Combining Different Marker Densities in Genomic Evaluation

Paul VanRaden[1], Jeff O'Connell[2], George Wiggans[1], Kent Weigel[3]

[1]Animal Improvement Programs Lab, USDA, Beltsville, MD, USA

[2]University of Maryland School of Medicine, Baltimore, MD, USA

[3]University of Wisconsin Dept. Dairy Science, Madison, WI, USA

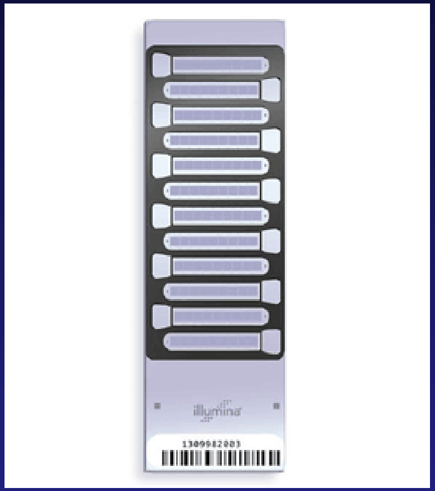
Paul.VanRaden@ars.usda.gov

2010

Topics

- Filling missing SNPs (imputation)
  - Find haplotypes from genotypes
  - Use lower density to track higher
  - Programs implemented April 2010
- Actual mixes of 3K with 50K
- Simulated mixes of 50K with 500K
- Calculating reliabilities

Mixing Different Chips



What is imputation?

- Genotypes indicate how many copies of each allele were inherited
- Haplotypes indicate which alleles are on which chromosome
- Use observed genotypes to impute unknown haplotypes
  - Pedigree haplotyping uses relatives
  - Population haplotyping finds matching allele patterns

Why impute haplotypes?

- Predict unknown SNP from known
  - Measure 3,000, predict 50,000 SNP
  - Measure 50,000, predict 500,000
  - Measure each haplotype at highest density only a few times
- Predict dam from progeny SNP
- Increase reliabilities for less cost

Haplotyping Program

findhap.f90

- Begin with population haplotyping
  - Divide chromosomes into segments, ~250 SNP / segment
  - List haplotypes by genotype match
  - Similar to FastPhase, IMPUTE
- End with pedigree haplotyping
  - Detect crossover, fix noninheritance
  - Impute nongenotyped ancestors

Recent Program Revisions

- Imputation and GEBV reliability are better than in 9WCGALP paper
- Changes since January 2010
  - Use known haplotype if second is unknown
  - Use current instead of base frequency
  - Combine parent haplotypes if crossover is detected
  - Begin search with parent or grandparent haplotypes

Most Frequent Haplotypes

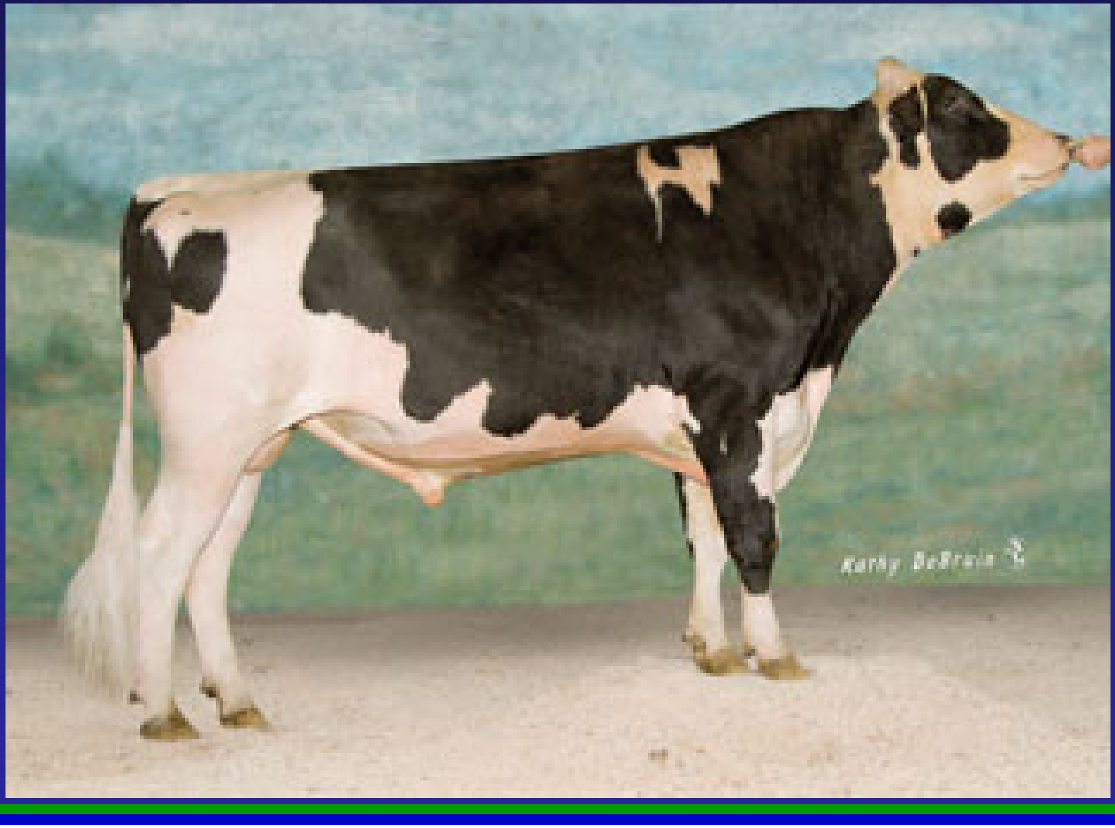
5.16%	022222222020020022002020200020000200202000022022222202220
4.37%	022020220202200020022022200002200200200000200222200002202
4.36%	022020022202200200022020220000220202200002200222200202220
3.67%	022020222020222002022022202020000202220000200002020002002
3.66%	022222220202220222020200220000020222220200002020220002022
3.65%	022020022202200200022020220000220202200002200222200202222
3.51%	022002222020222022202220220200222002200000002022220002220
3.42%	022002222002220022022020220020200202202000202020020002020
3.24%	022222222020200000022020220020200202202000202020020002020
3.22%	022002222002220022002020002220000020220000202022020202220

Most frequent haplotype in first segment of chromosome 15 for Holsteins had 4,316 copies = 41,822 * 2 * .0516

Example Bull: O-Style

USA137611441, Sire = O-Man

- Read genotypes, write haplotypes



Find Haplotypes – AB coding

Genotypes:

Oman BB,AA,AA,AB,AA,AB,AB,AA,AA,AB

Ostyle BB,AA,AA,AB,AB,AA,AA,AA,AA,AB

Haplotypes:

OStyle (pat)	B	A	A	_	A	A	A	A	A	_
OStyle (mat)	B	A	A	_	B	A	A	A	A	_

Find Haplotypes – 0,1,2 coding

Genotypes: codes 0 = BB, 1 = AB or BA, 2 = AA

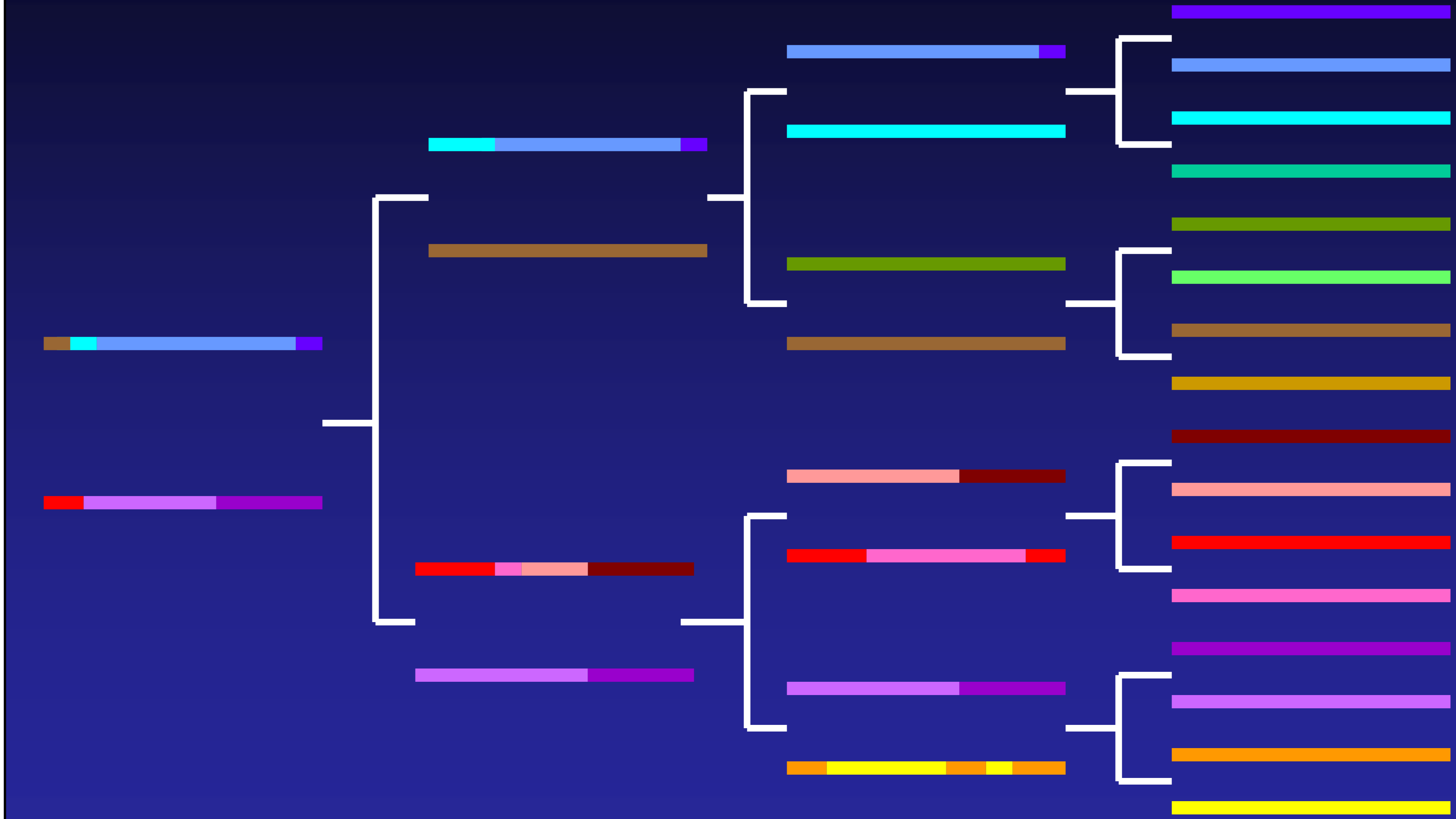
Oman	0 2 2 1 2 1 1 2 2 1
Ostyle	0 2 2 1 1 2 2 2 2 1

Haplotypes: codes 0 = B , 1 = unknown, 2 = A

OStyle (pat)	0 2 2 1 2 2 2 2 2 1
OStyle (mat)	0 2 2 1 0 2 2 2 2 1

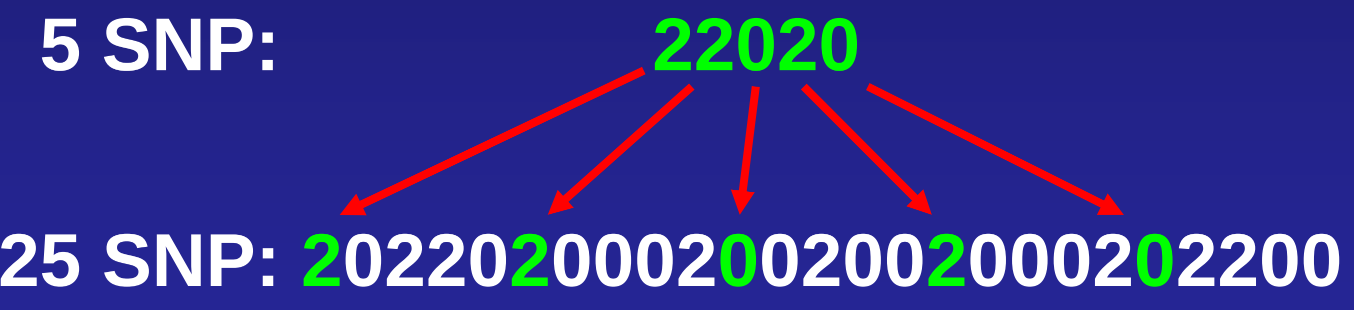
O-Style Haplotypes

chromosome 15



How does imputation work?

- Identify haplotypes in population using many markers
- Track haplotypes with fewer markers
- e.g., use 5 SNP to track 25 SNP



- 5 SNP: 22020
- 25 SNP: 2022020002002002000202200

Imputed Dams

- If progeny and sire both genotyped
  - First progeny inherits 1 of dam's 2 haplotypes
  - Second progeny has 50:50 chance to get same or other haplotype
  - Haplotypes known with 1, 2, 3, etc. progeny are ~50%, 75%, 87%, etc.

Better Communication is Needed

- **“Progeny genotypes should affect dam, but programs are not yet available” Jan 2009 USDA Changes Memo**
- **“Programs are available to impute 1300 dams” Oct 2009 USDA report to Council**
- **“Encourage USDA to use genotypes, derived by imputation, in genetic evaluation” Oct 2009 Holstein USA Board of Directors (in Holstein Pulse)**

Haplotyping Tests – Real Data

- Half of young animals assigned 3K
  - Proven bulls, cows all had 50K
  - Dams imputed using 50K and 3K
- Half of ALL animals assigned 3K
  - Could 3K reference animals help?
  - 10,000 proven bulls yet to genotype
  - Should cows with 3K be predictors?

Correlations^2 of 3K and PA with 50K

Half of YOUNG animals had 3K PTA, half 50K PTA

Trait	Corr(3K,50K)^2	Corr(PA,50K)^2	Gain
NM$	.899	.518	79%
Milk	.920	.523	83%
Fat	.920	.516	83%
Prot	.920	.555	82%
PL	.933	.498	87%
SCS	.912	.417	85%
DPR	.937	.539	86%

Using 3K as Reference Genotypes

Half of ALL animal NM$ were from 3K, half 50K

REL Gain as compared to all 50K

Breed	50K prog	3K prog	Imputed dams
HO	90%	73%	36%
JE	82%	56%	44%
BS	84%	72%	55%

Simulated 500K Genotypes

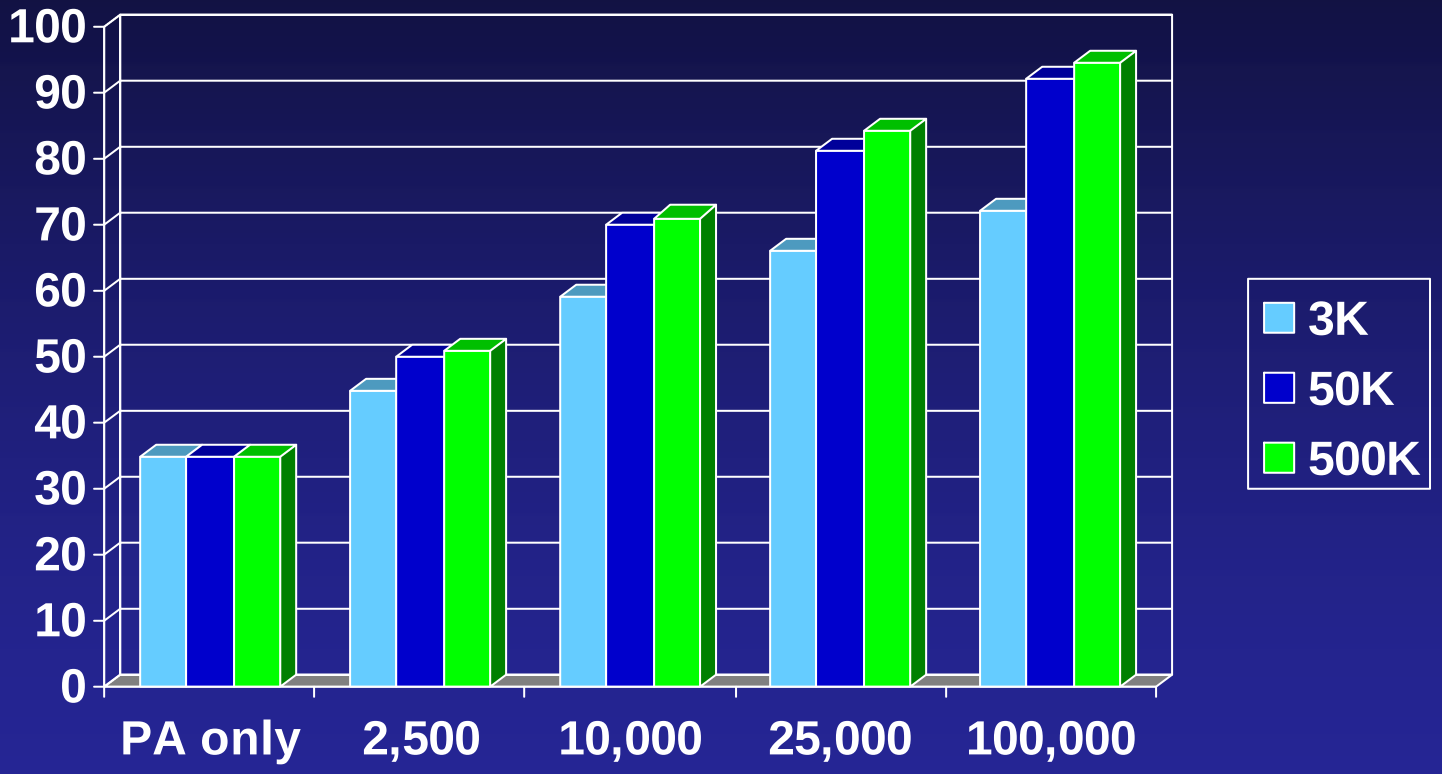
- Linkage in base population
  - Similar to actual linkage reported by:
    - De Roos et al, 2008 Genetics 179:1503
    - Villa-Angulo et al, 2009 BMC Genetics 10:19
  - Underlying linkage corresponds to D’
- Three subsets of mixed 50K and 500K:
  - Of 33,414, only 1,586 (young) had 500K
  - Also bulls > 99% REL, total 3,726
  - Also bulls > 90% REL, total 7,398

Results from 500K Simulation

Density	Single	Mixed			Single
Chips	50K	50K and 500K			500K
Missing	N = 0	1,586	3,726	7,398	33,414
Before	1%	88%	80%	70%	1%
After	.05%	5.3%	2.3%	1.5%	.05%
REL	82.6	83.4	83.6	83.7	84.0

REL Using Only 3K, 50K, or 500K

with increasing numbers of bulls



Conclusions

- Genomic evaluations can mix different chip densities to save $ (or € or ¥)
  - New programs implemented in April 2010
- Only a few thousand of highest density genotypes needed, and other animals imputed
- More animals can be genotyped to increase selection differential and size of reference population

Acknowledgments

- Curt Van Tassell of BFGL selected the 3,209 low density SNP
- Bob Schnabel of U. Missouri fixed map locations for several SNP
- Mel Tooker assisted with computation