Efficient identification of SNPs in high linkage disequilibrium in large genotype and sequence datasets

Mario Calus, **Jérémie Vandenplas**

EAAP Belfast; August 29, 2016

WAGENINGEN UR
For quality of life

BREED4FOOD

BREED4FOOD

Breed4Food is dedicated to be the leading research consortium in animal breeding, genetics and genomics enabling the Breed4Food partners to breed better products to benefit society's needs.

WAGENINGEN UR
For quality of life

HENDRIX GENETICS

Cobb

TOPIGS

CRV

Introduction

High density SNP & sequence data have:

- Many SNPs
- High co-linearity between loci due to high LD

=> A lot of redundant loci

Introduction

Reduced co-linearity by pruning SNPs based on LD:

- Ideally involves evaluating r^2 (LD) for all pairs of SNPs
- Which is too demanding for (very) large datasets
- So, r^2 are typically only evaluated for a sliding window

Objective

Develop an algorithm to prune for pairwise LD that does *not* require computation of all pairwise r^2 values

Algorithm for an r^2 threshold of 1

- r^2 can only be 1 when minor allele frequencies (MAF) of two loci are *the same*

- So, only r^2 values between pairs of SNPs with *the same MAF* need to be computed

Algorithm for an r^2 threshold of *close to* 1

- r^2 can only be close to 1 when minor allele frequencies (MAF) of two loci are *similar*
- So, only r^2 values between pairs of SNPs with *similar MAF* need to be computed

Software SNPrune

- Sorts loci based on MAF
- Implements the algorithms:
  - for an r^2 threshold of 1
  - for an r^2 threshold of *close to* 1
- Outputs list of removed SNPs & pruned data

- Input can be:
  - Allele counts (0,1,2)
  - Phased alleles (e.g. 0,1)

Data & analysis

Simulated sequence data:

- 10,812,225 segregating SNPs on 2500 individuals
- Phase assumed to be known

Prune data with $r^2 > 0.99$ using:

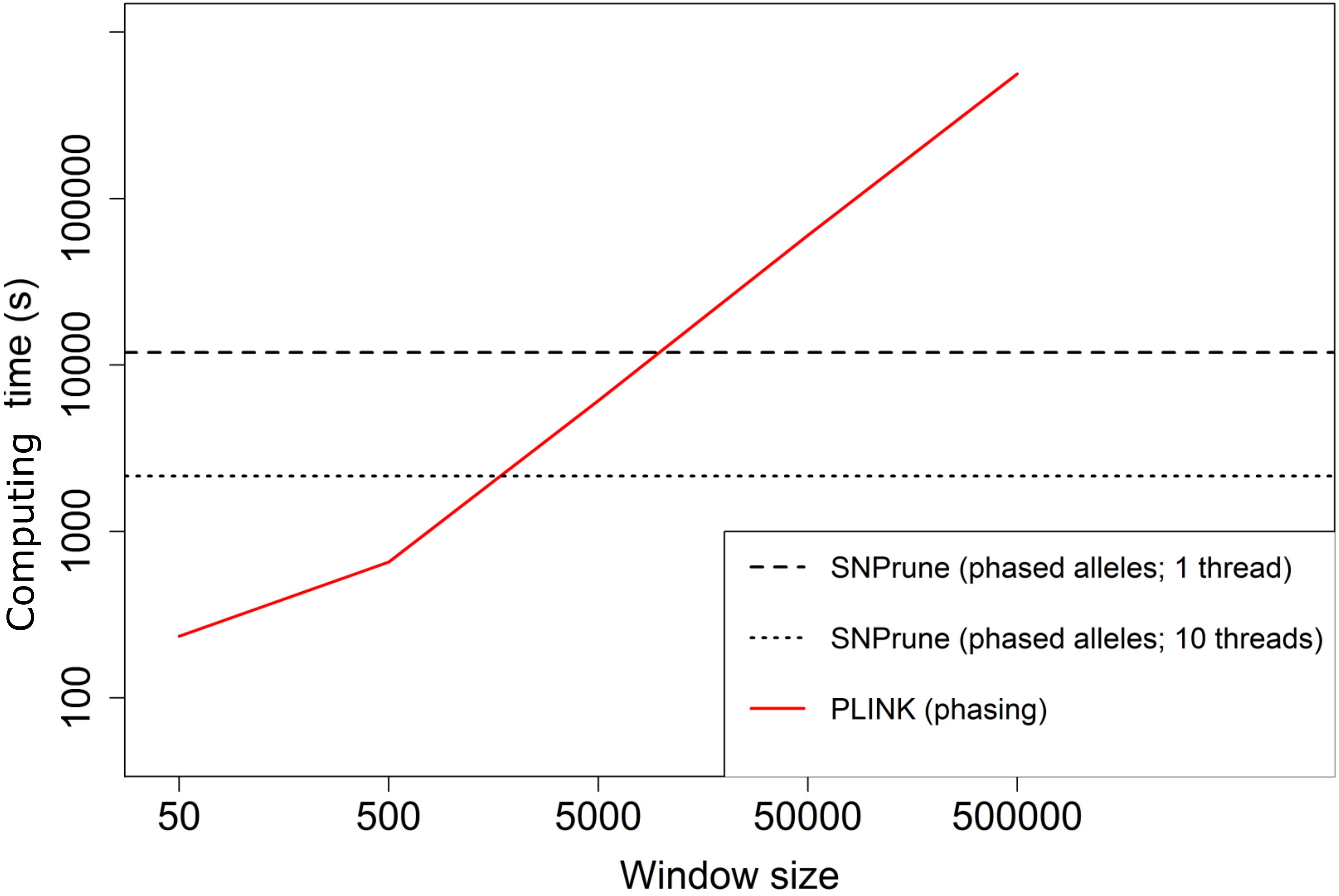
- SNPrune
- PLINK (v1.90 beta)
  - Different windows: 50-500,000 SNPs

Results – numbers of removed SNPs

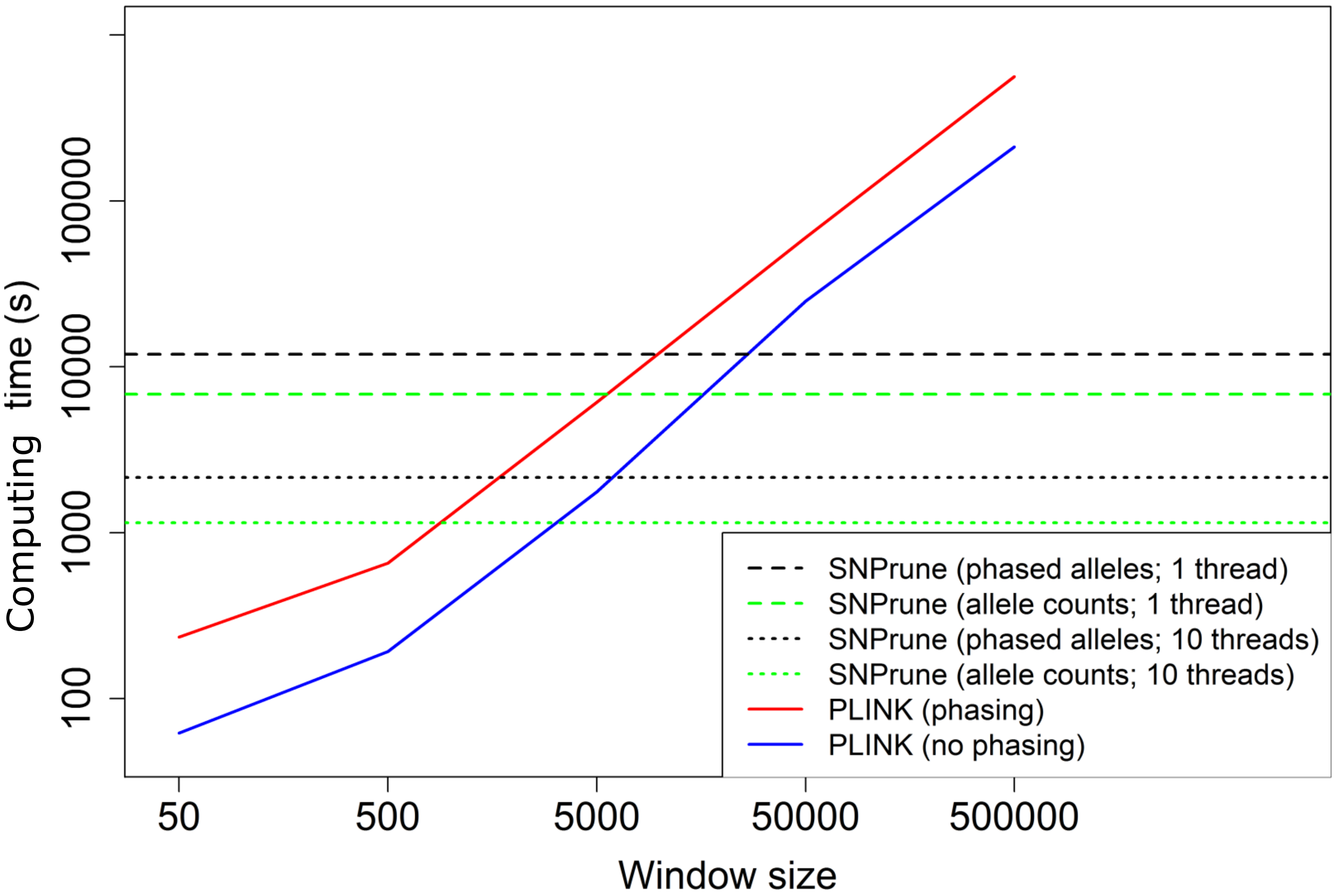
		#SNPs removed	
Software	Window size	Phased alleles	Allele counts
SNPrune	10 812 225	7 796 412	7 796 048
PLINK	500 000	7 752 485	7 752 008
PLINK	50 000	7 751 008	7 750 558
PLINK	5 000	7 741 279	7 740 937
PLINK	500	7 543 234	7 547 118
PLINK	50	5 401 527	5 401 197

- Large redundancy in sequence data
- Results are very similar with and without phasing
- SNPrune computed only 0.06% of all pairwise r^2 values

Results – computing time



Results – computing time



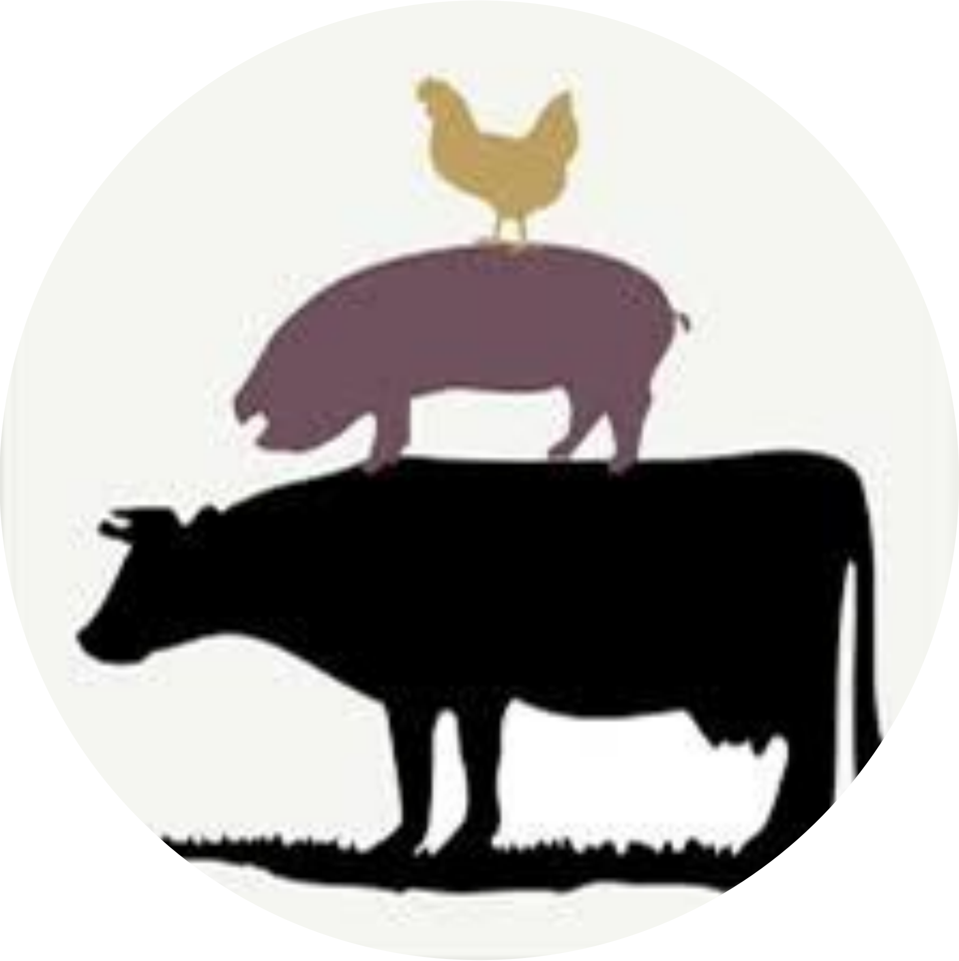
Conclusions

SNPrune is:

- Able to efficiently prune for LD across the genome
- By reducing the number of computed r^2 values
- Therefore feasible for (large) sequence datasets

Thank you!

BREED4FOOD



WAGENINGEN UR
For quality of life