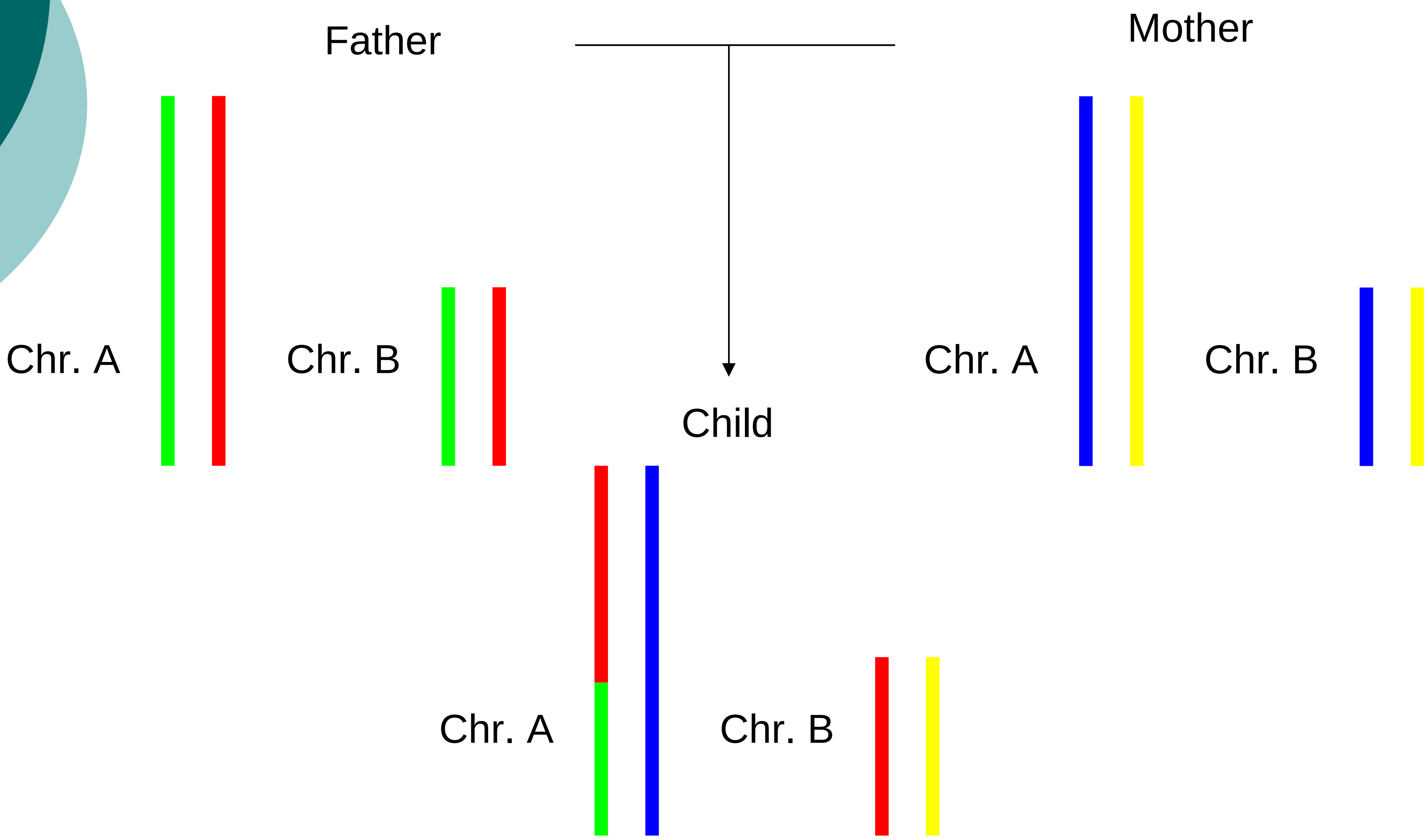
Bin analysis of genome-wide association study

N. Omont, K. Forner, M. Lamarine, G. Martin, F. Képès, J. Wojcik

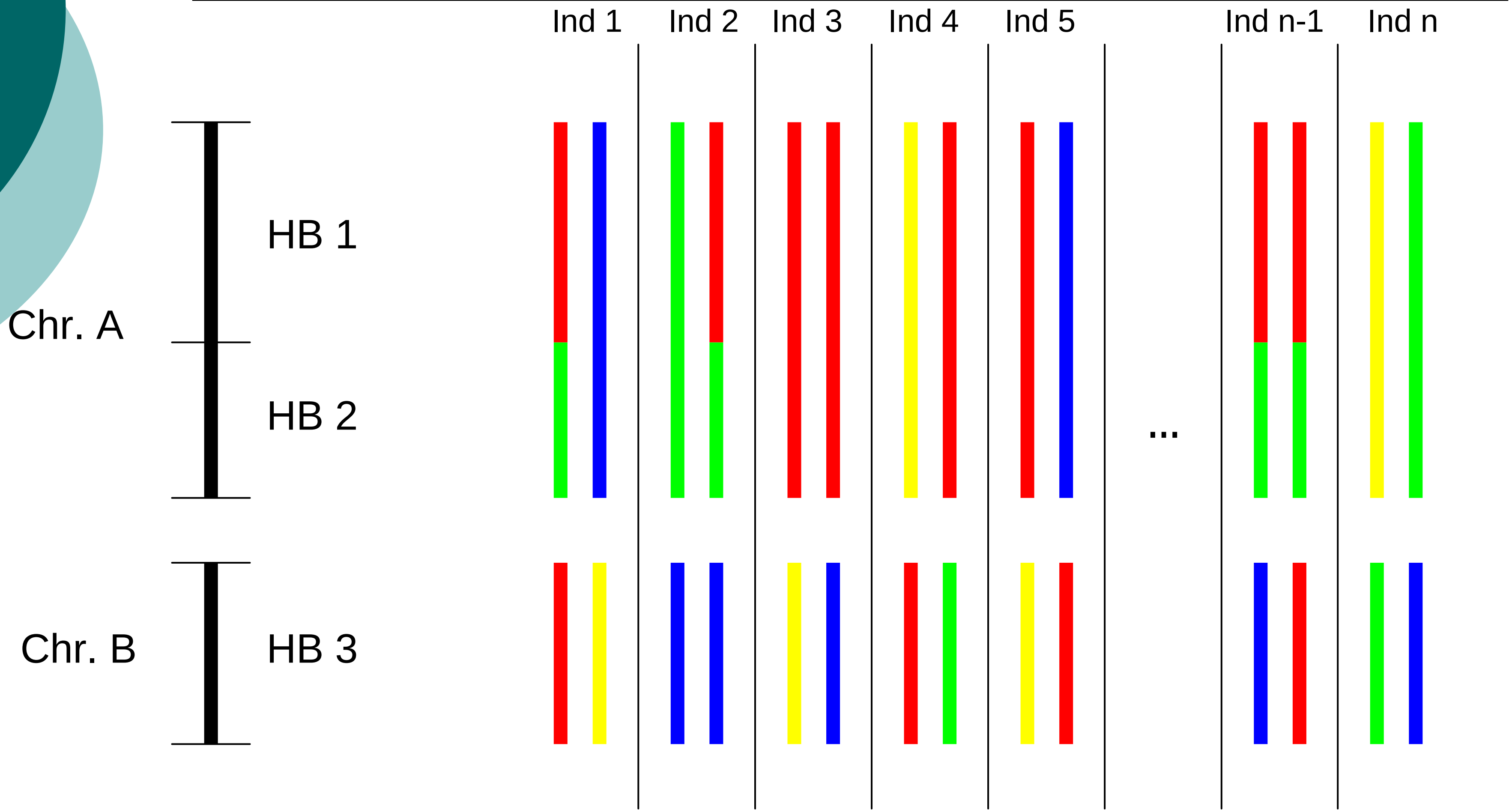
Bin analysis of genome-wide study

- Data
  - What is a Genome-wide association study
- Analysis
  - Multiple testing problem
  - Method
- Results

Transmission and recombination



Haplotype blocks (HB)



Ind 1 Ind 2 Ind 3 Ind 4 Ind 5 Ind n-1 Ind n

Chr. A

HB 1

HB 2

...

Chr. B

HB 3

Data – association study

Genetic disease

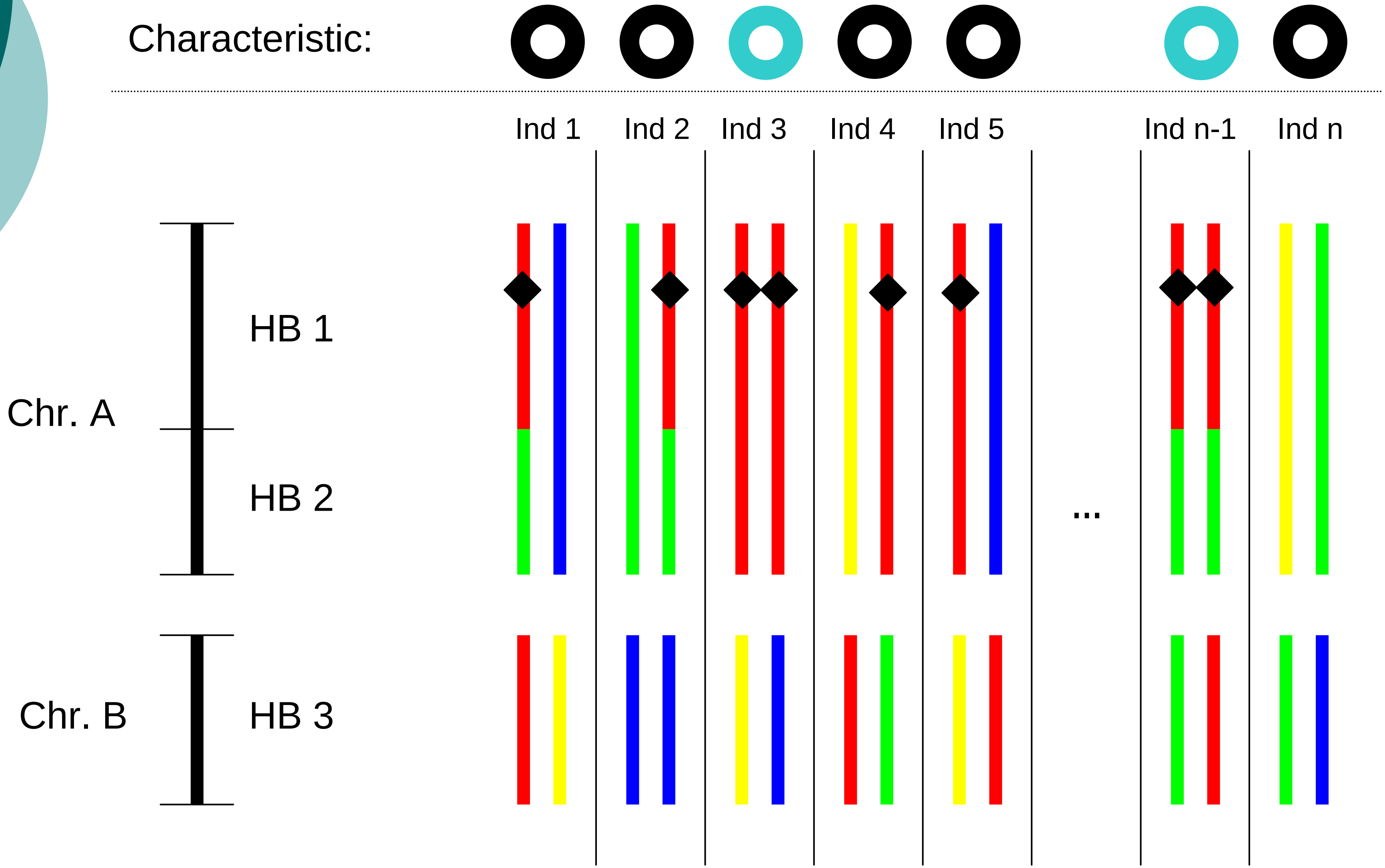
Variants of DNA causes disease:

- Simple case (« mendelian »):
  - One change in DNA
    - Simplest case: One letter change in DNA
- Complex case:
  - Many changes
  - Interaction of changes
  - Interaction with environment

Genetic disease

- How to find the variant(s) causing the disease? By looking for a correlation of a portion of DNA with a disease:
  - Linkage studies: whole families.
  - Association studies: independent individuals from the same population.

Association study: example



Association Study : cost problem

- Reading (sequencing) entirely the 2 DNA words of an individual is too expensive.

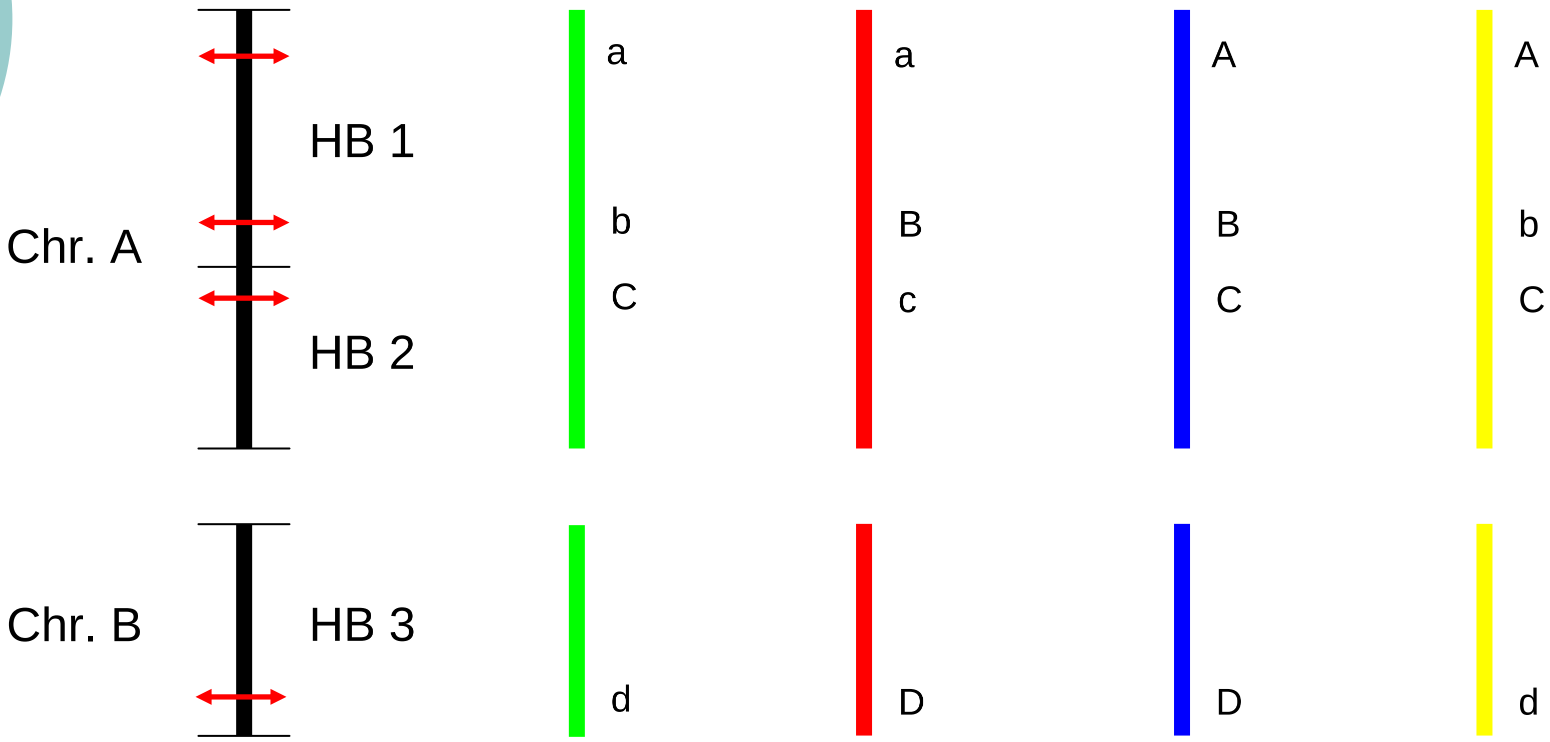
Single Nucleotide Polymorphism

- Predefined positions on DNA where different letters are found in a population.
  - For SNPs used, 2 letters among the 4 possible are found.
  - Letters are arbitrarily noted ‘a’ and ‘A’.

⇒ An individual holds either:

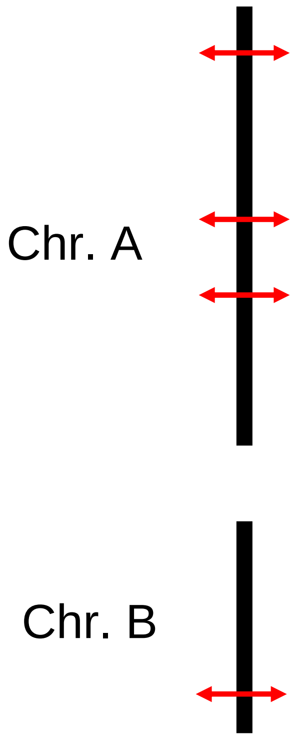
  - ‘aa’
  - ‘aA’ or ‘Aa’, but distinction is impossible
  - ‘AA’.

Association study: example



Association study: example

Characteristic:	●	●	○	●	●		○	●
	Ind 1	Ind 2	Ind 3	Ind 4	Ind 5		Ind n-1	Ind n
	aA	aa	aa	Aa	Aa		aa	Aa
	BB	Bb	BB	bB	BB		BB	bb
	cC	cc	cc	cC	cC	…	cc	Cc
	Dd	DD	dD	dD	dD		dD	dD



The Serono association study

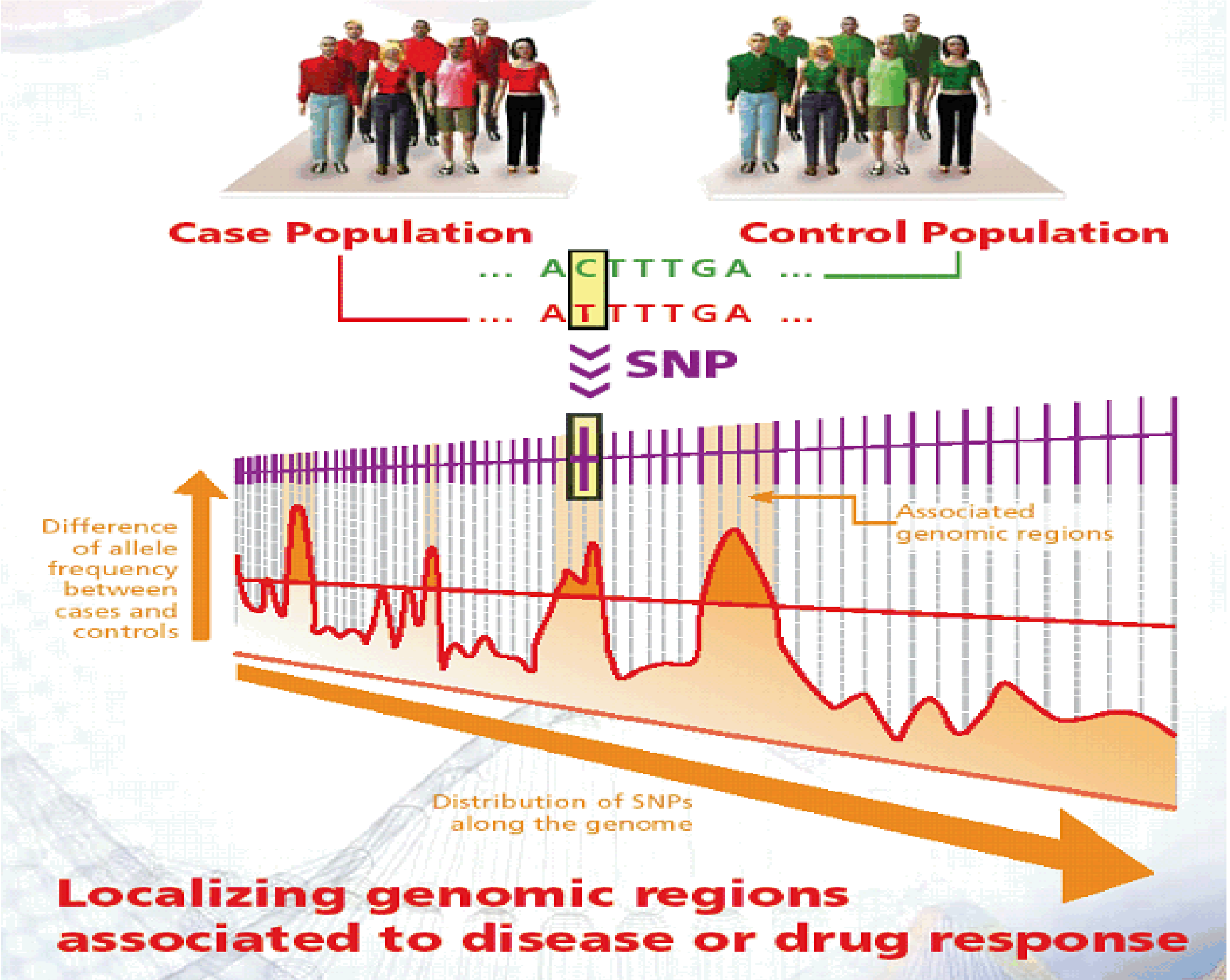
- Multiple Sclerosis: Complex disease
  - Concordance rate between twins: 15-20 %
- 3 collections of 300 cases/300 control
- 100,000 SNPs
- Cost: > 1,000 € per individual

Analysis

- Is there an association with the disease?
- If yes, where?

Method

The ideal vision


Case Population
Control Population
... ACTTTGA ...
... ATTTTGA ...
SNP
Difference of allele frequency between cases and controls
Associated genomic regions
Distribution of SNPs along the genome
Localizing genomic regions associated to disease or drug response

FDR estimation (no control)

- $\widehat{\pi}_0$: Proportion of bins under the null hypothesis assumed to be 1.0.
- B : Number of bins
- θ : Level at which FDR is computed
- π_b : P-value of bin b

$$\mathrm{FDR}(\theta) = \frac{\widehat{\pi}_0 \theta B}{\mathrm{card}(\{b|\pi_b < \theta\})}$$

Multiple testing problem

Assuming 1 association with p-value=1E-5

- Tested with 1,000 SNP under null hypothesis:

  FDR = 1 % *[= 1E-5 *1E3 / (1 + 1E-5*1E3)]*

  ⇒ **OK**

- Tested with 1,000,000 SNP under null hypothesis:

  FDR = 91 % *[= 1E-5 * 1E6 / (1 + 1E-5*1E6)]*

  ⇒ No association detected

Multiple testing problem

Linkage disequilibrium ⇒ 2 neighbour SNP truly associated: p-value=1E-5

- Independent testing:

  FDR = 83 % _[= 1E-5 * 1E6 / (2+1E-5*1E6)]_

  ⇒ No association detected

- Simultaneous testing:

  new p-value = c^2(2*invc^2(1E-5,1),2) = 3,4E-9

  FDR = 0,3% _[= 3,4E-9 * 1E6 / (1+3,4E-9 *1E6)]_

  ⇒ OK

Bin definition

- Haplotype blocks:
  - Unknown
  - Population dependent
  - Not adapted to functional analysis

⇒ Practically infeasible

Bin definition

- Gene:
  - (Relatively) well defined
  - Population independent
  - Adapted to functional analysis.

  But:
  - Generally larger than haplotype blocks
    - Loss of power
  - Boundary accross haplotype blocks
    - Not independent.

Bin definition : Loss of power example

- Too large bin definition: Assuming bin with 9 SNP:
  - 2 associated SNP: p-value=1E-5
  - 7 unassociated SNP: p-value=1
- Results:

⇒ *New p-value* = $\chi^2(2*\mathrm{inv}\chi^2(1E\text{-}5,1),9)$
= *1.1 E-5*

⇒ FDR = 92 %

⇒ No association detected

Bin definition : Loss of power example

- If all SNPs are tested by 9:
  - Only 1,000,000/9 = 111,111 tests
  - ⇒ FDR = 56 %

⇒ FDR reduced of 1/3.

⇒ Significant difference before starting costly experiments

Statistical test:

- Likelihood ratio test
  - Naive: SNPs are independent
  - Two-SNP: each SNP is dependent on the 2 SNPs directly on its sides.
- Collection design:
  - Each collection independently
  - Independence of each population

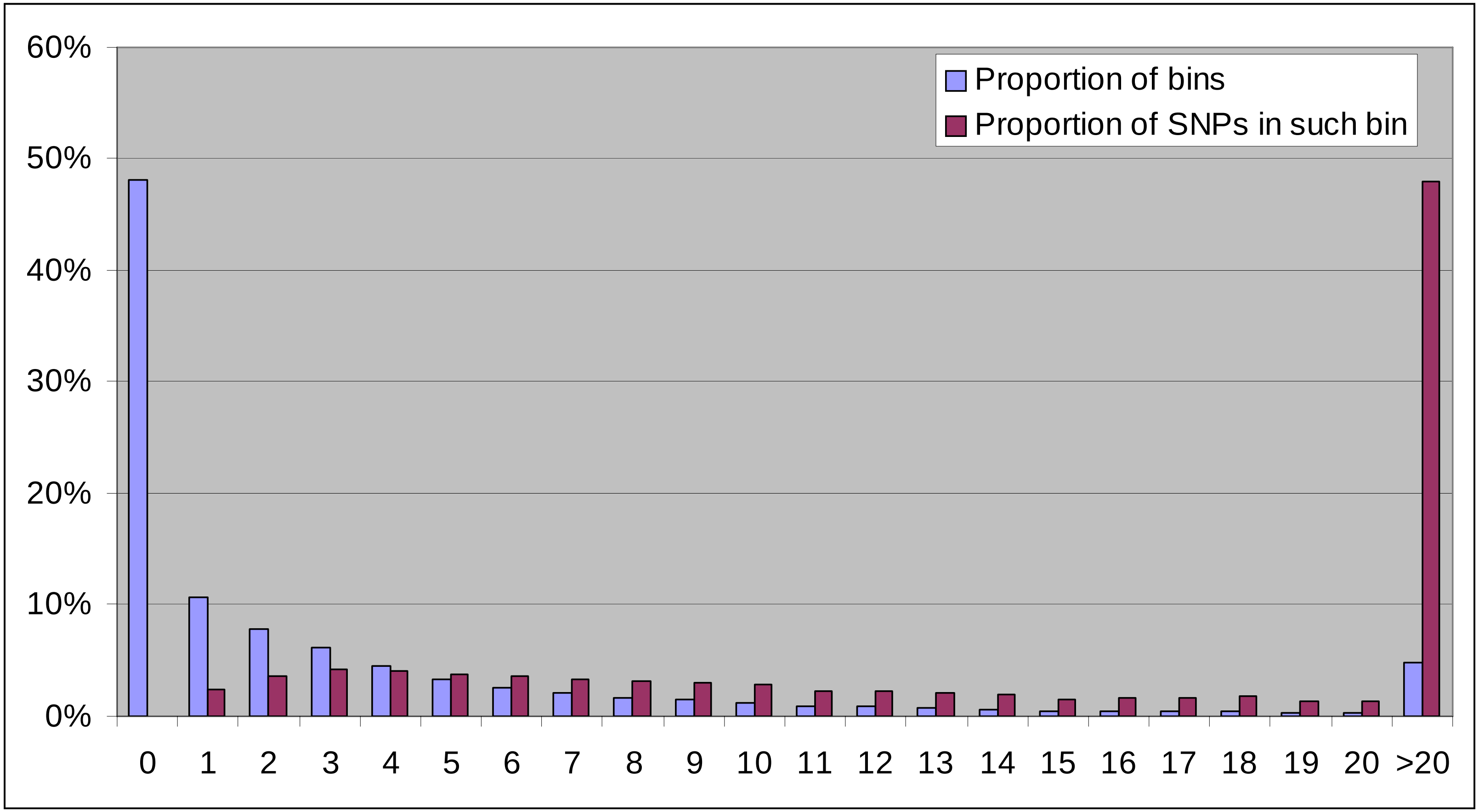
Estimation

- Asymptotic p-values:
  - Badly fit tables
  - Missing value and error model
- Exact p-values:
  - Not tractable given the model
- Empirical p-values:
  - Accurate control of error

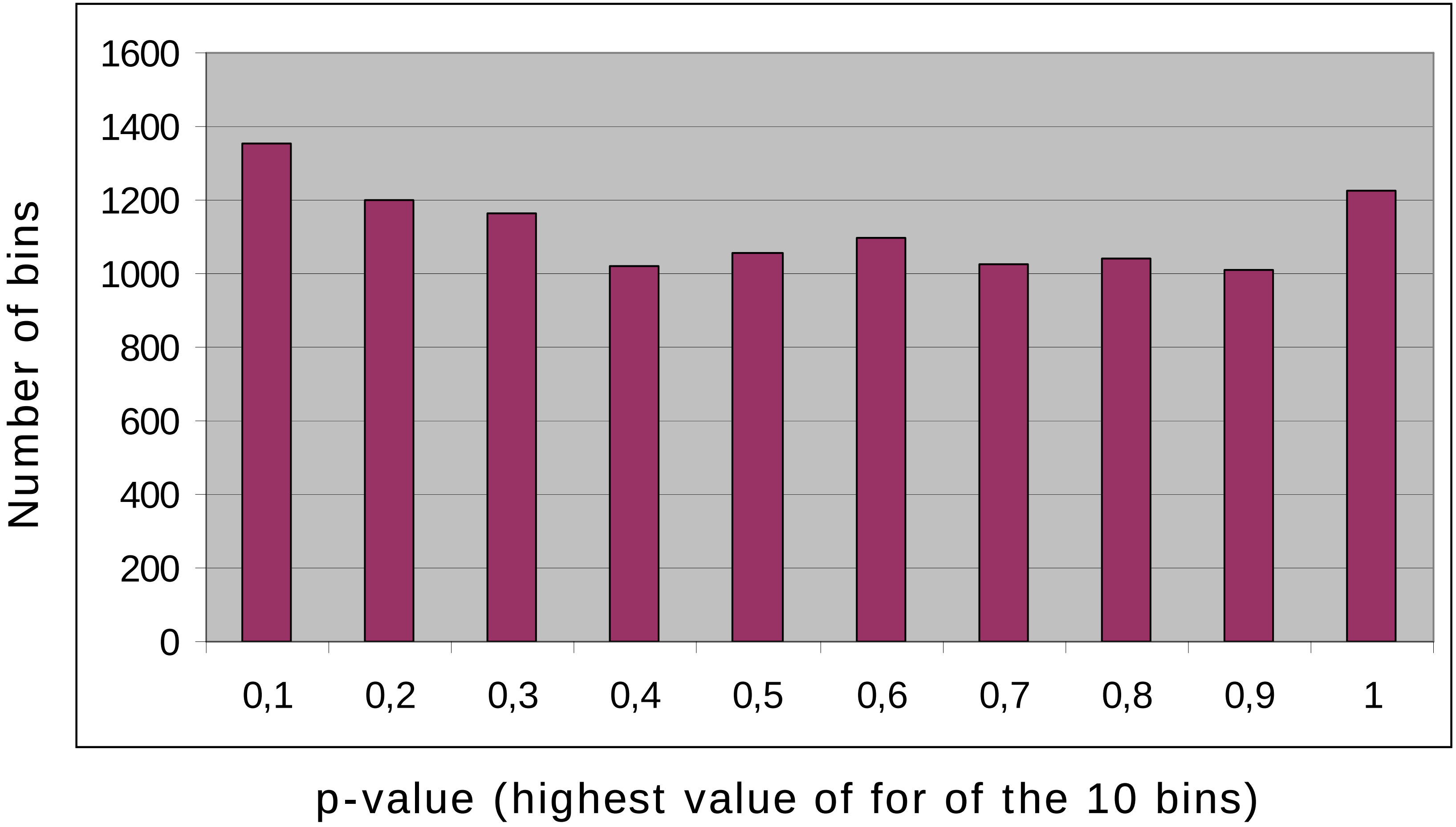
Results

Results: bins

Distribution of the number of SNP per bin:

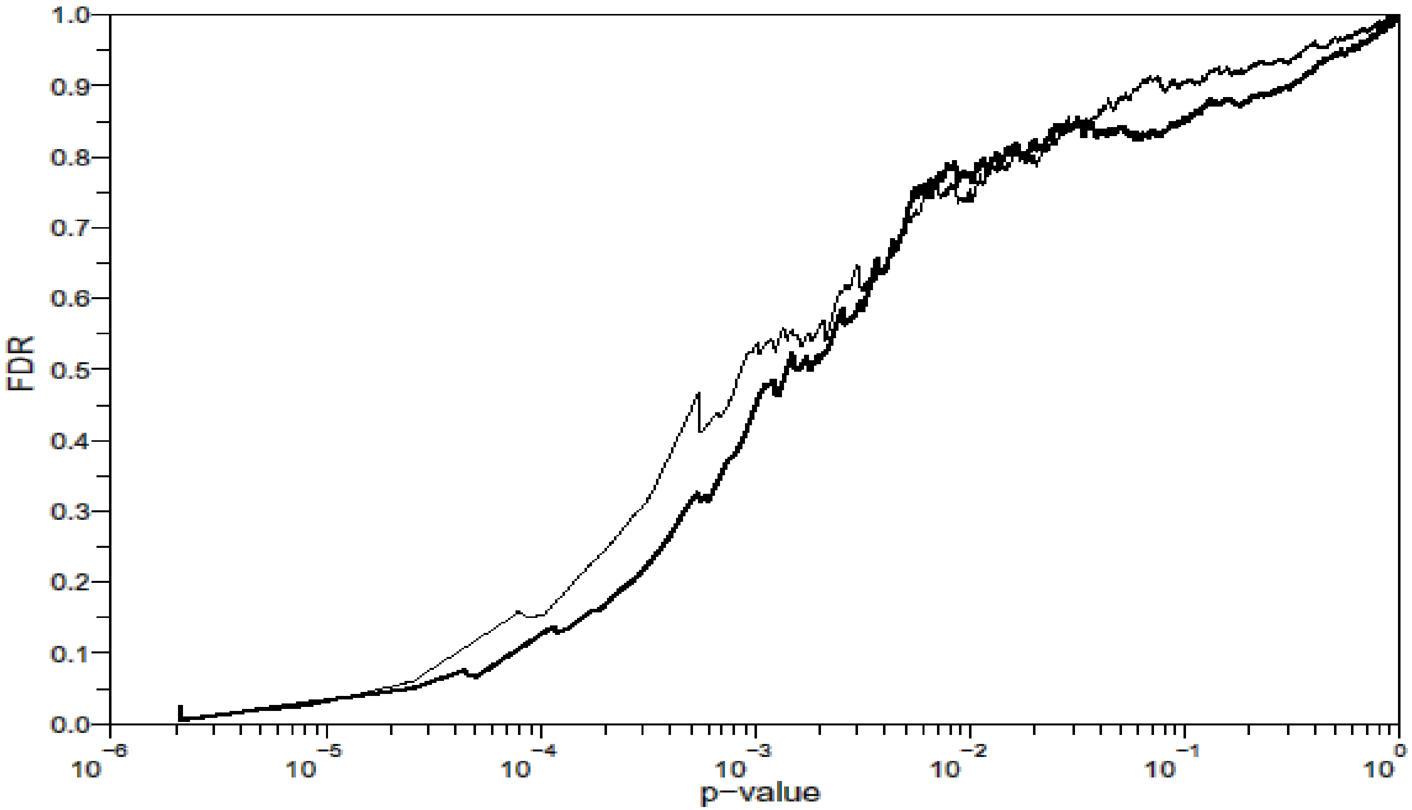


P-value distribution



3 collection design, two-marker

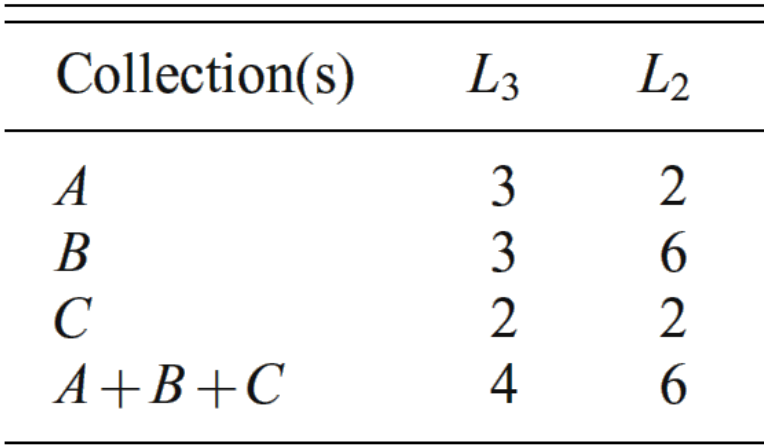
FDR: FDR vs p-value



(3 collection design, thick: naive, thin: two-SNP)

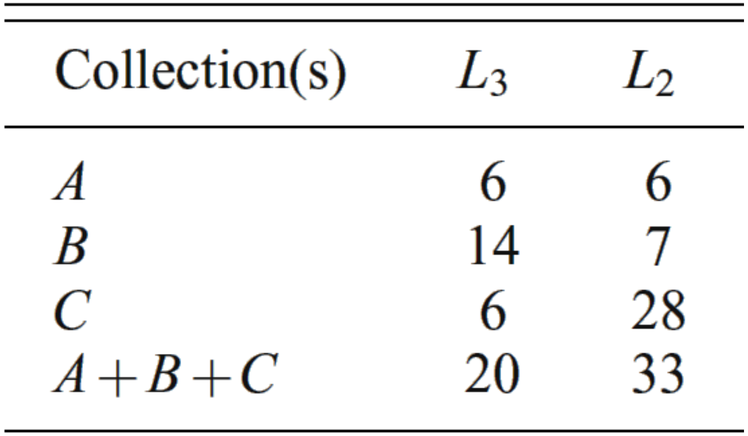
Number of bins selected

- FDR threshold 5%:



Collection(s)	L_3	L_2
A	3	2
B	3	6
C	2	2
$A+B+C$	4	6

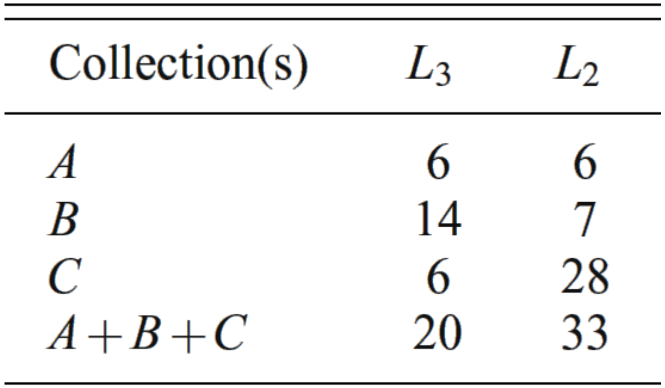
- FDR thres. 50%:



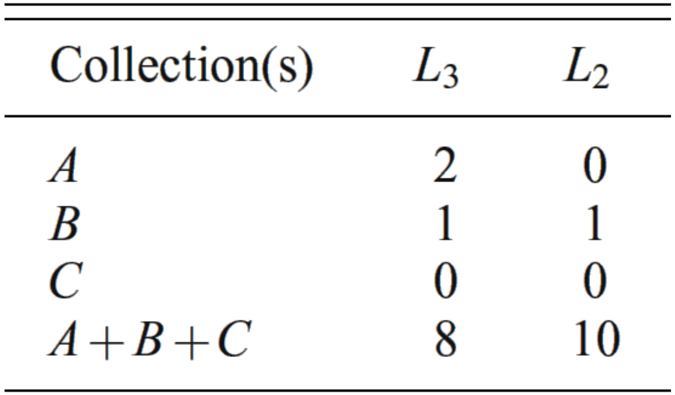
Collection(s)	L_3	L_2
A	6	6
B	14	7
C	6	28
$A+B+C$	20	33

FDR overestimation

- Known true positives
  - ⇒ FDR of subset of bins excluding the known true-positives is overestimated
  - ⇒ New estimation of FDR:



Collection(s)	L_3	L_2
A	6	6
B	14	7
C	6	28
$A+B+C$	20	33



Collection(s)	L_3	L_2
A	2	0
B	1	1
C	0	0
$A+B+C$	8	10

Conclusion

- Biological results:
    - Meaningful but insufficient compared to the investment
    - Complex diseases remain complex
        - Gene-gene interaction intractable
        - Heterogeneity of cases
        - Sample size problem

Conclusion

- A new method:
  - Computationally tractable
  - Rigorously estimating the FDR
  - Adapted to functional analysis
  - Taking advantage of the structure of the data

Bin analysis of genome-wide association study

N. Omont, K. Forner, M. Lamarine, G. Martin, F. Képès, J. Wojcik

Nicolas Omont
Decision Mathematics Consultant
nicolas.omont@artelys.com