Bin analysis of genome-wide association study

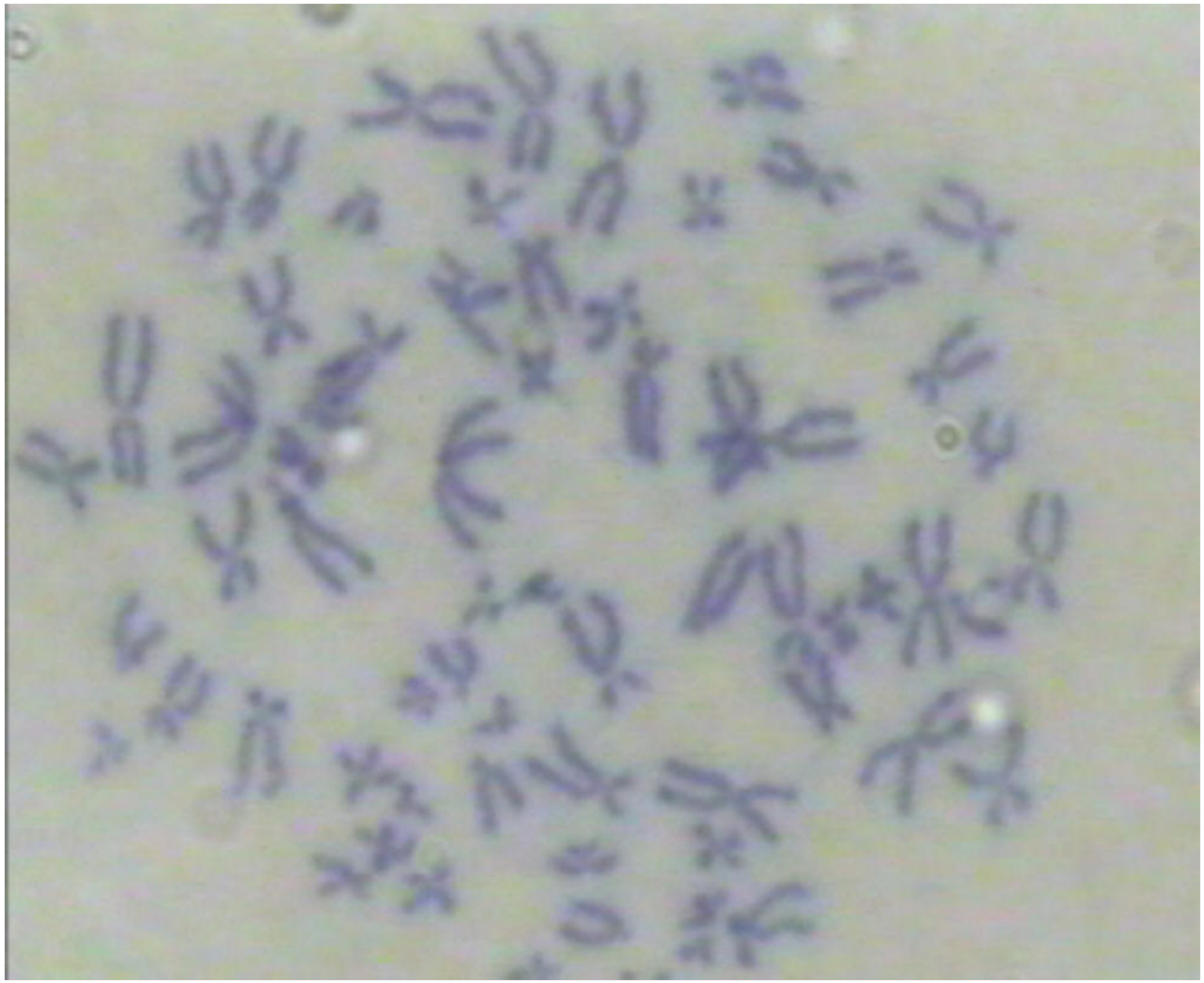
N. Omont, K. Forner, M. Lamarine, G. Martin, F. Képès, J. Wojcik

Bin analysis of genome-wide study

- Data
  - Biological primer
  - Genome-wide association study
- Analysis
  - Multiple testing problem
  - Method
- Results

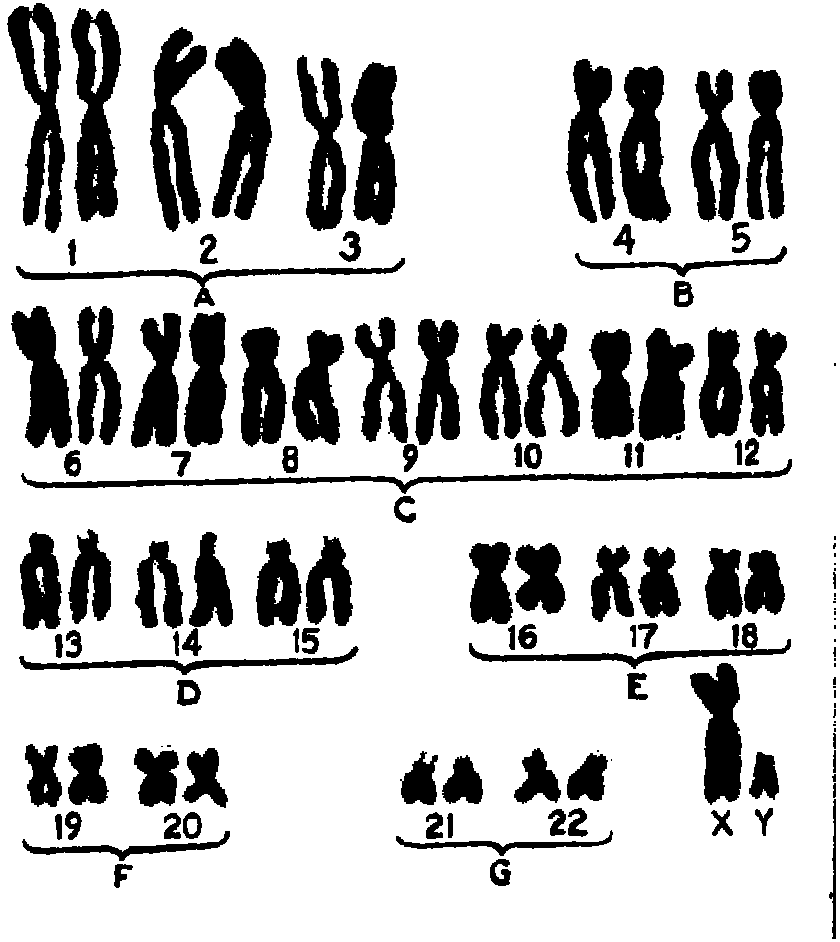
Data – Biological Primer

Nucleus of a dividing cell (x1000)



http://www.unm.edu/~vscience/microscopy.htm

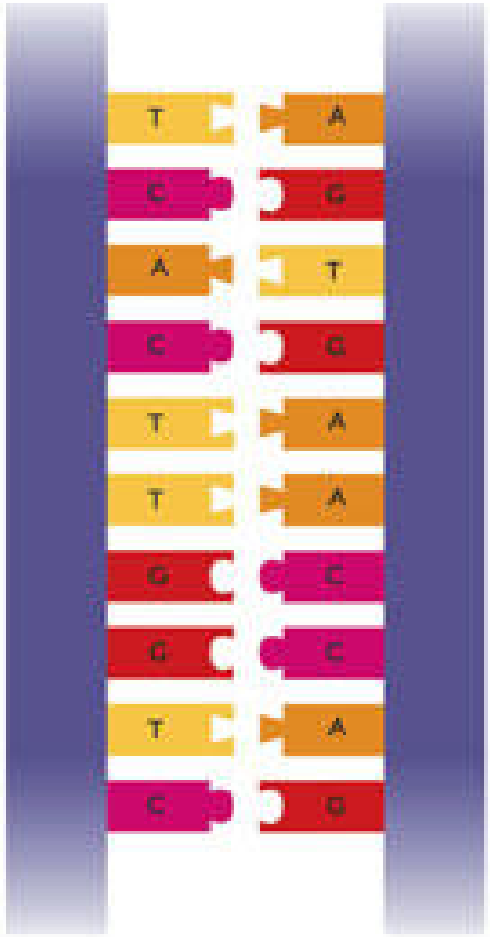
Human male karyotype



http://www.contexo.info/DNA_Basics/chromosomes.htm

DNA stores information

http://info.cancerresearchuk.org/youthandschools/latestfromthelab/howcellswork/cellsanddna/

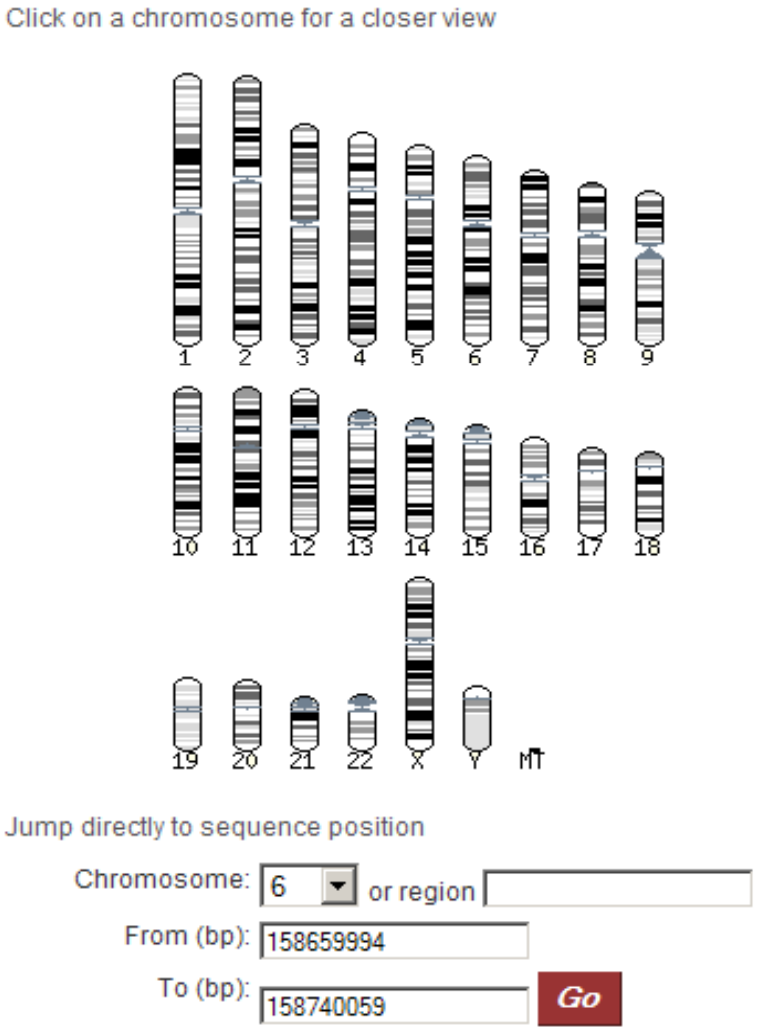


DNA stores information

- Human DNA molecule is a word:
  - Composed with 4 letters:
    - base pairs ATCG
  - Split in 22 chromosoms (+ sexual chromosoms)
  - 3,253,037,807 letter long

Browse the genome!

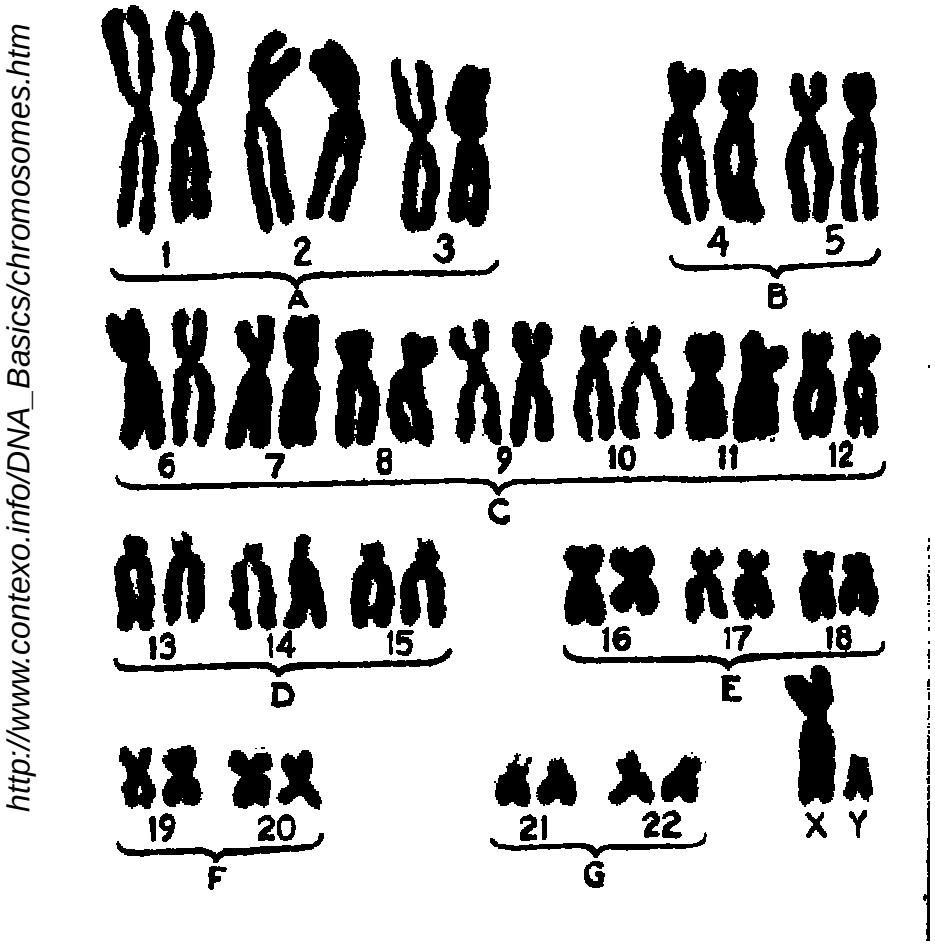
http://www.ensembl.org/Homo_sapiens/index.html



Browse the genome!

```
>6 dna:chromosome chromosome:NCBI36:6:158659994:158740059:1
AAACTTAAGTGATCCGCCCACCTCAGCCTCCCAAAGTGCTGGGATTATAGGTGAGAGCCA
CTGTGCCCAGCTCACGATTGTTAAATCTAAGGGCTCTGAAAATAGCAAGTTTTTGTATAA
TTGTTTCAATGGCAAGACCTGGCCTGGACTGATGTGAAGCCGCTGGTTGTTGTTATTCCA
TCACTTCAGCTGCAGAAATACTGTGTTTTATTATGGGGGCTGCCCAGACCTGTGGGTGGC
ACTAGATAATTGACAGGAGCTCCCCTTGTAGGACTTTGCTAAGATTTAAAAAATTCTGAA
TTCAAAATACTTGTCTCTAAGGATTCTCAATCAGGGGTTGTGAACTTGTGTCTTCATTTA
AGGGAATCATCAAAAGAAACCTTGGGTTTTATTTAATTTGGTTTTTCATTTCCGGGAAGG
CTGACATTTAACTCATCTCTGCCCTTACCTTATCTTCACTCCCTTCTACCACAAGAAGCA
GAAAAACCCTGTGTCCCCACCGGCCATCCCTTAAGAACACTACTGAAAGAACCATTGCAA
GATTTTATTTCTGGCCACGGAACTAACTAAATTGAAGGGGTCAGATCTAGTCGTCTGCTA
ATTTCACAACTGAATTAAAAAGGAAAAAAATCTGAATGAAAGATAATTTATTTCACCCTC
AGCTAGTATGTAAAATCTGTTTTATATTATGTAAGTATTTAACAAATCTTAACAGTTTTG
GAATTATAAATGTATTCAGTTAGATAAATTGAATCTGAAATTAAATACTGTGATAAATGT
TTGTCTAGTCTTTGAGACTAGTTTGCCTTTCTCACATTGTGTATTCTGTTGTTAGAAATT
TGTGATAAAATGGTATTAGCTATATTTGCCAGAGTTAAACATATAGTAGAGGAATTCTTA
TCCTGTGATACATAGCTTCATATTTGGTAAATATGTTAATGGTTCTAATCAGATGGTGAA
AATATATGCCCCCCTCAATCCTGAAAGCACCTTTCATAAAAAAGAACCACTACCACAAAA
…
```

Each cell hold 2 copies of DNA

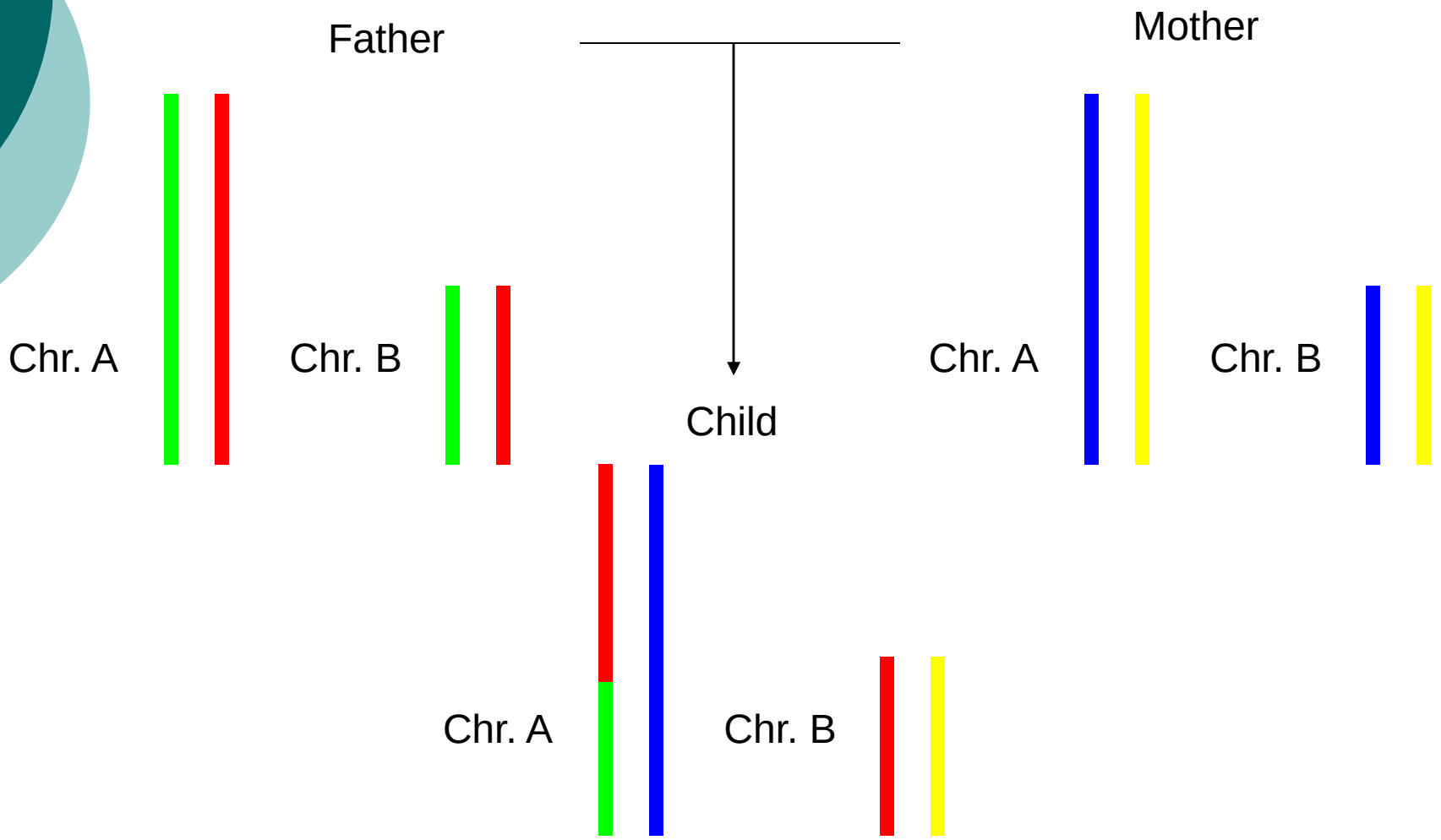


http://www.contexo.info/DNA_Basics/chromosomes.htm

The genetic material

- Each individual owns 2 templates of DNA (except for the male sexual chromosom).
- Each cell owns a copy of the 2 templates.
- The two templates are slightly different:
  - Approximately 1 difference every 1000 bp.

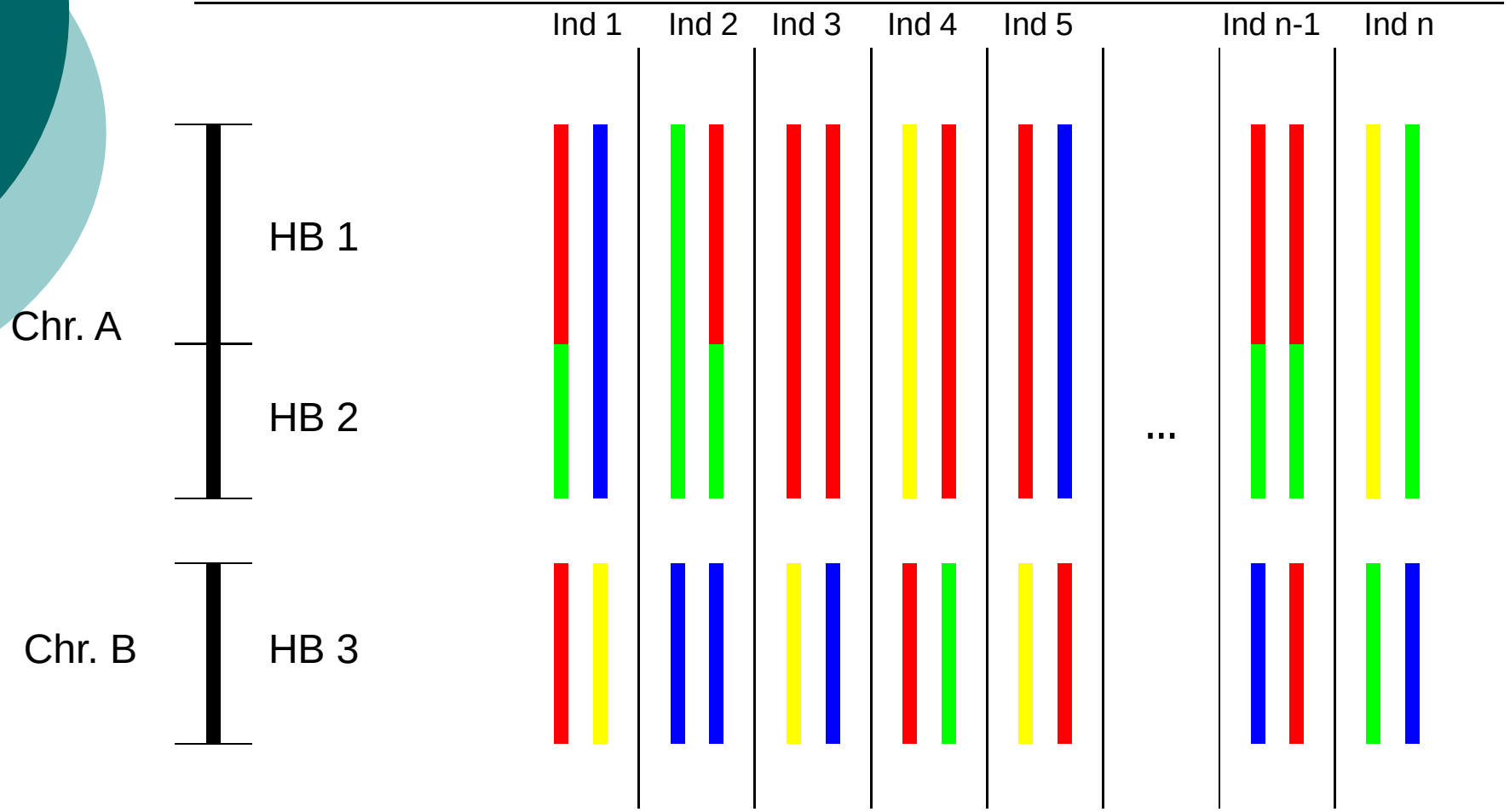
Transmission and recombination



Recombination: how often?

- Every 1E9 bp, i.e on average 3 events at each generation.
  - However: intensity of the Poisson process is variable from 1 to 10 given the portion of DNA.
- Population with near common ancesters:
  - ⇒ Few recombination events
  - ⇒ Finite set of recombination events

Haplotype blocks (HB)



Questions?

Data – association study

Genetic disease

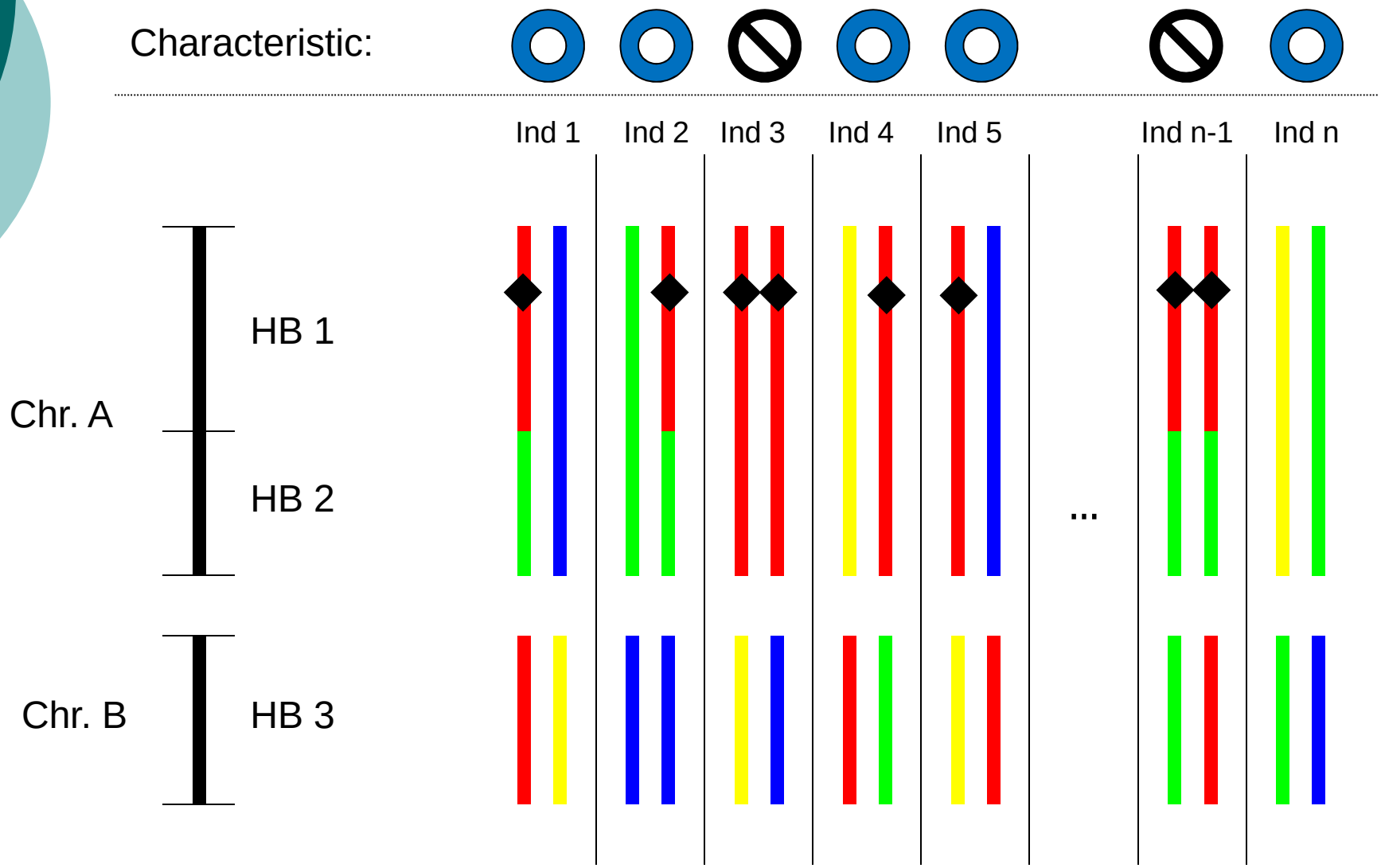
Variants of DNA causes disease:

- Simple case (« mendelian »):
  - One change in DNA
    - Simplest case: One letter change in DNA
- Complex case:
  - Variations at different locations
  - Interaction of variations
  - Interaction with environment

Genetic disease

- How to find the variant(s) causing the disease? By looking for an association 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:
  - done for only two (male) individuals…
- Current « affordable » technology:
  - reading 1 letter at around 100,000 predefined places on the 2 templates of DNA.

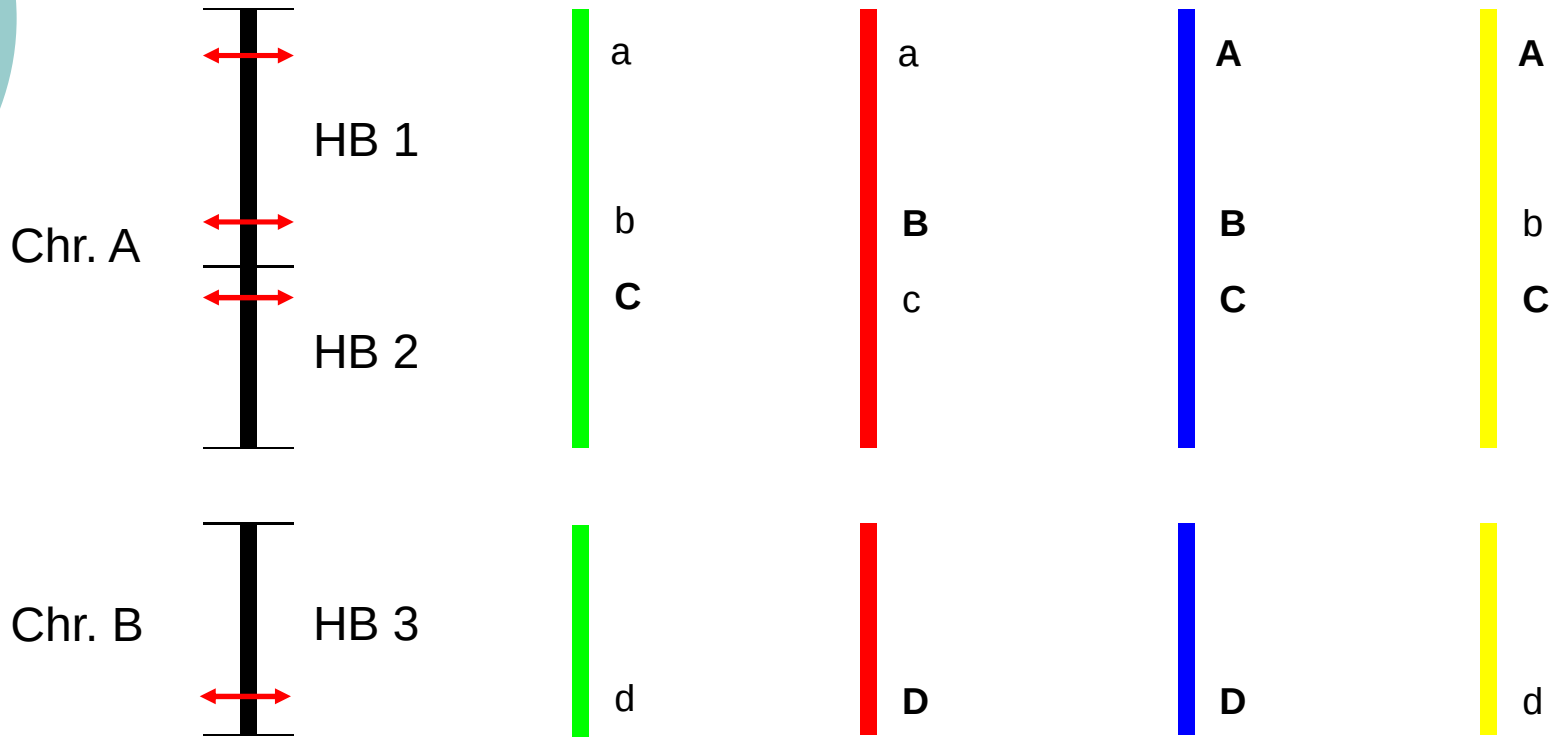
Single Nucleotide Polymorphism

- 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:

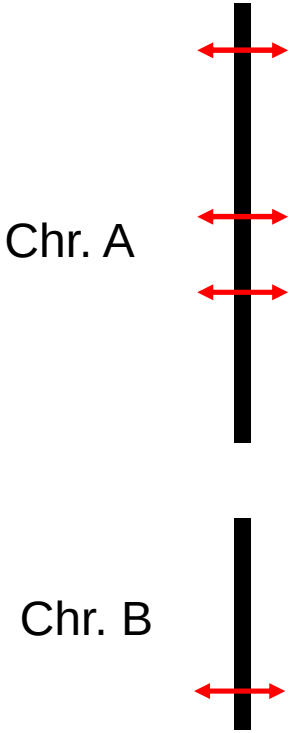
- ‘aa’
- ‘aA’ or ‘Aa’, but distinction is impossible
- ‘AA’.

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

Questions?

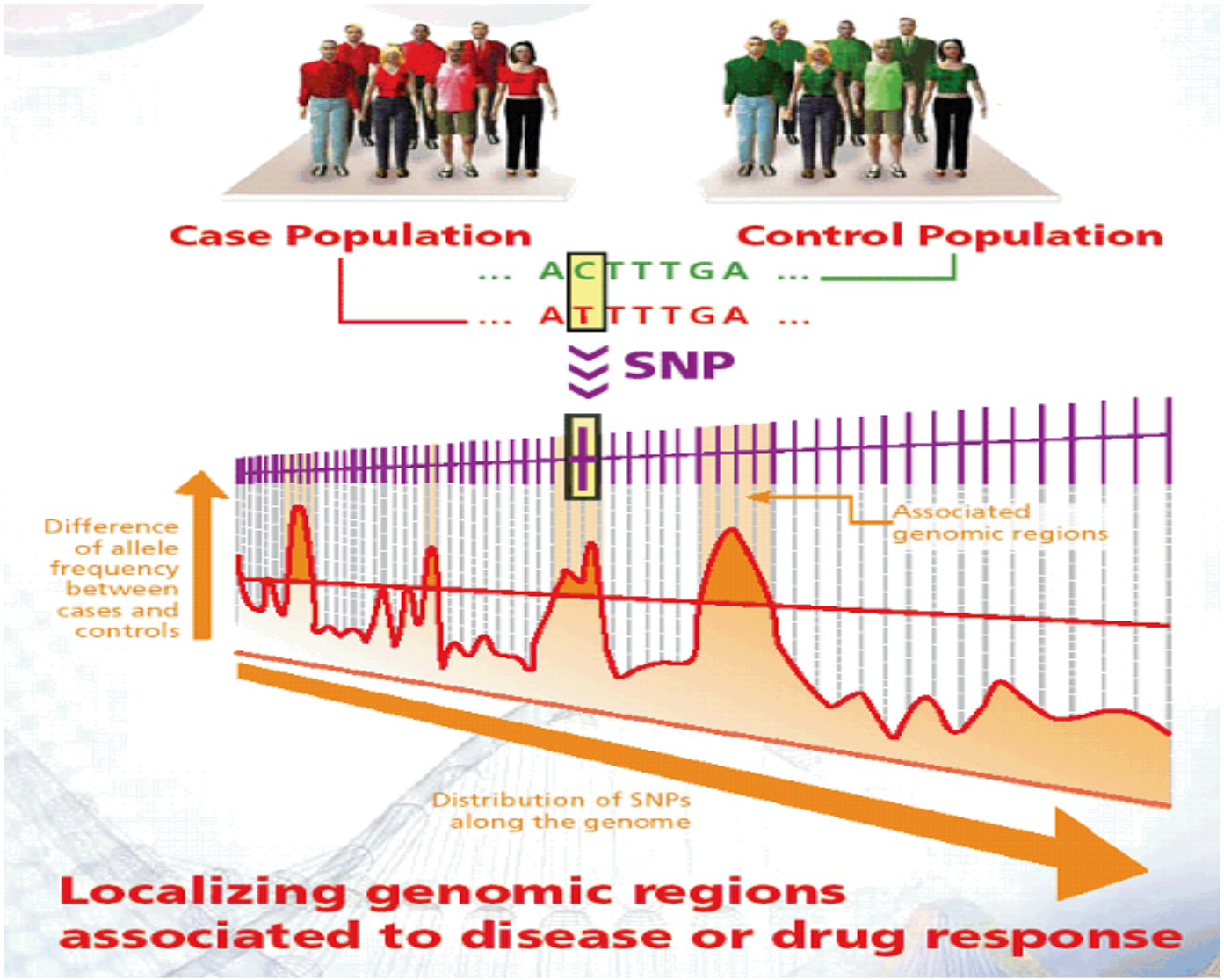
The Serono association study

- Multiple Sclerosis: Complex disease
  - Concordance rate between twins: 15-20 %
- Case/control design
- 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?

The “marketing” slide



Multiple testing problem

- **Test**: accepting or rejecting a null hypothesis given the value of a score computed on a dataset.
- **p-value**: probability of having a score value higher or equal to the observed value assuming that dataset follows the null hypothesis.
- **FDR** (False Discovery Rate): the proportion of tests that follows the null hypothesis among a subset of tests for which the null hypothesis has been rejected.

FDR estimation

- $\widehat{\pi_0}$: Proportion of tests under the null hypothesis *(assumed to be 1.0)*.
- B : Number of tests
- θ : Level at which FDR is computed
- π_b : P-value of test b

$$\mathrm{FDR}(\theta) = \frac{\widehat{\pi_0}\theta B}{\mathrm{card}(\{b|\pi_b < \theta\})}$$

Multiple testing problem

- *By definition, the distribution of p-values is uniform under the null hypothesis.*

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

Method

Bin definition

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

  $\Rightarrow$ Currently infeasible

Bin definition

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

  But:
  - Generally larger than haplotype blocks
    - Dilution of association signal
  - Boundary accross haplotype blocks
    - Not independent.

Multiple testing problem

Linkage disequilibrium $\Rightarrow$ 2 neighbour SNP truly associated (individual p-value=1E-5)

- Independent testing:

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

  $\Rightarrow$ No association detected

- Simultaneous testing:

  Assuming simple addition of χ^2 scores:

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

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

  $\Rightarrow$ OK

Bin definition : Dilution of signal

- Too large bin definition: Assuming bin with 9 SNP:
  - 2 associated SNP: p-value=1E-5
  - 7 unassociated SNP: p-value=1
- Results:
  - ⇒ *Assuming simple addition of χ^2 scores:*
    *new p-value = χ^2(2*invχ^2(1E-5,1),9) = 1.1 E-5*
  - ⇒ FDR = 92 %
  - ⇒ No association detected

Bin definition : Dilution of signal (2)

- Regrouping bins dilutes the signal but also decrease the number of tests.
  - If all SNPs are tested by 9:
    - Only 1E6/9 = 111,111 tests
    - ⇒ FDR = 56 %
  - ⇒ FDR reduced of 1/3.
    - ⇒ Significant difference before starting costly experiments

Statistical test of association

- One SNP:
  - Hardy-Weinberg:
    $P(aa)=p(a^2)$
    $P(aA)=2p(a)p(A)$
    $P(AA)=p(A)^2$
  - $P(a|case)=0.2$
    $P(a|control)=0.32$

	aa	aA	AA	Total
Case	12	96	192	300
Control	31	131	138	300
Total	43	227	330	600

- Statistics
  - $\chi^2(A) = 22.7$ *(p-value = 1.1E-5)*
  - $LR_3(A) = 2 * Log(L/L_0) = 23.1$
    (p-value = 9.6E-6)

Statistical test of association

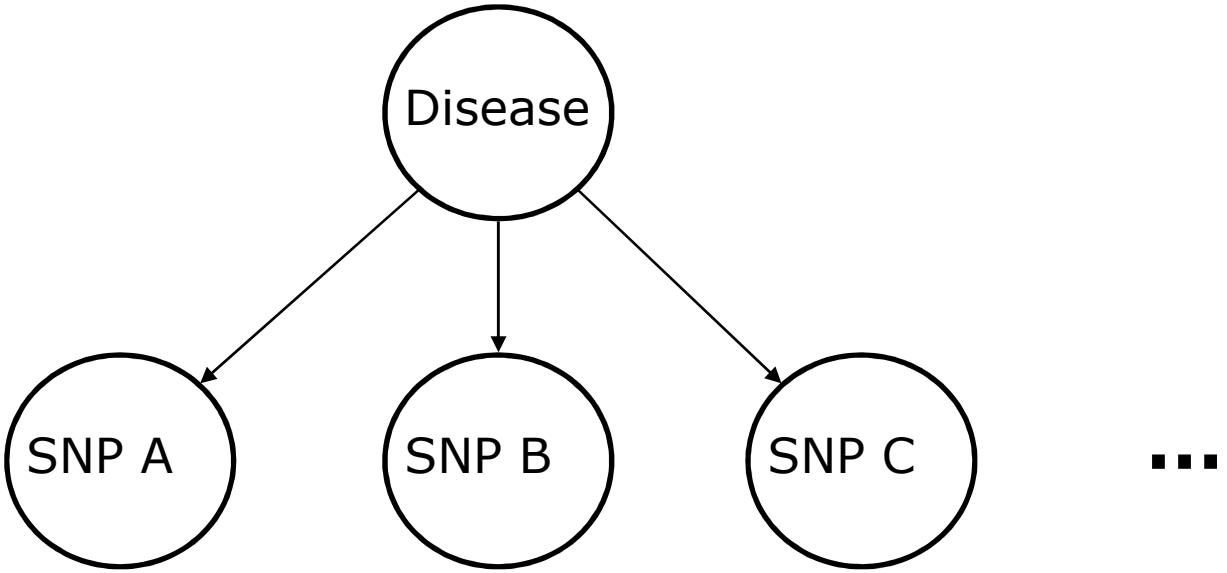
- Two SNPs - no linkage: p(aa,bb)=p(aa)p(bb):

SNP b	bb			bB			BB			
SNP a	aa	aA	AA	aa	aA	AA	aa	aA	AA	Total
Case	1	4	8	4	31	61	8	61	123	300
Control	3	14	14	14	57	60	14	60	63	300
Total	4	17	22	17	88	122	22	122	186	600

- Statistics
  - $\chi^2 = 44.8$ *(p-value = 5.7E-7)*
  - $LR_2(AB) = 2 * Log(L/L_0) = 45.4$ *(p-value = 3.1E-7)*
  - Asymptotic statistics inadequate

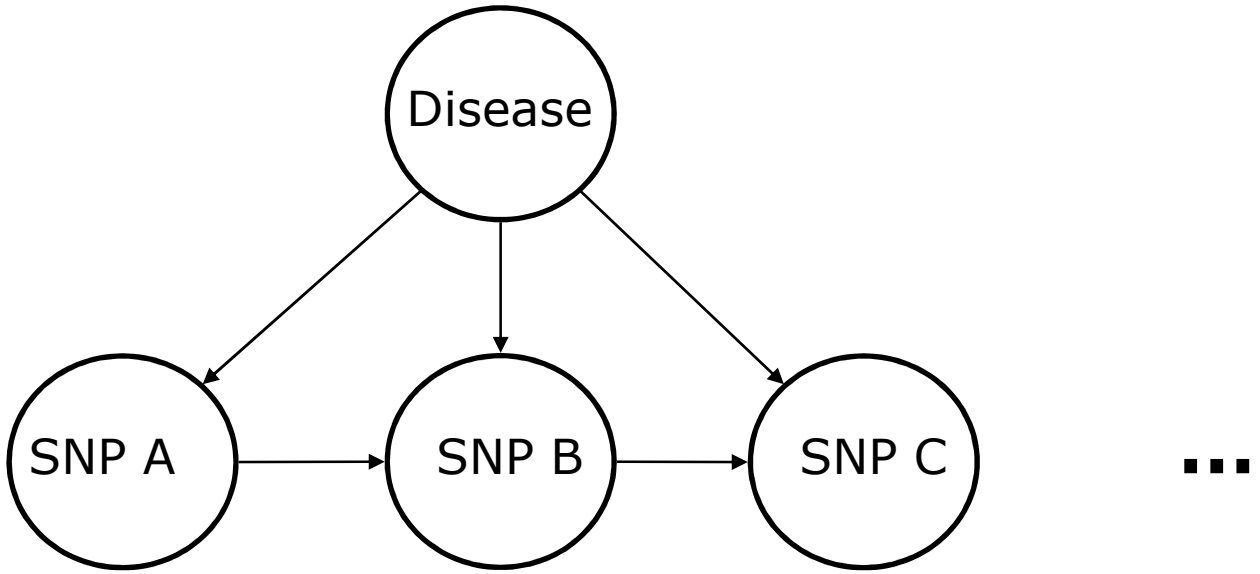
Statistical test: bin level

- Naive model: Adding the score of the single SNP tests
  - $LR_3(\text{bin}) = LR_3(A) + LR_3(B) + LR_3(C) + \ldots$



Statistical test: bin level

- Basic two-SNP model:
  - $LR_1(\text{bin}) = LR_3(AB) + LR_3(BC) + LR_3(C\ldots) + \ldots$



Statistical test: bin level

- One collection design:
  - LR(bin) = LR(bin,collection A)
- Three collection design:
  - LR(bin) = LR(bin,collection A)
    + LR(bin,collection B)
    + LR(bin,collection C)

Estimation

- Asymptotic p-values:
  - Badly filled tables
  - Missing value and error model
- Exact p-values:
  - Not tractable given the model
- Empirical p-values:
  - Flexible : any score can be used
  - Error is made but controlled
  - Computer intensive

Estimation: control of error

- Use computation power on smallest p-values:
  - θ : a-priori estimation of the p-value of the highest true-positive
  - π : estimation of the p-value of the bin
  - K : constant controlling the error level
  - N : number of tests
  - Number of permutations for the bin must be higher than:

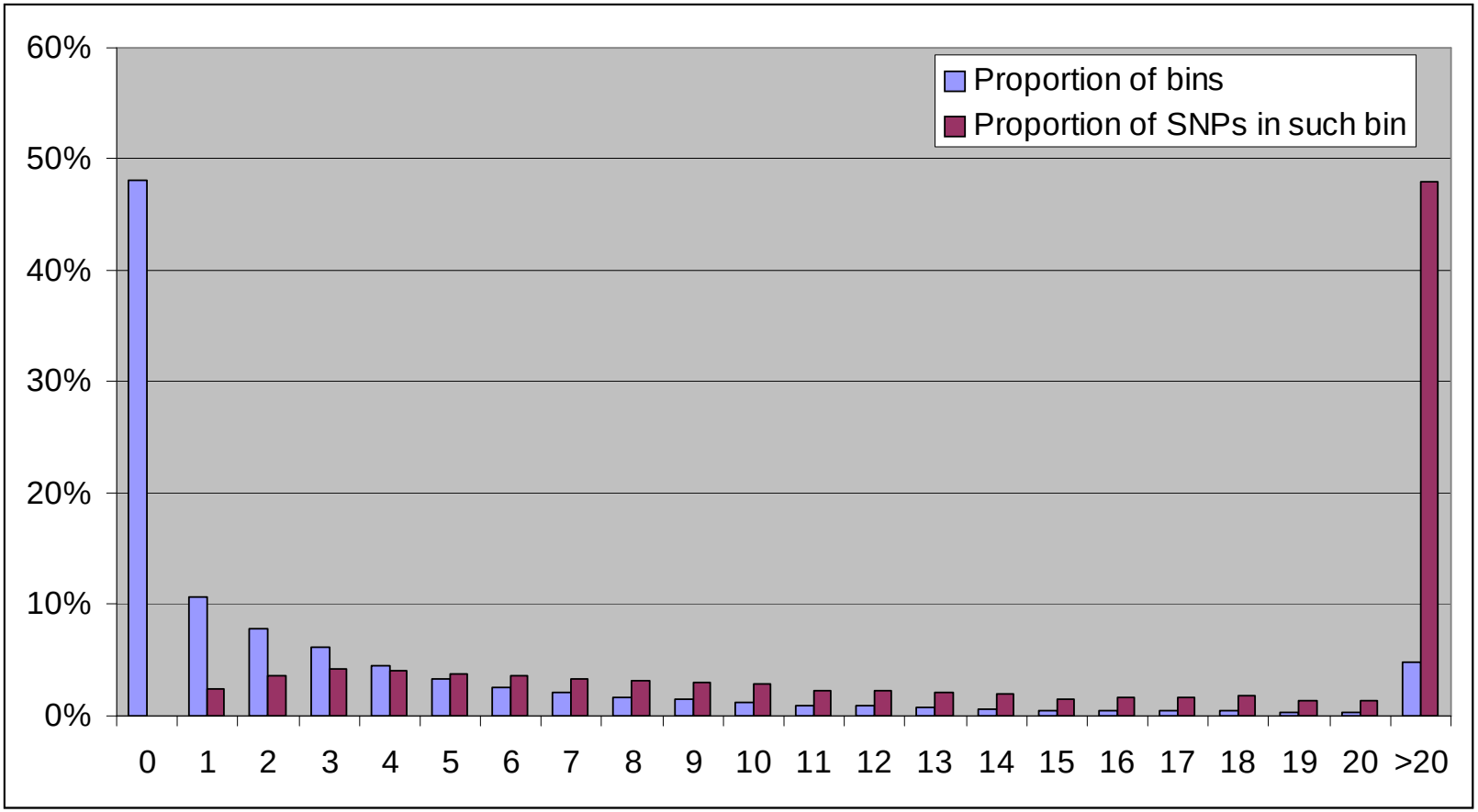
$$K \cdot N \cdot \text{Minimum}\left(1-\theta/\theta, 1-\pi/\pi \right)$$

Questions?

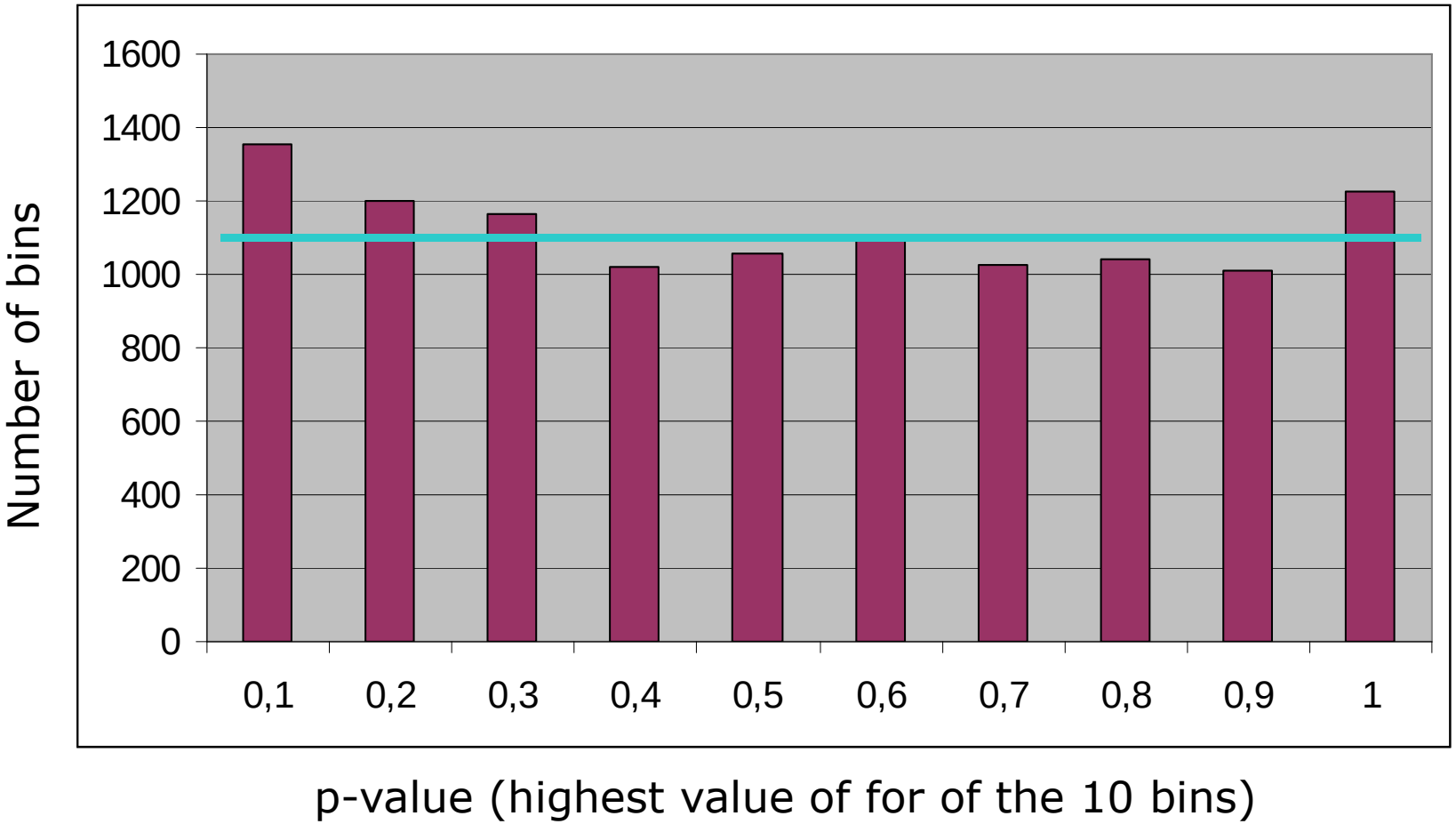
Results

Results: bins

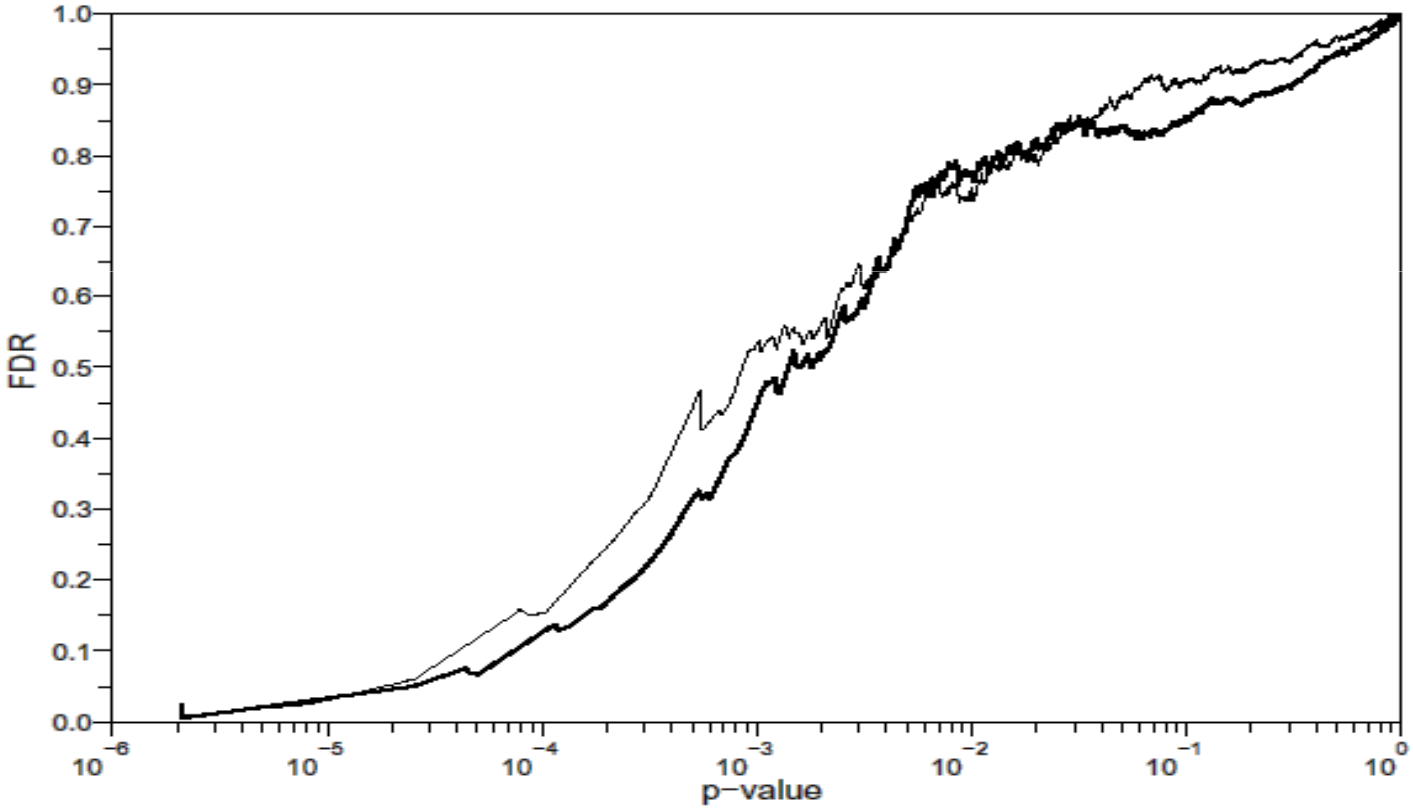
Distribution of the number of SNP per bin:



P-value distribution

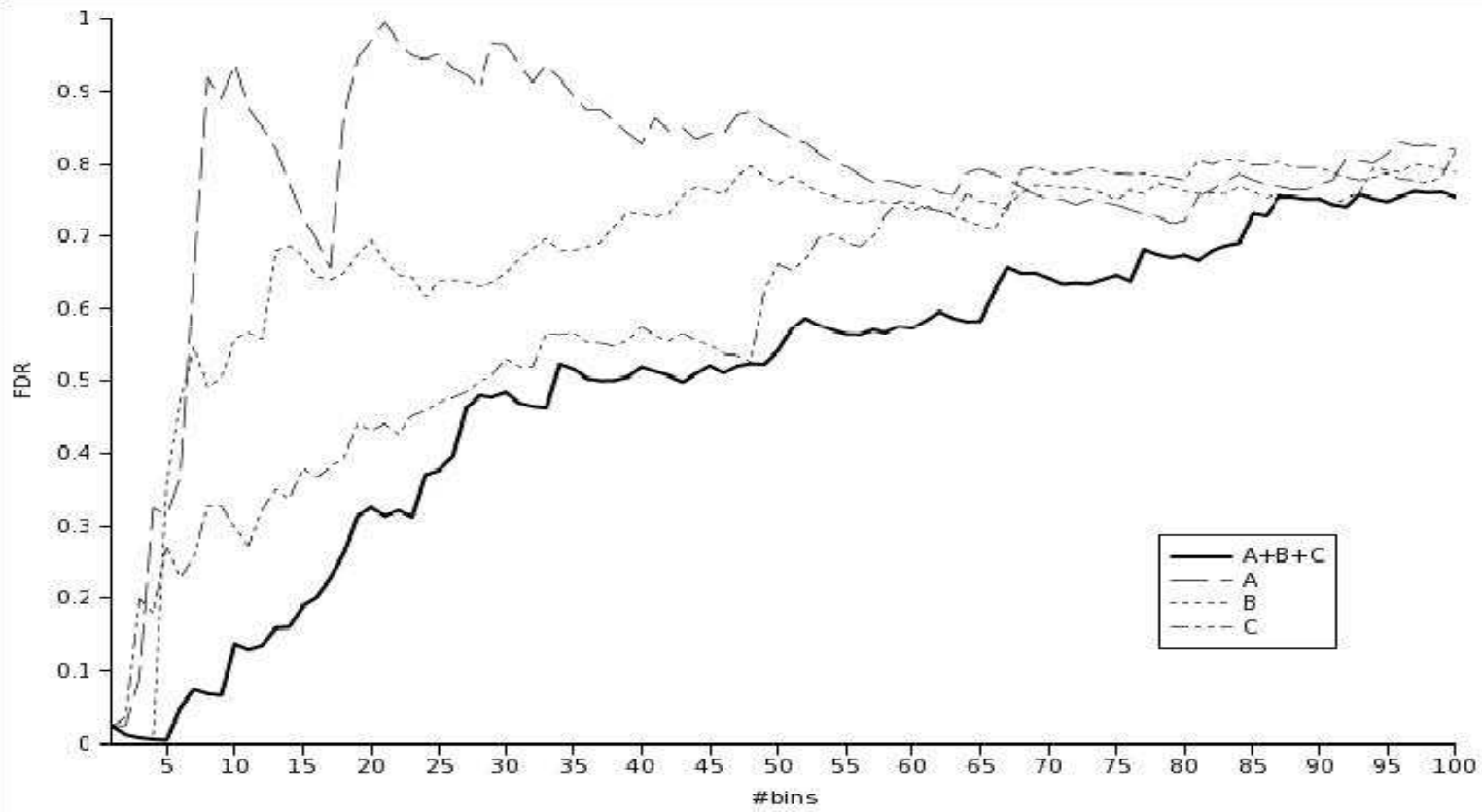


FDR: FDR vs p-value



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

FDR: FDR vs number of bins selected



Two-SNP likelihood

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

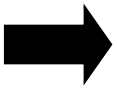
- FDR thres. 50%:

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