Learning Hierarchical Bayesian Networks for Genome-Wide Association Studies

Raphaël Mourad[1], Christine Sinoquet[2] and Philippe Leray[1]

KOD team (KnOwledge and Decision),
[1] LINA, UMR CNRS 6241, Ecole Polytechnique de l'Université de Nantes.
[2] LINA, UMR CNRS 6241, Université de Nantes.
FRANCE

Presented by Raphael Mourad
PhD student in Bioinformatics
raphael.mourad@univ-nantes.fr

Outline

1/ Introduction

2/ Fondamental concept of association genetics

3/ Presentation of genetic data

4/ Our approach

5/ Results and discussion

6/ Conclusion and outlooks

Introduction

- Context:

  Complex genetic diseases

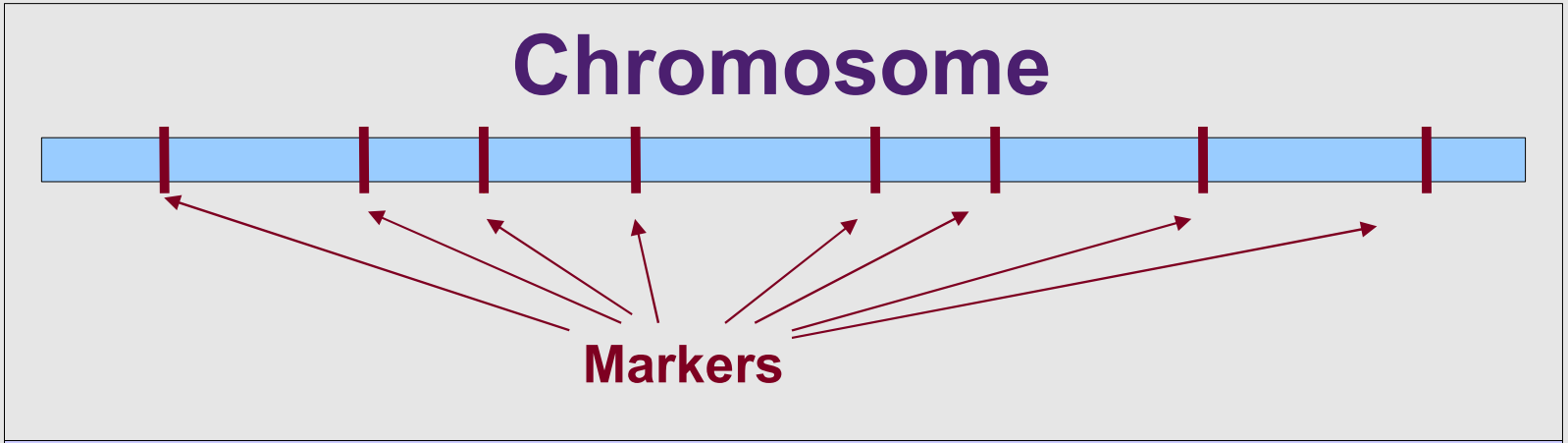
  = multifactorial genetic diseases caused by a combination of genetic factors (*eg* genes) and environmental factors (*eg* sex, age...).

  Examples: diabetes, asthma, hypertension, some cancers...

- Dissect the genetic basis of these diseases:

  Genome-wide association studies (GWAS)

  → identification of genetic markers associated with common, complex diseases.



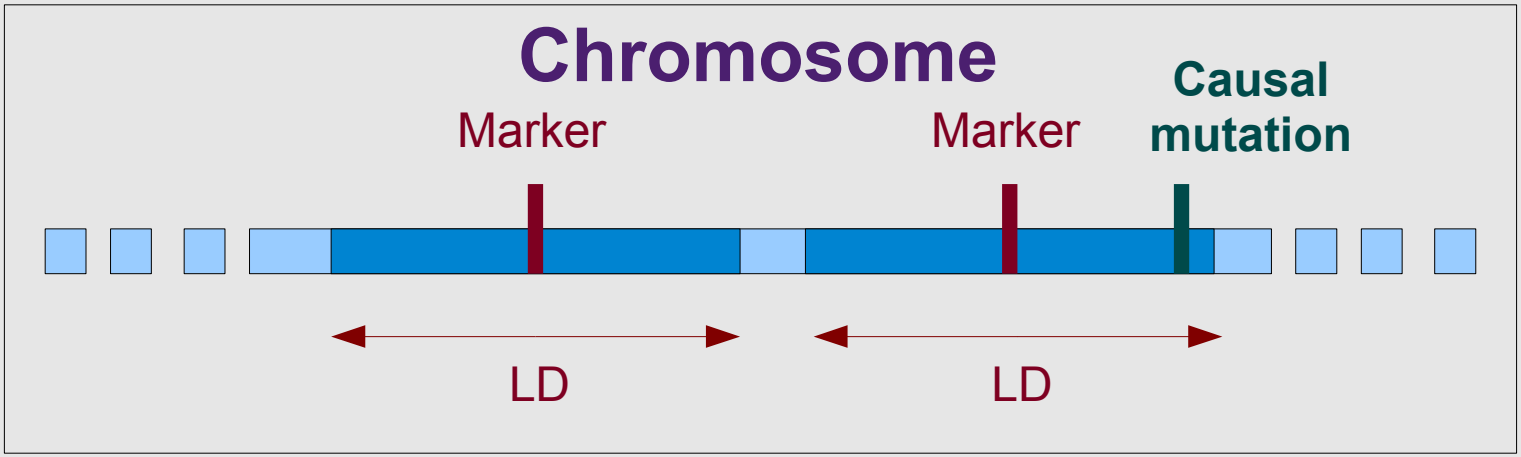
The human genome variability is covered by hundreds of thousands of markers.

Fondamental concept of association genetics

- Linkage disequilibrium (LD):

  → dependences generally observed between close SNPs on the chromosome,

  → at the basis of GWAS.



LD between markers and their surrounding area on the chromosome.

Presentation of genetic data

Phenotype

1 binary variable:
- 1000 non-affected individuals
- 1000 affected individuals

DNA

> 100k SNP
Ternary variables

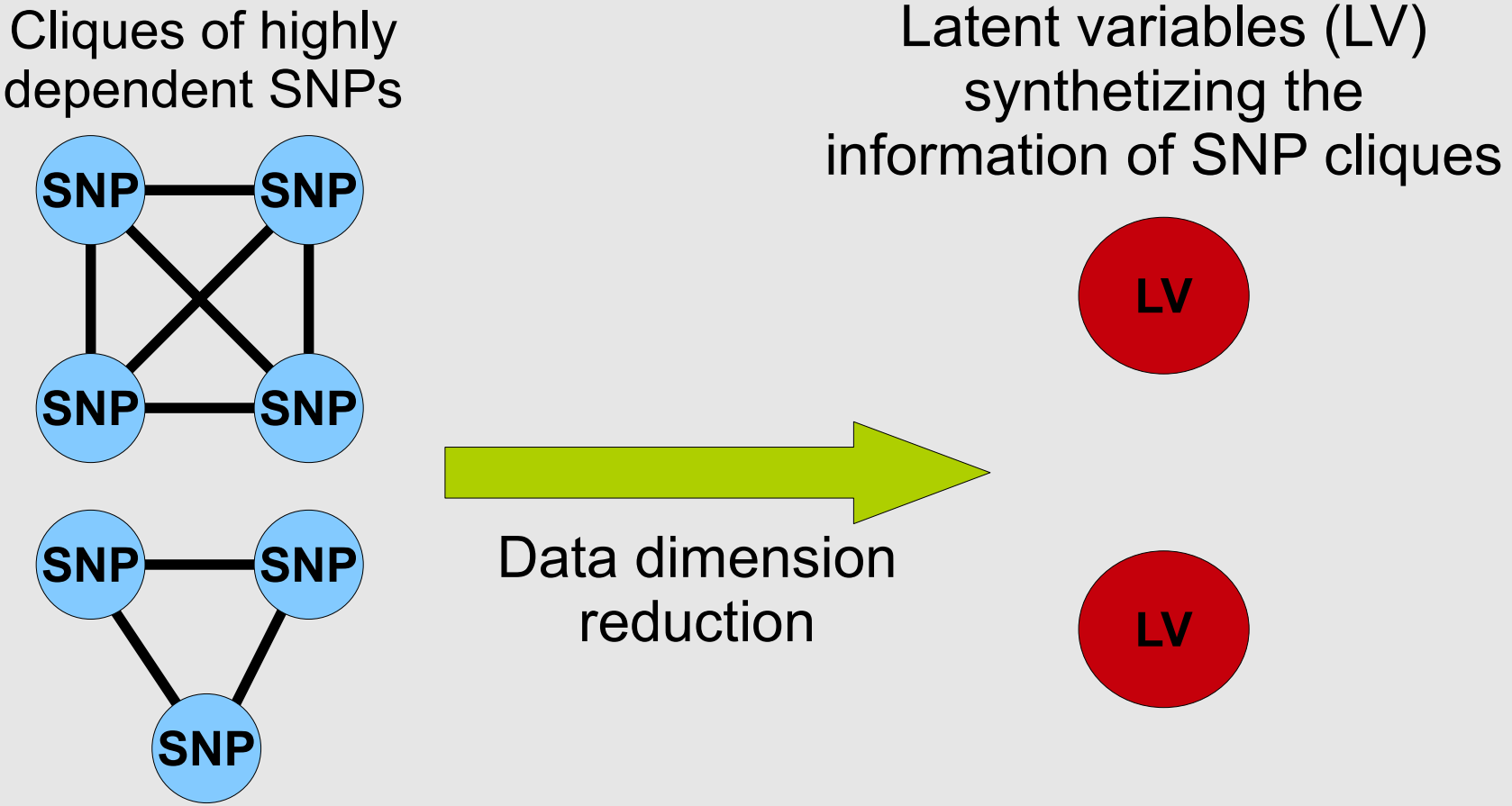
- Characteristics:

  → large number of genetic variables (SNP): combinatorial explosion

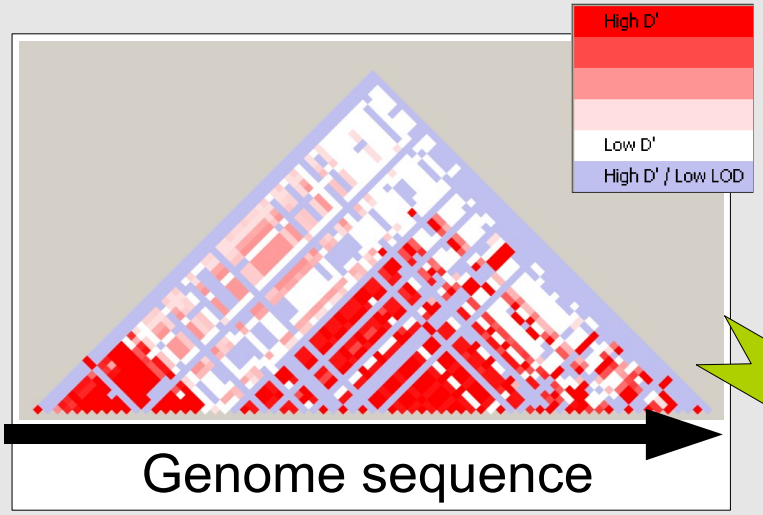
  → strong dependences among genetic variables

Our approach

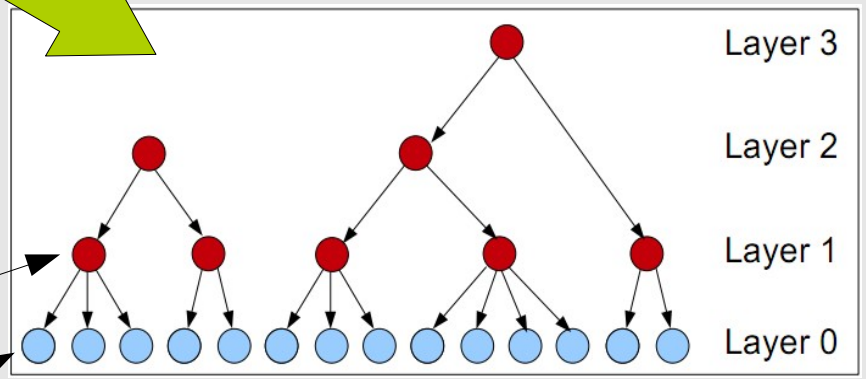
- Reduce the data dimension by synthetizing the information of highly dependent SNPs, due to LD.



- Provide a flexible and adapted probabilistic model to reduce dimension for genetic data.



Characteristics of data: dependences by blocs of SNPs



Latent variables

Observed variables (SNPs)

Proposed modelling

Forest of Hierarchical Latent Class models (FHLCMs)

- Advantages of this modelling:

→ hierarchical, thus :

- various degrees of dimension reduction,

- various degrees of LD strength,

→ each latent variable can reveal multiple-SNP patterns, potentially relevant to explain the disease,

→ contrary to Hierarchical Latent Class model, SNPs are not constrained to be dependent upon one another,

→ high-order interactions between SNPs can be taken into account.

- Proposed algorithm to learn both parameters and structure of FHLCMs from data:

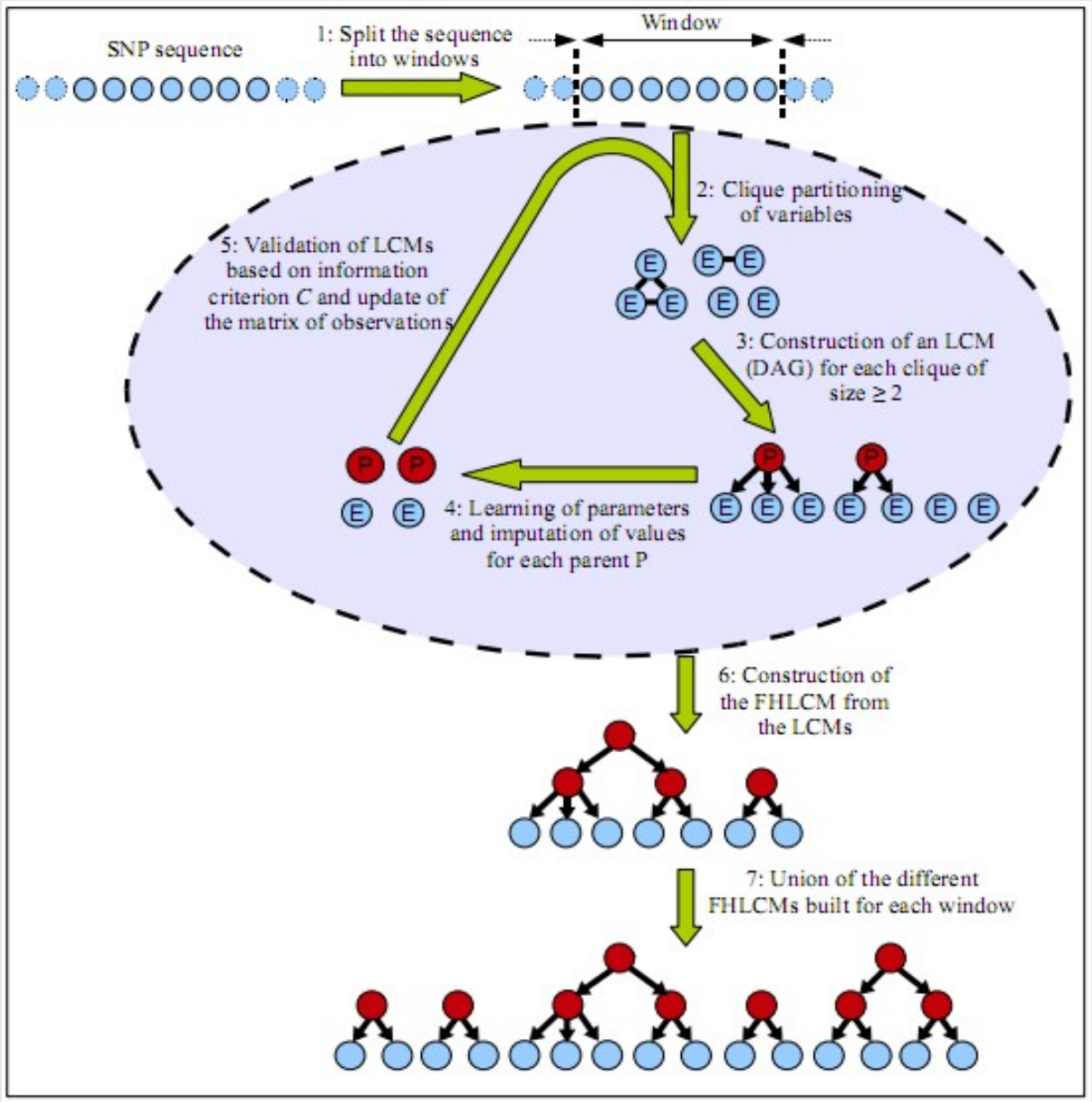
  CFHLC (Construction of Forests of Hierarchical Latent Class models).

  → based on an agglomerative hierarchical procedure to ensure scalability,

  → uses clique partitioning methods for an efficient discovery of non-overlapping cliques of dependent SNPs,

  → not restricted to binary variables and binary trees, as Hwang *et al*.'s algorithm.

Schema of the algorithm:



Results and discussion

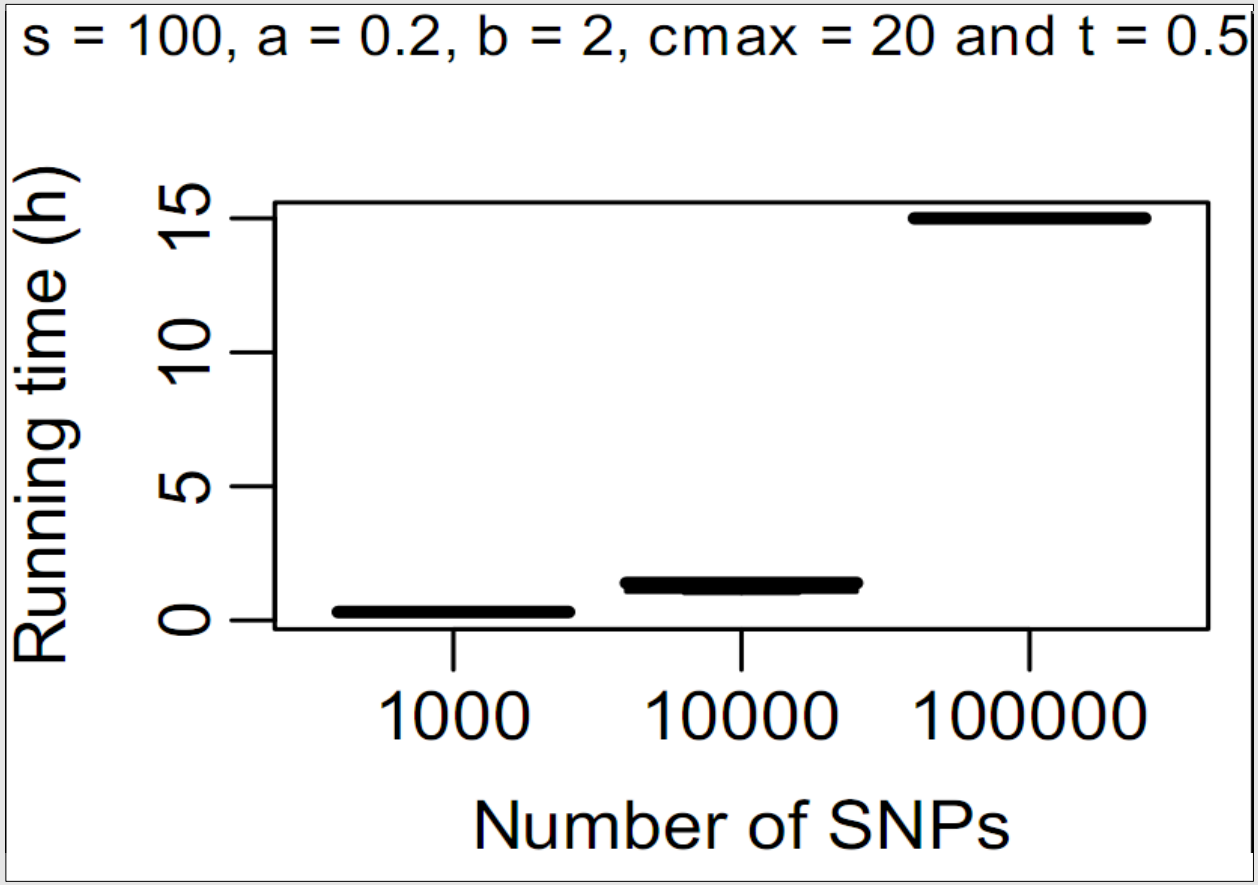
- Protocol testing:

  → C++ implementation,

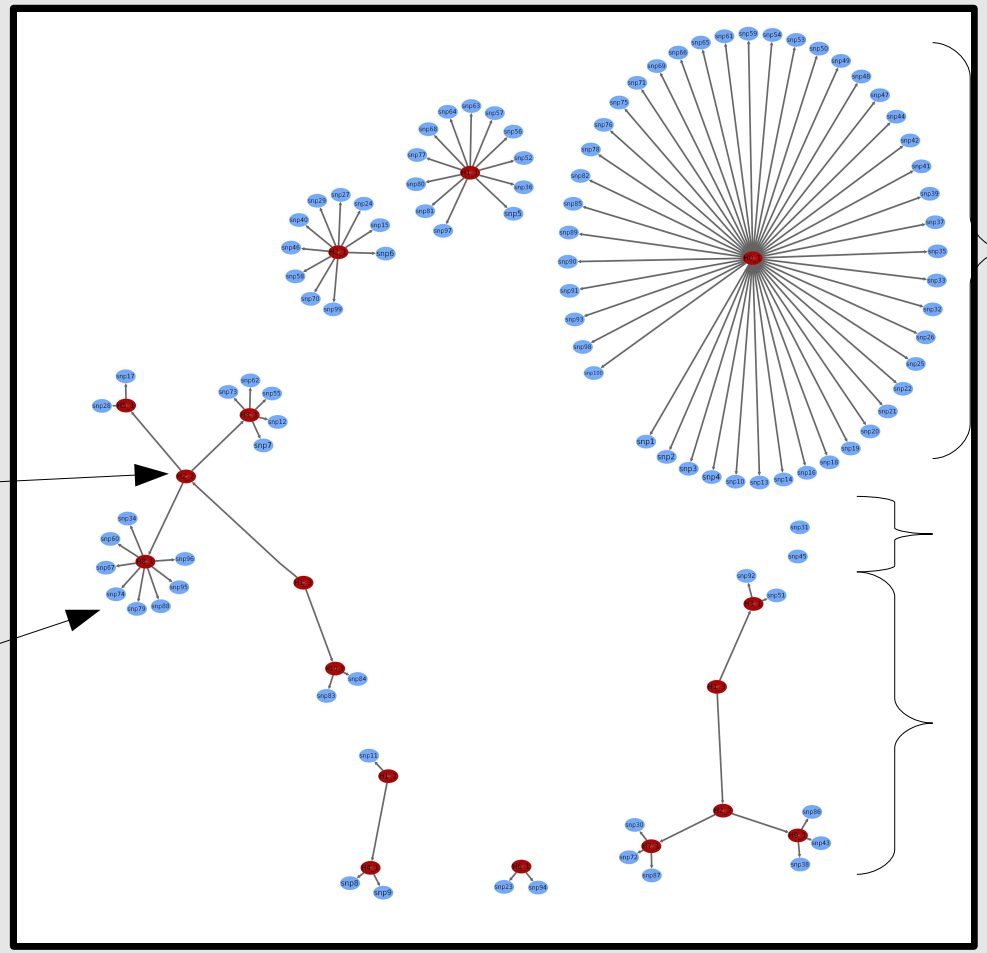
  → run on a standard pc (3.8 GHz, 3.3 Go RAM),

  → tested on simulated unphased genotypic data consisting of 2000 individuals and 1k, 10k or 100k SNPs, generated with the software Hapsimu.

Scalability



Visual display of a FHLCM: 100 snp sequence



Conclusion and outlooks

Conclusion:

- CFHLC algorithm have been shown to be efficient on genome-scaled data,
- Can provide a data dimension reduction of 80%.

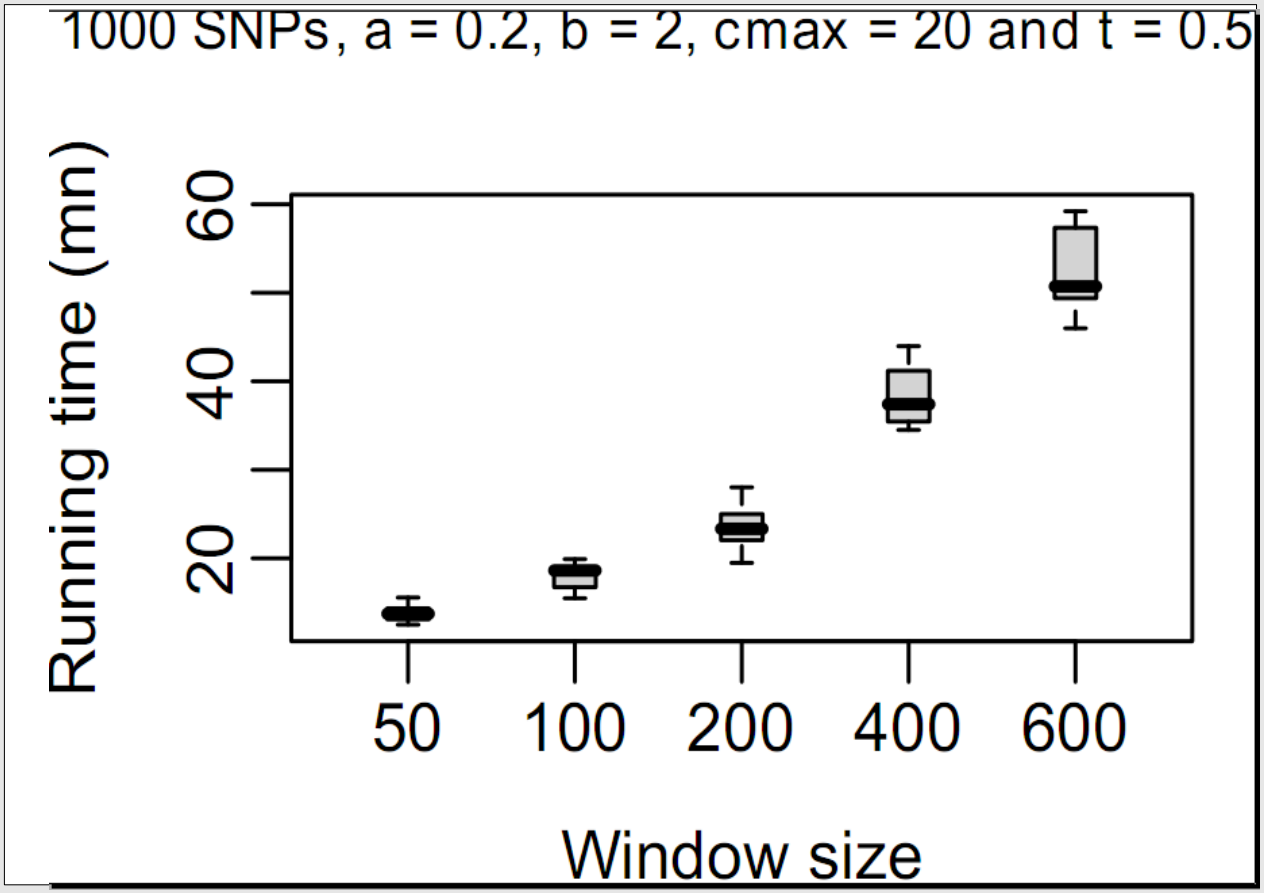
Perspectives:

- Application on the detection of genetic associations thanks to FHLCM's latent variables,
- Visualization of LD structure through the FHLCM's graph.

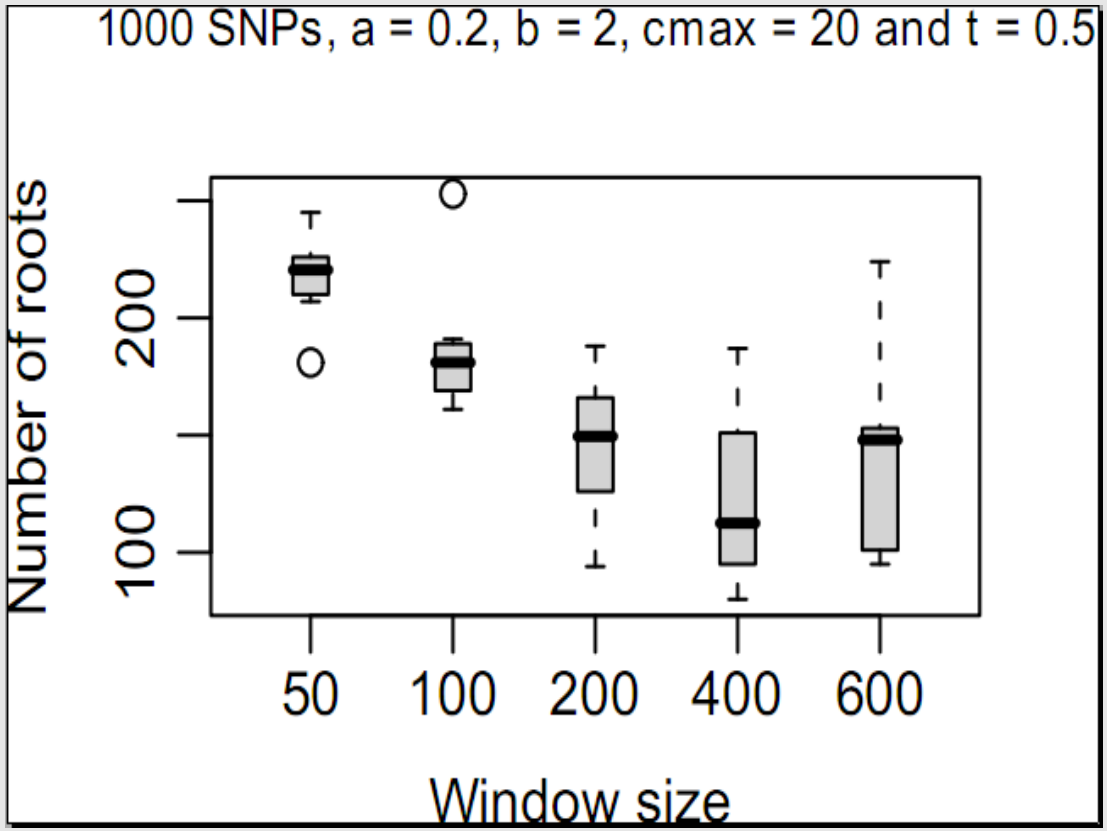
Thanks for your attention

Questions

Impact of window size on running time



Impact of window size on dimension reduction



Bibliography

General on GWASs:

- Balding D. (2006): a tutorial on statistical methods for population association studies.

Specific to probabilistic graphical models:

- Verzilli (2007): Bayesian graphical models for genome-wide association studies.

- Hwang (2006): learning hierarchical Bayesian networks for large-scale data analysis.