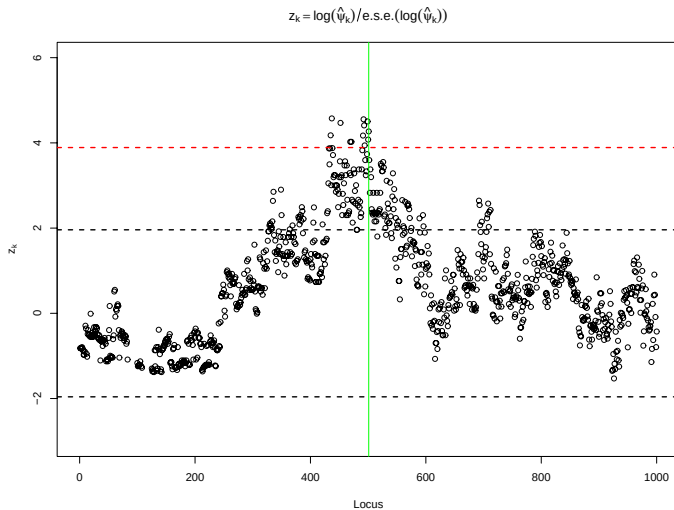
$$Z_k = \frac{\log \left(\frac{n_{k00} n_{k11}}{n_{k01} n_{k10}} \right)}{\sqrt{\frac{1}{n_{k00}} + \frac{1}{n_{k01}} + \frac{1}{n_{k10}} + \frac{1}{n_{k11}}}} \approx N(0, 1)$$

under null of no association.

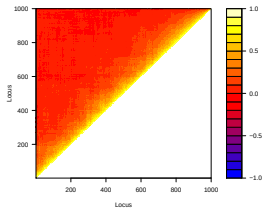
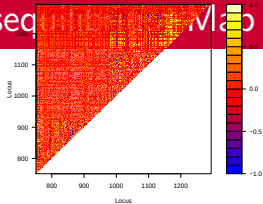
Results: Naive Threshold



Results: Bonferroni Threshold



Linkage Disequilibrium Heat Map (ENCODE YRI): r^2



Correcting for Confounding

Three related approaches:

1. Matching
2. Adjusting
3. Weighting (inverse probability/propensity scores)

These methods are simple, common in epidemiology/biostatistics, but not widely used in statistical genomics.

All (implicitly or explicitly) yield forms of **causal** inference

Matching

Suppose there are S strata of $X_{(k)} = (X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_K)$ within which the expected difference can be estimated in an unbiased fashion: for $s = 1, 2, \dots, S$,

$$\hat{\delta}_k(s) = \bar{y}_{x_k=1}(s) - \bar{y}_{x_k=0}(s)$$

where

$$\bar{y}_{x_k=0}(s) = \frac{1}{n_s} \sum_{i=1}^{n_s} Y_i I\{X_{ik} = 0, X_{i,(k)} \in s\}$$

$$\bar{y}_{x_k=1}(s) = \frac{1}{n_s} \sum_{i=1}^{n_s} Y_i I\{X_{ik} = 1, X_{i,(k)} \in s\}$$

where

$$n_s = |\{i : X_{i,(k)} \in s\}|$$

Then

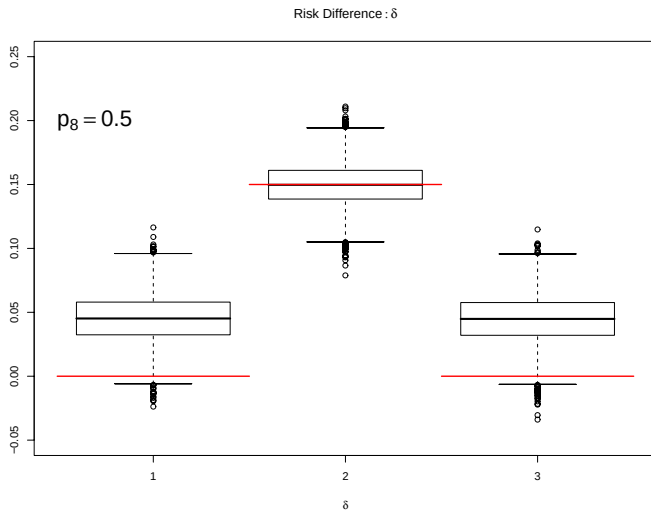
$$\hat{\delta}_k = \sum_{s=1}^S w_s \hat{\delta}_k(s)$$

where $w_s = n_s/n$ is an unbiased estimator of δ_k .

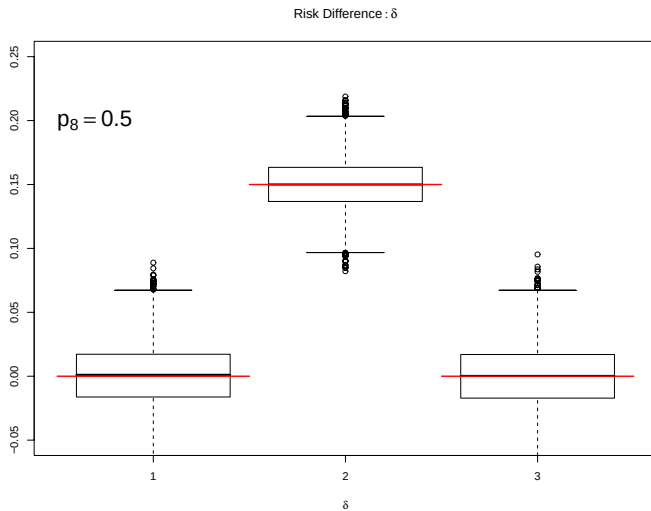
By iterated expectation,

$$\begin{aligned}\delta_k &= E[Y = 1|X_k = 1] - E[Y = 1|X_k = 0] \\ &= \int (E[Y = 1|X_k = 1, X_{(k)}] - E[Y = 1|X_k = 0, X_{(k)}])p(X_{(k)}) dX_{(k)} \\ &= \sum_s (E[Y = 1|X_k = 1, X_{(k)} \in s] - E[Y = 1|X_k = 0, X_{(k)} \in s])p(X_{(k)} \in s)\end{aligned}$$

Results: Strong Dependence



Results: Strong Dependence after Matching



Causal inference

Let $Y_i(d)$ denote the **potential outcome** (change in visual acuity) that would be observed in individual i if dose $D = d$ was received, where $d \in \mathcal{D}$ (the set of possible doses).

The average causal effect of dose is measured via

$$E[Y_i(D = d_0) - Y_i(D = d_1)],$$

where the average is across individuals in the population.

The average causal effect can also be measured at a fixed covariate level $\mathbf{X} = \mathbf{x}$ by

$$E[Y_i(D = d_0) - Y_i(D = d_1) | \mathbf{X}_i = \mathbf{x}].$$

Causal inference

Let $Y_i(d)$ denote the **potential outcome** (change in visual acuity) that would be observed in individual i if dose $D = d$ was received, where $d \in \mathcal{D}$ (the set of possible doses).

The average causal effect of dose is measured via

$$E[Y_i(D = d_0) - Y_i(D = d_1)],$$

where the average is across individuals in the population.

The average causal effect can also be measured at a fixed covariate level $\mathbf{X} = \mathbf{x}$ by

$$E[Y_i(D = d_0) - Y_i(D = d_1) | \mathbf{X}_i = \mathbf{x}].$$

Causal inference

Let $Y_i(d)$ denote the **potential outcome** (change in visual acuity) that would be observed in individual i if dose $D = d$ was received, where $d \in \mathcal{D}$ (the set of possible doses).

The average causal effect of dose is measured via

$$E[Y_i(D = d_0) - Y_i(D = d_1)],$$

where the average is across individuals in the population.

The average causal effect can also be measured at a fixed covariate level $\mathbf{X} = \mathbf{x}$ by

$$E[Y_i(D = d_0) - Y_i(D = d_1) | \mathbf{X}_i = \mathbf{x}].$$

Propensity Score

A method for estimating causal effects in observational studies when treatment is binary is the **propensity score**, $\psi()$

- For covariate(s) X and treatment Z , ψ is defined by

$$\psi_X(z) = p(Z = z|X = x)$$

that is, the conditional pmf for $Z = z$, given that $X = x$.

- $\psi_X(z)$ is a **balancing** score: the distribution of the response is independent of the covariates, given the treatment and the propensity function

$$p(Y|Z = z, \psi_X(z), X) \equiv p(Y|Z = z, \psi_X(z)).$$

Propensity Score

A method for estimating causal effects in observational studies when treatment is binary is the **propensity score**, $\psi()$

- For covariate(s) X and treatment Z , ψ is defined by

$$\psi_X(z) = p(Z = z|X = x)$$

that is, the conditional pmf for $Z = z$, given that $X = x$.

- $\psi_X(z)$ is a **balancing** score: the distribution of the response is independent of the covariates, given the treatment and the propensity function

$$p(Y|Z = z, \psi_X(z), X) \equiv p(Y|Z = z, \psi_X(z)).$$

Linkage Disequilibrium (LD) is the correlation structure under which an individual who carries a particular allele at one site often predictably carries specific alleles at nearby variant sites. Thus, we often come across particular combination of alleles, i.e. haplotypes, across the chromosome [?].

The problem:

Let us denote the SNP at the i th position of the dataset by SNP_i . We test whether SNP_i, SNP_j are independent, i.e.

$$H_{ij}^o : \phi_{ij} = \phi_i \cdot \phi_j, \text{ holds for all } i, j$$

where ϕ_{ij} are the observed probabilities of the inferred haplotypes.

In this case, all the pairwise comparisons are examined by simply referring to the joint distribution of alleles at the two loci of interest, through 2×2 contingency tables. We consider the sample size of each of these contingency tables to be fixed.

		<i>SNP</i> ₂		
		0	1	Total
<i>SNP</i> ₁	0	n_{00}	n_{01}	$n_{0.}$
	1	n_{10}	n_{11}	$n_{1.}$
Total		$n_{.0}$	$n_{.1}$	$n_{..}$

If we can accurately characterize the dependence structure between markers and a trait or a disease in the genome, we are able to trace a small group of markers responsible for the observed variation.

- We aim to identify **haplotype blocks**
 - ▶ regions of high SNP dependence in the observed haplotypes
 - ▶ different measures of dependence
- Perform segmentation of haplotypes
- Search for SNPs associated with the disease accounting for block-like dependence structure.

If we can accurately characterize the dependence structure between markers and a trait or a disease in the genome, we are able to trace a small group of markers responsible for the observed variation.

- We aim to identify **haplotype blocks**
 - ▶ regions of high SNP dependence in the observed haplotypes
 - ▶ different measures of dependence
- Perform segmentation of haplotypes
- Search for SNPs associated with the disease accounting for block-like dependence structure.

If we can accurately characterize the dependence structure between markers and a trait or a disease in the genome, we are able to trace a small group of markers responsible for the observed variation.

- We aim to identify **haplotype blocks**
 - ▶ regions of high SNP dependence in the observed haplotypes
 - ▶ different measures of dependence
- Perform segmentation of haplotypes
- Search for SNPs associated with the disease accounting for block-like dependence structure.

Given the sample size $n_{..}$, the probability of a specific sample haplotype configuration is a **multinomial distribution** given by the formula

$$p(\mathbf{n}) \propto \prod_{i=0}^1 \prod_{j=0}^1 \phi_{ij}^{n_{ij}}$$

where $\mathbf{n} = \{n_{ij}, i, j = 0, 1\}$, ϕ_{ij} denote the observed counts and the probabilities of the inferred haplotypes respectively.

LD reduces when a new mutation occurs and with increasing distance, or equivalently, with high recombination.

HB Hypothesis

Most human recombination occurs in particular places of the genome, causing discontinuous LD patterns and often breaking the DNA data into block structures.

Under an ideal scenario

- Blocks are long DNA regions.
- Blocks are separated by **recombination hotspots**.
- There is no association between blocks.
- All SNPs within a block are highly correlated.
- Few haplotypes account for the majority of variation in each block.

Most human recombination occurs in particular places of the genome, causing discontinuous LD patterns and often breaking the DNA data into block structures.

Under an ideal scenario

- Blocks are long DNA regions.
- Blocks are separated by **recombination hotspots**.
- There is no association between blocks.
- All SNPs within a block are highly correlated.
- Few haplotypes account for the majority of variation in each block.

HB Hypothesis

Most human recombination occurs in particular places of the genome, causing discontinuous LD patterns and often breaking the DNA data into block structures.

Under an ideal scenario

- Blocks are long DNA regions.
- Blocks are separated by **recombination hotspots**.
- There is no association between blocks.
- All SNPs within a block are highly correlated.
- Few haplotypes account for the majority of variation in each block.

HB Hypothesis

Most human recombination occurs in particular places of the genome, causing discontinuous LD patterns and often breaking the DNA data into block structures.

Under an ideal scenario

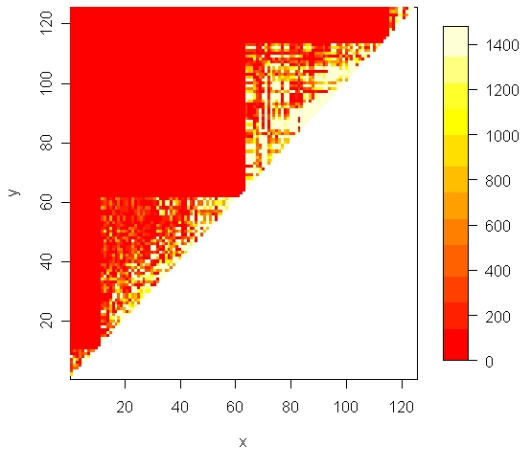
- Blocks are long DNA regions.
- Blocks are separated by **recombination hotspots**.
- There is no association between blocks.
- All SNPs within a block are highly correlated.
- Few haplotypes account for the majority of variation in each block.

Most human recombination occurs in particular places of the genome, causing discontinuous LD patterns and often breaking the DNA data into block structures.

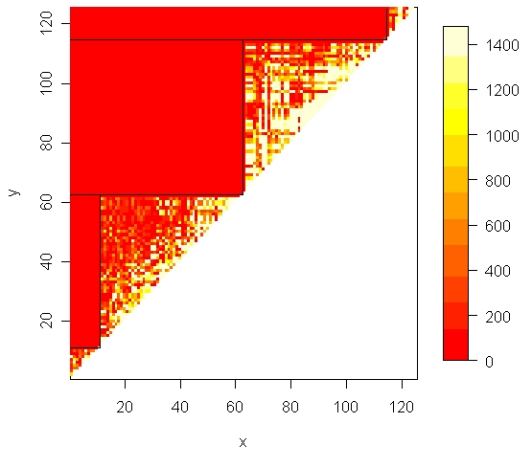
Under an ideal scenario

- Blocks are long DNA regions.
- Blocks are separated by **recombination hotspots**.
- There is no association between blocks.
- All SNPs within a block are highly correlated.
- Few haplotypes account for the majority of variation in each block.

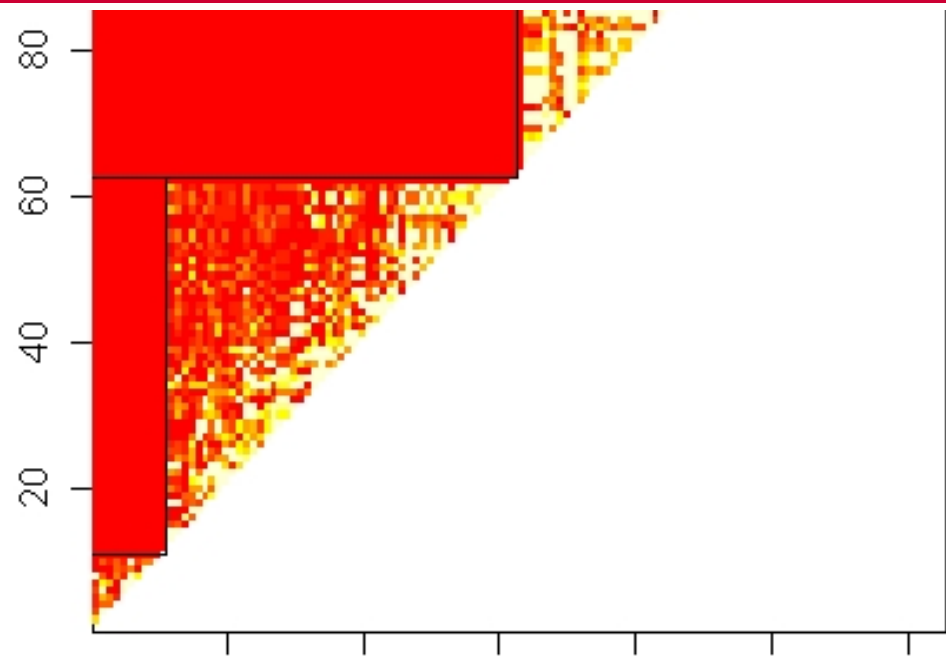
Example: χ^2 statistics



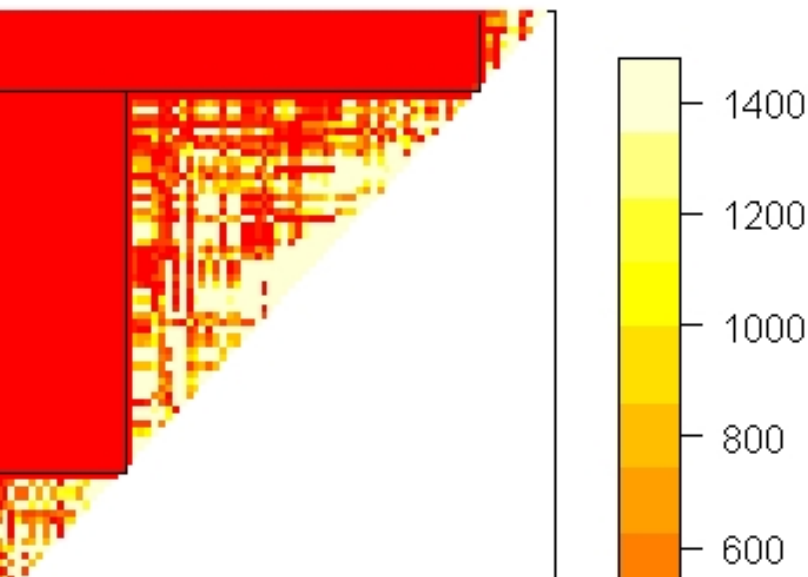
Example: χ^2 statistics



Example: χ^2 statistics



Example: χ^2 statistics



We estimate the extent of block structure in the data by considering the problem of identifying the **number** and the **position** of the distinct blocks in the region under study, as a clustering problem.

In order to test the null hypothesis of LE for all the pairwise comparisons we employ Pearson's χ^2 goodness-of-fit statistic.

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}},$$

where $\hat{m}_{ij}$ is the mle for the expected number m_{ij} of haplotypes. Under H_{ij}^o , the χ^2 statistics follow a χ_1^2 distribution.

Some Assumptions:

- All the x^2 statistics belonging to **independent regions, (IR)** are χ^2_1 asymptotically distributed.
- On the contrary, the x^2 statistics belonging to **blocks** where the alternative hypothesis of LD stands, are dependent and derived from a non-central $\chi^2_{1(\lambda)}$ distribution with different **non-central parameter** λ per block.
- Blocks and IR's are completely independent, which causes the full likelihood of the model to be the product of the likelihoods of all blocks and all IR's.

Some Assumptions:

- All the x^2 statistics belonging to **independent regions, (IR)** are χ^2_1 asymptotically distributed.
- On the contrary, the x^2 statistics belonging to **blocks** where the alternative hypothesis of LD stands, are dependent and derived from a non-central $\chi^2_{1(\lambda)}$ distribution with different **non-central parameter** λ per block.
- Blocks and IR's are completely independent, which causes the full likelihood of the model to be the product of the likelihoods of all blocks and all IR's.

Some Assumptions:

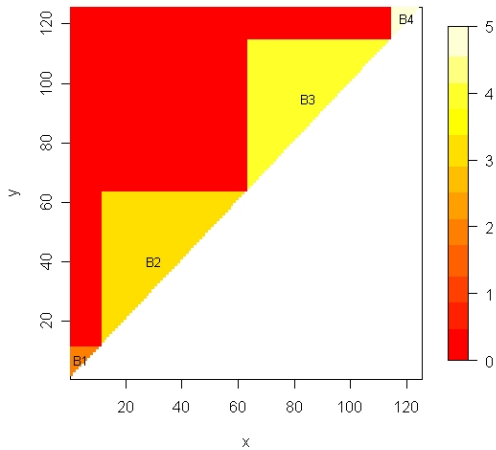
- All the x^2 statistics belonging to **independent regions, (IR)** are χ^2_1 asymptotically distributed.
- On the contrary, the x^2 statistics belonging to **blocks** where the alternative hypothesis of LD stands, are dependent and derived from a non-central $\chi^2_{1(\lambda)}$ distribution with different **non-central parameter** λ per block.
- Blocks and IR's are completely independent, which causes the full likelihood of the model to be the product of the likelihoods of all blocks and all IR's.

- If N is the number of SNPs considered and k is the number of pairwise comparisons between SNPs, the statistics calculated are $\mathbf{y} = (y_1, y_2, \dots, y_k)$.
- The possible **change-points** are $\mathbf{x} = (x_5, \dots, x_{N-5}) \subset \mathbf{y}$, which are the x^2 -statistics that lay along the upper diagonal of the $x^2_{N \times N}$ matrix.
- The change-point vector, $\mathbf{b}$, can be of size $0, \dots, (N - 8)$, after eliminating the first and last 4 SNPs. Blocks are at least 4 SNPs wide, according to HapMap results [?].

- If N is the number of SNPs considered and k is the number of pairwise comparisons between SNPs, the statistics calculated are $\mathbf{y} = (y_1, y_2, \dots, y_k)$.
- The possible **change-points** are $\mathbf{x} = (x_5, \dots, x_{N-5}) \subset \mathbf{y}$, which are the x^2 -statistics that lay along the upper diagonal of the $x^2_{N \times N}$ matrix.
- The change-point vector, $\mathbf{b}$, can be of size $0, \dots, (N - 8)$, after eliminating the first and last 4 SNPs. Blocks are at least 4 SNPs wide, according to HapMap results [?].

- If N is the number of SNPs considered and k is the number of pairwise comparisons between SNPs, the statistics calculated are $\mathbf{y} = (y_1, y_2, \dots, y_k)$.
- The possible **change-points** are $\mathbf{x} = (x_5, \dots, x_{N-5}) \subset \mathbf{y}$, which are the x^2 -statistics that lay along the upper diagonal of the $x^2_{N \times N}$ matrix.
- The change-point vector, $\mathbf{b}$, can be of size $0, \dots, (N - 8)$, after eliminating the first and last 4 SNPs. Blocks are at least 4 SNPs wide, according to HapMap results [?].

the model schematically



The model:

The full likelihood of the model is considered for various cases, namely when length of **b** varies in (0,5).

$$L(\mathbf{y}|\mathbf{b}) = \prod_{l=2}^v L_{\text{BLOCKS}}(\mathbf{y}_l|b_{l-1}, b_l) \times L_{\text{IR}}(\mathbf{y}'_{l-1}|b_{l-1}, b_l)$$

where $L_{\text{BLOCKS}} = \prod_l L_{\text{BLOCK}l}$ and $L_{\text{IR}} = \prod_l L_{\text{IR}l}$.

The likelihood is optimised for various **b** via stochastic search methods, namely **simulated annealing** methods, so that an optimal **b** is chosen for every model considered. Then, we accept the best fit model by means of **AIC or BIC**.

The model is employed to both simulated
www.birc.dk/Software/CoaSim and real data www.hapmap.org.

The model:

The full likelihood of the model is considered for various cases, namely when length of **b** varies in (0,5).

$$L(\mathbf{y}|\mathbf{b}) = \prod_{l=2}^v L_{\text{BLOCKS}}(\mathbf{y}_l|b_{l-1}, b_l) \times L_{\text{IR}}(\mathbf{y}'_{l-1}|b_{l-1}, b_l)$$

where $L_{\text{BLOCKS}} = \prod_l L_{\text{BLOCK}l}$ and $L_{\text{IR}} = \prod_l L_{\text{IR}l}$.

The likelihood is optimised for various **b** via stochastic search methods, namely **simulated annealing** methods, so that an optimal **b** is chosen for every model considered. Then, we accept the best fit model by means of **AIC or BIC**.

The model is employed to both simulated

www.birc.dk/Software/CoaSim and real data www.hapmap.org.

...for simulated datasets

We applied the model to 5 different datasets with well defined blocks generated by CoaSim software. All datasets are similar, namely:

1. 2 change-points, thus the correct model is model 3 (i.e. 3 blocks),
2. change-points are placed in various positions
3. on average 200 SNPs
4. 1,000 individuals
5. span across a 300kb region
6. small recombination rates

...for simulated datasets

We applied the model to 5 different datasets with well defined blocks generated by CoaSim software. All datasets are similar, namely:

1. 2 change-points, thus the correct model is model 3 (i.e. 3 blocks),
2. change-points are placed in various positions
3. on average 200 SNPs
4. 1,000 individuals
5. span across a 300kb region
6. small recombination rates

...for simulated datasets

We applied the model to 5 different datasets with well defined blocks generated by CoaSim software. All datasets are similar, namely:

1. 2 change-points, thus the correct model is model 3 (i.e. 3 blocks),
2. change-points are placed in various positions
3. on average 200 SNPs
4. 1,000 individuals
5. span across a 300kb region
6. small recombination rates

...for simulated datasets

We applied the model to 5 different datasets with well defined blocks generated by CoaSim software. All datasets are similar, namely:

1. 2 change-points, thus the correct model is model 3 (i.e. 3 blocks),
2. change-points are placed in various positions
3. on average 200 SNPs
4. 1,000 individuals
5. span across a 300kb region
6. small recombination rates

...for simulated datasets

We applied the model to 5 different datasets with well defined blocks generated by CoaSim software. All datasets are similar, namely:

1. 2 change-points, thus the correct model is model 3 (i.e. 3 blocks),
2. change-points are placed in various positions
3. on average 200 SNPs
4. 1,000 individuals
5. span across a 300kb region
6. small recombination rates

...for simulated datasets

We applied the model to 5 different datasets with well defined blocks generated by CoaSim software. All datasets are similar, namely:

1. 2 change-points, thus the correct model is model 3 (i.e. 3 blocks),
2. change-points are placed in various positions
3. on average 200 SNPs
4. 1,000 individuals
5. span across a 300kb region
6. small recombination rates

...for simulated datasets

In the following table we can observe the results obtained in 5 different datasets, where in all cases the correct model is the 3rd one. In particular, for each dataset we consider 100 different individual samples, keeping the SNPs unaltered. We report the number of the **correct positives** and **false positives**.

<i>model</i>	0	1	2	3	4	5
<i>dataset1</i>	0	0	0	87%	11%	2%
<i>dataset2</i>	0	0	0	75%	22%	3%
<i>dataset3</i>	0	0	0	98%	2%	0
<i>dataset4</i>	0	0	0	97%	3%	0
<i>dataset5</i>	0	0	0	76%	21%	3%

...for simulated datasets

However we are interested not only in detecting the correct model, but also in how close are the change-points to the estimated ones. The following table shows the distance in *bp* between the detected change-points and the estimated ones, for each dataset.

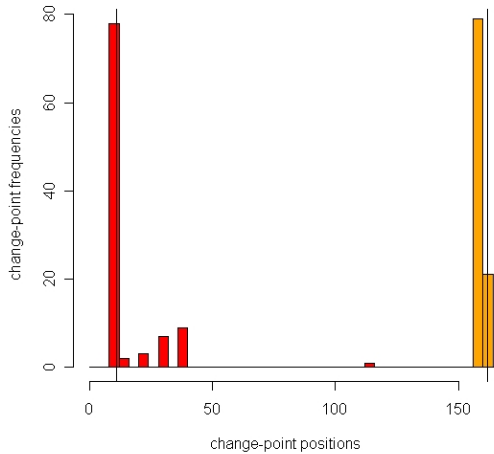
	1st change-point					2nd change-point						
	-	-20	-10	10	20	+	-	-20	-10	10	20	+
<i>dataset1</i>	0	0	99%	1%	0		11%	10%	78%	1%	0	
<i>dataset2</i>	0	0	79%	4%	17%		4%	13%	83%	0	0	
<i>dataset3</i>	0	0	100%	0	0		0	0	100%	0	0	
<i>dataset4</i>	0	0	92%	6%	2%		6%	2%	92%	0	0	
<i>dataset5</i>	0	0	97%	0	3%		2%	8%	90%	0	0	

...for simulated datasets

However we are interested not only in detecting the correct model, but also in how close are the change-points to the estimated ones. The following table shows the distance in *bp* between the detected change-points and the estimated ones, for each dataset.

	1st change-point					2nd change-point						
	-	-20	-10	10	20	+	-	-20	-10	10	20	+
<i>dataset1</i>	0	0	99%	1%	0		11%	10%	78%	1%	0	
<i>dataset2</i>	0	0	79%	4%	17%		4%	13%	83%	0	0	
<i>dataset3</i>	0	0	100%	0	0		0	0	100%	0	0	
<i>dataset4</i>	0	0	92%	6%	2%		6%	2%	92%	0	0	
<i>dataset5</i>	0	0	97%	0	3%		2%	8%	90%	0	0	

Histogram for Dataset2



...for real datasets

We also applied the model to regions of HapMap ENCODE datasets of 4 different populations (Yoruba in Ibadan-Nigeria (YRI), Utah-USA (CEU), Han Chinese in Beijing-China (HCB), Japanese in Tokyo-Japan (JPT)).

1. we consider chunks of the datasets with less than 5 change-points
2. on average 300 SNPs
3. span approximately 300kb (the data supplied from HapMap is 500 kb)
4. estimated recombination rates vary in $[0.19\text{cM}, 1.25\text{cM}]$

...for real datasets

We also applied the model to regions of HapMap ENCODE datasets of 4 different populations (Yoruba in Ibadan-Nigeria (YRI), Utah-USA (CEU), Han Chinese in Beijing-China (HCB), Japanese in Tokyo-Japan (JPT)).

1. we consider chunks of the datasets with less than 5 change-points
2. on average 300 SNPs
3. span approximately 300kb (the data supplied from HapMap is 500 kb)
4. estimated recombination rates vary in $[0.19\text{cM}, 1.25\text{cM}]$

...for real datasets

We also applied the model to regions of HapMap ENCODE datasets of 4 different populations (Yoruba in Ibadan-Nigeria (YRI), Utah-USA (CEU), Han Chinese in Beijing-China (HCB), Japanese in Tokyo-Japan (JPT)).

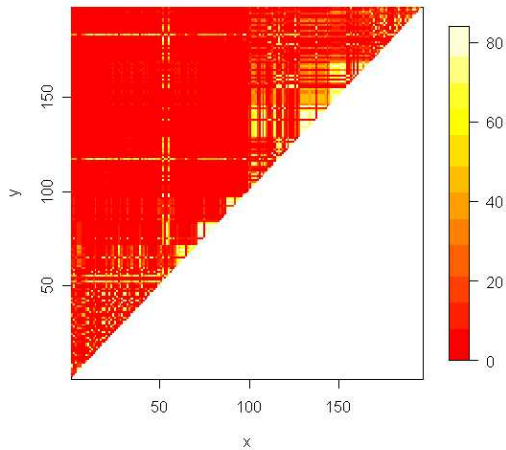
1. we consider chunks of the datasets with less than 5 change-points
2. on average 300 SNPs
3. span approximately 300kb (the data supplied from HapMap is 500 kb)
4. estimated recombination rates vary in $[0.19\text{cM}, 1.25\text{cM}]$

...for real datasets

We also applied the model to regions of HapMap ENCODE datasets of 4 different populations (Yoruba in Ibadan-Nigeria (YRI), Utah-USA (CEU), Han Chinese in Beijing-China (HCB), Japanese in Tokyo-Japan (JPT)).

1. we consider chunks of the datasets with less than 5 change-points
2. on average 300 SNPs
3. span approximately 300kb (the data supplied from HapMap is 500 kb)
4. estimated recombination rates vary in $[0.19\text{cM}, 1.25\text{cM}]$

χ^2 for Japanese population- chromosome band 7q31



...for real datasets

In the following table we can observe the results obtained from 4 different datasets. We employ the use of bootstrapping techniques in order to obtain some reasonable measure of uncertainty due to relatively small number of individuals.

<i>model</i>	0	1	2	3	4	5
<i>YRI/ENr131.2q37.1</i>	0	2%	13%	72%	13%	0
<i>CEU/ENm014.7q31.33</i>	0	2%	83%	16%	0	0
<i>JPT/ENm014.7q31.33</i>	0	0	0	9%	11%	80%
<i>HCb/ENr113.4q26</i>	0	0	0	24%	71%	5%

Some Conclusions:

- We are able to detect the correct model and the estimated change-points with high accuracy.
- Have similar results when **Hypergeometric distribution** is used to construct the likelihood.
- Comparison with other software (Haploview, Hotspotter).
- Extend the model to a Bayesian framework.

Some Conclusions:

- We are able to detect the correct model and the estimated change-points with high accuracy.
- Have similar results when **Hypergeometric distribution** is used to construct the likelihood.
- Comparison with other software (Haploview, Hotspotter).
- Extend the model to a Bayesian framework.

Some Conclusions:

- We are able to detect the correct model and the estimated change-points with high accuracy.
- Have similar results when **Hypergeometric distribution** is used to construct the likelihood.
- Comparison with other software (Haploview, Hotspotter).
- Extend the model to a Bayesian framework.

Some Conclusions:

- We are able to detect the correct model and the estimated change-points with high accuracy.
- Have similar results when **Hypergeometric distribution** is used to construct the likelihood.
- Comparison with other software (Haploview, Hotspotter).
- Extend the model to a Bayesian framework.

Acknowledgements:

- Georgia Tsiliki and Maria De Iorio (Imperial College London)
- Rob Sladek, Ghislain Rocheleau (Genome Innovation Centre, McGill)
- Erica Moodie (McGill) and Jon Wakefield (UW)
- Genome Quebec/Genome Canada
- NSERC

References

-  Agresti, A., *Categorical Data Analysis*, John Wiley & Sons, 1990.
-  Weir, B.S., *Genetic Data Analysis II*, Sinauer Associates, Inc. Publishers, 1996.
-  Daly et al., *High-resolution haplotype structure in the human genome*, Nature Genetics, vol 29, 2001.
-  Gabriel et al., *The Structure of Haplotype Blocks in the Human Genome*, Science Magazine, vol 296, 2002.
-  HapMap Consortium, *A haplotype map of the human genome*, Nature, vol 437, 2005.