Genotype imputation for the prediction of genomic breeding values in non-genotyped and low-density genotyped individuals

M.A. Cleveland[1*], J.M. Hickey[2] and B.P. Kinghorn[2]

[1]Genus/PIC, Hendersonville, TN, USA and [2]University of New England, Armidale, NSW, Australia

14th QTL-MAS Workshop, Poznań
18 May 2010

Approach

- Generally assume high-density genotypes for all individuals
  - i.e., training and prediction (old and young)
  - May not be the case in practice

- Predict genomic breeding values when individuals are genotyped, genotyped at low-density or not genotyped
  - Pedigree may be sparsely genotyped
  - “Unknown” genotypes are imputed using segregation analysis and a long haplotype library

- **Objective:**
  - Evaluate accuracy of genotype imputation
  - Evaluate accuracy of GEBVs when training at high density and predicting at high density versus predicting using imputed values

Imputation Background

- Haplotype library
  - Long range phasing (Kong *et al*., 2008)
    - Rule-based method using information from related and unrelated individuals
  - Recursive long range phasing and long haplotype imputation (Hickey *et al*., 2009)
    - Library of long haplotypes
  - Construct library based on training individuals

- Segregation analysis
  - Algorithm described by Kerr and Kinghorn, 1996
  - Relies on pedigree information
  - Genotype probabilities (geneprobs) for un-genotyped loci, with a measure of reliability (information content; GPI)

Imputation

Steps (for each individual):

1. Compare most probable genotype based on geneprob at each homozygous locus to corresponding locus in each haplotype (above GPI minimum)
   - Exclude haplotypes with opposing homozygotes

Most Prob Geno	Haplotype Library
2012000121010 10	1010100101101 0001 hap1
	1011000100010 01 hap2 ←
	1011000110010 00 hap3 ←
	…

Imputation

Steps (for each individual):

2. Compare most probable genotype based on geneprobs at each locus to remaining haplotype pairs (above GPI minimum)
   - Exclude pairs with opposing genotypes
   - Repeat until single pair remains or GPI minimum

Most Prob Geno
2012000121010 10

Haplotype Pairs	
10110001000 1001	hap2
10110001100 1000	hap3
20220002100 2001	geno

Imputation

Steps (for each individual):

3. Identify the most probable pair of remaining haplotypes using geneprobs for all loci, scaled by GPI
   - For any remaining haplotype pairs

Data

- 14th QTL-MAS Workshop
- N=2,326 training (4 generations)
- N=900 prediction
- M=10,031 SNP markers (5 chr. ~100 mbp each)
- Phenotype: Trait Q

Data

- BASE
  Training:
  – all have HD genotypes
  Prediction:
  – all have HD genotypes

- S1
  Training:
  – males have HD genotypes, females have HD genotypes imputed
  Prediction:
  – all genotypes imputed or all imputed except SNPs spaced at 2, 5 and 10mbp

- S2
  Training:
  – all have HD genotypes
  Prediction:
  – all genotypes imputed or all imputed except SNPs spaced at 2, 5 and 10mbp

Methods

- S1
  1. Create haplotype library using training males
     - 12 cores per chromosome (~10mbp) → 1 long core
  2. Calculate geneprobs for training females and prediction individuals
  3. Impute missing HD genotypes
  4. Estimate marker effects
     - BayesA
  5. Calculate GEBVs for prediction set, using imputed and low-density genotypes

Methods

- S2
  1. Create haplotype library using full training set
     - 12 cores per chromosome (~10mbp) → 1 long core
  2. Calculate geneprobs for prediction set
  3. Impute missing HD genotypes
  4. Estimate marker effects
     - BayesA
  5. Calculate GEBVs for prediction set, using imputed and low-density genotypes

Results

- Computation speed
  - LRPLHI (AlphaPhase): ~5 minutes / chromosome
  - Genotype probabilities: ~12 hours / chromosome
    - I/O not yet optimized
  - Imputation: ~30 minutes / chromosome

Results

- Evaluate accuracy of imputation

Percentage of genotypes correctly imputed in the two scenarios, considering alternative low-density genotyping

	% correctly imputed	
	S1	S2
training females - all genotypes imputed	69	
prediction - all genotypes imputed (m=0)	64	68
prediction - all genotypes imputed, except every 10mbp (m=55)	65	73
prediction - all genotypes imputed, except every 5mbp (m=105)	65	75
prediction - all genotypes imputed, except every 2mbp (m=251)	68	78

Results

- Evaluate GEBV prediction using imputed values compared to high-density genotyping – loss of accuracy

Correlation between GEBVs calculated when high density genotypes are known in the prediction set and GEBVs calculated using imputed genotypes, in the two scenarios

	correlation[a]	
prediction set	**S1**	**S2**
all genotypes imputed (m=0)	0.48	0.48
all genotypes imputed, except every 10mbp (m=55)	0.57	0.71
all genotypes imputed, except every 5mbp (m=105)	0.58	0.77
all genotypes imputed, except every 2mbp (m=251)	0.62	0.77

[a]GEBVs for each scenario correlated with GEBVs from BASE

Conclusions

- Imputation and correlation improved when sire and dams genotyped, and when including low-density genotypes
  - Implication: implement a “cost-effective” genomic selection strategy in systems where individuals can’t be HD genotyped
- Scale well to existing livestock datasets
  - 50-60k SNP chips
- Need to evaluate impact of accuracy loss from imputing
  - What is cost in practice?
  - Loss compared to true BV?
- Development/improvement of algorithm in progress

Thank you.