

Computation of Large-Scale Genomic Evaluations

Paul VanRaden

Animal Improvement Programs Laboratory
Agricultural Research Service, USDA
Beltsville, MD

Paul.vanraden@ars.usda.gov

Early genomic theory

- Nejat-Javaremi et al (1997) tested use of genomic relationship matrix in BLUP
- Meuwissen et al (2001) tested linear and nonlinear estimation of haplotype effects
- Both studies assumed that few (<1,000) markers could explain all genetic variance (no polygenic effects in model)
- Polygenic variance was only 5% with 50,000 SNP (VanRaden, 2008), but 50% with 1,000

Multi-step genomic evaluations

- **Traditional evaluations computed first and used as input data to genomic equations**
- **Allele effects estimated for 45,187 markers by multiple regression, assuming equal prior variance**
- **Polygenic effect estimated for genetic variation not captured by markers, assuming pedigree covariance**
- **Selection index step combines genomic info with traditional info from non-genotyped parents**
- **Applied to 30 yield, fitness, calving and type traits**

Single-step genomic evaluation

- **Benefits of 1-step genomic evaluation**
 - **Account for genomic pre-selection**
 - **Expected Mendelian Sampling $\neq 0$**
 - **Improve accuracy and reduce bias**
 - **Include many genotyped animals**
- **Redesign animal model software used since 1989**

Pedigree: Parents, Grandparents, etc.



O-Style

O-Man

Manfred

Jezebel

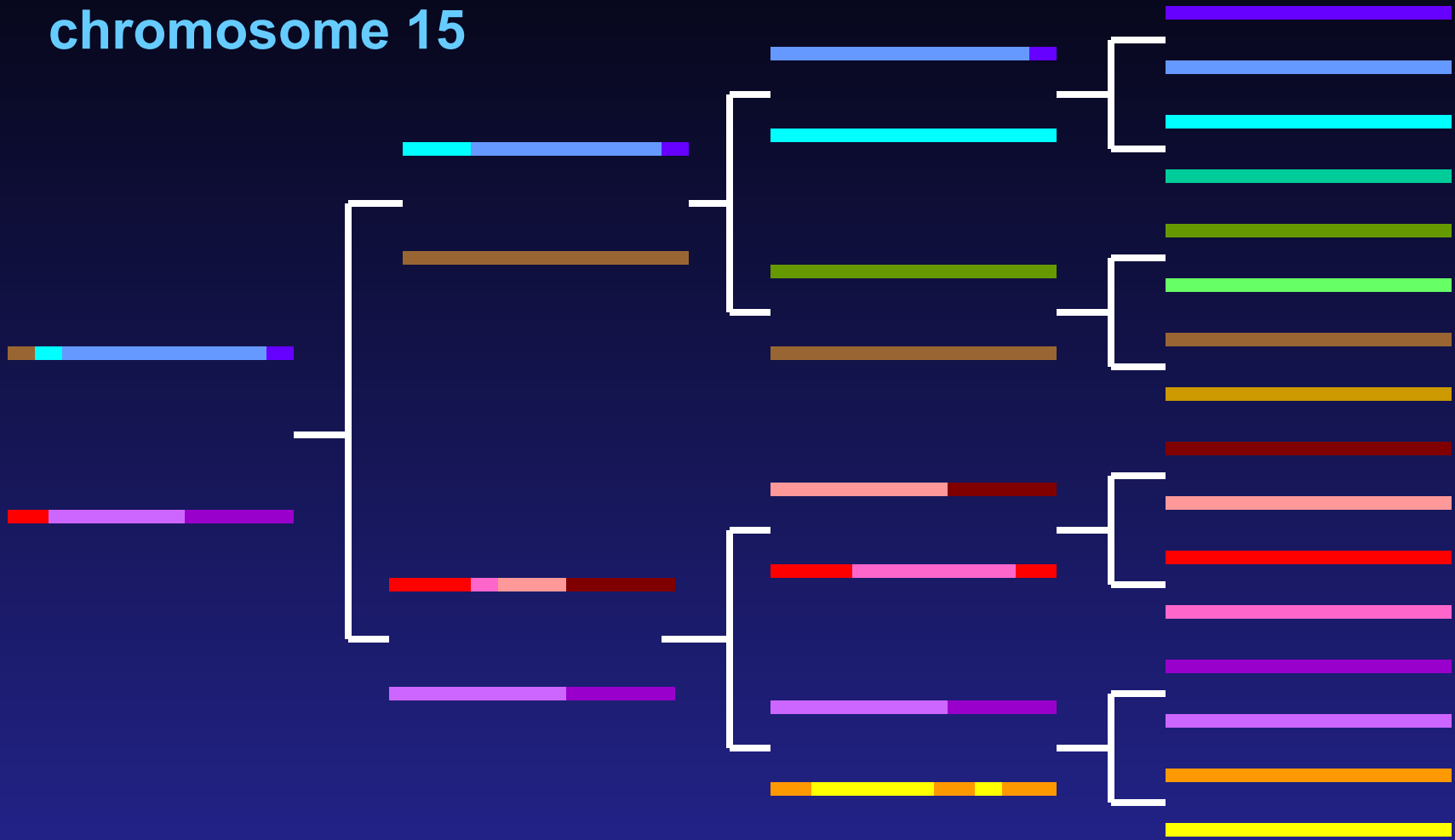
Deva

Teamster

Dima

O-Style Haplotypes

chromosome 15



Expected Relationship Matrix¹

1HO9167 O-Style

| | PGS | PGD | MGS | MGD | Sire | Dam | Bull |
|-----------------|-----|-----|-----|-----|------|-----|------|
| Manfred | 1.0 | .0 | .0 | .0 | .5 | .0 | .25 |
| Jezebel | .0 | 1.0 | .0 | .0 | .5 | .0 | .25 |
| Teamster | .0 | .0 | 1.0 | .0 | .0 | .5 | .25 |
| Dima | .0 | .0 | .0 | 1.0 | .0 | .5 | .25 |
| O-Man | .5 | .5 | .0 | .0 | 1.0 | .0 | .5 |
| Deva | .0 | .0 | .5 | .5 | .0 | 1.0 | .5 |
| O-Style | .25 | .25 | .25 | .25 | .5 | .5 | 1.0 |

¹Calculated assuming that all grandparents are unrelated

Pedigree Relationship Matrix

1HO9167 O-Style

| | PGS | PGD | MGS | MGD | Sire | Dam | Bull |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Manfred | 1.053 | .090 | .090 | .105 | .571 | .098 | .334 |
| Jezebel | .090 | 1.037 | .051 | .099 | .563 | .075 | .319 |
| Teamster | .090 | .051 | 1.035 | .120 | .071 | .578 | .324 |
| Dima | .105 | .099 | .120 | 1.042 | .102 | .581 | .342 |
| O-Man | .571 | .563 | .071 | .102 | 1.045 | .086 | .566 |
| Deva | .098 | .075 | .578 | .581 | .086 | 1.060 | .573 |
| O-Style | .334 | .319 | .324 | .342 | .566 | .573 | 1.043 |

Genomic Relationship Matrix

1HO9167 O-Style

| | PGS | PGD | MGS | MGD | Sire | Dam | Bull |
|----------|-------|-------|-------|-------|-------|-------|-------|
| Manfred | 1.201 | .058 | .050 | .093 | .609 | .054 | .344 |
| Jezebel | .058 | 1.131 | .008 | .135 | .618 | .079 | .357 |
| Teamster | .050 | .008 | 1.110 | .100 | .014 | .613 | .292 |
| Dima | .093 | .135 | .100 | 1.139 | .131 | .610 | .401 |
| O-Man | .609 | .618 | .014 | .131 | 1.166 | .080 | .626 |
| Deva | .054 | .079 | .613 | .610 | .080 | 1.148 | .613 |
| O-Style | .344 | .357 | .292 | .401 | .626 | .613 | 1.157 |

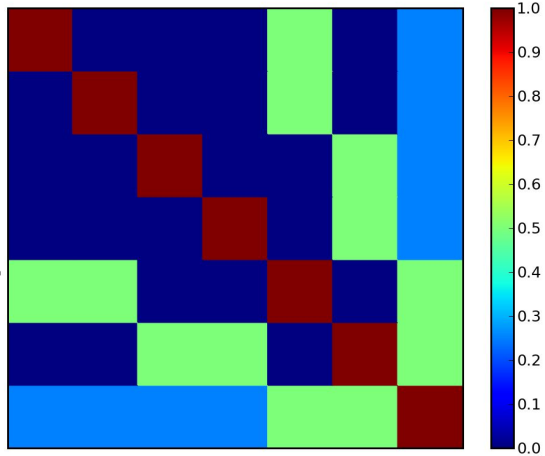
Difference (Genomic – Pedigree)

1HO9167 O-Style

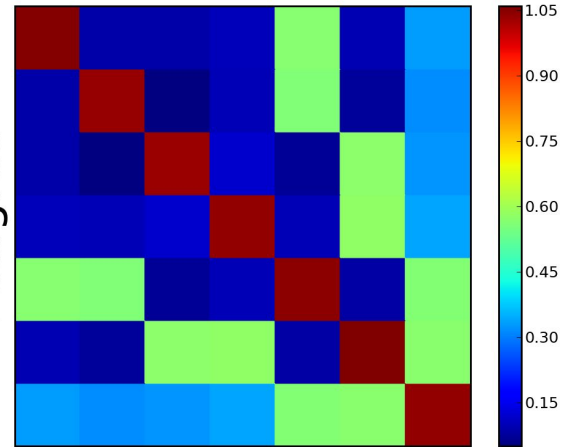
| | PGS | PGD | MGS | MGD | Sire | Dam | Bull |
|-----------------|-------|-------|-------|-------|-------|-------|-------|
| Manfred | .149 | -.032 | -.040 | -.012 | .038 | -.043 | .010 |
| Jezebel | -.032 | .095 | -.043 | .036 | .055 | .004 | .038 |
| Teamster | -.040 | -.043 | .075 | -.021 | -.057 | .035 | -.032 |
| Dima | -.012 | .036 | -.021 | .097 | .029 | .029 | .059 |
| O-Man | .038 | .055 | -.057 | .029 | .121 | -.006 | .060 |
| Deva | -.043 | .004 | .035 | .029 | -.006 | .087 | .040 |
| O-Style | .010 | .038 | -.032 | .059 | .060 | .040 | .114 |

Pseudocolor Plots — O-Style

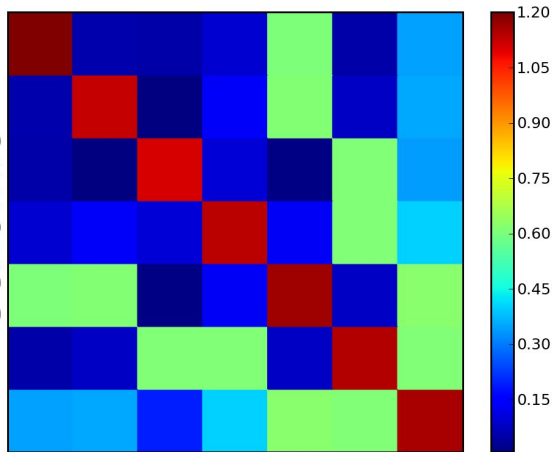
Expected



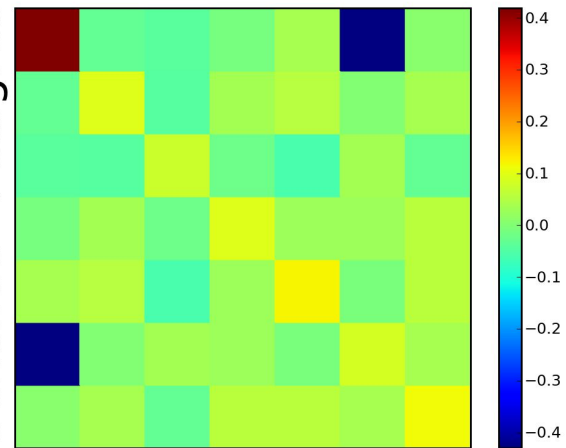
Pedigree



Genomic



Genomic - Pedigree



1 – Step Equations

Aguilar et al., 2010

Model: $y = X b + W u + e$

+ other random effects not shown

$$\begin{bmatrix} X' R^{-1} X & X' R^{-1} W \\ W' R^{-1} X & W' R^{-1} W + H^{-1} k \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X' R^{-1} y \\ W' R^{-1} y \end{bmatrix}$$

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

Size of G and A_{22} >300,000 and doubling each year
Size of A is 60 million animals

Modified 1-Step Equations

Legarra and Ducrocq, 2011

To avoid inverses, add equations for γ , ϕ
 Use math opposite of absorbing effects

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}W & 0 & 0 \\ W'R^{-1}X & W'R^{-1}W+A^{-1}k & Q & Q \\ 0 & Q' & -G/k & 0 \\ 0 & Q' & 0 & A_{22}/k \end{bmatrix} \begin{bmatrix} b \\ u \\ \gamma \\ \phi \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ W'R^{-1}y \\ 0 \\ 0 \end{bmatrix}$$

Iterate for γ using $G = Z Z' / [2 \Sigma p(1-p)]$

Iterate for ϕ using A_{22} multiply (Colleau)

$Q' = [0 \ I]$ (I for genotyped animals)

Genomic Algorithms Tested

- **1-step** genomic model
 - Add extra equations for γ and ϕ (Legarra and Ducrocq)
 - Converged ok for JE, bad for HO
 - Extended to MT using block diagonal
 - Invert 3x3 $A^{-1}u$, $G\gamma$, $-A_{22}\phi$ blocks? **NO**
 - PCG iteration (hard to debug) **Maybe**

Genomic Algorithms (continued)

- **Multi-step** insertion of GEBV
 - $[W'R^{-1}W + A^{-1}k] u = W'R^{-1}y$ (without G)
 - Previous studies added genomic information to $W'R^{-1}W$ and $W'R^{-1}y$
 - Instead: insert GEBV into u , iterate
- **1-step** genomic model using DYD
 - Solve SNP equations from DYD & YD
 - May converge faster, but approximate

Data for 1-Step Test

- **National U.S. Jersey data**
 - **4.4 million lactation phenotypes**
 - **4.1 million animals in pedigree**
 - **Multi-trait milk, fat, protein yields**
 - **5,364 male, 11,488 female genotypes**
- **Deregressed MACE evaluations for 7,072 bulls with foreign daughters (foreign dams not yet included)**

Jersey Results

New = 1-step GPTA milk, Old = multi-step GPTA milk

| Statistic | Animals | |
|------------------------------|-----------------|-------|
| Corr(New, Old) | All bulls | 0.994 |
| Corr(New, Old) | Genotyped bulls | 0.992 |
| Corr(DYD _g , DYD) | Genotyped bulls | 0.999 |
| Corr(New, Old) | Young genomic | 0.966 |
| SD old PTA milk | Young genomic | 540 |
| SD new PTA milk | Young genomic | 552 |
| Old milk trend | 1995-2005 cows | 1644 |
| New milk trend | 1995-2005 cows | 1430 |

1-Step vs Multi-Step: Results

Data cutoff in August 2008

| Evaluation | Regression | Squared Correlation |
|-------------------|-------------------|--------------------------------|
| Parent Average | .73 | .436 |
| Multi-Step GPTA | .75 | .520 |
| 1-Step GPTA | .85 | .520 |
| Expected | .93 | |

Multi-step regressions also improved by modified selection index weights

Computation Required

- CPU time for 3 trait ST model
 - JE took **11 sec** / round including G
 - HO took **1.6 min** / round including G
 - JE needed **~1000** rounds (3 hours)
 - HO needed **>5000** rounds (>5 days)
- Memory required for HO
 - **30 Gigabytes** (256 available)

Remaining Issues

- **Difficult to match G and A across breeds**
- **Nonlinear model (Bayes A) possible with SNP effect algorithm**
- **Interbull validation not designed for genomic models**
- **MACE results may become biased**

Steps to prepare genotypes

- **Nominate animal for genotyping**
- **Collect blood, hair, semen, nasal swab, or ear punch**
 - **Blood may not be suitable for twins**
- **Extract DNA at laboratory**
- **Prepare DNA and apply to BeadChip**
- **Do amplification and hybridization, 3-day process**
- **Read red/green intensities from chip and call genotypes from clusters**

Ancestor Validation and Discovery

- **Ancestor discovery can accurately confirm, correct, or discover parents and more distant ancestors for most dairy animals because most sires are genotyped.**
- **Animal checked against all candidates**
- **SNP test and haplotype test both used**
- **Parents and MGS are suggested to breed associations and breeders since December 2011 to improve pedigrees.**

Ancestor Discovery Results by Breed

| Breed | SNP Test | Haplotype Test | |
|-------------|----------------------|----------------|-------------|
| | MGS | MGS | MGGS |
| | % Confirmed* | % Confirmed | % Confirmed |
| Holstein | 95 (98) [†] | 97 | 92 |
| Jersey | 91 (92) | 95 | 95 |
| Brown Swiss | 94 (95) | 97 | 85 |

***Confirmation = top MGS candidate matched true pedigree MGS.**

[†]50K genotyped animals only.

Data (Yield and Health)

- **One step model includes:**
 - **72 million lactation phenotypes**
 - **50 million animals in pedigree**
 - **29 million permanent environment**
 - **7 million herd mgmt groups**
 - **11 million herd by sire interactions**
 - **7 traits: Milk, Fat, Protein, SCS, longevity, fertility**
 - **Genotypes not yet included**

New Features Added

- **Model options now include:**
 - **Multi-trait models**
 - **Multiple class and regress variables**
 - **Suppress some factors / each trait**
 - **Random regressions**
 - **Foreign data**
 - **Parallel processing**
 - **Genomic information**
- **Renumber factors in same program**

Computation Required: Evaluation

- CPU for all-breed model (7 traits)
 - ST: **4 min** / round with 7 processors and **~1000** rounds (2.8 days)
 - MT: **15 min** / round and **~1000** rounds
 - **~200** rounds for updates using priors
 - Little extra cost to include foreign
- Memory required
 - ST or MT: **32 Gbytes** (**256** available)

Computation Required: Imputation

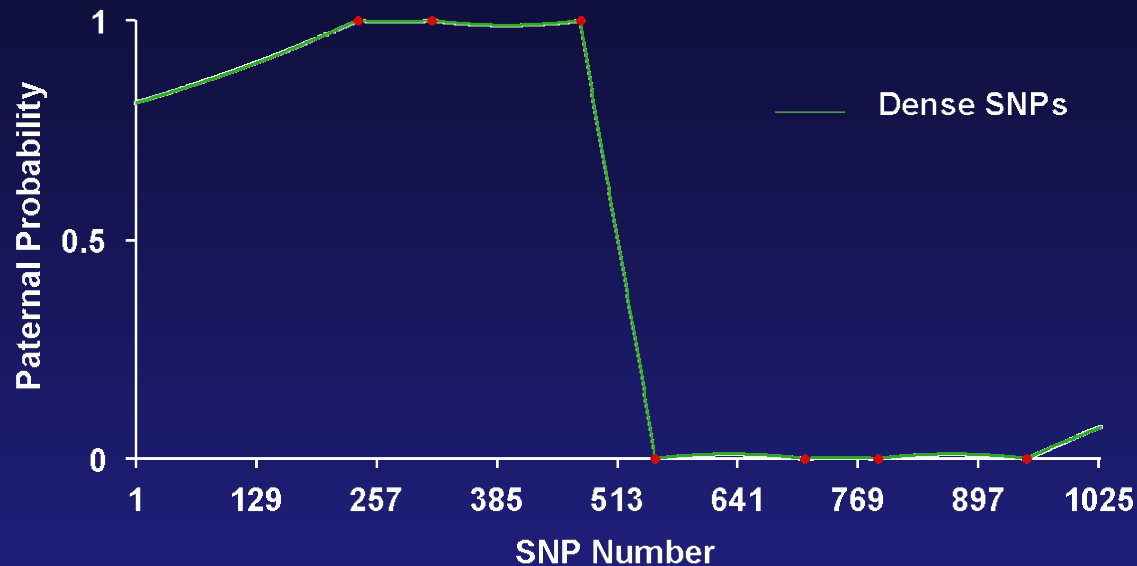
- **Impute 636,967 markers for 103,070 animals**
 - **Required 10 hours with 6 processors (**findhap**)**
 - **Required 50 Gbytes memory**
 - **Program **FImpute** from U. Guelph slightly better**
- **Impute 1 million markers on 1 chromosome (sequences) for 1,000 animals**
 - **Required 15 minutes with 6 processors**
 - **Required 4 Gbytes memory**

Methods to Trace Inheritance

- **Few** markers
 - Pedigree needed
 - Prob (paternal or maternal alleles inherited) computed within families
- **Many** markers
 - Can find matching DNA segments without pedigree
 - Prob (haplotypes are identical) mostly near 0 or 1 if segments contain many markers

Haplotype Probabilities

with Few Markers (12 SNP / chromosome)



Haplotype Probabilities

with More Markers (50 SNP / chromosome)



Haplotyping Program: findhap.f90

- **Population haplotyping**
 - Divide chromosomes into segments
 - List haplotypes by genotype match
 - Similar to FastPhase, IMPUTE, or long range phasing
- **Pedigree haplotyping**
 - Look up parent or grandparent haplotypes
 - Detect crossovers, fix noninheritance
 - Impute nongenotyped ancestors

Coding of Alleles and Segments

- **Genotypes**

- 0 = **BB**, 1 = **AB** or **BA**, 2 = **AA**, 5 = **__** (missing)
- Allele frequency used for missing

- **Haplotypes**

- 0 = **B**, 1 = **not known**, 2 = **A**

- **Segment inheritance (example)**

- Son has haplotype numbers **5** and **8**
- Sire has haplotype numbers **8** and **21**
- Son got haplotype number **5** from dam

Population Haplotyping Steps

- Put first genotype into haplotype list
- Check next genotype against list
 - Do any homozygous loci conflict?
 - If haplotype conflicts, continue search
 - If match, fill any unknown SNP with homozygote
 - 2nd haplotype = genotype minus 1st haplotype
 - Search for 2nd haplotype in rest of list
 - If no match in list, add to end of list
- Sort list to put frequent haplotypes 1st

Check New Genotype Against List

1st segment of chromosome 15

Search for 1st haplotype that matches genotype:

022112222011221022021110220010110212202000102020120002021

5.16% 022222222020020022002020200020000200202000022022222202220
4.37% 022020220202200020022022200002200200200000200222200002202
4.36% 02202022202200200022020220000220202200002200222200202220
3.67% 02202022202022200202202220202000022220000200002020002002
3.66% 0222222220202220222020200220000020222202000002020220002022

Get 2nd haplotype by removing 1st from genotype:

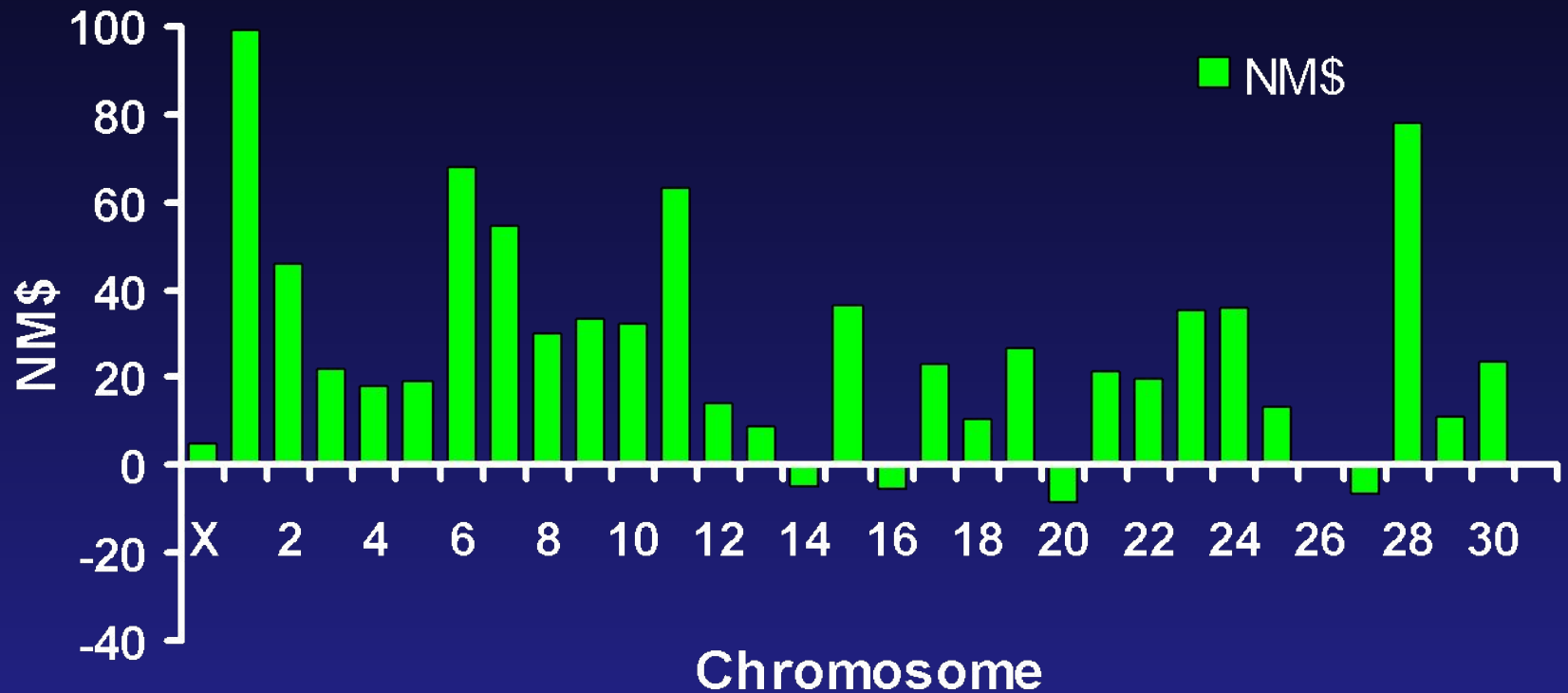
02200222200222002202202022002020020220200020202002000202

0
3.65%
022020022202200200022020220000220202200002200222200202222
3.51%
02200222202022202202022020200222002200000002022220002220
3.42%
022002222002220022022020220020200202202000202020020002020
3.24%
022222222020200000022020220020200202202000202020020002020



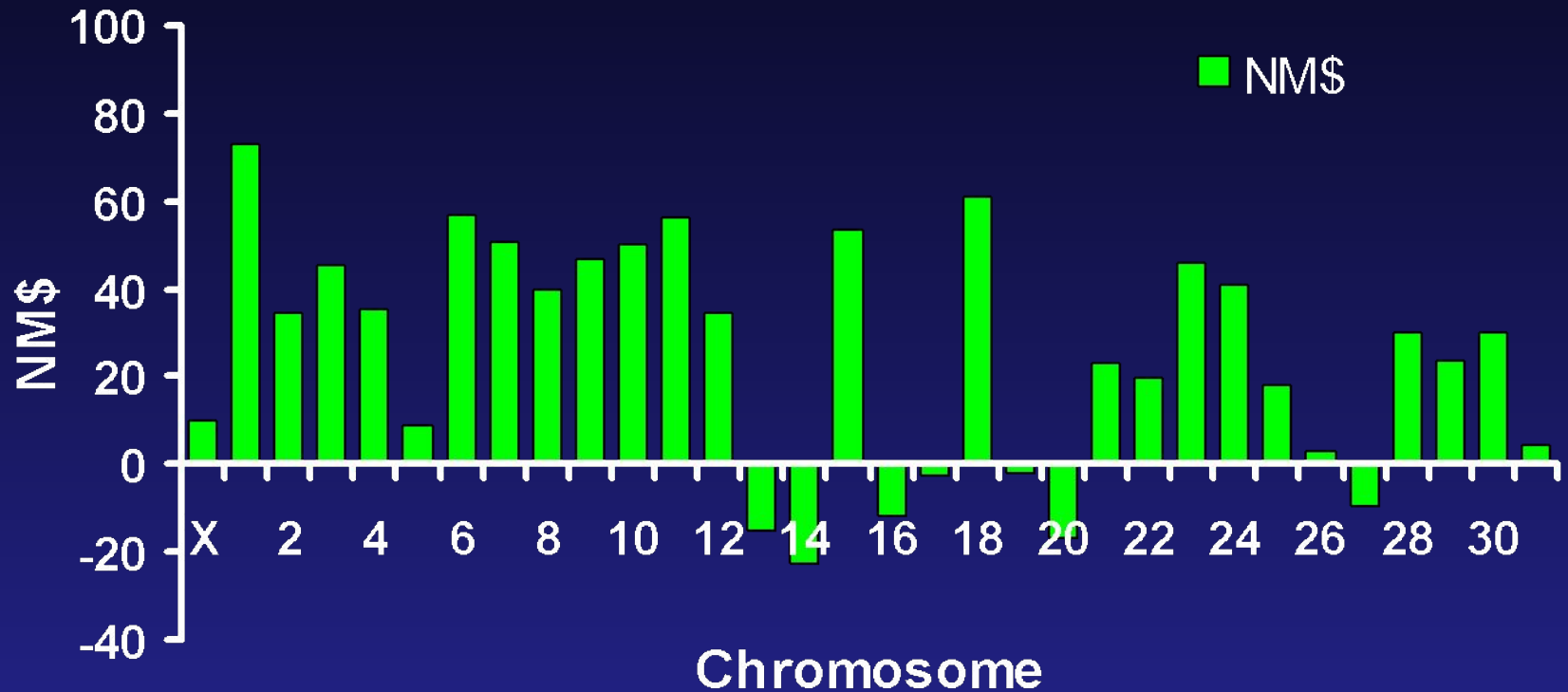
Net Merit by Chromosome

Freddie - highest Net Merit bull



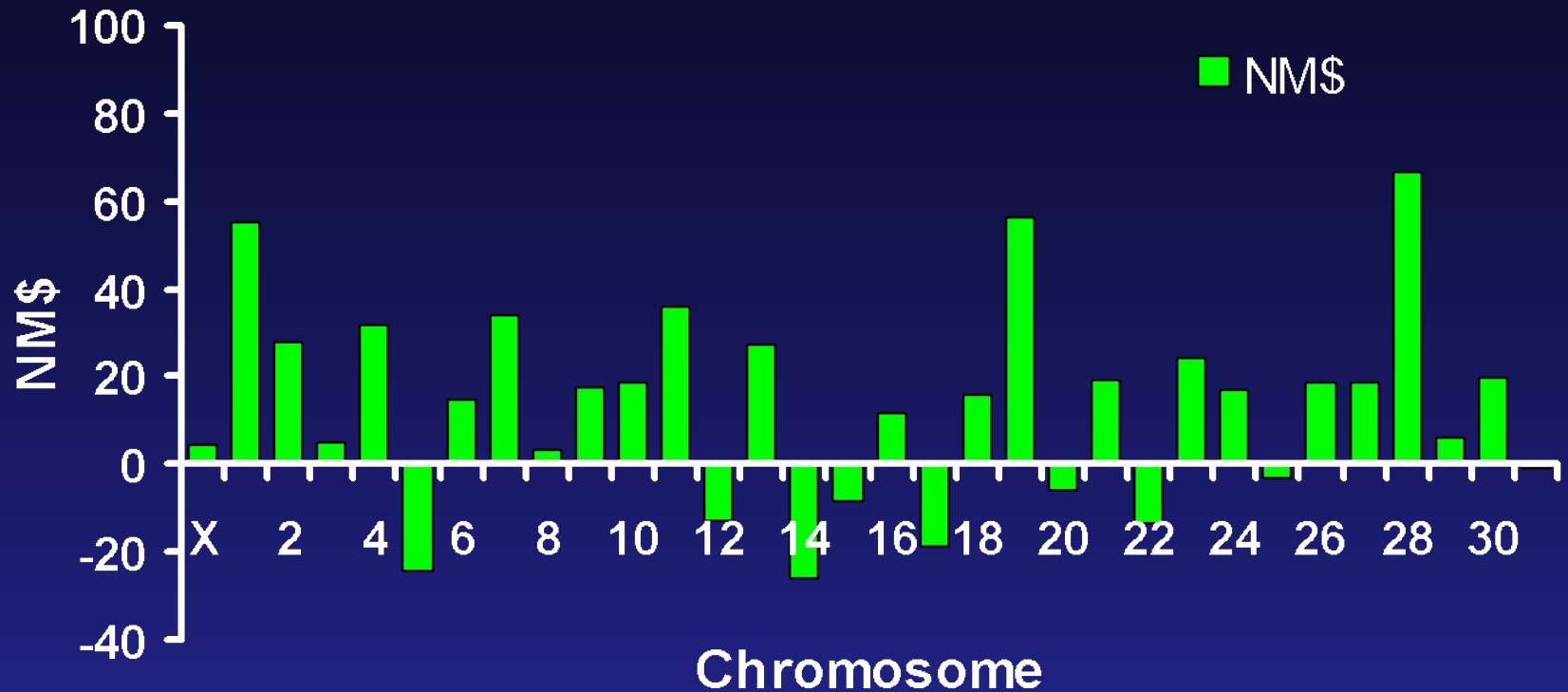
Net Merit by Chromosome

Man – Sire of Freddie



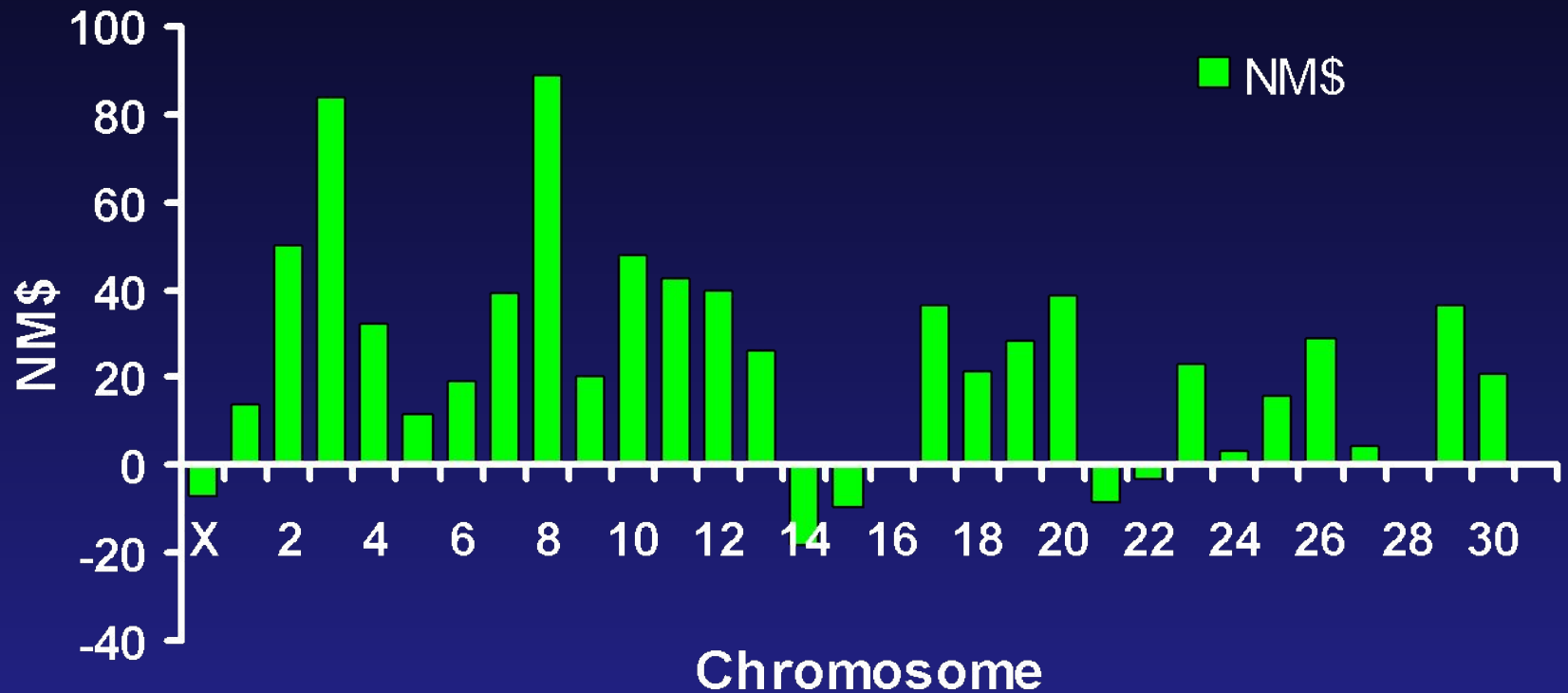
Net Merit by Chromosome

Die-Hard - maternal grandsire

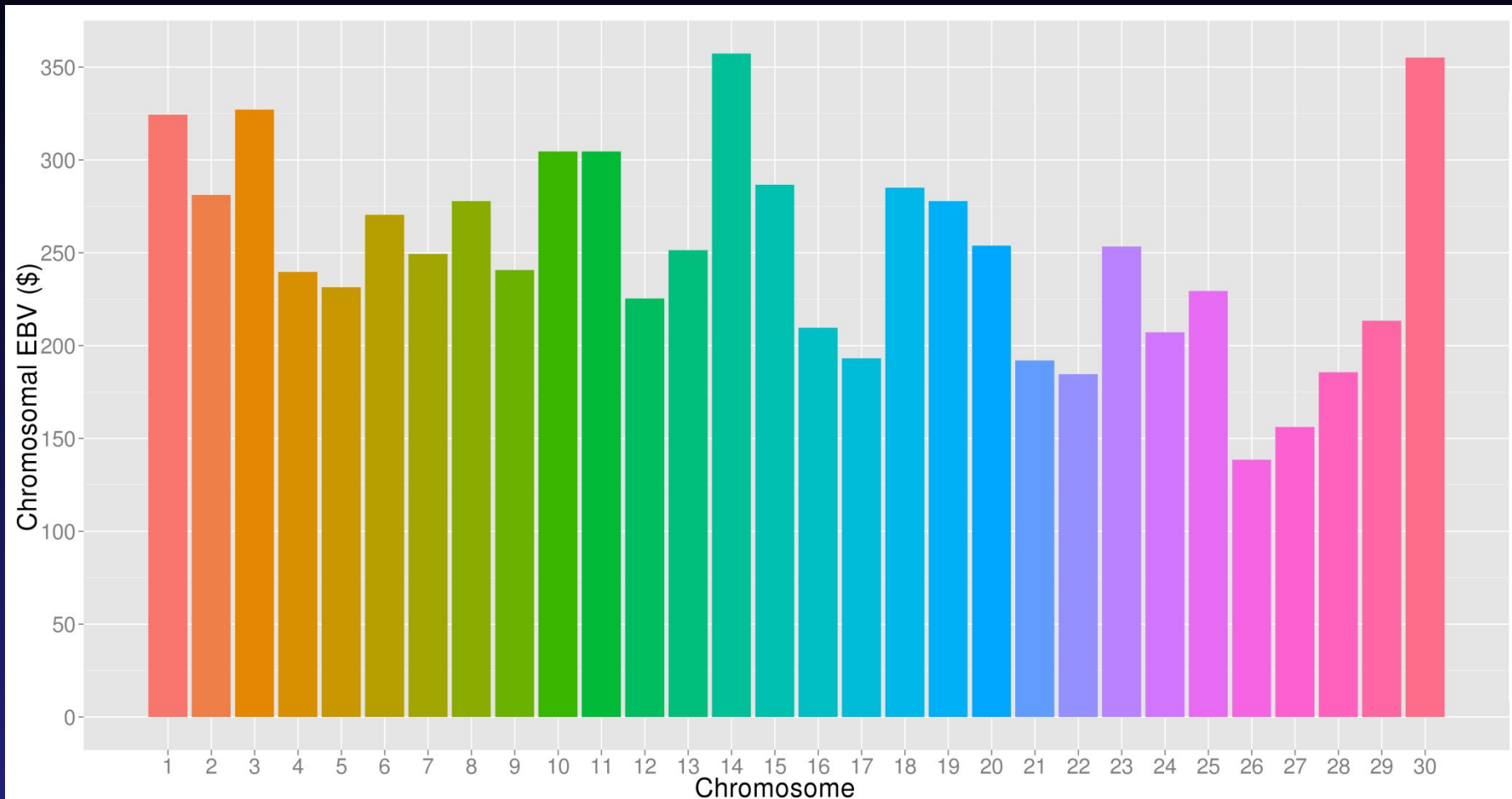


Net Merit by Chromosome

Planet – high Net Merit bull



What's the best cow we can make?



A “Supercow” constructed from the best haplotypes in the Holstein population would have an EBV(NM\$) of **\$7515**

Conclusions

- **1-step genomic evaluations tested**
 - **Inversion avoided using extra equations**
 - **Converged well for JE but not for HO**
 - **Same accuracy, less bias than multi-step**
 - **Foreign data from MACE included**
- **Further work needed on algorithms**
 - **Including genomic information**
 - **Extending to all-breed evaluation**

Conclusions

- **Foreign data can add to national evaluations**
 - **In one step model instead of post-process**
 - **High correlations of national with MACE**
- **Multi-trait all-breed model developed**
 - **Replace software used since 1989**
 - **Many new features added**
 - **Correlations $\sim .99$ with traditional AM**
 - **Tested with 7 yield and health traits**
 - **Also tested with 14 JE conformation traits**

Acknowledgments

- **George Wiggans, Ignacy Misztal, and Andres Legara provided advice on algorithms**
- **Mel Tooker, Tabatha Cooper, and Jan Wright assisted with computation, program design, and ancestor discovery**
- **Members of the Council on Dairy Cattle Breeding provided data**