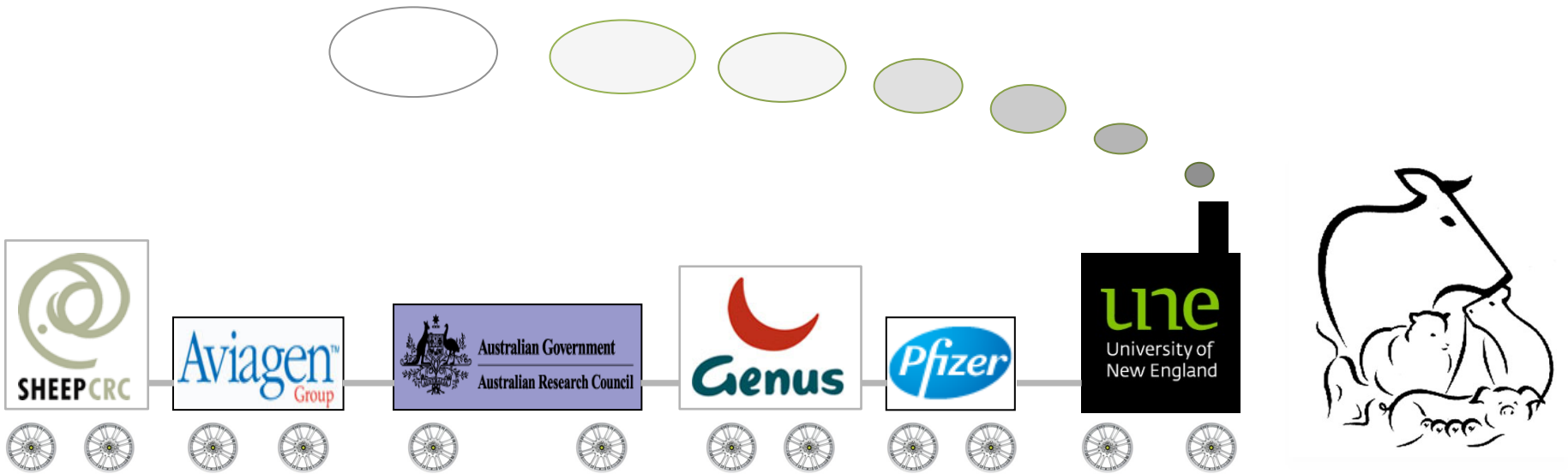
Genotype imputation

In populations that are not inbred!!

J.M. Hickey

The Roslin Institute



Genotype imputation

- Why we do it
- How we do it
- How it performs
- By way of background
  - AlphaImpute is an imputation software package that is used in the commercial pigs and poultry sector
  - It has been used to impute genotypes for more than 150,000 individuals (each breeding program adds 2,000 individuals each week)
  - We needed to make it work for $20 per individual

Why we do it

- To empower GS and GWAS at low cost

- Response to selection of GS affected by the number of
  - Genotyped animals in the training set
    - e.g. 20,000 individuals
    - Accuracy of selection

  - Genotyped candidates in the selection set
    - e.g. 100,000 individuals per year
    - Intensity and accuracy of selection

- Costs
  - 50,000 markers = $120
  - DNA extraction = $3
  - 50 markers = $2

Phasing/Imputation algorithms

- Two groups of imputation algorithms
  - Hidden markov based models
  - Heuristic methods

- Hidden markov based models
  - Probabilistic / pedigree free
  - Model linkage disequilibrium/short haplotypes
  - e.g. fastPHASE, Beagle, Impute2, Shape-IT, MaCH, minimach

- Heuristic methods
  - Tend not to be probabilistic / use pedigree information
  - Model linkage/long haplotypes
  - e.g. AlphaPhase/AlphaImpute (Long-range phasing and haplotype libraries), fimpute, findhap

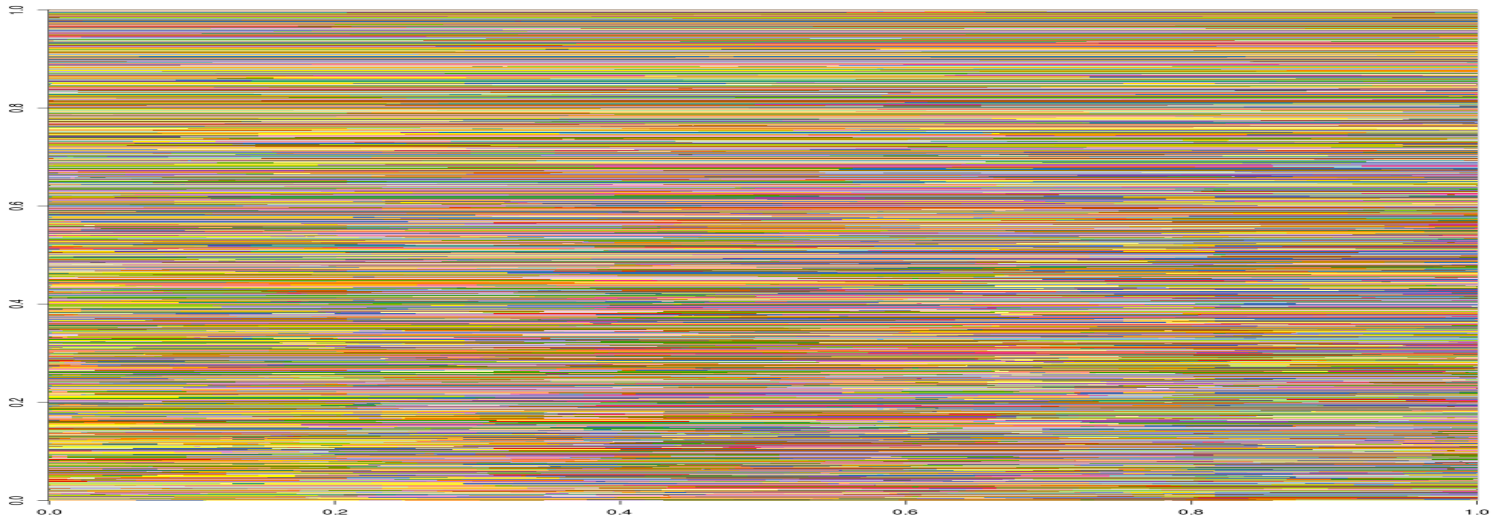
- Combined methods
  - Phasebook
  - AlphaPhase

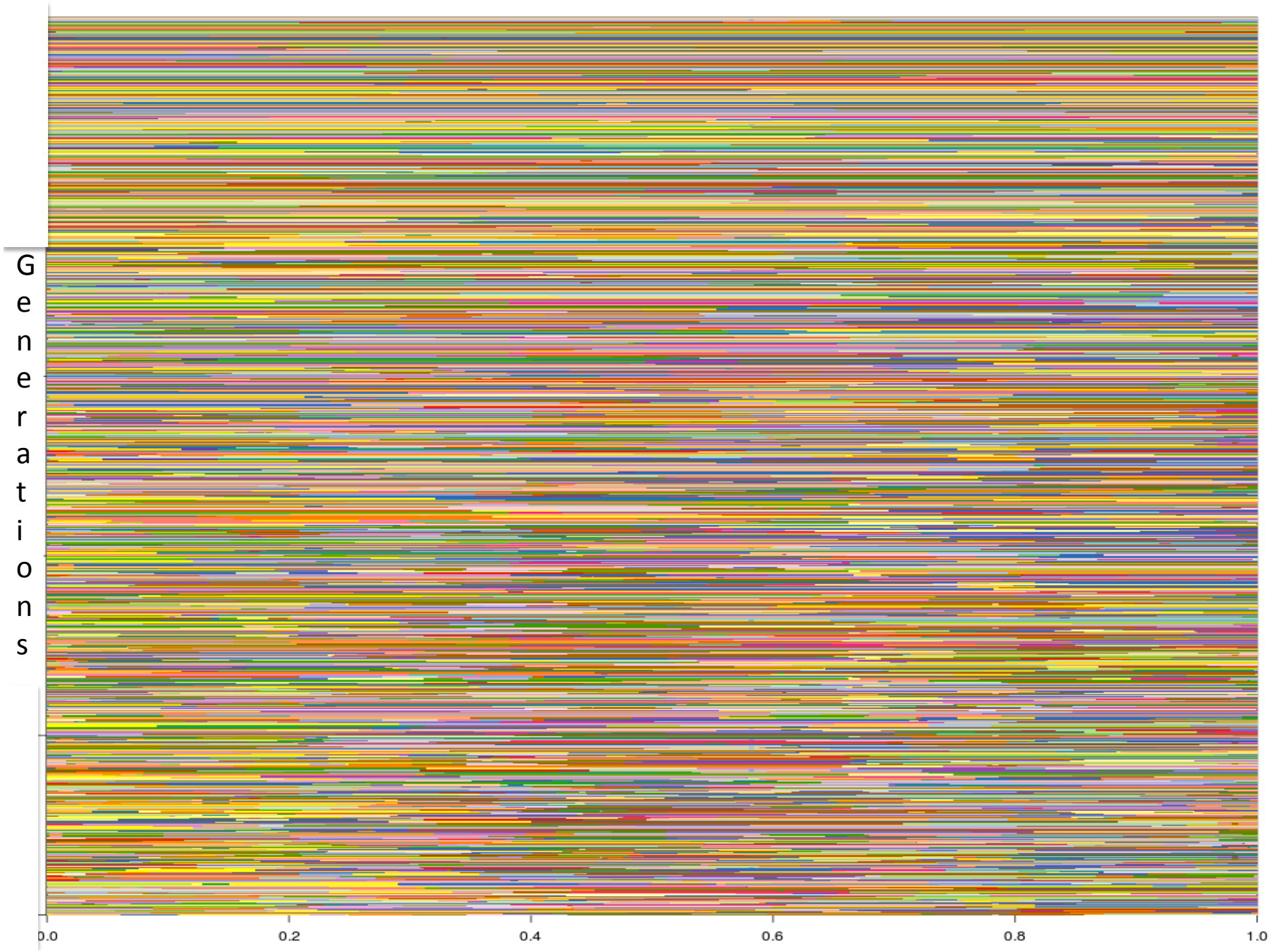
How we do it

- Two steps
  - Phase the genotypes of the high density individuals and identify the haplotypes
  - Choose which combination of these haplotypes is carried by the individuals genotyped at low density
    - Underlying assumption is that haplotypes are preserved and recombination's can be modelled
- In other words
  - Alleles coded as 0/1 genotypes as 0/1/2
  - My father is heterozygous
    - Is the 0 on his paternal gamete or his maternal gamete?
  - Did I receive my fathers paternal gamete or maternal gamete at this location and at the next location?
    - 0 or a 1 from my father

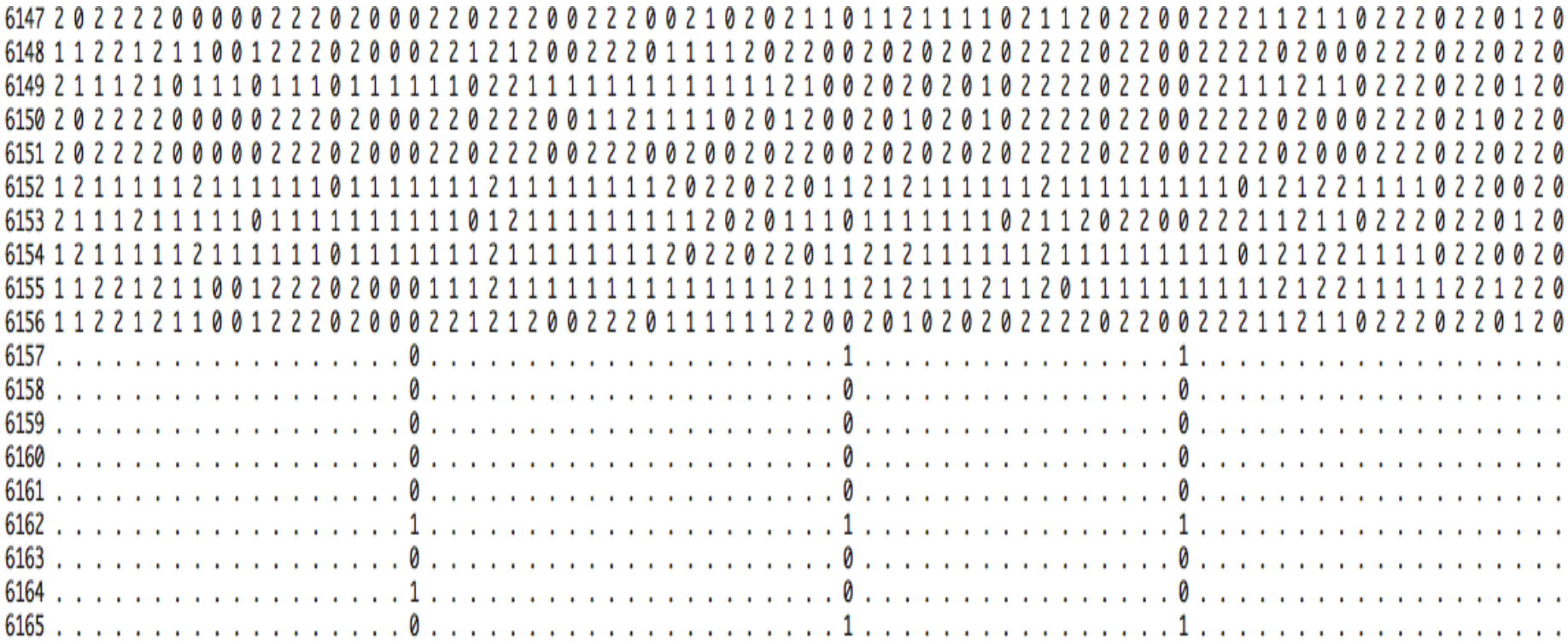
What an animal pedigree looks like

- 6 generations random mating followed by 3 generations selection
- 200 individuals per generation
- As we get further from the base
  - The haplotypes get smaller
  - The more markers needed

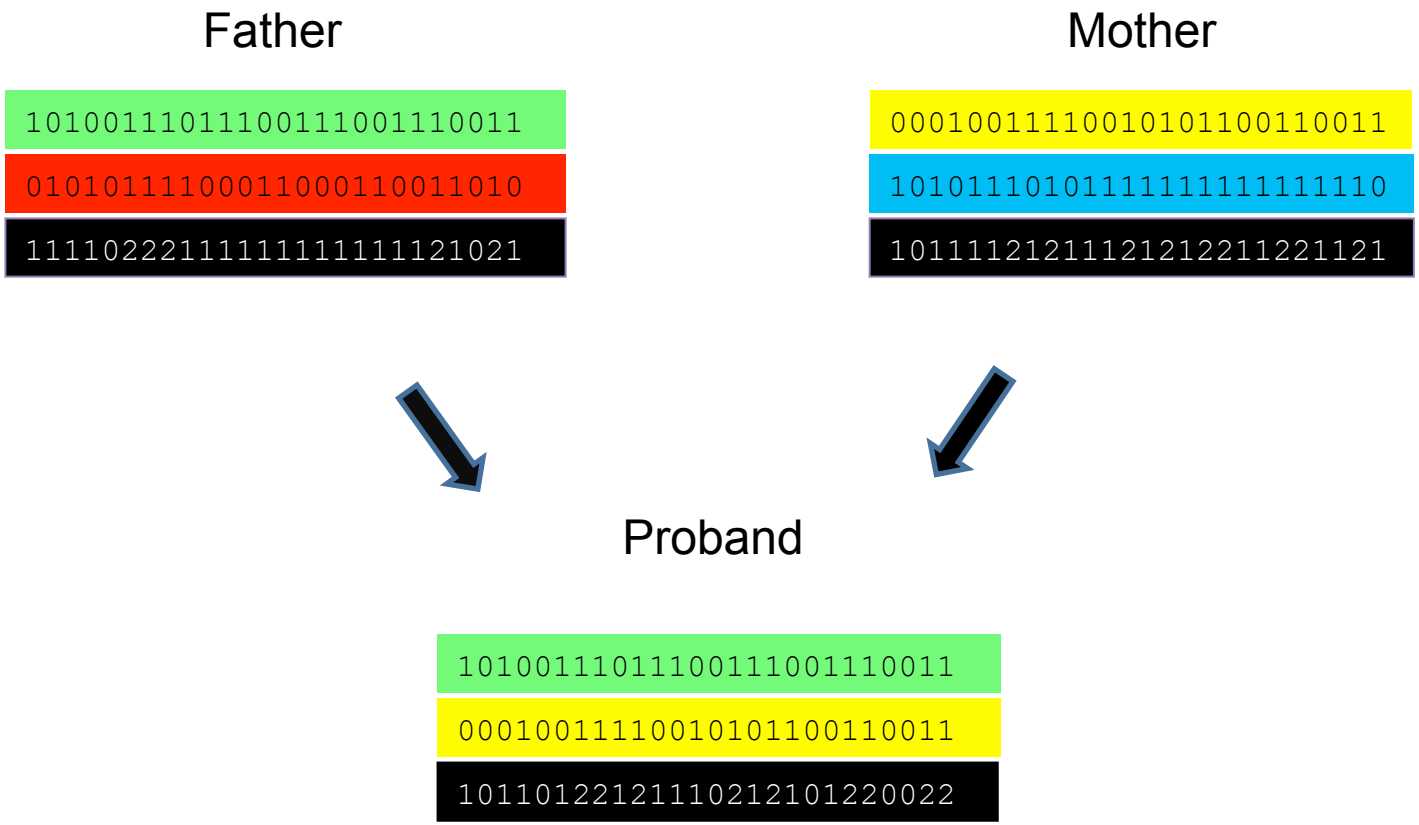



Generations
0.0
0.2
0.4
0.6
0.8
1.0

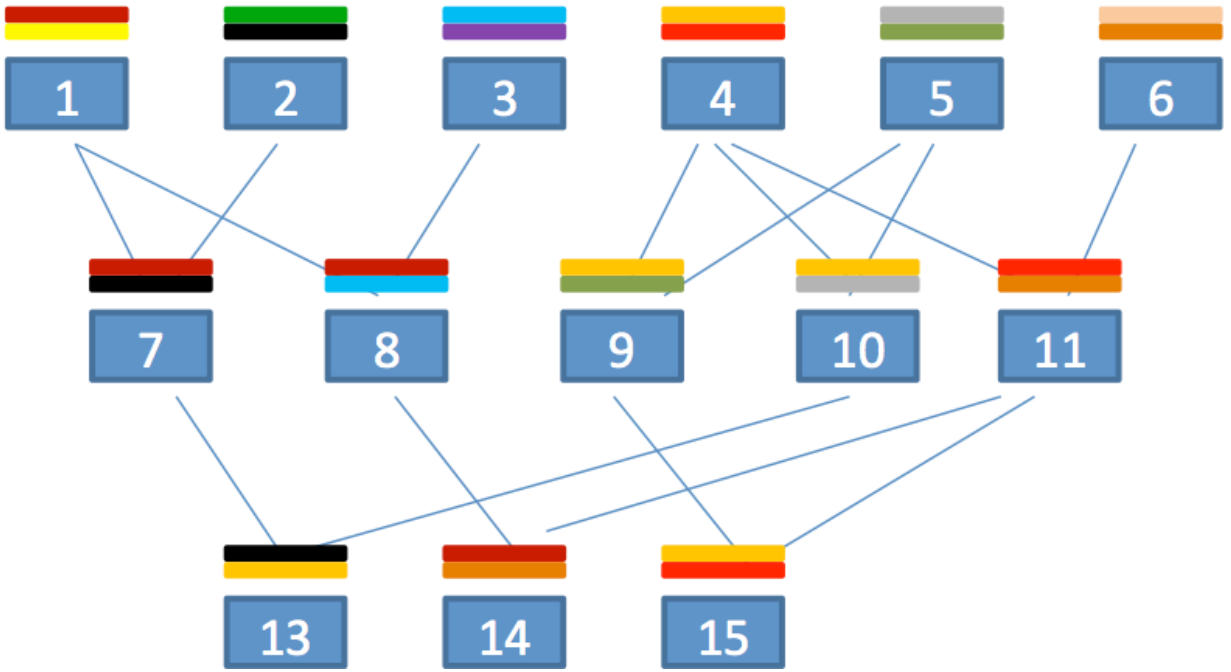
The imputation problem for a 2.5k low density chip in pigs



What underlies a genotype?

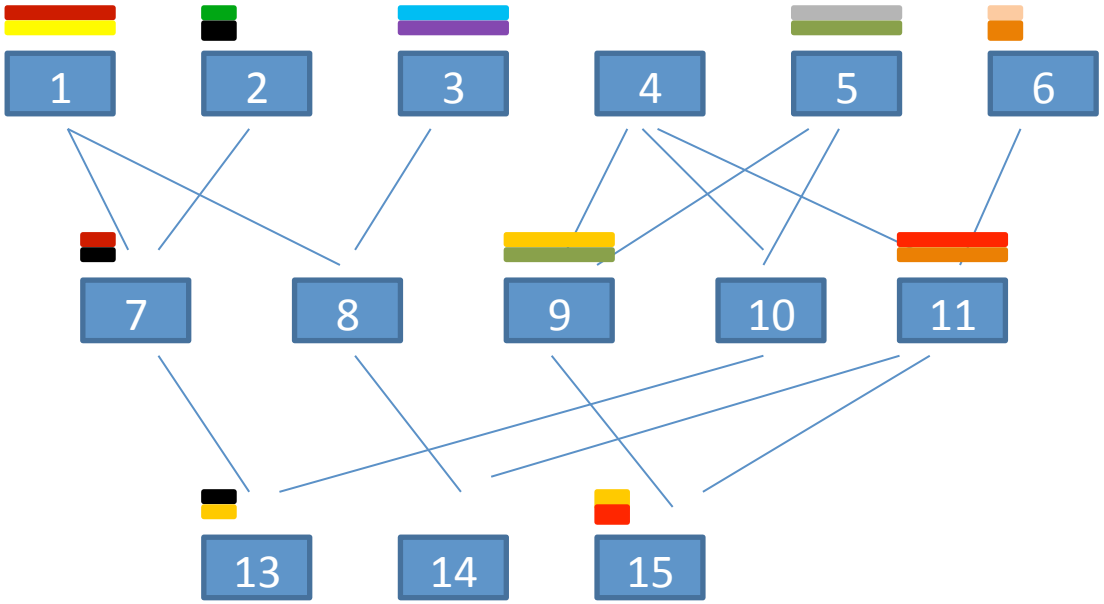


The basic idea of imputation



General pedigree with its haplotypes represented

Segregation analysis and haplotype library imputation

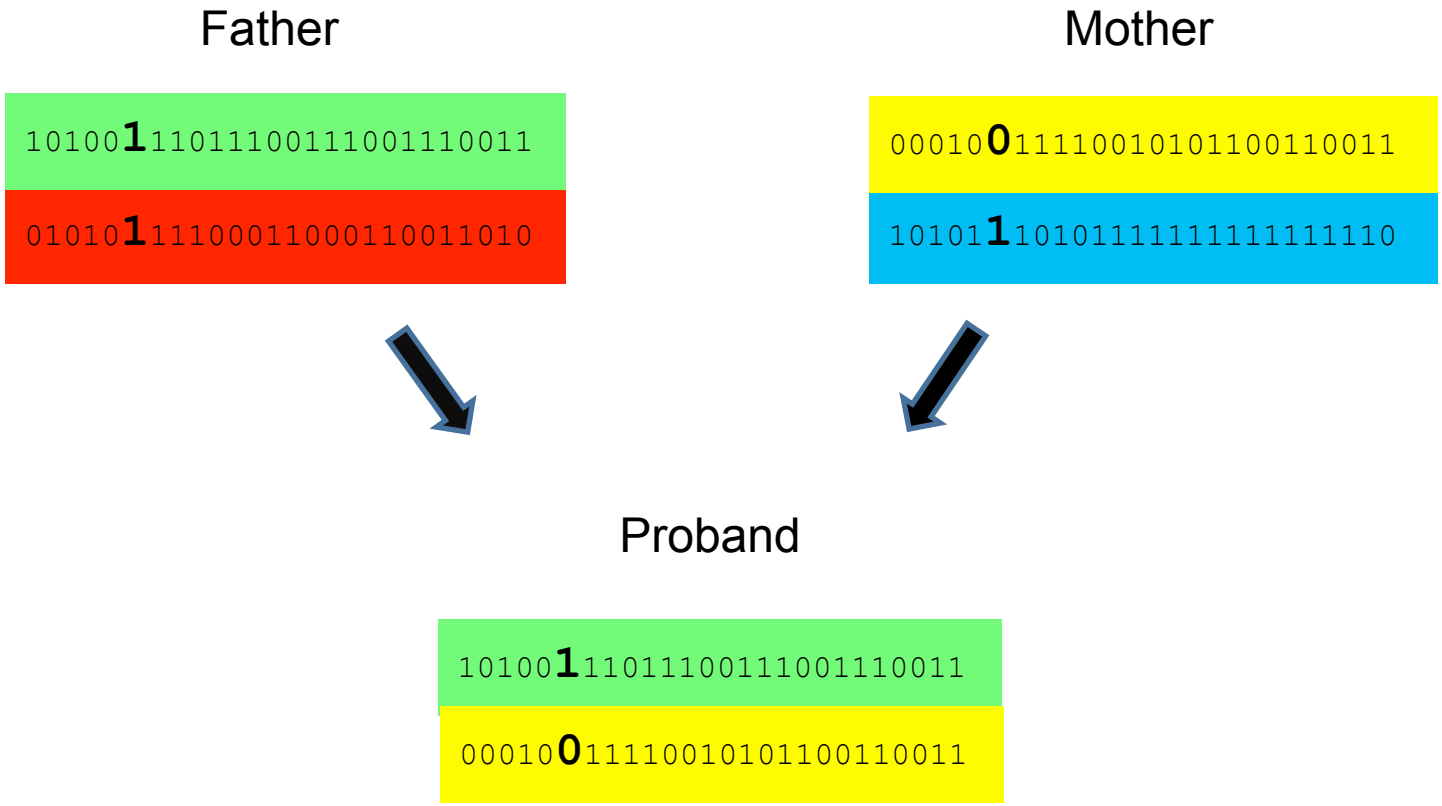


- Individual's are densely, sparsely, or not genotyped
- Pedigree information available
- Single locus segregation analysis for each SNP
- Match each pair of haplotypes with low density genotypes and genotype probabilities

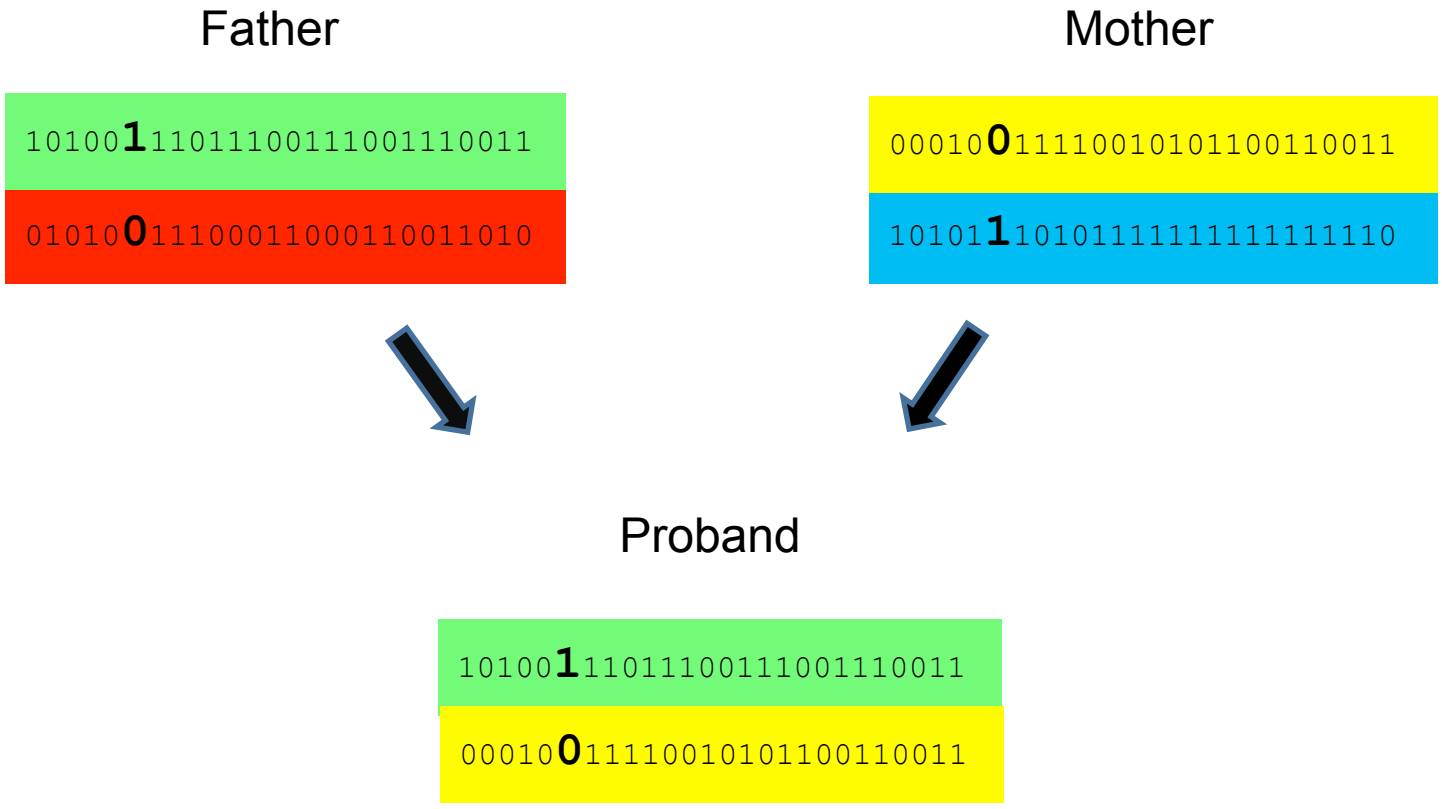
Genotyping strategy in terms of high density, low density and not genotyped

Haplotype library for population

Phasing a Trio

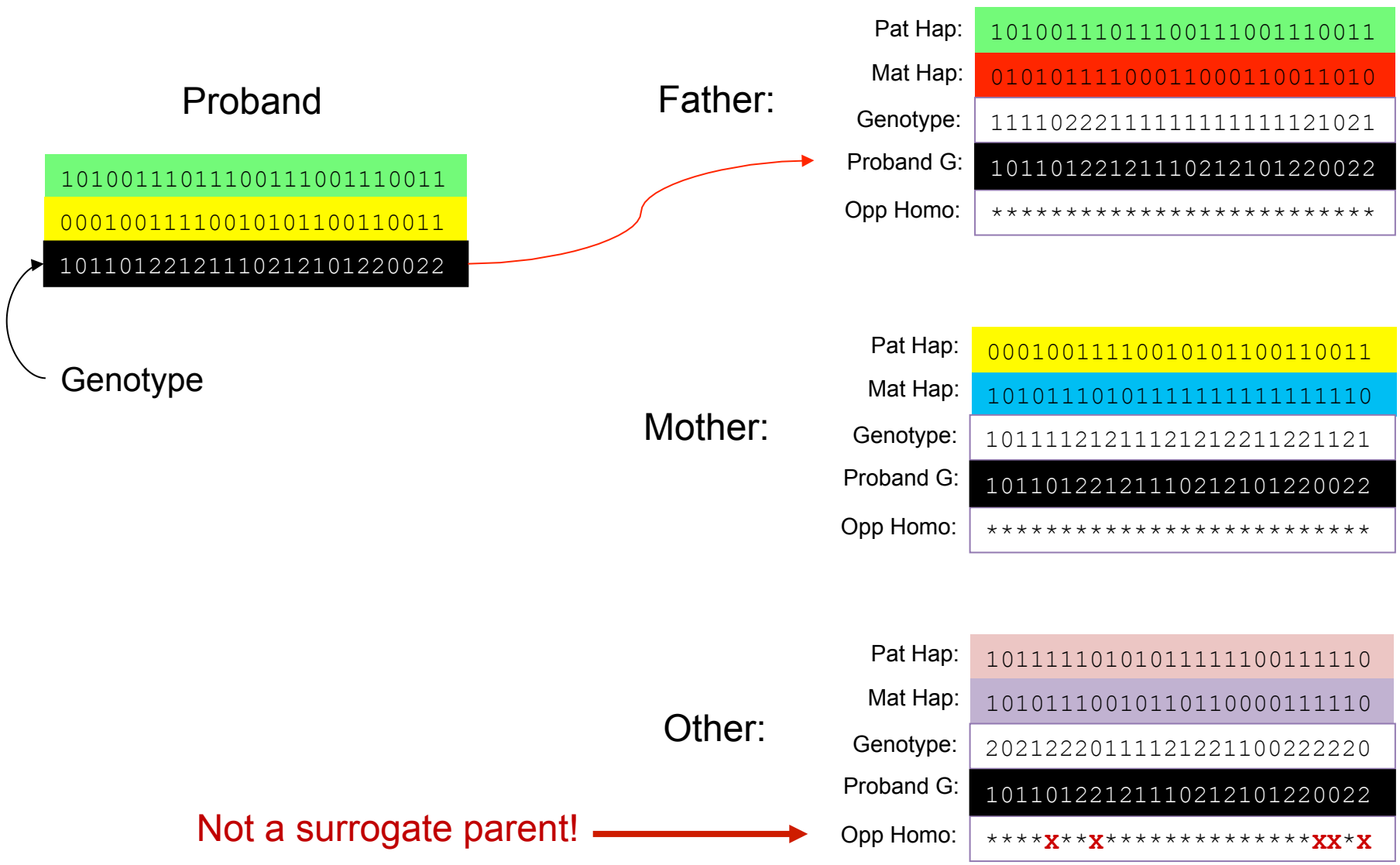


Phasing a Trio

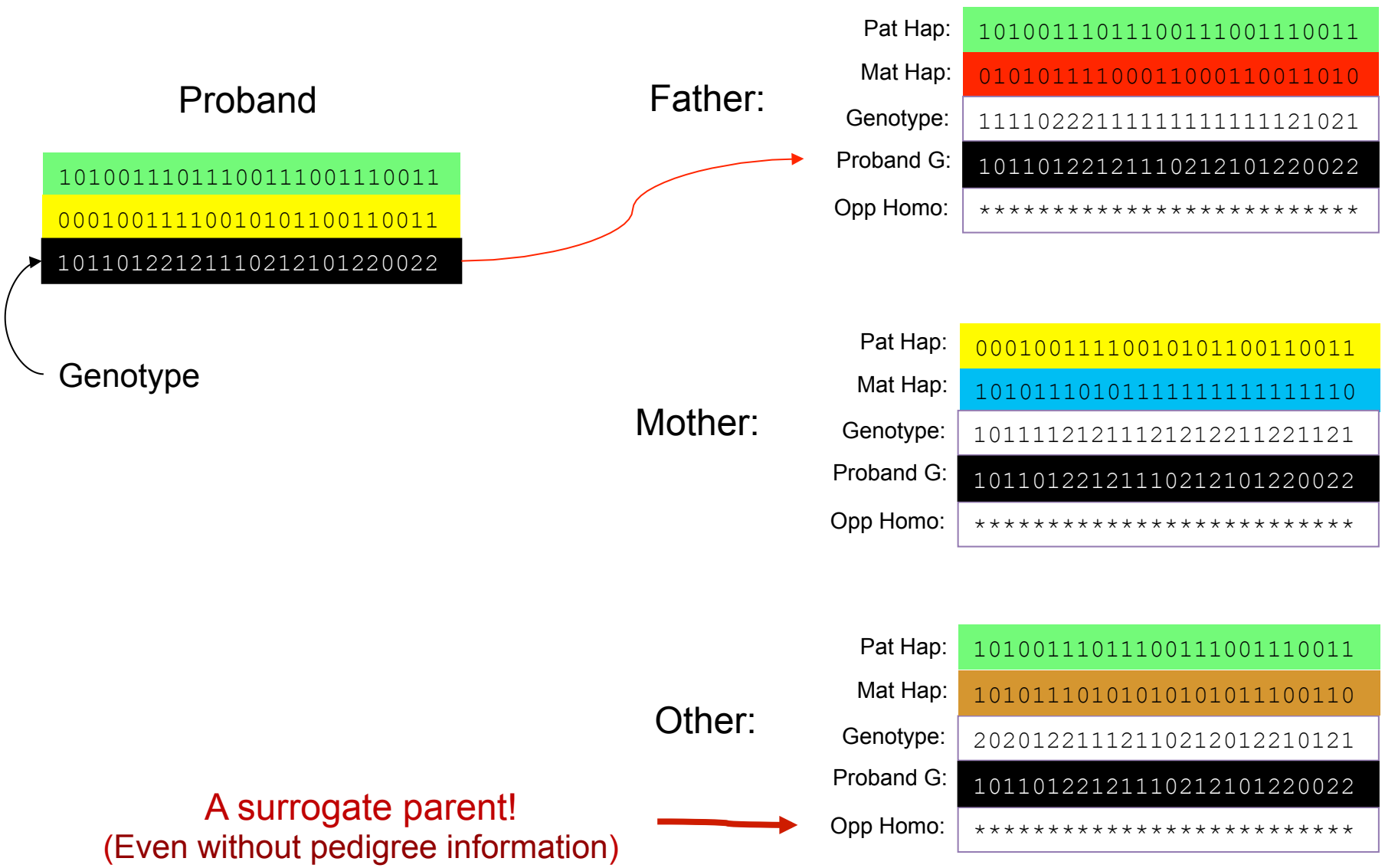


Cannot phase this locus!!

Surrogate parents are the driver of long range phasing



Surrogate parents are the driver of long range phasing



Phasing a Trio

Could be a female
Could be a descendant
Could be many generations distant
Can be 'unrelated'

Surrogate Father

10100**1**1101110011100111 0011
10101**1**1010101010101110 0110

Mother

00010**0**1111001010110011 0011
10101**1**1010111111111111 1110

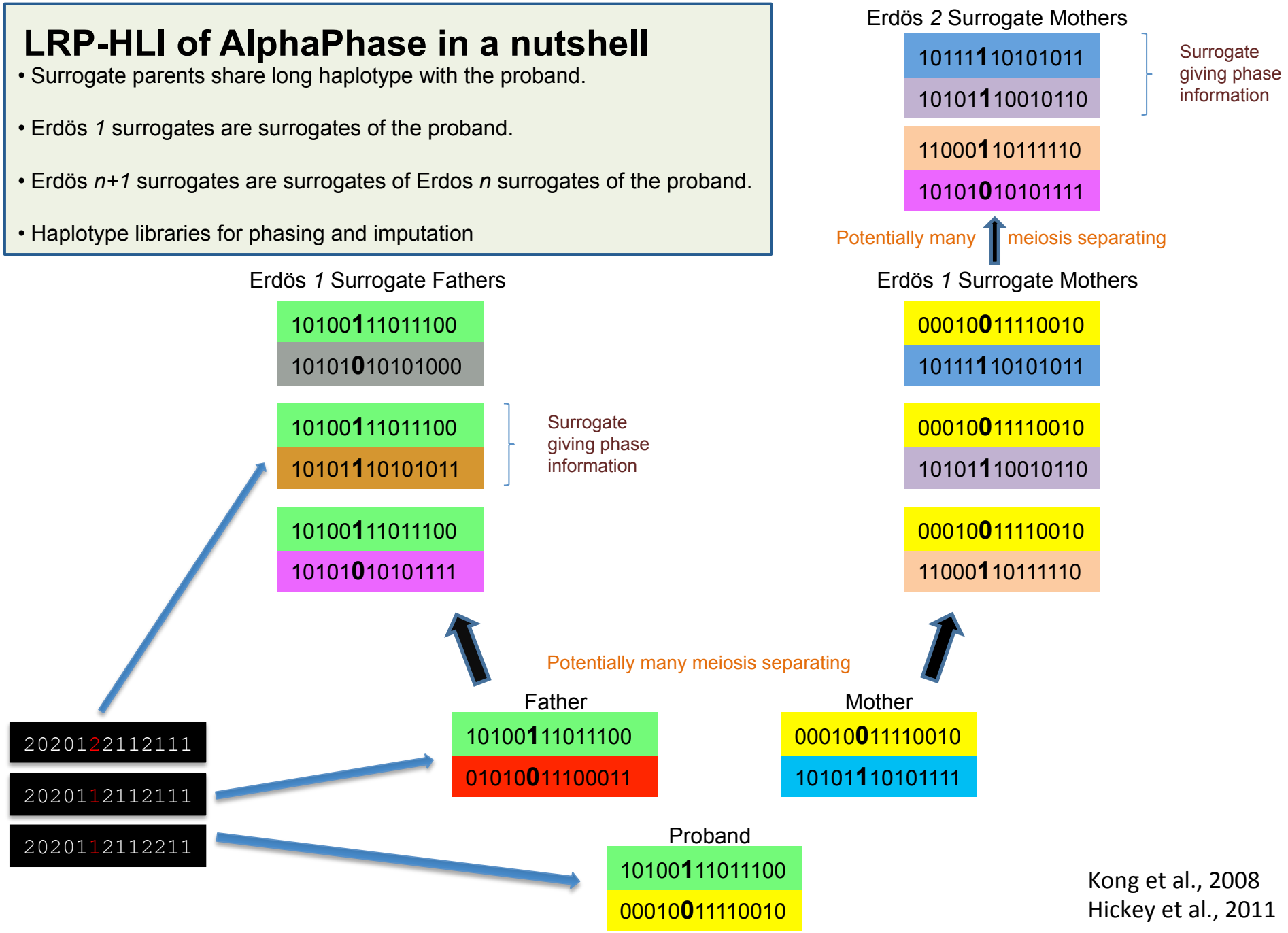
Proband

10100**1**1101110011100111 0011
00010**0**1111001010110011 0011

Can now phase this locus!!

LRP-HLI of AlphaPhase in a nutshell

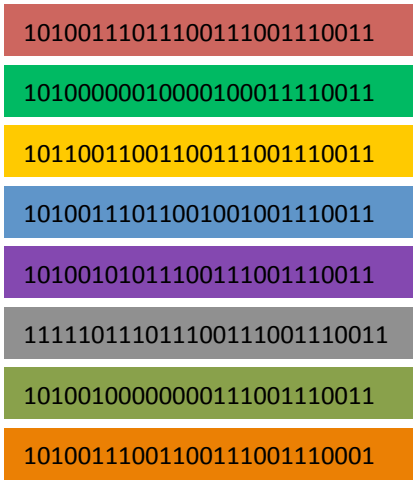
- Surrogate parents share long haplotype with the proband.
- Erdös *1* surrogates are surrogates of the proband.
- Erdös *n+1* surrogates are surrogates of Erdos *n* surrogates of the proband.
- Haplotype libraries for phasing and imputation



Kong et al., 2008
Hickey et al., 2011

Haplotype library imputation

- Build library of all completely phased haplotypes
- Find haplotypes in the library which can explain an individuals genotype
- Low error rates
- Computationally fast
- Useful for extremely large data sets
  - Strategic use



Phasing results simulated data

Percentage of alleles correctly phased / incorrectly phased by the most optimal core and CplusT lengths[1] for each pedigree and data scenario and for both effective population sizes.

	Ne 100		Ne 1000	
	With pedigree	Without pedigree	With pedigree	Without pedigree
Pedigree 1	99.11 / 0.30	NR	99.11 / 0.30	NR
Pedigree 2 without parents genotyped	97.85 / 0.43	98.49 / 0.63	97.73 / 0.29	97.88 / 0.39
Pedigree 2 with parents genotyped	98.85 / 0.49	99.03 / 0.42	99.70 / 0.17	99.48 / 0.17
Pedigree 3 without parents genotyped	98.35 / 0.38	98.61 / 0.63	99.23 / 0.14	99.14 / 0.27
Pedigree 3 with parents genotyped	99.23 / 0.37	99.05 / 0.41	99.76 / 0.16	99.58 / 0.13
Pedigree 4 without parents genotyped	98.20 / 0.41	98.61 / 0.63	98.19 / 0.41	98.61 / 0.63
Pedigree 4 with parents genotyped	99.35 / 0.31	99.29 / 0.32	99.74 / 0.20	99.59 / 0.15
Pedigree 5	97.59 / 0.42	98.28 / 0.60	99.30 / 0.30	99.31 / 0.22
Pedigree 6 sires	97.05 / 0.45	98.40 / 0.62	99.05 / 0.17	99.25 / 0.20
Pedigree 6 last 2000	98.24 / 0. 39	98.24/ 0.39	99.34 / 0.20	99.42 / 0.26
Pedigree 7 sires	97.56 / 0.40	98.71 / 0.50	98.98 / 0.20	99.15 / 0.29
Pedigree 7 last 2000	96.86 / 0.46	98.40 / 0.66	98.85 / 0.20	99.34 / 0.26
Pedigree 8	95.01 / 1.10	96.67 / 1.39	96.02 / 0.57	96.36 / 1.01

[1]Core length was 100 SNPs, CplusT length varied between 300 and 500 SNPs

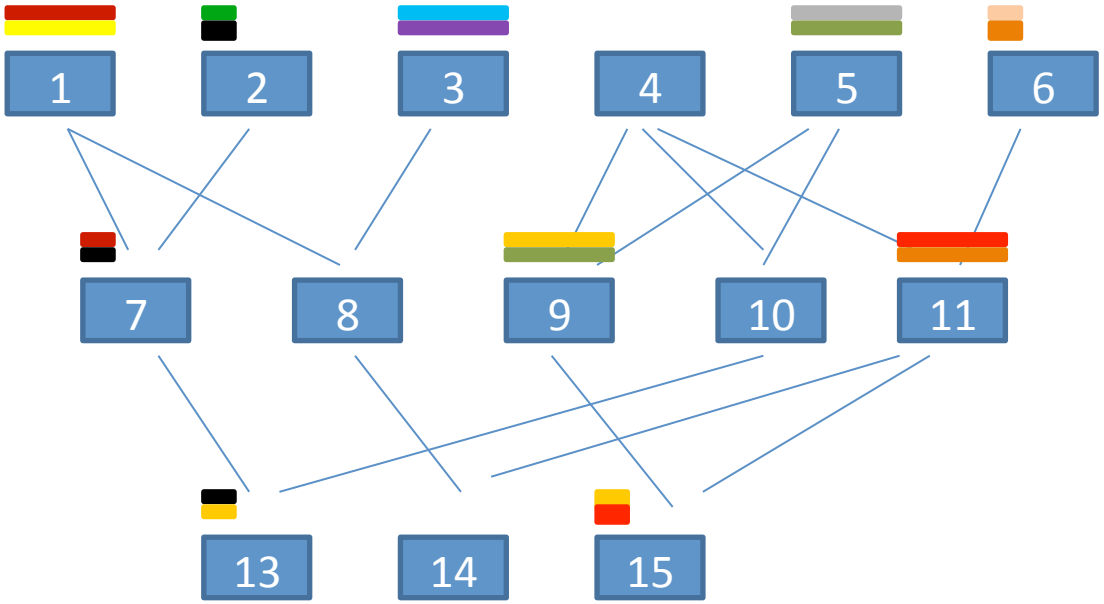
Phasing results real data

Table 2. Numbers of individuals in the data set, numbers of SNPs to be phased, core and CplusT length and missing genotype / genotype error % (M/E%) error threshold parameters, computation time, and percentage of alleles phased for 6 real data sets.

Data set	# Individuals	# SNPs	Core / CplusT length	M/E%	[1]Time	% Phased
Sheep Chr. 4	1019	2278	100 / 300	1.00	3 min. 39 sec.	98.17
Sheep Chr. 5	1016	1927	100 / 400	1.00	5 min. 1 sec.	97.62
Pig Chr. 1	2723	3999	100 / 500	0.00	364 min.	96.87
Beef Chr. 24	2171	874	100 / 300	0.00	17 min. 8 sec.	98.42
Dairy Chr. 1	5057	2296	100 / 400	0.00	456 min.	97.99
Human Chr. 1	879	4472	100 / 300	1.00	3 min. 29 sec.	93.73

[1]A 64 bit desktop with an Intel i7 3.07 GHz quad core processor running Linux was used to measure computation time. Computation time includes time required to parse and summarise the data and write out the results.

Segregation analysis and haplotype library imputation



- Individual's are densely, sparsely, or not genotyped
- Pedigree information available
- Single locus segregation analysis for each SNP
- Match each pair of haplotypes with low density genotypes and genotype probabilities

Genotyping strategy in terms of high density, low density and not genotyped

Haplotype library for population

Results PIC pig data set

		0.5k LD		2.5k LD		5k LD		7.5k LD	
Category	Count	AlphaImpute	IMPUTE2	AlphaImpute	IMPUTE2	AlphaImpute	IMPUTE2	AlphaImpute	IMPUTE2
BothParents	51	0.98	0.77	0.99	0.92	1.00	0.96	1.00	0.96
SireMGS	62	0.93	0.80	0.98	0.92	0.99	0.94	0.99	0.96
DamPGS	47	0.96	0.79	0.98	0.92	0.99	0.95	0.99	0.96
Sire	45	0.89	0.78	0.97	0.92	0.99	0.95	0.99	0.97
Dam	13	0.90	0.76	0.96	0.89	0.98	0.93	0.98	0.95
Other	291	0.86	0.79	0.94	0.91	0.97	0.95	0.97	0.96

Correlation is the statistic that matters

The cost and accuracy of sensible strategies

Scenarios	Other	MGS + PGS	MGD + PGD	Sire	Dam	Candidates	Individual cost	Accuracy of Imputation R2
SC1	60k	60k	0	60k	0	384	**!**	**0.878**
SC2	60k	60k	384	60k	384	384	**$20.58**	**0.929**
SC3	60k	60k	3k	60k	3k	384	**$24.74**	**0.950**
SC4	60k	60k	6k	60k	6k	384	**$26.28**	**0.944**
SC5	60k	60k	60k	60k	60k	384	**$34.84**	**0.964**
SC6	60k	60k	0	60k	0	3k	**!**	**0.968**
SC7	60k	60k	384	60k	384	3k	**!**	**0.972**
SC8	60k	60k	3k	60k	3k	3k	**$35.58**	**0.984**
SC9	60k	60k	6k	60k	6k	3k	**$41.28**	**0.983**
SC10	60k	60k	60k	60k	60k	3k	**$49.84**	**0.993**
SC11	60k	60k	0	60k	0	6k	**!**	**0.982**
SC12	60k	60k	384	60k	384	6k	**!**	**0.983**
SC13	60k	60k	3k	60k	3k	6k	**!**	**0.986**
SC14	60k	60k	6k	60k	6k	6k	**$48.58**	**0.991**
SC15	60k	60k	60k	60k	60k	6k	**$62.84**	**0.996**
SC16	60k	60k	60k	60k	60k	60k	**$120.00**	**1.000**

nSires = 480
nDams = 11884
nCandidates = 100000

60k chip = $120
6k chip = $48
3k chip = $35
384 chip = $20

Effect of imputation on GEBV accuracy

		Genotyping Scenario						Imputed gEBV Accuracy		
	N HD Geno.	Other	PGS+ MGS	PGD+ MGD	Sire	Dam	Progeny	450	3k	6k
S1	2519	H	H	H	H	H	L	**0.94**	0.97	0.97
S2	2344	H	0	0	H	H	L	**0.89**	0.95	0.96
S3	2318	H	H	0	H	0	L	**0.87**	0.92	0.93
S4	2318	H	H	L	H	L	L	**0.90**	0.96	0.97

Another thing one can do:

- Scenario 1 - Train using the 3200 HD animals
- Scenario 2 - Train using the 3200 HD and 2764 completely ungenotyped animals
- Predict in the 509 testing animals
- Growth trait with h2 of 0.61
- Results - correlation with progeny test EBV from BLUP

	Accuracy
Scenario 1	0.51
Scenario 2	0.62

Routine use

- Phasing on a monthly basis
  - Every time significant amount of high density genotypes are added
- Imputation on a weekly basis
  - Each week approximately 2000 low-density genotypes are added

Phasing/Imputation algorithms

- Two groups of imputation algorithms
  - Hidden markov based models
  - Heuristic methods

- Hidden markov based models
  - Probabilistic / pedigree free
  - Model linkage disequilibrium/short haplotypes
  - e.g. fastPHASE

- Heuristic methods
  - Tend not to be probabilistic / use pedigree information
  - Model linkage/long haplotypes
  - e.g. AlphaPhase (Long-range phasing and haplotype libraries)

Strengths, weaknesses, and solution

- HMM
  - Computationally intensive
  - In some cases less accurate
  - Flexible to data set design
  - Probabilistic

- LRP-HLI (Heuristic methods)
  - Faster and more accurate than HMM
  - Needs linkage blocks
  - Lacks probabilistic basis

- Because these methods access different information we should combine:

  LRP-HLI of AlphaPhase

  and HMM of fastPHASE

  to make a genomic relationship matrix directly

  and account for the uncertainty in genomic information

HMM of fastPHASE in an nutshell

Theta

0.0001	0.9999	0.0001	0.9999	0.0001	0.9999	0.0001	0.0001	0.0001	0.0001
0.0001	0.9568	0.0001	0.9485	0.0001	0.9999	0.0801	0.5076	0.0001	0.2348
0.0001	0.0279	0.0001	0.0301	0.1109	0.0325	0.0001	0.0981	0.0001	0.0984
0.9544	0.9999	0.9999	0.9999	0.9999	0.9999	0.8485	0.9999	0.5999	0.5538

Output probabilities for individual 29

29	1	1	0.0
29	1	2	0.0105
29	1	3	0.0
29	1	4	0.0
29	2	1	0.0105
29	2	2	0.9718
29	2	3	0.0036
29	2	4	0.0
29	3	1	0.0
29	3	2	0.0036
29	3	3	0.0
29	3	4	0.0
29	4	1	0.0
29	4	2	0.0
29	4	3	0.0
29	4	4	0.0

Column Index
ID
Paternal gamete
Maternal gamete
Probabilities for marker 1

For individual 29 it is highly probable that its two gametes derive from founder haplotype 2

Scheet and Stephens, 2006

Founder Haps

0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1

Genotypes for 30 Inds

1	1	1	2	2	2	2	2	2	2	1
2	0	0	1	1	1	1	1	1	1	0
3	0	0	0	0	0	0	0	0	0	0
4	0	1	0	1	0	1	0	0	0	0
5	0	2	0	2	0	2	0	1	0	1
6	2	2	2	2	2	2	2	2	1	1
7	0	1	0	1	0	1	0	1	0	1
8	0	2	0	2	0	2	0	1	0	0
9	0	2	0	1	0	1	0	0	0	0
10	0	1	1	1	1	1	1	1	1	0
11	0	1	0	1	0	1	0	1	0	1
12	0	1	0	1	0	1	0	0	0	0
13	1	1	1	1	1	1	1	1	1	1
14	0	1	0	2	0	1	0	0	0	0
15	0	1	0	1	0	1	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0
17	1	1	1	1	1	1	1	1	0	0
18	0	0	0	0	0	0	0	0	0	0
19	1	2	1	2	1	2	1	2	1	1
20	0	1	0	1	0	1	0	1	0	1
21	1	2	1	2	1	2	1	2	0	1
22	0	0	0	1	0	1	0	1	0	1
23	1	2	1	2	1	2	0	1	0	0
24	0	1	0	0	0	2	1	1	0	0
25	2	2	2	2	2	2	1	2	1	2
26	1	1	1	1	1	1	1	1	0	0
27	0	1	0	1	0	1	1	2	0	1
28	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	1	0	1
30	0	2	0	2	0	2	0	1	0	0

The tabular method for computing genetic relationships

<Momentary digression>
<How to unify this information>

- Recall tabular method for computing the numerator relationship matrix:
  - Wright
  - Henderson, 1976.

A = $\{a_{ij}\}$

a_{ij} is the genetic relationship between animals *i* and *j*.

Let parents of *j* be d_j and s_j.

$$a_{ij} = 0.5a_{i,d_j} + 0.5a_{i,s_j}$$

Diagonal elements follow similar logic

Rules of Dempfle

- When parents are uncertain we can also make A
  - Dempfle, 1987
  - Henderson, 1988

  - Suppose dam of *j* be known to be d_j

  - But there are v_j different candidate sires ($s_1, s_2, \ldots s_{vj}$)
    - With probabilities ($p_1, p_2, \ldots p_{vj}$) of being the true sire

$$a_{ij} = 0.5a_{i,d_j} + 0.5\left(p_{s_1} a_{i,s_1} + p_{s_2} a_{i,s_2} + \ldots + p_{v_j} a_{i,s_{v_j}} \right)$$

$$a_{jj} = 1 + 0.5\left(p_{s_1} a_{s_1,d_j} + p_{s_2} a_{s_2,d_j} + \ldots + p_{v_j} a_{v_j,d_j} \right) = 1 + F_j$$

How does this relate to HMM

- Founder haplotypes can be treated as parents of each individual with a certain probability
  - Probability is the output probability of HMM
- Assume founder haplotypes (theta) are:
  - Fully inbred diploid individuals
  - Completely unrelated to each other (could soften this)
- Apply the rules of Dempfle to the output probabilities of HMM to make G at each marker position
- Whole genome G is the average of G at each marker

Combined in LHH algorithm as implemented in AlphaPhase

- Run LRP-HLI on genotypic data
  - Data is phased blocks or genomic regions
  - Overlapping blocks
- For each block partition the resulting data into
  - Phased for majority of markers
  - Not phased for majority of markers
- Run haploid HMM on phased group
  - Fast and accurate
- Run diploid HMM on unphased group
- Pass output probabilities to Dempfle algorithm to make G

Results - Genomic prediction

G method	Simulated Normal Gen 6	Simulated Normal Gen 10	Simulated Gamma Gen 6	Simulated Gamma Gen 10	Maize
LHH	0.47	0.33	0.56	0.38	0.52
VanRaden	0.47	0.35	0.55	0.36	0.51
True G from QTL	0.50	0.40	0.73	0.69	------