Genotypic data densification for genomic selection in the black poplar

Thesis director : Léopoldo SANCHEZ, Supervisor : Véronique JORGE

Three main questions

Thesis project focus on best way to implement genomic evaluation in the French black poplar breeding program with three main questions

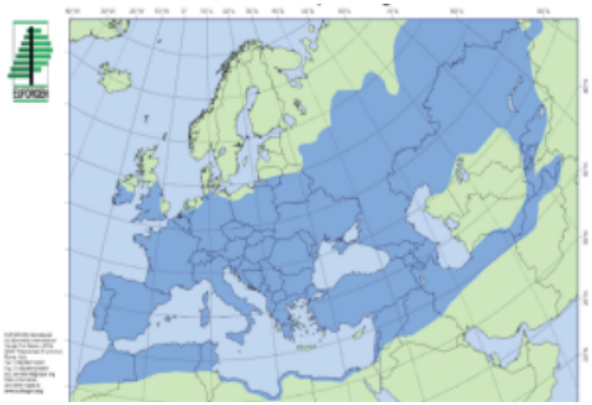
1. How to improve the prediction of crosses performance?
2. How best manage the genetic diversity?
3. How and where to integrate genomic evaluation in a breeding program to increase genetic gain?

$\Rightarrow$ This study deals with the first question by providing a first assessments of the high density genotype imputation accuracy in a factorial mating design

Populus nigra

Main species of the riparian forest

Wide distribution area



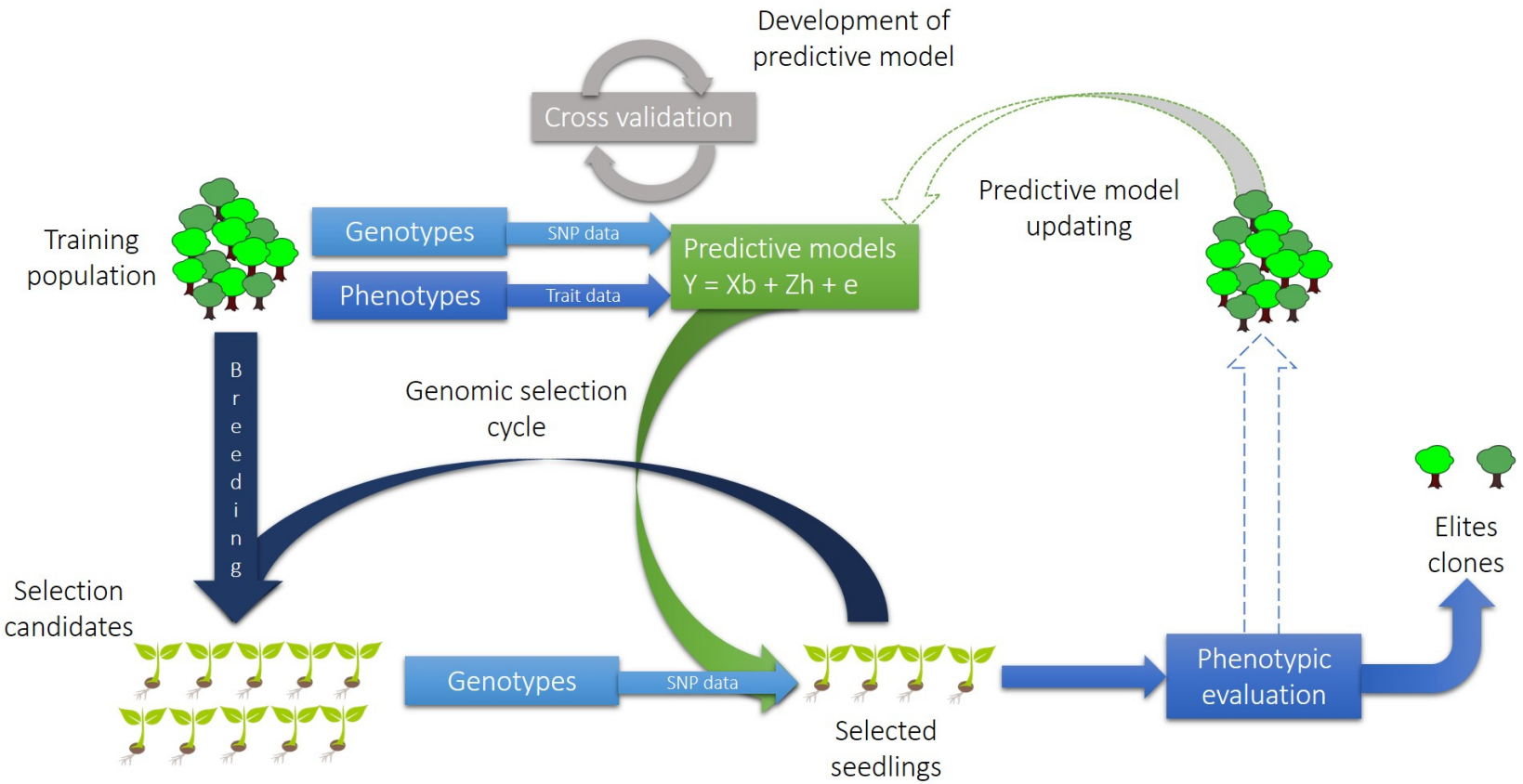
P. deltoides *P. nigra*

One of the most used trees in wood industry

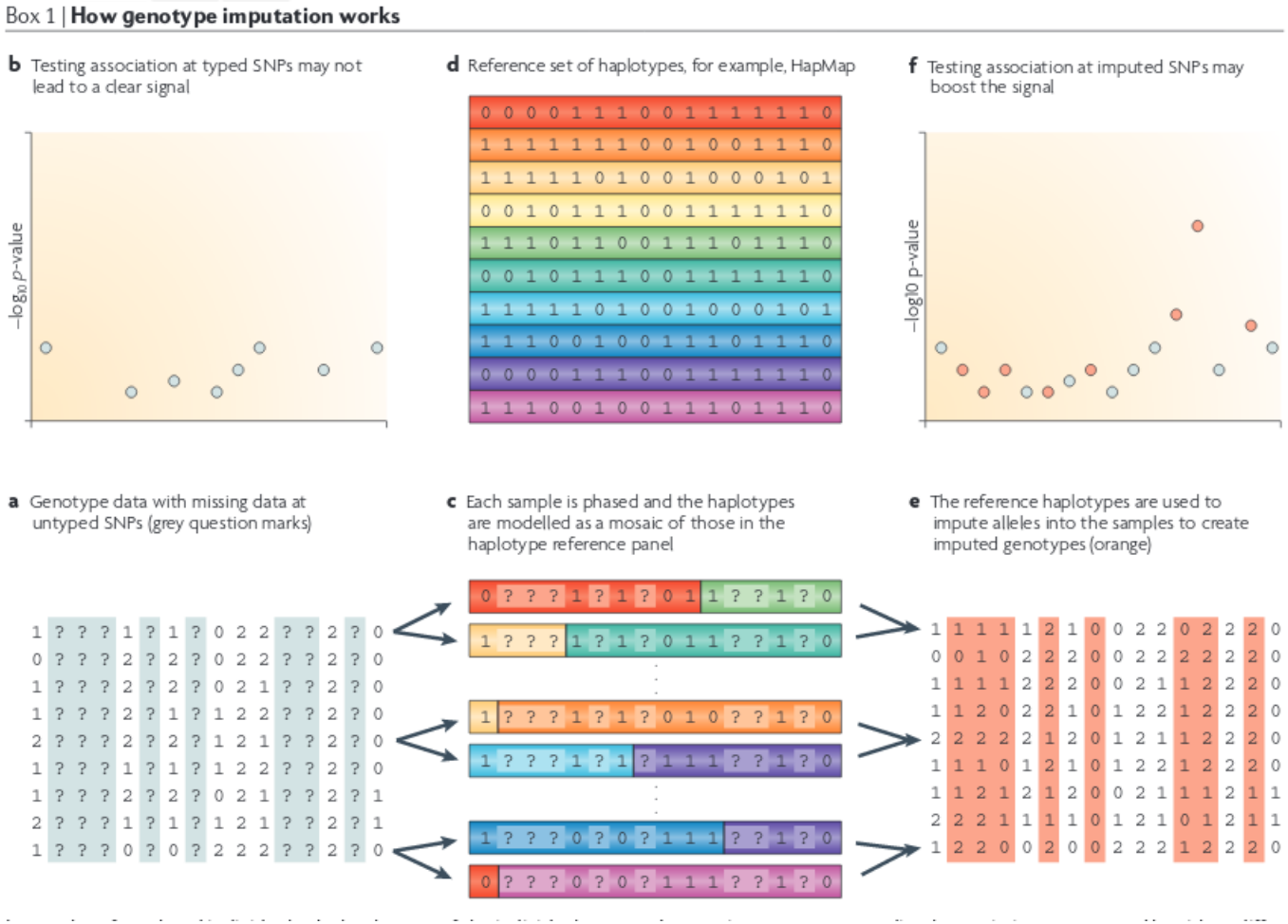
monoclonal plantations

Fast-growing tree

Genomic Selection



What is genotype imputation ?



Marchini et Howie (2010) Genotype imputation for genome-wide association studies (Nature reviews)

Material & Methods

Material & Methode : Factorial mating design

	SRZ	BDG	71077-308	
VGN-CZB25	52	49 & 2	48	Sequenced & genotyped
71041-3-402		22	10	genotyped
71072-501	21	28	29	

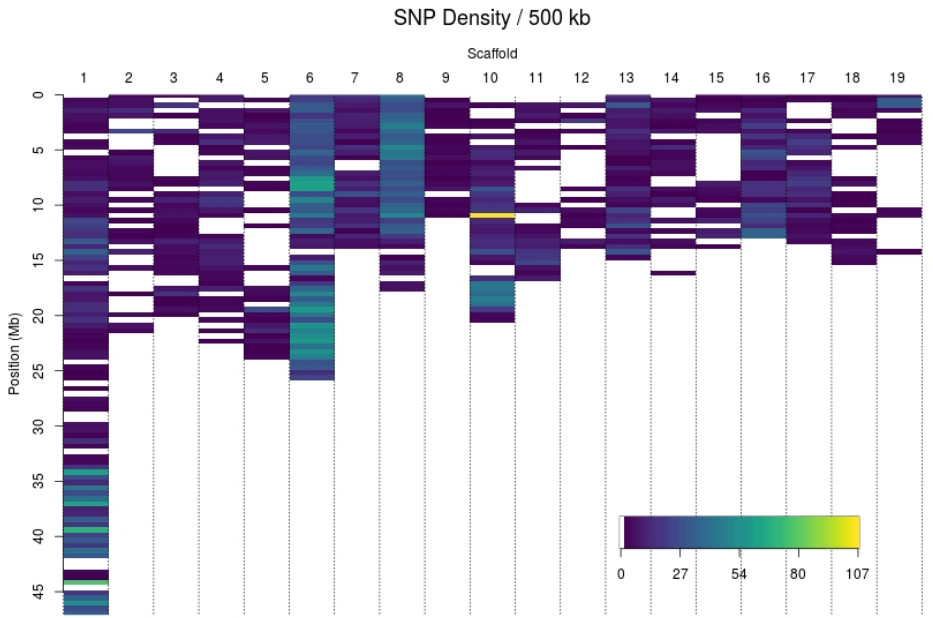
Indivisuals	Coverage
SRZ	10X
BDG	57X
71077-308	17X
VGN-CZB25	15X
71041-3-402	5X
71072-501	1X
662200037	6X
66220216	4X

2 636 075 detected variant
$\Rightarrow$ We try to impute 97% of genotyping data

Material & methods : Genotypic Data

Infinium BlackPoplar 10 332 SNP (Faivre-Rampant et al., 2016)

- non homogeneous marker density (denser candidate regions)
- 8259 SNPs available, several filters have been applied and :
  - Markers with more than 90% of missing data
  - Monomorphic markers
  - Not consistent markers after imputation
- 7755 markers remain
- Variable distribution on the 19 chromosomes
  - High density : 80 SNPs / Mb
  - Medium density : 20 SNPs /Mb
  - Low density : 5 SNPs / Mb



Material & methods : Sequence Data

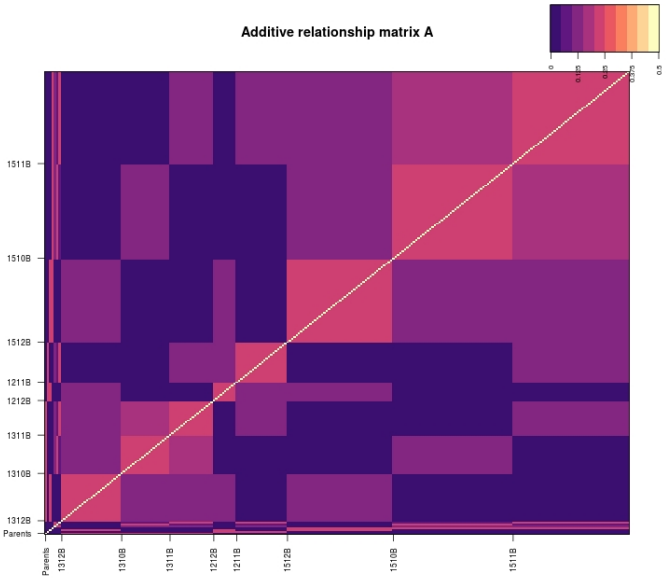
Whole-genome re-sequencing was performed at the INRA-EPGV/CEA-IG/CNG using Hiseq 2000 Illumina platforms. Options are the same as Faivre-Rampant et al., 2016

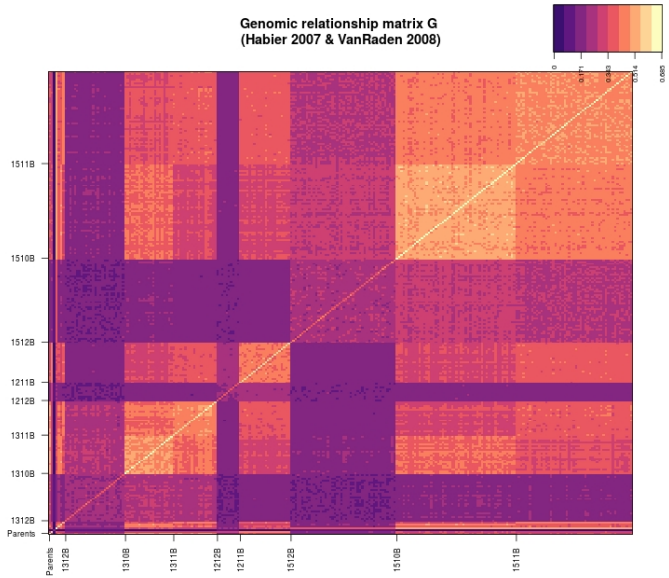
Import	Trimming
▶ Performed with CLC workbench genomics 9.0 ▶ Distance between two paired-end reads = 250 and 800 bases ▶ Range of illumina quality score version 1,8	▶ Performed with CLC workbench genomics 9.0 ▶ Phred score quality = 30 ▶ Small reads and ambigous sequences
Mapping	**Variant detection**
▶ Performed with CLC workbench genomics 9.0 ▶ Unique matches ▶ Length fraction >= 0.9 ▶ Similarity >= 0.9	▶ Performed with Freebayes ▶ min alternate count = 2 ▶ min alternate qsum = 40 ▶ genotype variant threshold = 2

Material & Methode : Population Structure

The population structure via relatedness heatplot

- pedigree-based relationship matrix (A-matrix)
- marker-based genomic relationship matrix following Habier (2007) and VanRaden (2008) methods

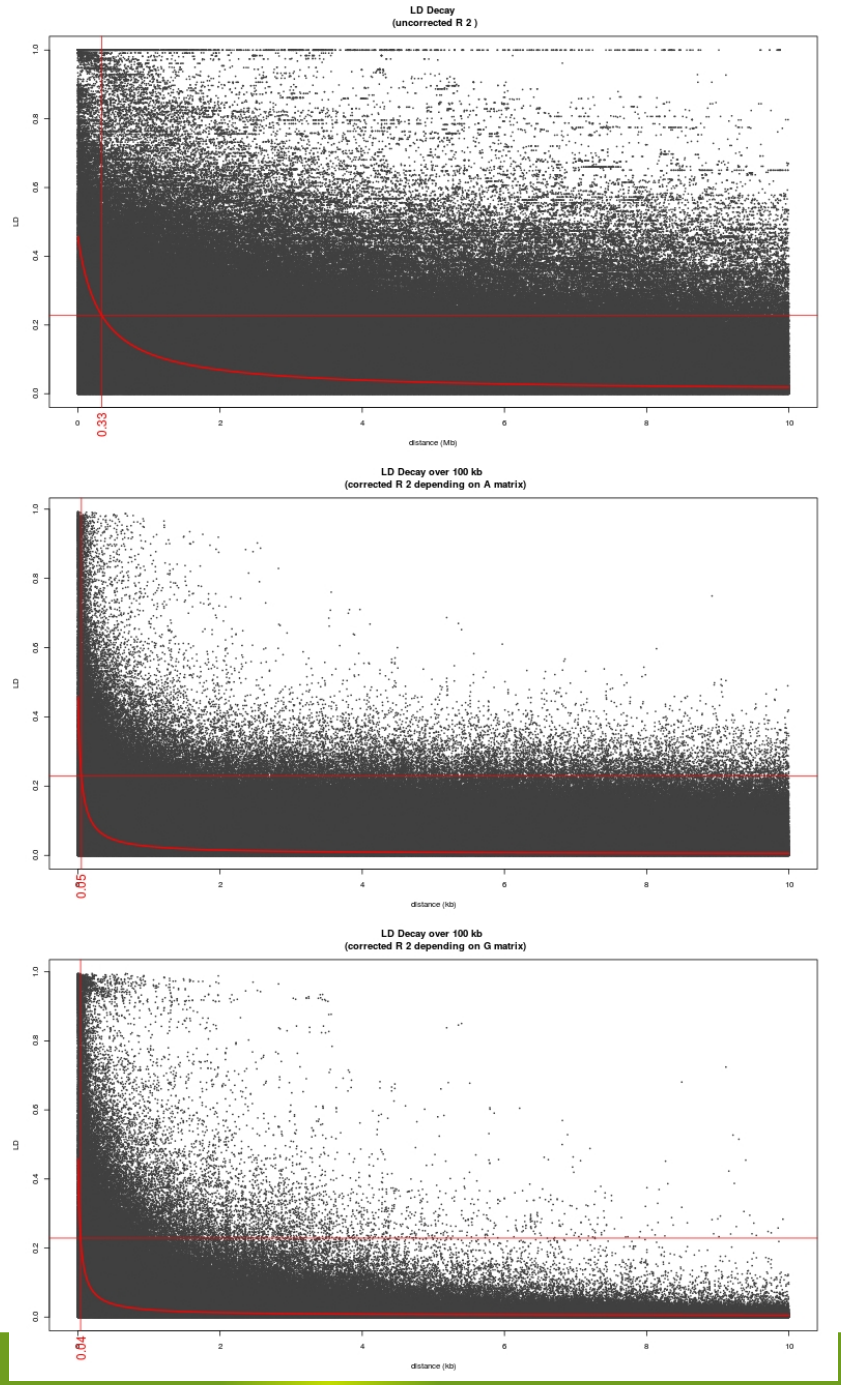




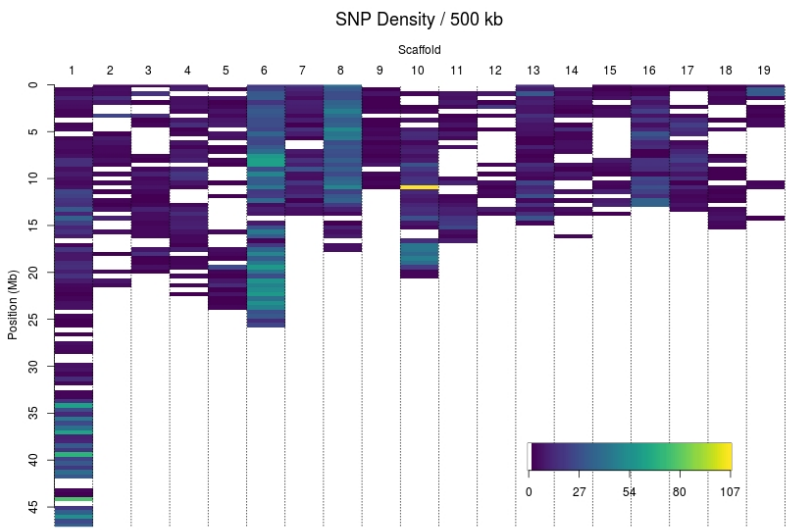
Material & Methode : Data description

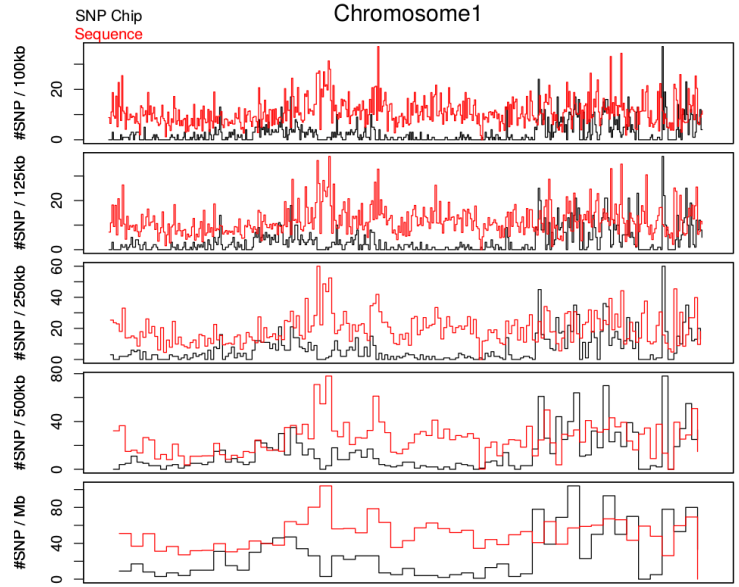
Linkage disequilibrium representation corrected by A and G (Mangin *et al.*, 2012 & lin *et al.*, 2012 . LD decay was modelized by Hill & Weir 1988 methods.

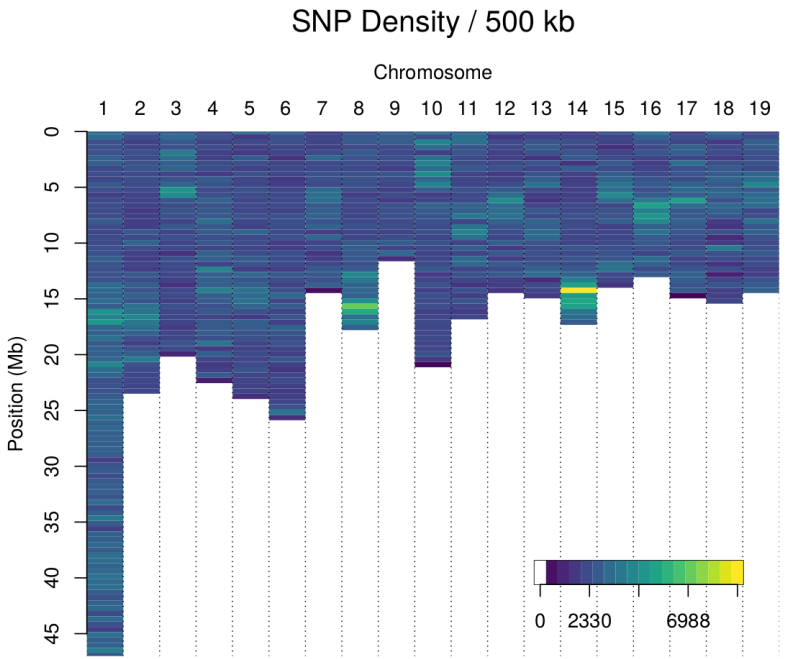
- 330 kb 1/2 LD decay without correction
- 50 kb 1/2 LD decay with A correction
- 40 kb 1/2 LD decay with G correction

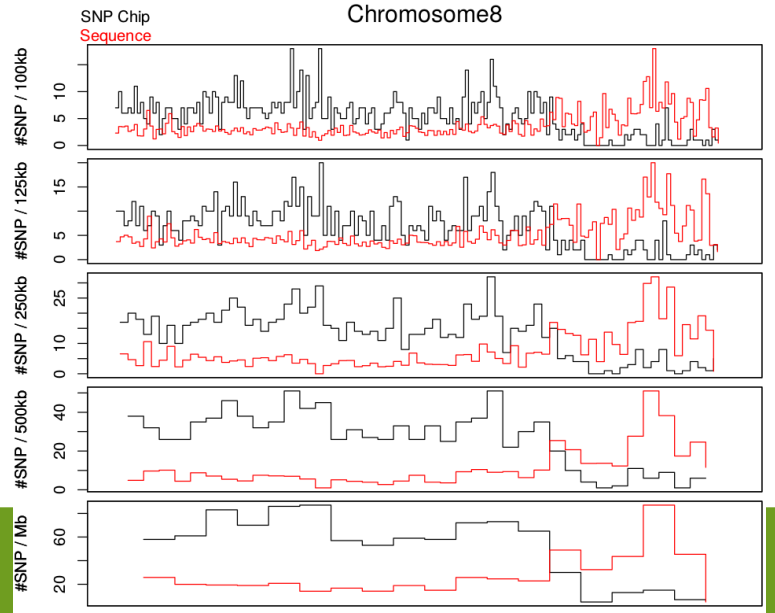


Material & Methode : SNP density









Material & Methode : FImpute

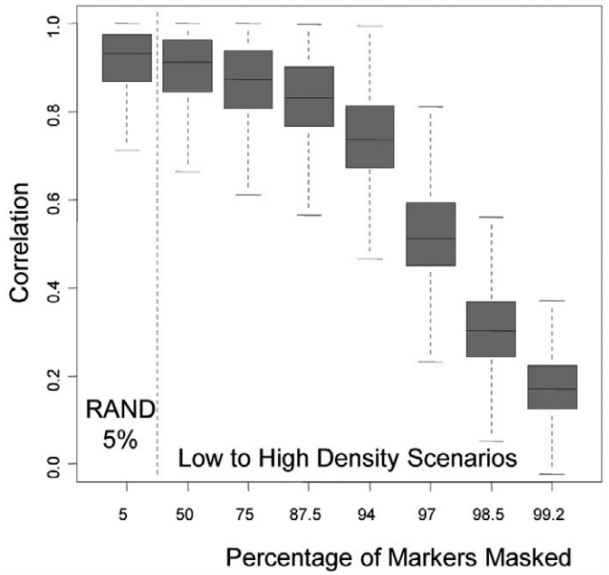
FImpute (Sargolzaei *et al.* 2014) :

- mainly developed for large scale genotype imputation in livestock
- uses an overlapping sliding window approach to efficiently exploit relationships or haplotype similarities between target and reference individuals
- makes use of pedigree information for more accurate imputation mostly with low density panel
- Easy to used with control, genotype, pedigree, haplotype files

Preliminary results & Discussion

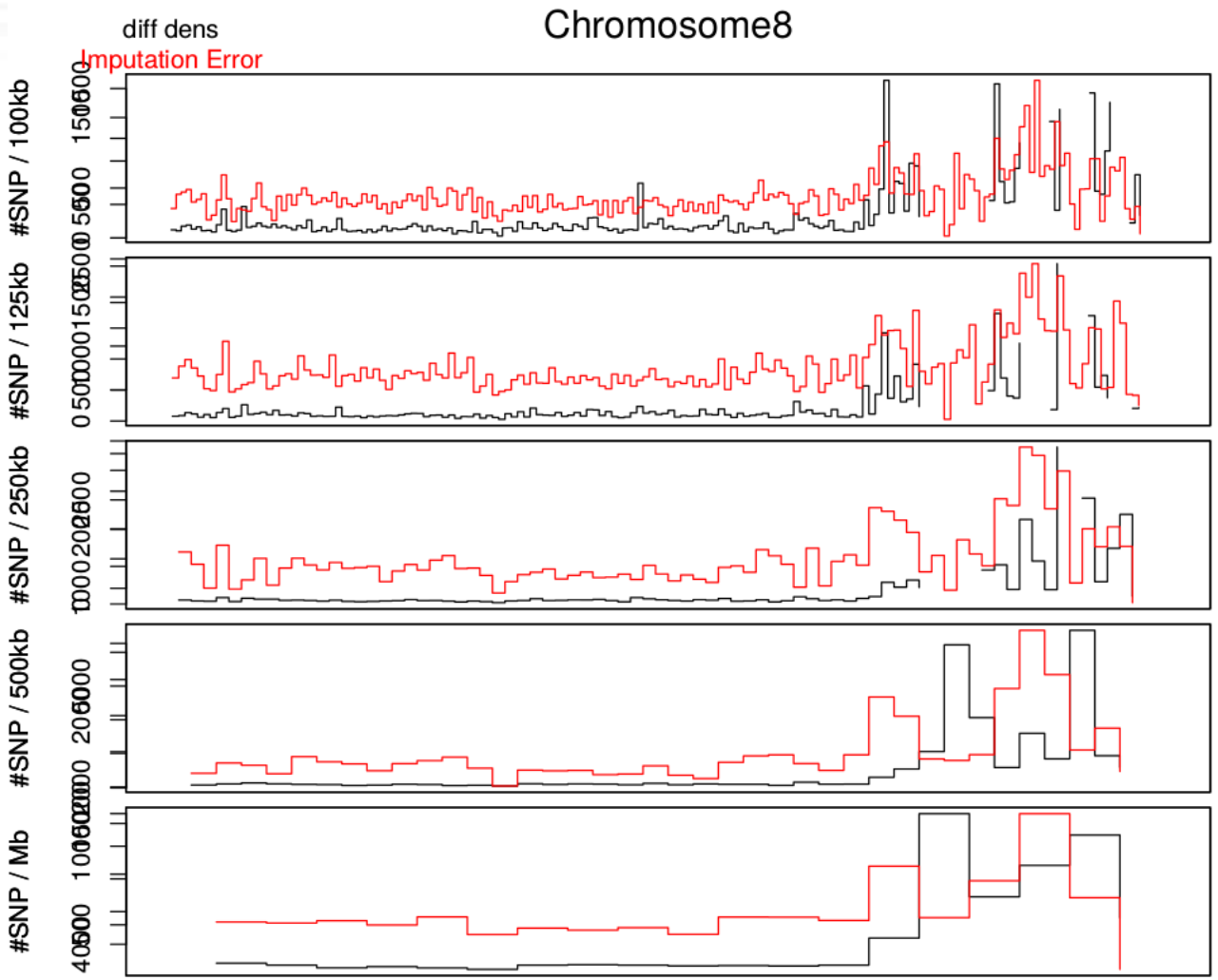
Results : Correlation between imputed genotypes and detected variants

	DV 662200037	DV 662200216
Imp with 1 FS	0.59	0.59
Imp with 0 FS	0.60	0.63



Hickey *et al.*, 2012 Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs

Results : Imputation errors



Conclusions & Propect

Conclusions

- This preliminary results are not accurate
  - Increase the number of cross-validation trying to impute each parent independently
  - Trying different densities to improve imputation accuracy
- Some adjustments are necessary

Prospect

- This pineplines will be used with an extend factorial mating design (25 sequences WGS)
- The impact of denser genotype information will be tested with predictions models
- Genotype imputation may allow to optimize genotyping efforts to decrease costs

Acknowledgment to

- Léopoldo Sanchez UR AGPF
- Véronique Jorge UR AGPF
- Catherine Bastien UR AGPF
- Marie-Christine Le Paslier US EPGV/CEA/CNG
- Patricia Faivre-Rampant US EPGV/CEA/CNG
- Odile Rogier UR AGPF
- Vincent Segura UR AGPF
- Facundo Munoz UR AGPF
- Support team EPGV/CEA/CNG

Thank you for your attention