Introduction to GWAS

Imputation of Missing Genotypes

Christian Werner

(Quantitative geneticist and biostatistician) **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini HerrFalloppio

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

Oscar González-Recio OscarGenomics

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

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), median imputation
   - K-Nearest Neighbour Imputation (KNNI)
   - many more ...

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

   Two groups (and combinations of them):

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

Imputation of missing genotypes

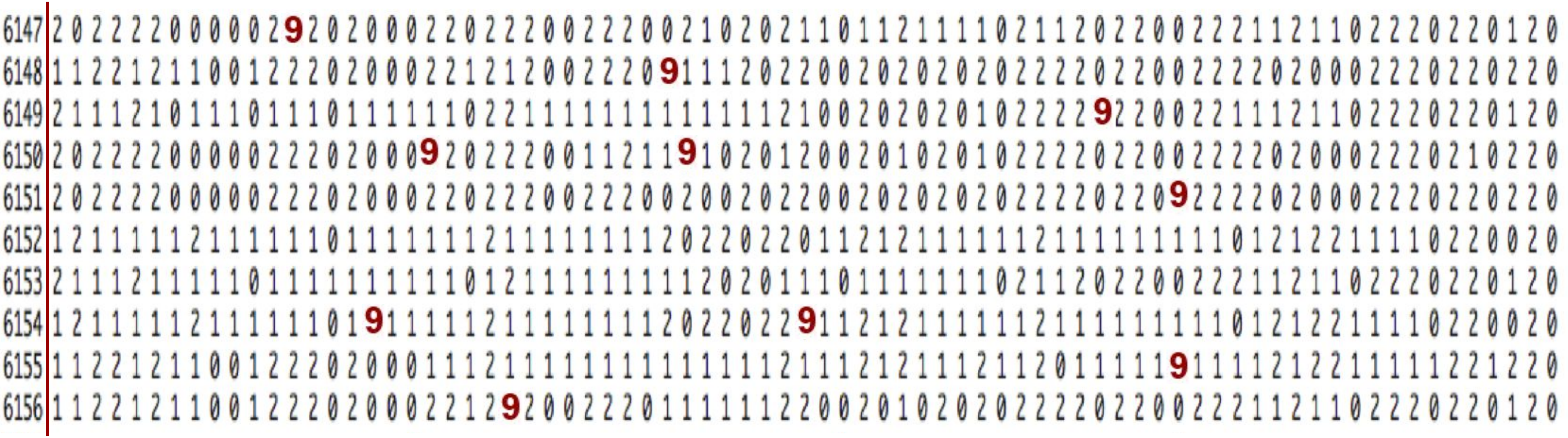
Pedigree imputation uses linkage

- Family statistic
- Correlation between adjacent markers within a family
- Fast and simple, but limitations when inheritance is unclear

Haplotype library imputation uses LD

- Population statistic
- Correlation between adjacent markers within a population
- Very powerful, but computationally demanding

Imputation of missing genotypes – filling the gaps



```
6147|2 0 2 2 2 2 0 0 0 0 0 2 9 2 0 2 0 0 0 2 2 0 2 2 2 0 0 2 2 2 0 0 2 1 0 2 0 2 1 1 0 1 1 2 1 1 1 1 0 2 1 1 2 0 2 2 0 0 2 2 2 1 1 2 1 1 0 2 2 2 0 2 2 0 1 2 0
6148|1 1 2 2 1 2 1 1 0 0 1 2 2 2 0 2 0 0 0 2 2 1 2 1 2 0 0 2 2 2 0 9 1 1 1 2 0 2 2 0 0 2 0 2 0 2 0 2 0 2 2 2 2 0 2 2 0 0 2 2 2 2 0 2 0 0 0 2 2 2 0 2 2 0 2 2 0
6149|2 1 1 1 2 1 0 1 1 1 0 1 1 1 0 1 1 1 1 1 1 0 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 0 0 2 0 2 0 2 0 1 0 2 2 2 2 9 2 2 0 0 2 2 1 1 1 2 1 1 0 2 2 2 0 2 2 0 1 2 0
6150|2 0 2 2 2 2 0 0 0 0 0 2 2 2 0 2 0 0 0 9 2 0 2 2 2 0 0 1 1 2 1 1 9 1 0 2 0 1 2 0 0 2 0 1 0 2 0 1 0 2 2 2 2 0 2 2 0 0 2 2 2 2 0 2 0 0 0 2 2 2 0 2 1 0 2 2 0
6151|2 0 2 2 2 2 0 0 0 0 0 2 2 2 0 2 0 0 0 2 2 0 2 2 2 0 0 2 2 2 0 0 2 0 0 2 0 2 2 0 0 2 0 2 0 2 0 2 0 2 2 2 2 0 2 2 0 9 2 2 2 2 0 2 0 0 0 2 2 2 0 2 2 0 2 2 0
6152|1 2 1 1 1 1 1 2 1 1 1 1 1 1 0 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 0 2 2 0 2 2 0 1 1 2 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 0 1 2 1 2 2 1 1 1 1 0 2 2 0 0 2 0
6153|2 1 1 1 2 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 2 1 1 1 1 1 1 1 1 1 2 0 2 0 1 1 1 0 1 1 1 1 1 1 1 0 2 1 1 2 0 2 2 0 0 2 2 2 1 1 2 1 1 0 2 2 2 0 2 2 0 1 2 0
6154|1 2 1 1 1 1 1 2 1 1 1 1 1 1 0 1 9 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 0 2 2 0 2 2 9 1 1 2 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 0 1 2 1 2 2 1 1 1 1 0 2 2 0 0 2 0
6155|1 1 2 2 1 2 1 1 0 0 1 2 2 2 0 2 0 0 0 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 2 1 1 1 2 1 1 2 0 1 1 1 1 1 9 1 1 1 1 2 1 2 2 1 1 1 1 1 2 2 1 2 2 0
6156|1 1 2 2 1 2 1 1 0 0 1 2 2 2 0 2 0 0 0 2 2 1 2 9 2 0 0 2 2 2 0 1 1 1 1 1 1 2 2 0 0 2 0 1 0 2 0 2 0 2 2 2 2 0 2 2 0 0 2 2 2 1 1 2 1 1 0 2 2 2 0 2 2 0 1 2 0
```

9 = NA

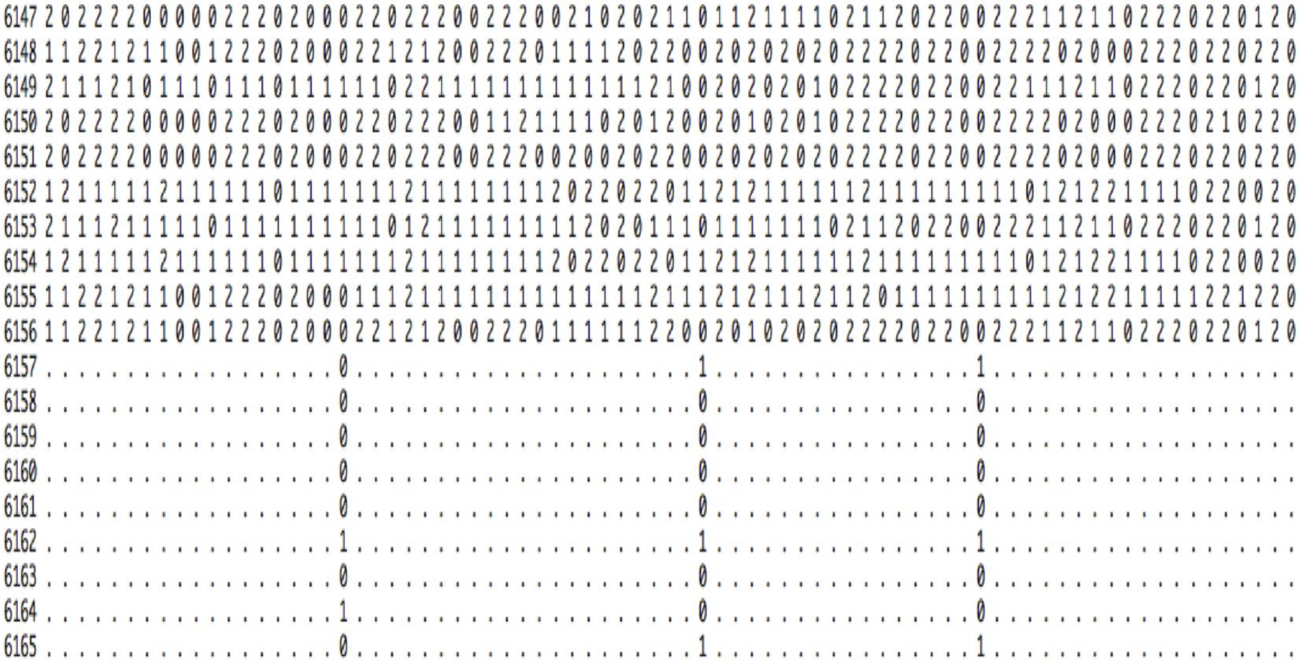
Imputation of missing genotypes – filling the gaps

Sequenced genotypes

- full profile
- no missing data

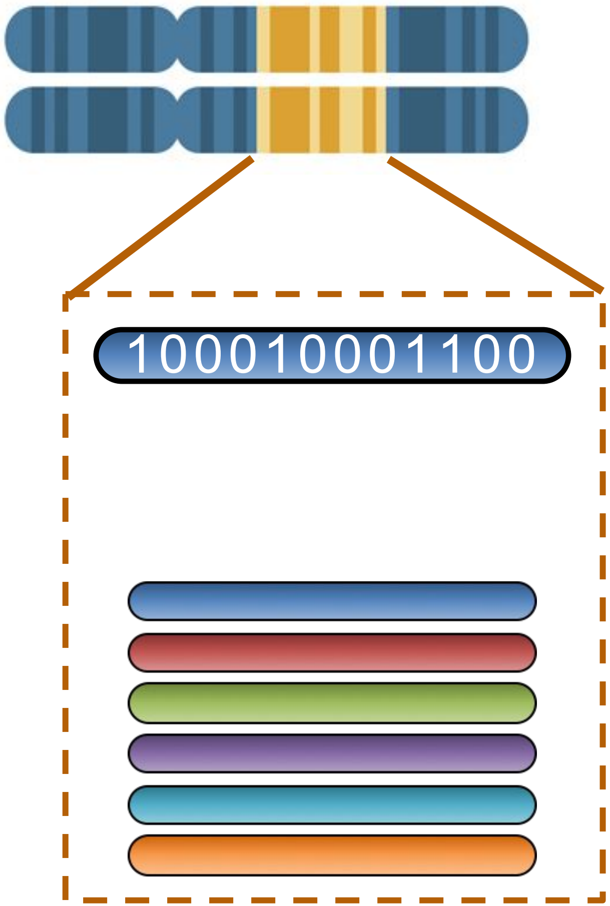
SNP genotypes / skim sequencing

- landmarks along the genome
- empty slots to be filled in



```
6147 2 0 2 2 2 2 0 0 0 0 0 2 2 2 0 2 0 0 0 2 2 0 2 2 2 0 0 2 2 2 0 0 2 1 0 2 0 2 1 1 0 1 1 2 1 1 1 1 0 2 1 1 2 0 2 2 0 0 2 2 2 1 1 2 1 1 0 2 2 2 0 2 2 0 1 2 0
6148 1 1 2 2 1 2 1 1 0 0 1 2 2 2 0 2 0 0 0 2 2 1 2 1 2 0 0 2 2 2 0 1 1 1 1 2 0 2 2 0 0 2 0 2 0 2 0 2 0 2 2 2 2 0 2 2 0 0 2 2 2 2 0 2 0 0 0 2 2 2 0 2 2 0 2 2 0
6149 2 1 1 1 2 1 0 1 1 1 0 1 1 1 0 1 1 1 1 1 1 0 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 0 0 2 0 2 0 2 0 1 0 2 2 2 2 0 2 2 0 0 2 2 1 1 1 2 1 1 0 2 2 2 0 2 2 0 1 2 0
6150 2 0 2 2 2 2 0 0 0 0 0 2 2 2 0 2 0 0 0 2 2 0 2 2 2 0 0 1 1 2 1 1 1 1 0 2 0 1 2 0 0 2 0 1 0 2 0 1 0 2 2 2 2 0 2 2 0 0 2 2 2 2 0 2 0 0 0 2 2 2 0 2 1 0 2 2 0
6151 2 0 2 2 2 2 0 0 0 0 0 2 2 2 0 2 0 0 0 2 2 0 2 2 2 0 0 2 2 2 0 0 2 0 0 2 0 2 2 0 0 2 0 2 0 2 0 2 0 2 2 2 2 0 2 2 0 0 2 2 2 2 0 2 0 0 0 2 2 2 0 2 2 0 2 2 0
6152 1 2 1 1 1 1 1 2 1 1 1 1 1 1 0 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 0 2 2 0 2 2 0 1 1 2 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 0 1 2 1 2 2 1 1 1 1 0 2 2 0 0 2 0
6153 2 1 1 1 2 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 2 1 1 1 1 1 1 1 1 1 2 0 2 0 1 1 1 0 1 1 1 1 1 1 1 0 2 1 1 2 0 2 2 0 0 2 2 2 1 1 2 1 1 0 2 2 2 0 2 2 0 1 2 0
6154 1 2 1 1 1 1 1 2 1 1 1 1 1 1 0 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 0 2 2 0 2 2 0 1 1 2 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 0 1 2 1 2 2 1 1 1 1 0 2 2 0 0 2 0
6155 1 1 2 2 1 2 1 1 0 0 1 2 2 2 0 2 0 0 0 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 2 1 1 1 2 1 1 2 0 1 1 1 1 1 1 1 1 1 1 2 1 2 2 1 1 1 1 1 2 2 1 2 2 0
6156 1 1 2 2 1 2 1 1 0 0 1 2 2 2 0 2 0 0 0 2 2 1 2 1 2 0 0 2 2 2 0 1 1 1 1 1 1 2 2 0 0 2 0 1 0 2 0 2 0 2 2 2 2 0 2 2 0 0 2 2 2 1 1 2 1 1 0 2 2 2 0 2 2 0 1 2 0
6157 . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . .
6158 . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . .
6159 . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . .
6160 . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . .
6161 . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . .
6162 . . . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . .
6163 . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . .
6164 . . . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . .
6165 . . . . . . . . . . . . . . . . . . 0 . . . . . . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . .
```

Haplotypes



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 can exhibit **four allelic combinations**
- Label a=0 and A=1

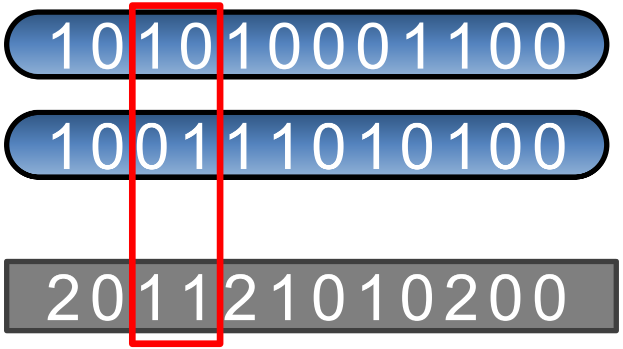
Thus the dosage is:

AA = 2

Aa = 1

aA = 1

aa = 0



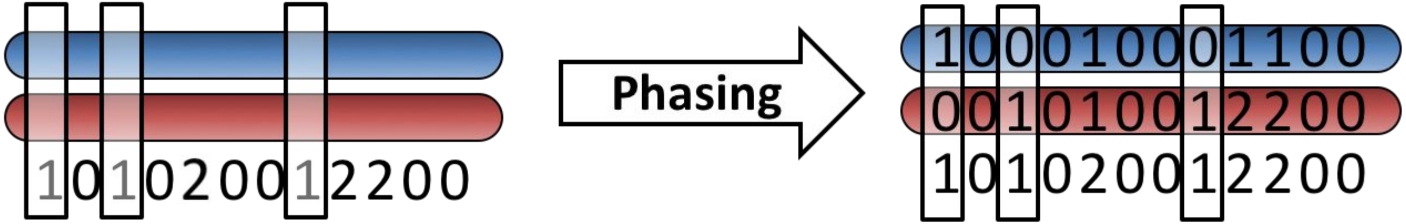
1 0 1 0 1 0 0 0 1 1 0 0 Maternal chromosome

1 0 0 1 1 1 0 1 0 1 0 0 Paternal chromosome

2 0 1 1 2 1 0 1 0 2 0 0 Progeny SNP profile

Haplotype phasing

- Phasing
  - Determining the haplotype of origin for heterozygotic loci



1010200122 00

Phasing

100010001100 Maternal chromosome

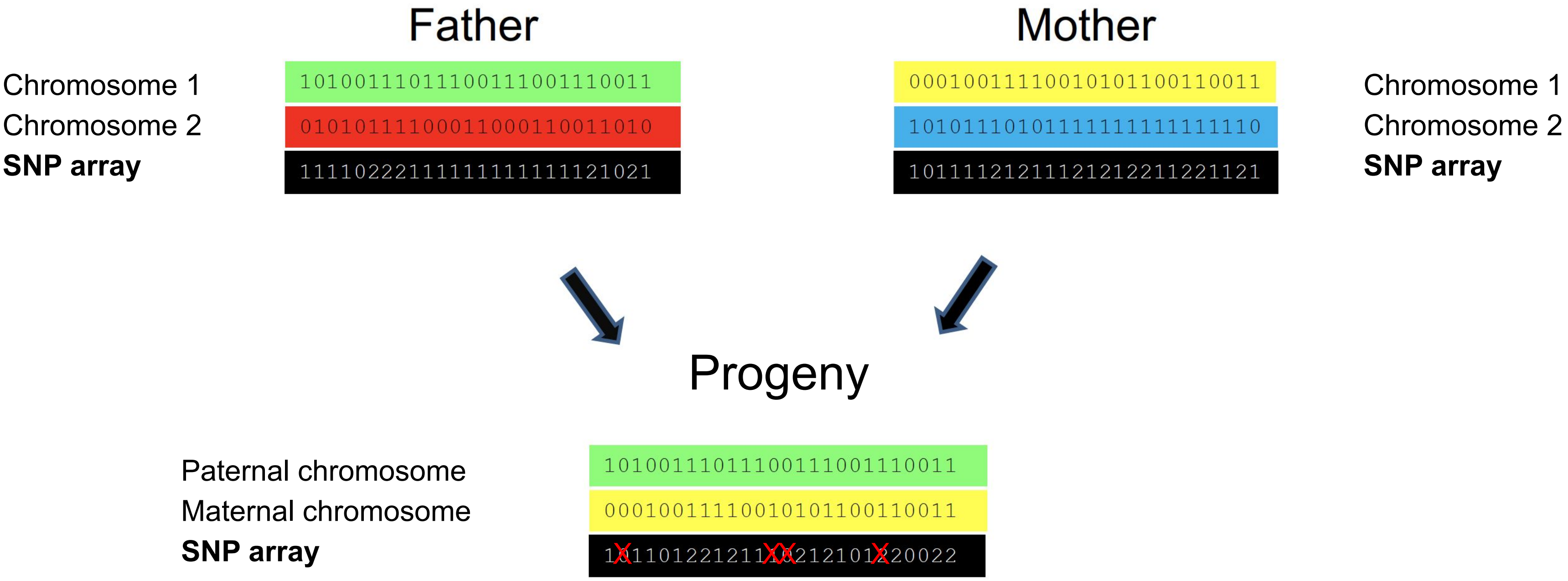
001010012200 Paternal chromosome

101020012200 Progeny SNP profile

Imputation using pedigree information

(expected inheritance patterns)

Inheritance of genotypes – filling NAs



Father

Chromosome 1 10100111011100111001110011

Chromosome 2 01010111100011000110011010

SNP array 11110222111111111111121021

Mother

00010011110010101100110011 Chromosome 1

10101110101111111111111110 Chromosome 2

10111121211121212211221121 **SNP array**

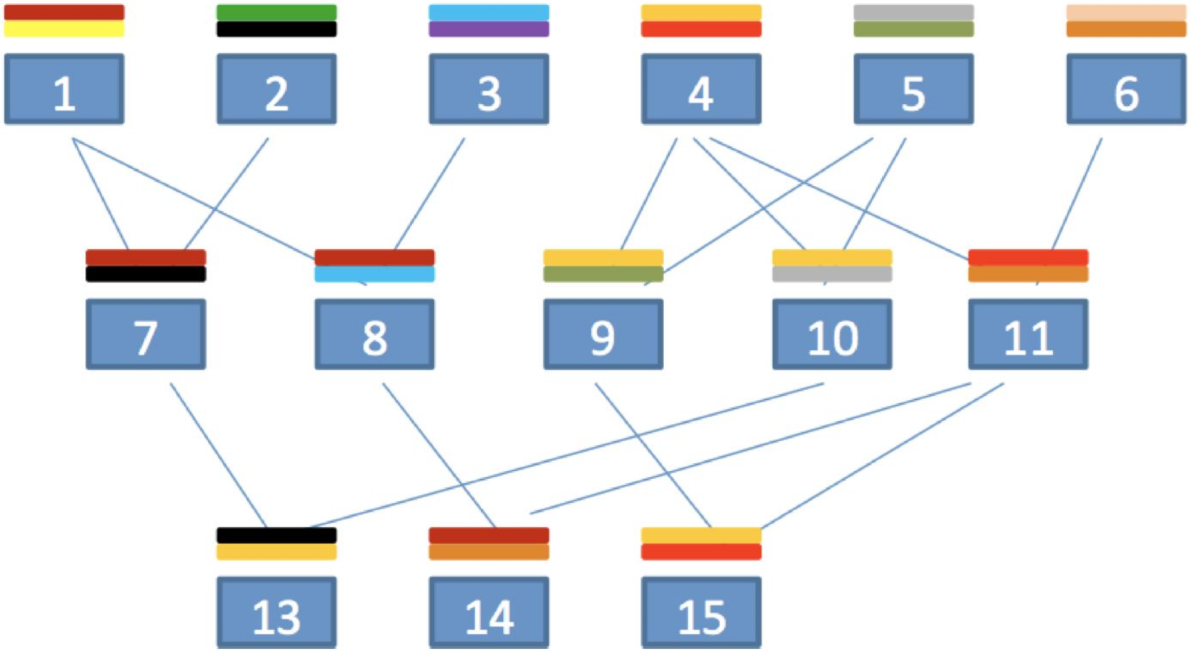
Progeny

Paternal chromosome 10100111011100111001110011

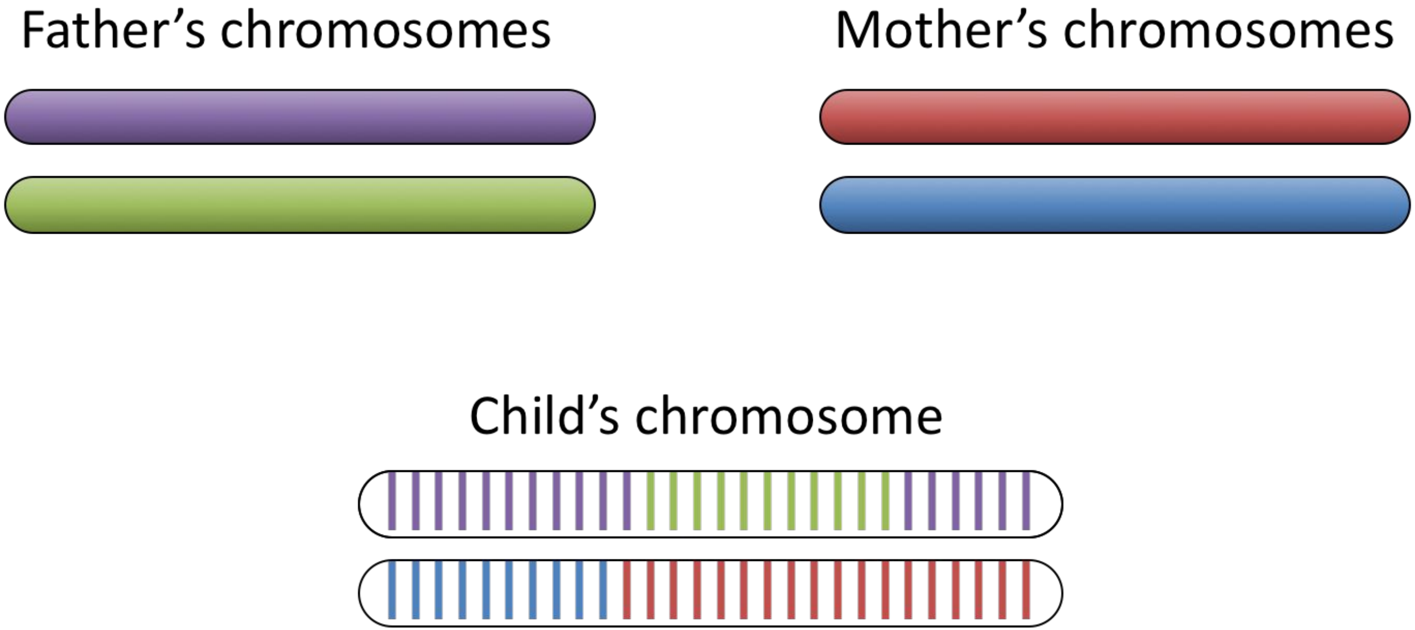
Maternal chromosome 00010011110010101100110011

SNP array 1~~0~~1101221211~~10~~212101~~2~~20022

Inheritance of genotypes – Pedigree



Imputation of sequence data into low-density genotypes - scaling up

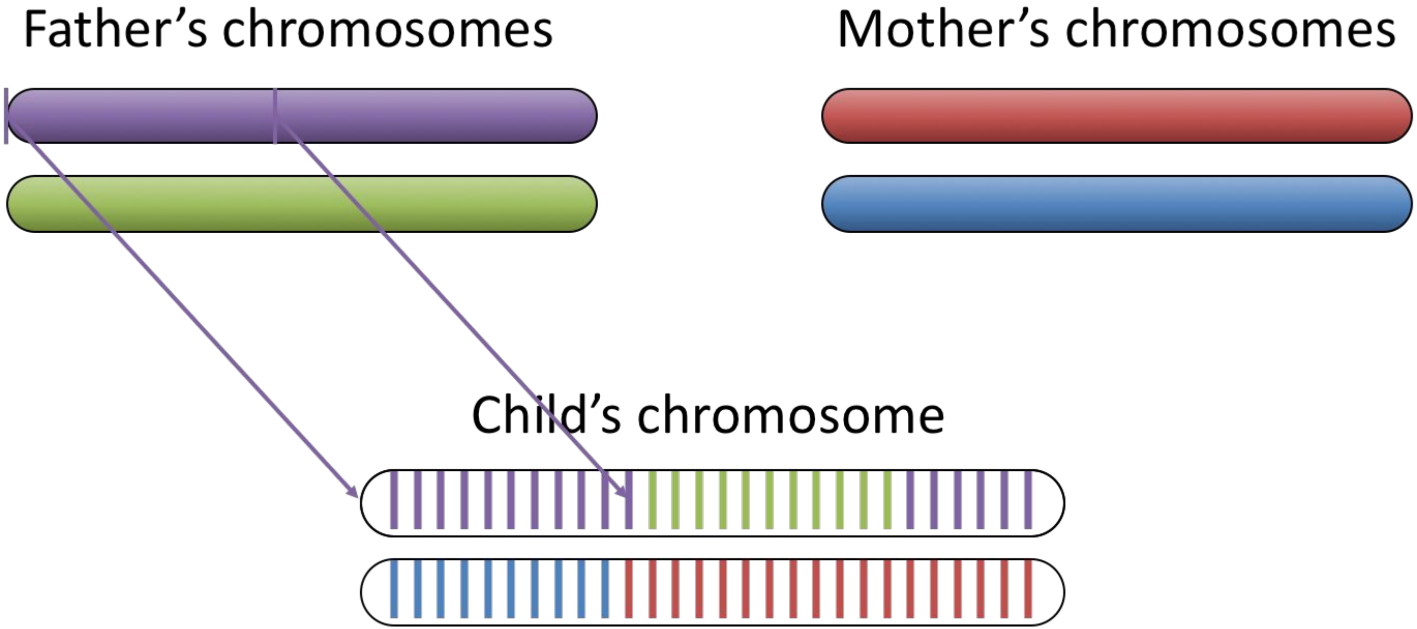


Father’s chromosomes

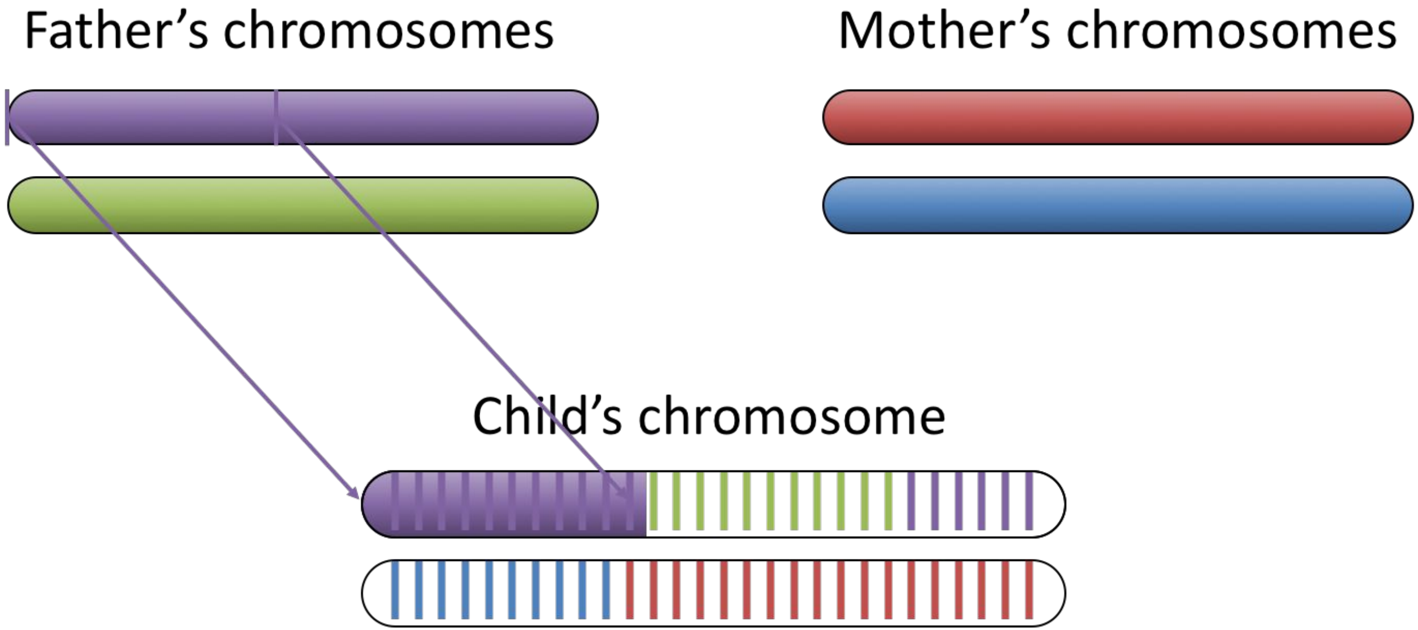
Mother’s chromosomes

Child’s chromosome

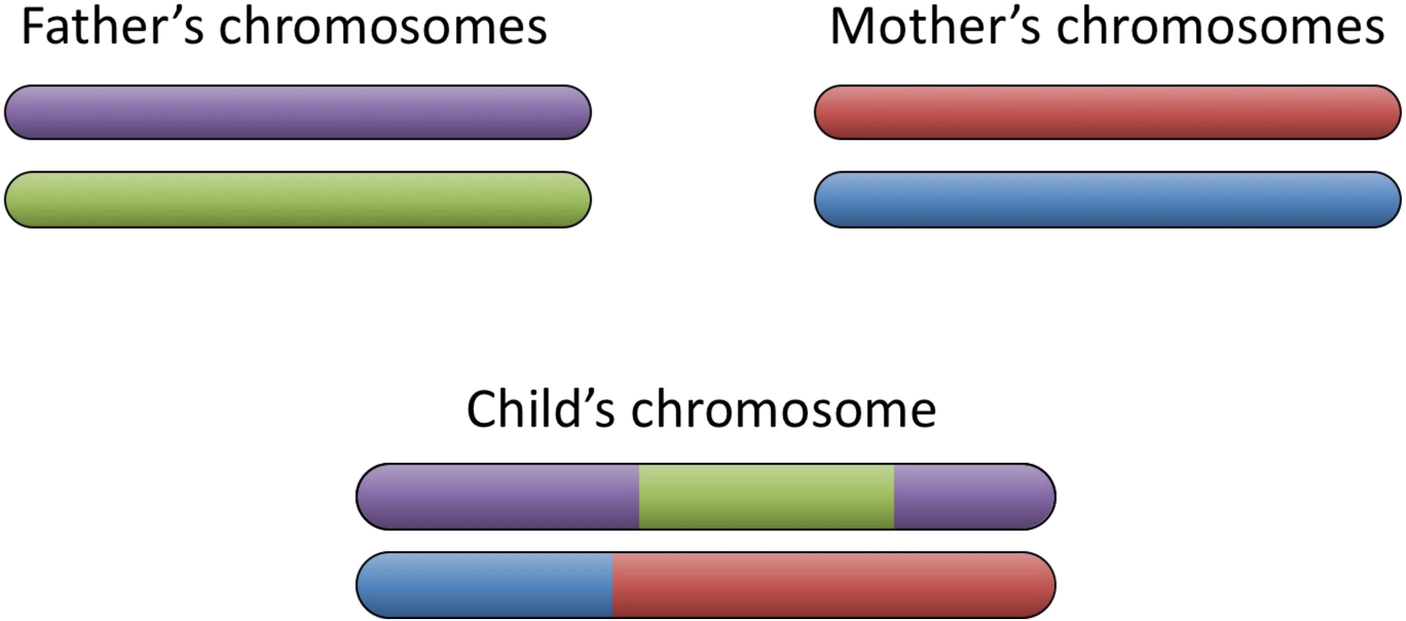
Imputation of sequence data into low-density genotypes - scaling up



Imputation of sequence data into low-density genotypes - scaling up



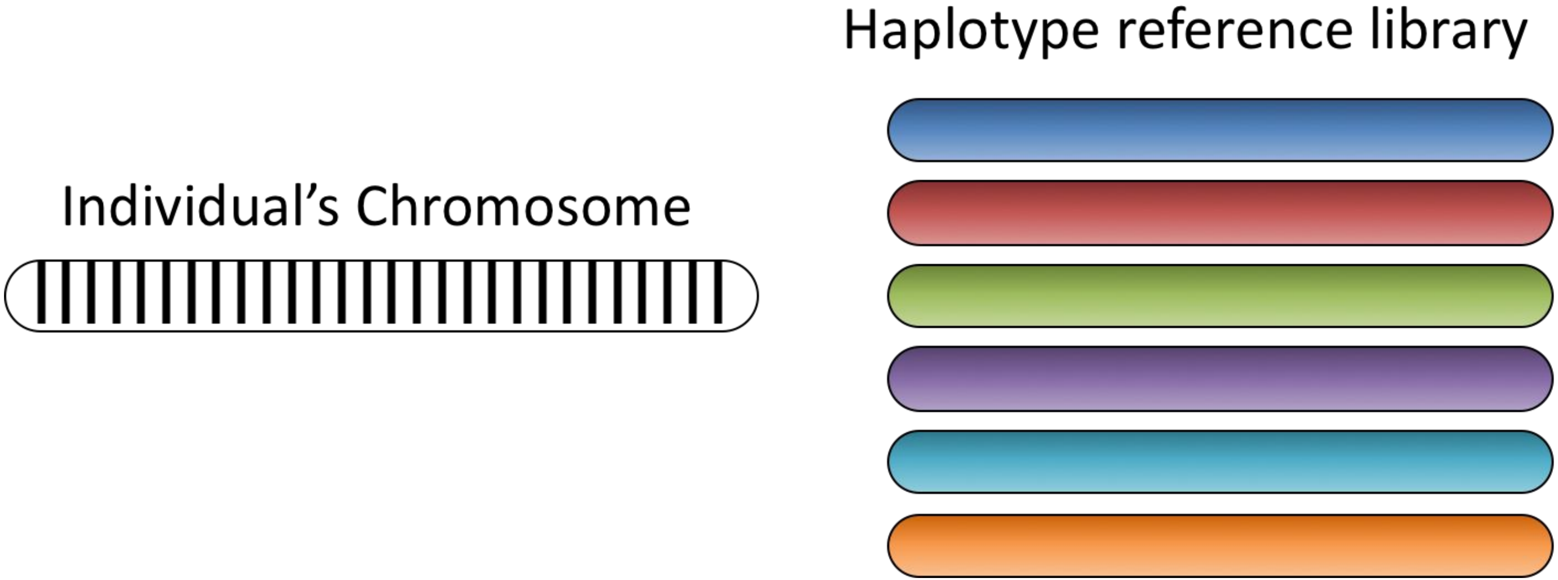
Imputation of sequence data into low-density genotypes - scaling up



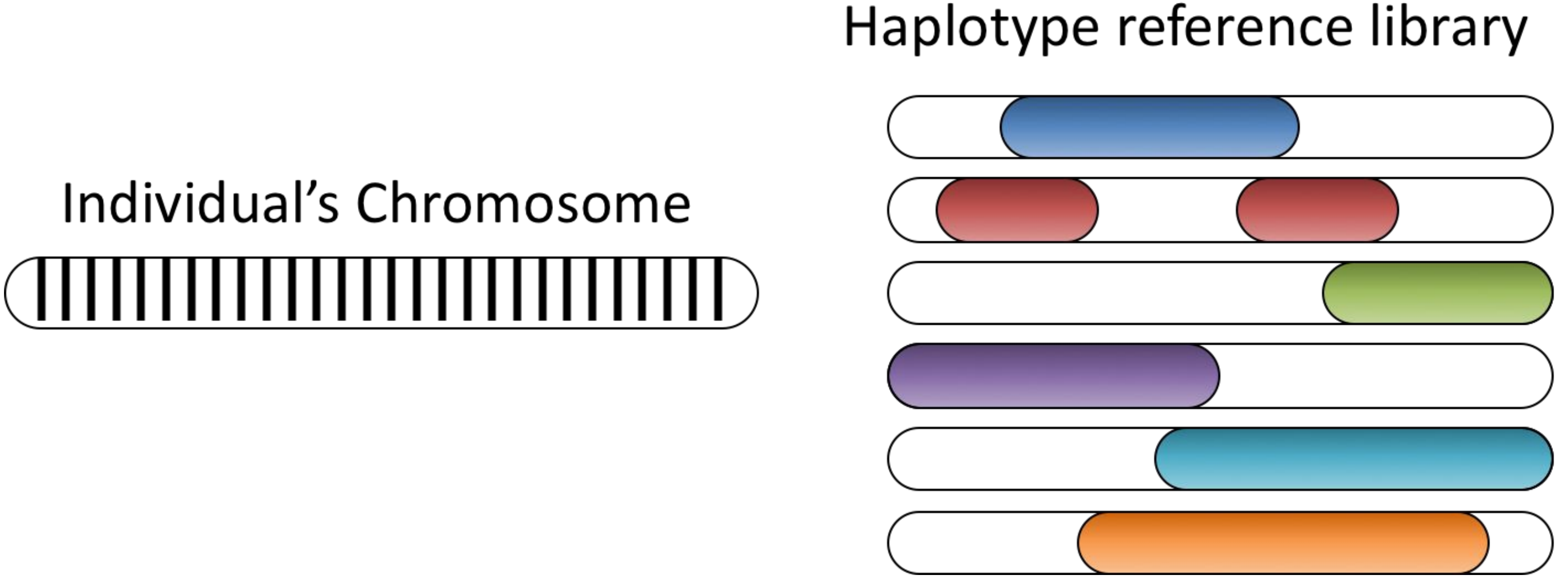
Imputation based on LD

(haplotype patterns)

Imputing from sequenced parents using haplotype libraries

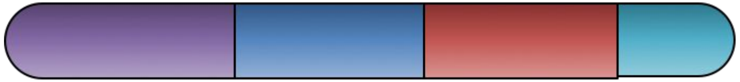


Imputing from sequenced parents using haplotype libraries



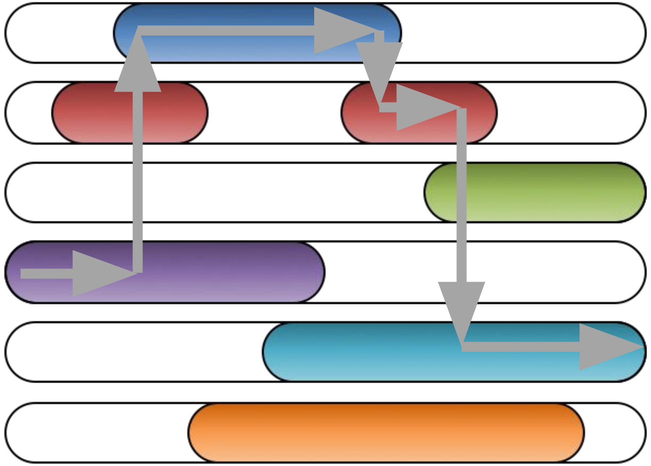
Imputing from sequenced parents using haplotype libraries

Individual’s Chromosome



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

Haplotype reference library



Imputation of missing genotypes - 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 in diploid genomes
- 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

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