

Introduction to **GWAS**

Imputation of Missing Genotypes

Christian Werner

(Computer biologist and quantitative geneticist) **EiB**, **CIMMYT**, Texcoco (Mexico)

Filippo Biscarini

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)



HerrFaloppio

Oscar González-Recio

(Computer biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)



OscarGenomics



Imputation of missing genotypes – **why?**

Imputation - the process of replacing missing data with substituted values

Preliminary step for a wide range of genetic analyses

Most models and software for population genetics, genomic selection (GS) and genome-wide association studies (GWAS) **do not handle missing data** by default and require complete datasets

1. Genotyping techniques generate a proportion of missing data (uncalled genotypes)

- SNP arrays ~5%
- RAD-Seq (e.g. GBS) ~50%

2. Optimization/efficiency of genotyping strategies (low → high density data)

scaling-up: **low** → **high density** (mixed genotyping strategies)

whole-genome sequence imputation

Imputation of missing genotypes – **methods**

1. **General methods for the imputation of any type of data**

- mean substitution (replacing missing values with the mean of the SNP across the population)
- K-Nearest Neighbour Imputation (KNNI)
- many more ...

2. **Methods specific for the imputation of missing genotypes**

Two groups (and combinations of them):

- based on LD and allele frequency
- based on pedigree information

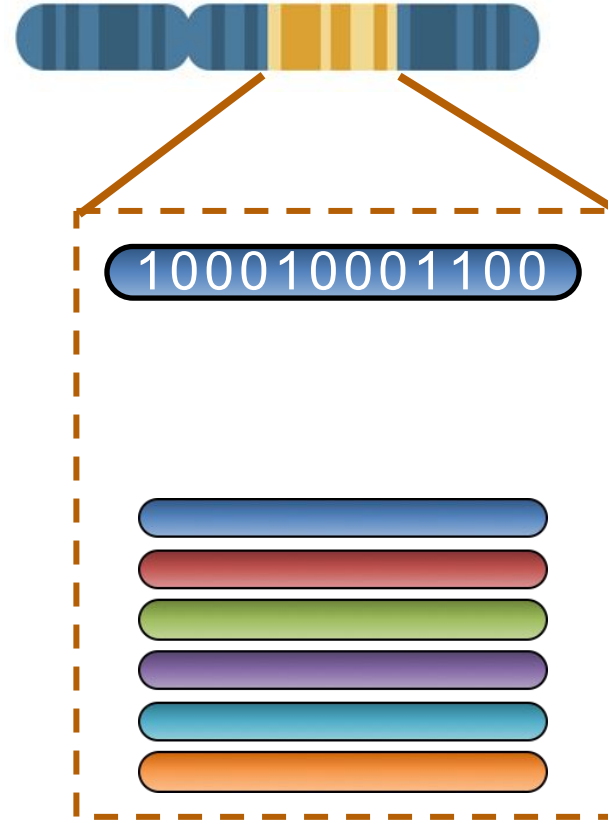
Haplotype libraries

Haplotype

A (section of) a single chromosome with known sequence (phased)

Haplotype Library

A collection of haplotypes



Allele dosage – prerequisite for imputation and regression

- Diploid genomes (or diploid-like meiotic behaviour)
- A single locus exhibits **four allelic combinations**
- Label $a=0$ and $A=1$

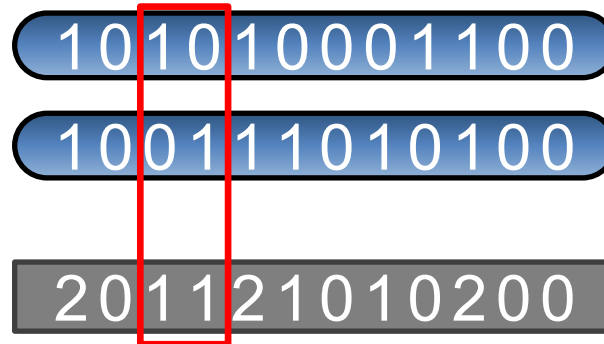
Thus the dosage is:

$$AA = 2$$

$$Aa = 1$$

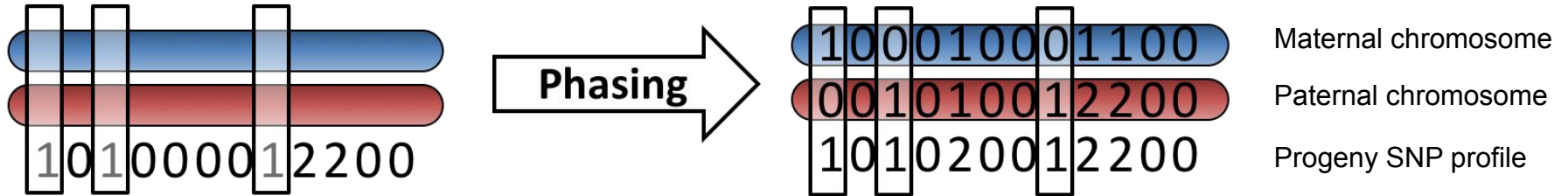
$$aA = 1$$

$$aa = 0$$



Haplotype phasing

- Phasing
 - Determining the haplotype of origin for heterozygotic loci



Inheritance of genotypes – Pedigree

Father

10100111011100111001110011
010101111100011000110011010
11110222111111111111121021

Mother

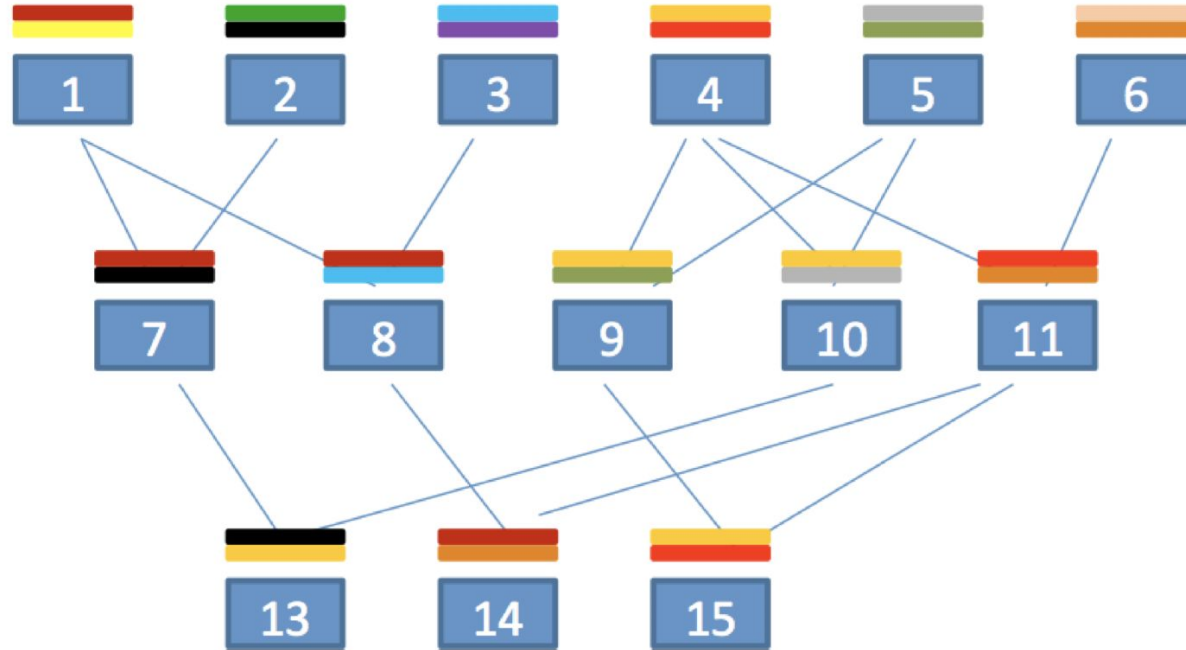
000100111110010101100110011
10101110101111111111111110
10111121211121212211221121



Progeny

10100111011100111001110011
000100111110010101100110011
10110122121110212101220022

Inheritance of genotypes – Pedigree

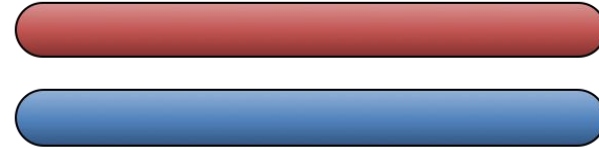


Imputing from sequenced parents using **pedigree** information

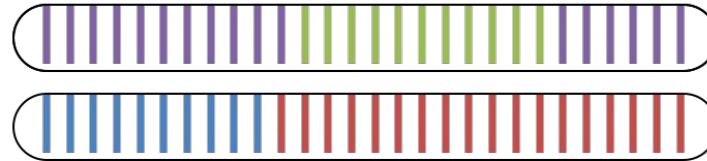
Father's chromosomes



Mother's chromosomes

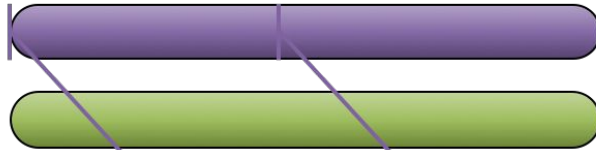


Child's chromosome

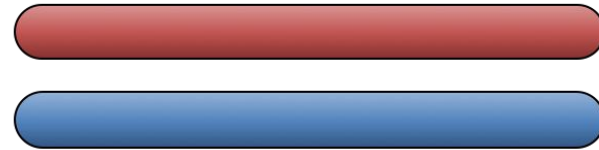


Imputing from sequenced parents using **pedigree** information

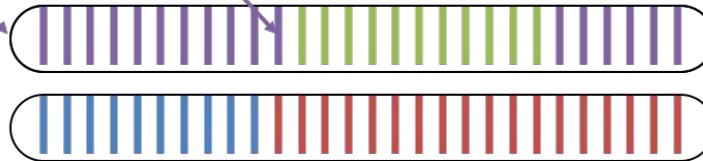
Father's chromosomes



Mother's chromosomes

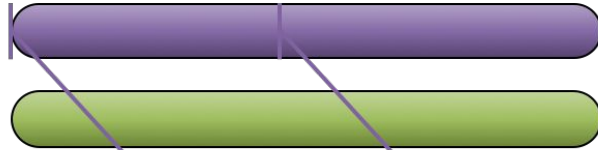


Child's chromosome

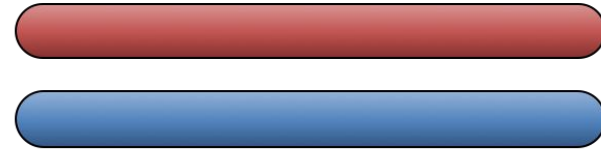


Imputing from sequenced parents using **pedigree** information

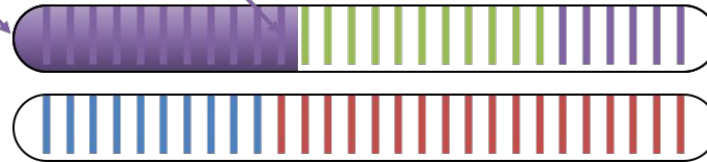
Father's chromosomes



Mother's chromosomes



Child's chromosome

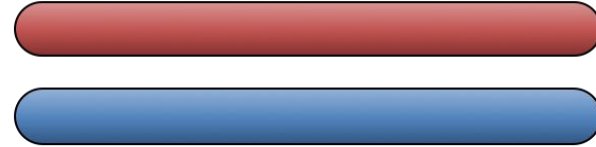


Imputing from sequenced parents using **pedigree** information

Father's chromosomes



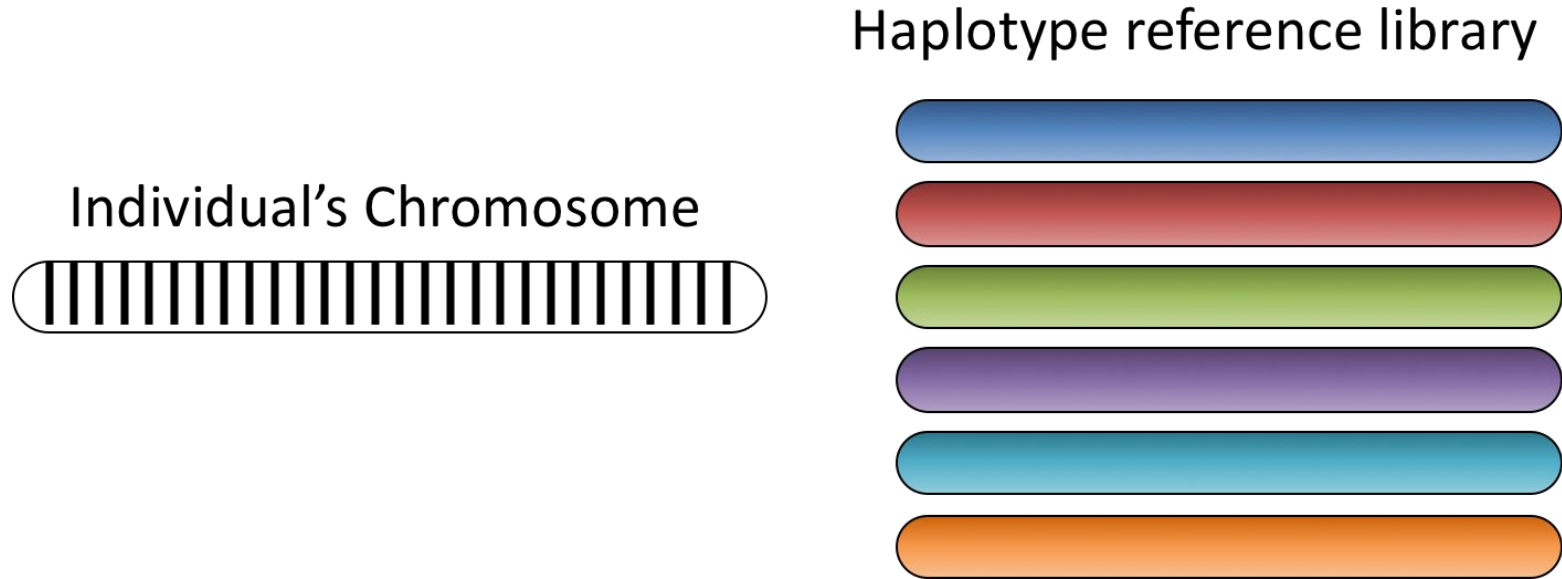
Mother's chromosomes



Child's chromosome



Imputing from sequenced parents using **haplotype libraries**

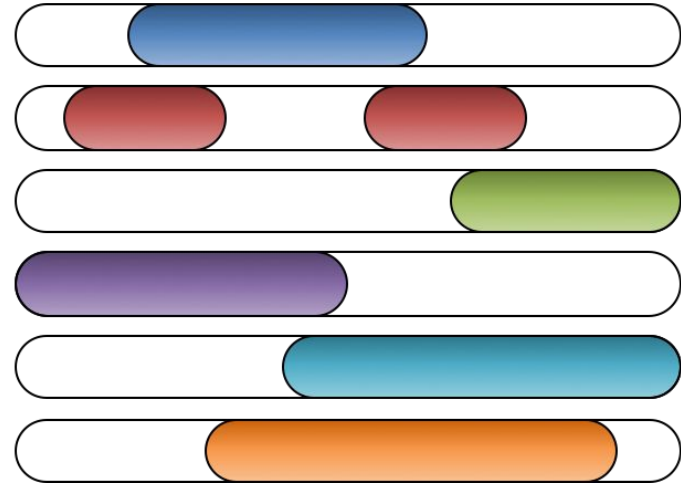


Imputing from sequenced parents using **haplotype libraries**

Individual's Chromosome



Haplotype reference library



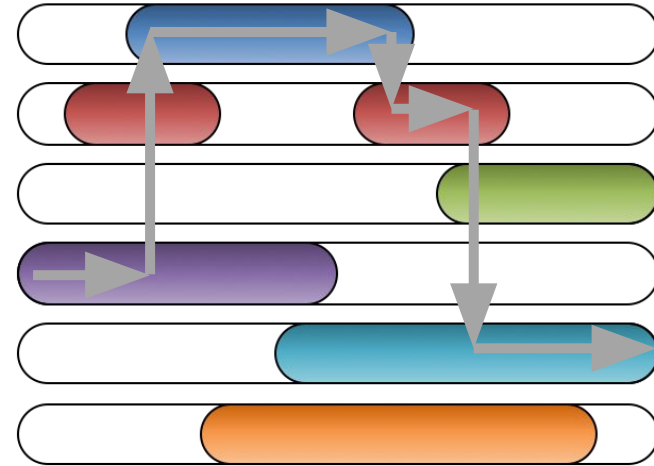
Imputing from sequenced parents using **haplotype libraries**

Individual's Chromosome



An individual's haplotype is a mosaic of haplotypes from a reference library.

Haplotype reference library



Imputation of missing genotypes

Pedigree imputation uses linkage

- Family statistic
- Correlation between adjacent markers within a family

Haplotype library imputation uses LD

- Population statistic
- Correlation between adjacent markers within a population

Imputation of missing genotypes - Which approach?

Beagle uses an LD-based approach (Hidden Markov Model; HMM) which in general does a good job using default settings.

- relatively user-friendly
- widely used in the literature
- also does phasing of your data

There are other HMM-based algorithms which show comparable imputation accuracies and computational efficiency. Some of them, however, might not phase your genotypes.

- In this case you need another software to perform phasing before imputation.

A stand-alone pedigree imputation approach should not be used as it is less accurate than algorithms using an HMM.

- However, some algorithms combine pedigree information and an HMM. This might increase accuracy and / or computational efficiency.

Imputation with BEAGLE

Localised haplotype clustering imputation – **LHCI**

- **popular method** for the imputation of missing genotypes
- developed originally for **humans**, has since found wide application also in animals and plants
- makes use solely of **genomic information** (LD, allele frequency etc.) - no pedigree!
- **haplotypes** are inferred (reconstructed), their frequency estimated, and are **clustered** “locally”

Detailed introduction how BEAGLE works

<https://www.youtube.com/watch?v=-oUvXXg6tl8>

Localised haplotype clustering imputation – LHCI

- Hidden Markov Model (HMM)
- Find the most likely haplotype pair for each individual given the genotype data for that individual and the haplotype frequency model
- genotypes are then **imputed** based on probabilities from the last fitted model (iterative algorithm)
- **LHCI** is implemented in the software “**BEAGLE**” (Browning and Browning 2007: <https://faculty.washington.edu/browning/beagle>)
- LHCI is the method
- Beagle is the software that implements it

K-nearest neighbour Imputation (KNNI)

K-nearest neighbor imputation – **KNNI**

- general imputation method
- applicable to any type of data (including genotypes)
- **similarity matrix** between samples from a **distance function** based on available data
e.g. Euclidean distances or Hamming distances based on SNP genotypes

Genotype imputation – **measuring accuracy**

Imputation accuracy of **all genotype classes** (total, AA, AB, BB)

Why is this important?

- Data are usually **unbalanced** (major/minor alleles)
- Rare allele (1%) → a **naive classifier that always predicts the major allele** would be correct 99 times out of 100

99% accuracy overall

100% accuracy for the major allele

but **0%** accuracy in the minor allele!



Genotype imputation – **measuring accuracy**

Imputation accuracy of **all genotype classes** (total, AA, AB, BB)

Why is this important?

- Data are usually **unbalanced** (major/minor alleles)
- Rare allele (1%) → a **naive classifier that always predicts the major allele** would be correct 99 times out of 100

Key message

Check the accuracy in the different genotype classes, not the total accuracy

