Labelling haplotypes when genotypes are missing

C. Stricker

stricker@genetics-network.ch
agn Genetics GmbH, Davos

Quick Reminder of Genomic Prediction

- 1k,... >>100k SNPs per animal
- 1k, ... 100k animals
- Meuwissen et al. (2001): Marker Effects Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\alpha + \mathbf{e} \tag{1}$$

  - $\mathbf{y}$ vector of trait phenotypes
  - $\mathbf{X}$ incidence matrix relating non-genetic, fixed effects $\boldsymbol{\beta}$ to $\mathbf{y}$
  - $\mathbf{Z}$ matrix of SNP genotype covariates,
  - α vector of random, partial-regression coefficients for SNPs
  - $\mathbf{e}$ is vector of residuals
- Bayesian alphabet is based on this model
- can show that BayesCπ=0 is GBLUP (see below)

Quick Reminder of Genomic Prediction, cont.

- Nejati-Javaremi et al. (1997): Animal Effect Model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad (2)$$

  - $\mathbf{y}$, $\mathbf{X}\beta$, $\mathbf{e}$ as above in model (1)
  - $\mathbf{Z}$ incidence matrix relating animal effects to $\mathbf{y}$,
  - $\mathbf{u}$ vector of random animal effects
- Solutions usually by MMEs
- use $\mathbf{G} = \mathbf{Z}\mathbf{Z}'/2\sum pq$ instead of $\mathbf{A}$
- requires $\mathbf{G}^{-1}$ (instead of $\mathbf{A}^{-1}$)
- $\mathbf{G}$ is dense, cannot be inverted, when >50k animals genotyped....
- Legarra et al. (2009), Christensen & Lund (2010) combined $\mathbf{A}$ for ungenotyped animals with $\mathbf{G}$ for genotyped animals
- requires $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$, dense and not easy!
- need alternatives when $>$ 50k animals genotyped

Problems in Genomic Prediction

- growing number of genotyped animals >> 50'000
  - MEM using Bayesian regression applicable, but MCMC is CPU/Memory-intensive! 'Currently' only for genotyped animals
  - ssAEM unfeasible, due to $\mathbf{G}^{-1}$
  - ssMEM using the approach of Fernando et al. (2013)
  - MEM more efficient when $n > k$
- growing number of markers >> 50'000
  - single locus LD to causal mutations only for very high marker densities to be expected
  - increasing the marker density requires more data
  - highly redundant marker information
  - variable number of markers per genotyped animal, imputation to use available software. No additional information through imputing
  - expensive variable reduction methods (Bayes B) applied to imputed data

Genomic Prediction with Haplotypes

- why haplotypes?
  - high single locus LD is not common in 54k/880K SNP arrays
  - the signal we pick up with GP is mostly cosegregation (multi-locus LD)
  - haplotypes pick up signal from LD and cosegregation, depending on segment size
  - haplotypes for small segments $\rightarrow$ powerfull if strong single locus LD
  - haplotypes for large segments $\rightarrow$ powerful if cosegregation (within-family LD)
  - data reduction: 1 cM Intervall, 54K SNPs, about 10-20 haplotypes segregating, but $>$60'000 single locus allele combinations
  - most of these combinations do not exist in the data, additivity of marker effects assumed...

Genomic Prediction with Haplotypes

- Defining segment size
- Phasing of segment based on surrogacy, 'genetic' distance
  - calculate number of incompatible genotypes between individuals in given segment
  - for each animal (pivot), all compatible animals (distance small) are considered its surrogate offspring
  - find the two most distant offspring among the surrogates, apply k-medoids clustering to identify two clusters carrying the two haplotypes of the pivot animal
  - assign the label of the corresponding haplotype of the pivot to all surrogate offspring in a cluster

Genomic Prediction with Haplotypes

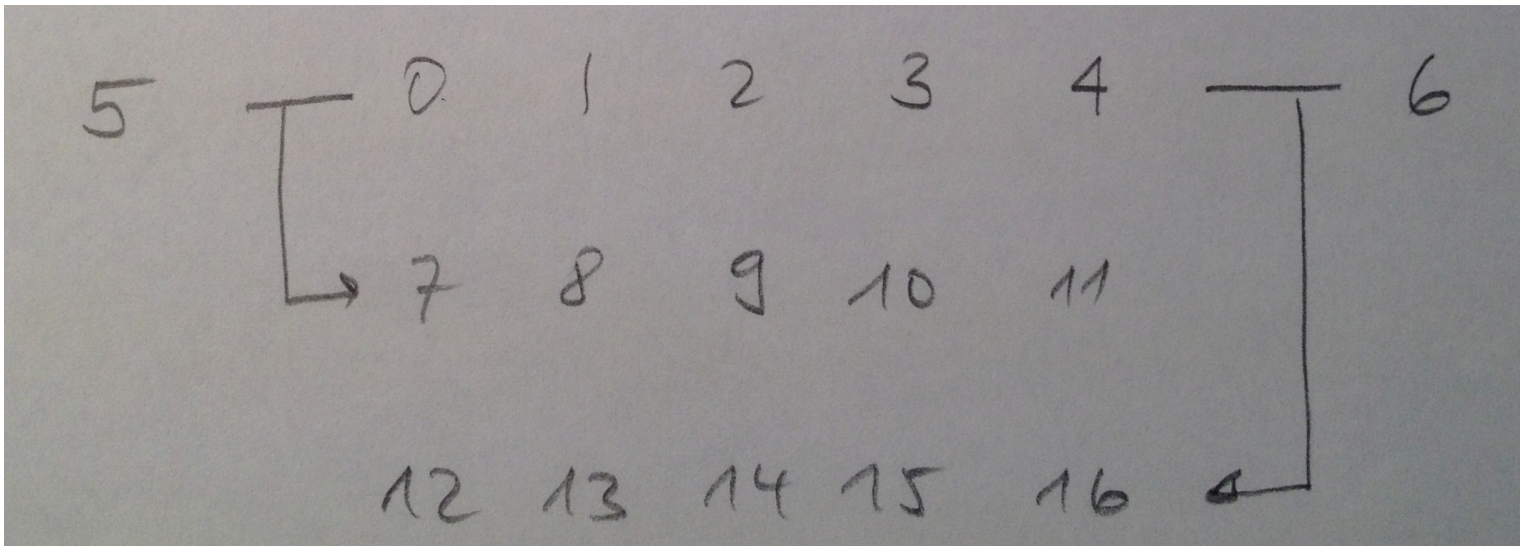
- each haplotype has a single cluster of surrogates, but an animal will be present in many pivots surrogate clusters
- recombination and mutation is noise, if haplotype is not passed onto offspring
- currently implemented for a single marker density, extension to missing markers (different densities) conceptually trivial
- no imputation: hapLabels are associated with a unique allelic state combination across all loci of a segment
- segment size can be a parameter in Gensel $\rightarrow$ MPI

Genomic Prediction with Haplotypes

- if imputation wanted, then
  - establish haplotype labels through phasing above
  - estimate allelic vectors corresponding to labels based on rules or regression haplotypes onto genotypes
  - store allelic states for haplotypes in library
  - join adjacent segments by overlapping
- haplotypes can be inferred based on different SNP densities
- (ss)MEM oder (ss)AEM

Example

- segment size of 1 cM with 1000 Markers, Rec at 7, 8, 12



Example

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	0	76	135	135	0	126	123	0	77	139	127	0	0	125	123	127	0
1	76	0	118	110	76	131	77	122	0	134	122	76	79	0	69	122	76
2	135	118	0	125	132	110	109	124	107	0	111	113	120	109	0	111	113
3	135	110	125	0	0	135	127	144	131	128	0	130	141	131	131	0	130
4	0	76	132	0	0	138	127	133	79	140	124	0	37	127	126	124	0
5	126	131	110	135	138	0	0	0	0	0	0	0	0	0	132	0	0
6	123	77	109	127	127	0	0	96	0	0	0	0	0	0	0	0	0
7	0	122	124	144	133	0	96	0	88	79	88	87	87	92	134	88	87
8	77	0	107	131	79	0	0	88	0	0	0	0	0	0	124	0	0
9	139	134	0	128	140	0	0	79	0	0	0	0	0	0	128	0	0
10	127	122	111	0	124	0	0	88	0	0	0	0	0	0	129	0	0
11	0	76	113	130	0	0	0	87	0	0	0	0	0	0	121	0	0
12	0	79	120	141	37	0	0	87	0	0	0	0	0	0	129	0	0
13	125	0	109	131	127	0	0	92	0	0	0	0	0	0	68	0	0
14	123	69	0	131	126	132	0	134	124	128	129	121	129	68	0	129	121
15	127	122	111	0	124	0	0	88	0	0	0	0	0	0	129	0	0
16	0	76	113	130	0	0	0	87	0	0	0	0	0	0	121	0	0

Example

haplotypeLabels

posInPedVec	hapID1	hapID2
0	1	2
1	10	12
2	3	4
3	5	6
4	1	5
5	7	8
6	7	9
7	2	8
8	7	10
9	3	7
10	6	7
11	1	7
12	7	11
13	7	12
14	4	9
15	6	7
16	1	7

Furutre directions

- fitting haplotypes instead of single marker covariables $\rightarrow$ data reduction, picks up signal from LD and cosegregation
- modelling LD (in the founder population) and cosegregation (from there onwards) explicitely, could be done with haplotypes
- ssMEM using haplotypes looks most promising
- efficient computing strategies become important with #markers$\uparrow$ and #animals genotyped$\uparrow$