# Detecting relationships among genotypes in a rapidly growing collection

**G.R. Wiggans,** G.B. Jansen, L.R. Bacheller, and J.A. Carrillo

Council on Dairy Cattle Breeding
Bowie, Maryland, USA

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

# Background

- CDCB has >6 million genotypes with >1 million added each year
- On receipt, extensive checking done to validate pedigree and discover unreported relatives
- Time required for checks and validating pedigree corrections steadily increases
- Discovered parents, grandsires, and great-grandsires can augment pedigree

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

# Speeding detection of parent-progeny relationships

- Selected 3552 of SNPs present on nearly all chips based on
  - Call rate
  - Mendelian consistency
- Store 3552 SNPs for all genotype (eliminates locating SNPs in common)
- End comparisons at 96 or 1000 SNPs if percentage of opposite homozygous above threshold (12.0, 3.1, 0.5)
- Store discovered genotype pairs

# Details of SNP management

- Store genotypes in 2 bits to minimize storage
- Use SNP values as subscripts in matrix with 0/1 values, where 1 indicates conflict
- Use memory-mapped genotype file to minimize access time

# Determining which genotypes to search

- Include all animals with genotyped progeny
- Exclude non-AI bulls without progeny born in the past 5 years if bull born >5 years ago and genotype not loaded recently
- If both parents confirmed, exclude animals born >5 years before animal
- Otherwise exclude animals born >12 years before animal
- Include only 1 genotype/animal

# Storing discovered genotype pairs

- Usability of genotype determined from conflicts
- Member of a conflicting pair with fewer confirmations typically designated not usable
- Unreported parent-progeny relationships or identical genotype pairs designated as conflicts if not present in pedigree data
- Determination of usability relies on accessing table of stored relatives to check for discovered relationships in place of genotype comparison

# Grandsire unlikely

- Same 3552 SNP set used to determine if MGS or PGS unlikely
- If other parent genotyped, heterozygous call designated as conflict if parent and grandparent are same homozygous call
- Percentage conflict thresholds are 8.0% without other parent; otherwise, 13%
- Unlikely determination removed if haplotype-based discovery confirms pedigree grandsire

# MGS and MGGS discovery

- Imputation provides maternal and paternal haplotypes
- If parent not confirmed, haplotypes compared with those of possible male ancestors to discover MGS and MGGS
- To be designated as discovered, portion of haplotypes in common must be higher by 15% than for bull with next highest value
- Age at birth of progeny must also be reasonable

# Updating pedigree

- Discovered MGS and MGGS added to pedigree of dam and MGD if no pedigree submitted
- Discovered MGS and MGGS reported to nominator and pedigree source otherwise
- If dam or MGD unknown, constructed IDs proposed to enable those ancestors to be stored
- Dams proposed based on herd, calving date, sire, and service sire

# Further steps to speed genotype validation

- Discontinue loading all genotypes in memory
  - Fetch only parent and same animal genotypes from database
  - Memory-mapped file provides genotypes for discovery
- Store genotypes compressed (4 SNPs/byte)
- Develop a program for faster preliminary checks on submissions

# Speed up from changes in discovery

- Adding 3251 new genotypes

| Level | Old (min) | New (min) | Speed up (old/new) |
|---|---|---|---|
| 1 | 110.7 | 22.2 | 5.0 |
| 2 | 35.3 | 8.6 | 4.1 |
| 3 | 22.4 | 0.9 | 26.3 |

- Reprocessing 1461 genotypes

| Old (min) | New (min) | Speed up (old/new) |
|---|---|---|
| 44.0 | 1.9 | 23.8 |

# Summary

- New genotypes compared with existing genotypes to discover parent-progeny and identical relationships
- Pairs identified by genotype IDs (thus independent of animal ID)
- Improved speed of determination of usable genotype when pedigree corrected or genotype reassignment
- MGS and MGGS discovered based on haplotypes in common

# Acknowledgments and disclaimers

- Participating dairy producers supplied pedigree and genomic data
- Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by CDCB
- CDCB is an equal opportunity provider and employer

CDCB
COUNCIL ON DAIRY CATTLE BREEDING