Animal breeding software lessons: 1926-2026

Paul VanRaden, retired

Previously: USDA, Agricultural Research Service, Animal Genomics and Improvement Laboratory, Beltsville, MD 20705

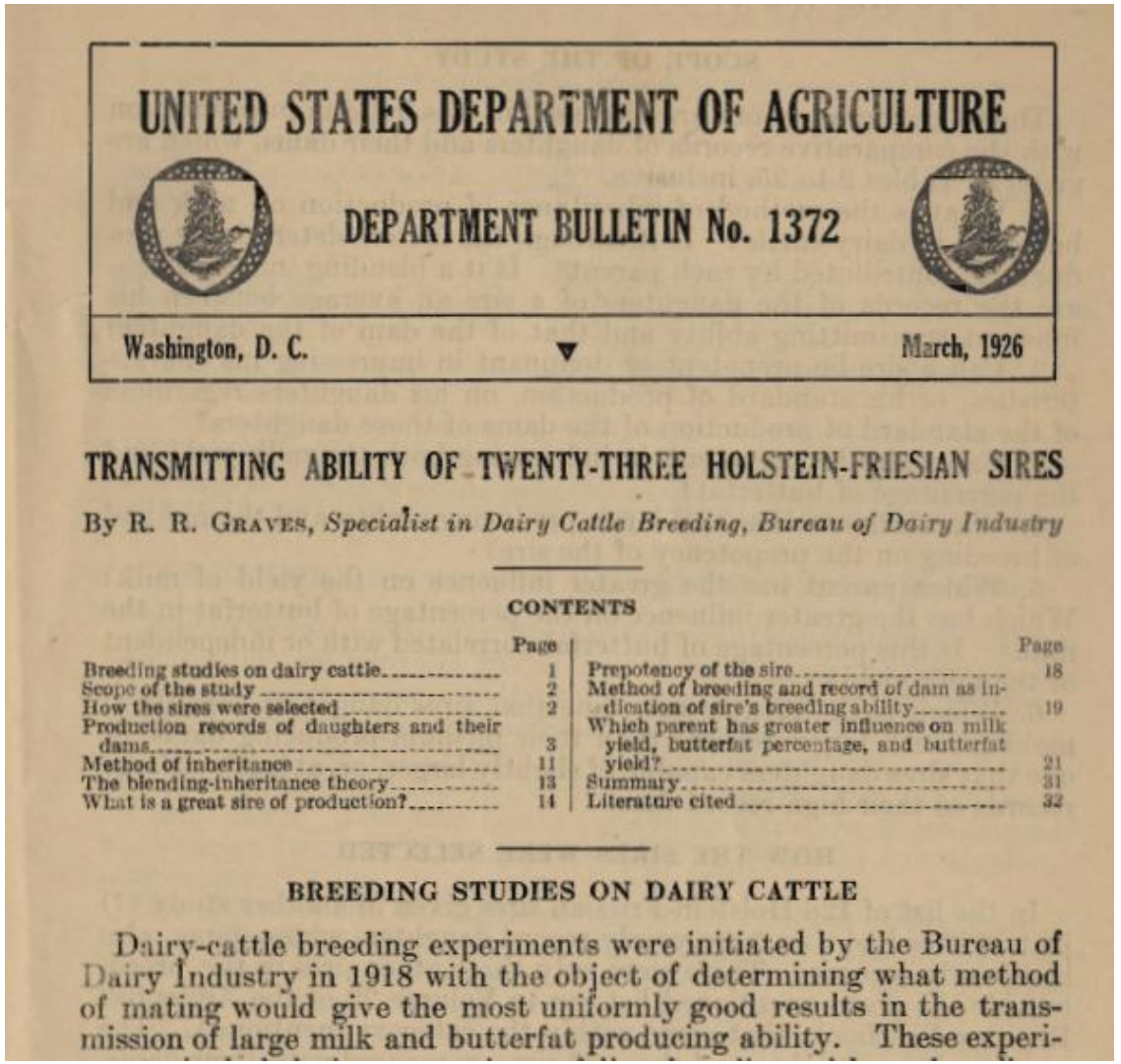
vanraden@aol.com

Introduction

- In past century, genetic research used mainly pedigrees
  - Using math derived for cows in 1922 at USDA-Beltsville
  - Scientists could write their own code or use packages
- DNA can be read very affordably (~$30 for 50,000 loci)
  - Chips developed after human genome project
  - Methods transferable to almost all species
  - Cows were second species in 2007
- Now most genetic research and selection reads DNA directly

USDA PTAs calculated since 1926

- Graves, R. R. 1926.
- Transmitting ability of twenty-three Holstein-Friesian sires.
- USDA Dept. Bull. 1372.
- Daughter-dam comparison
  - Mean yield of daughter's minus dam's yield
  - Method used until 1962



UNITED STATES DEPARTMENT OF AGRICULTURE

DEPARTMENT BULLETIN No. 1372

Washington, D. C. ▼ March, 1926

TRANSMITTING ABILITY OF TWENTY-THREE HOLSTEIN-FRIESIAN SIRES

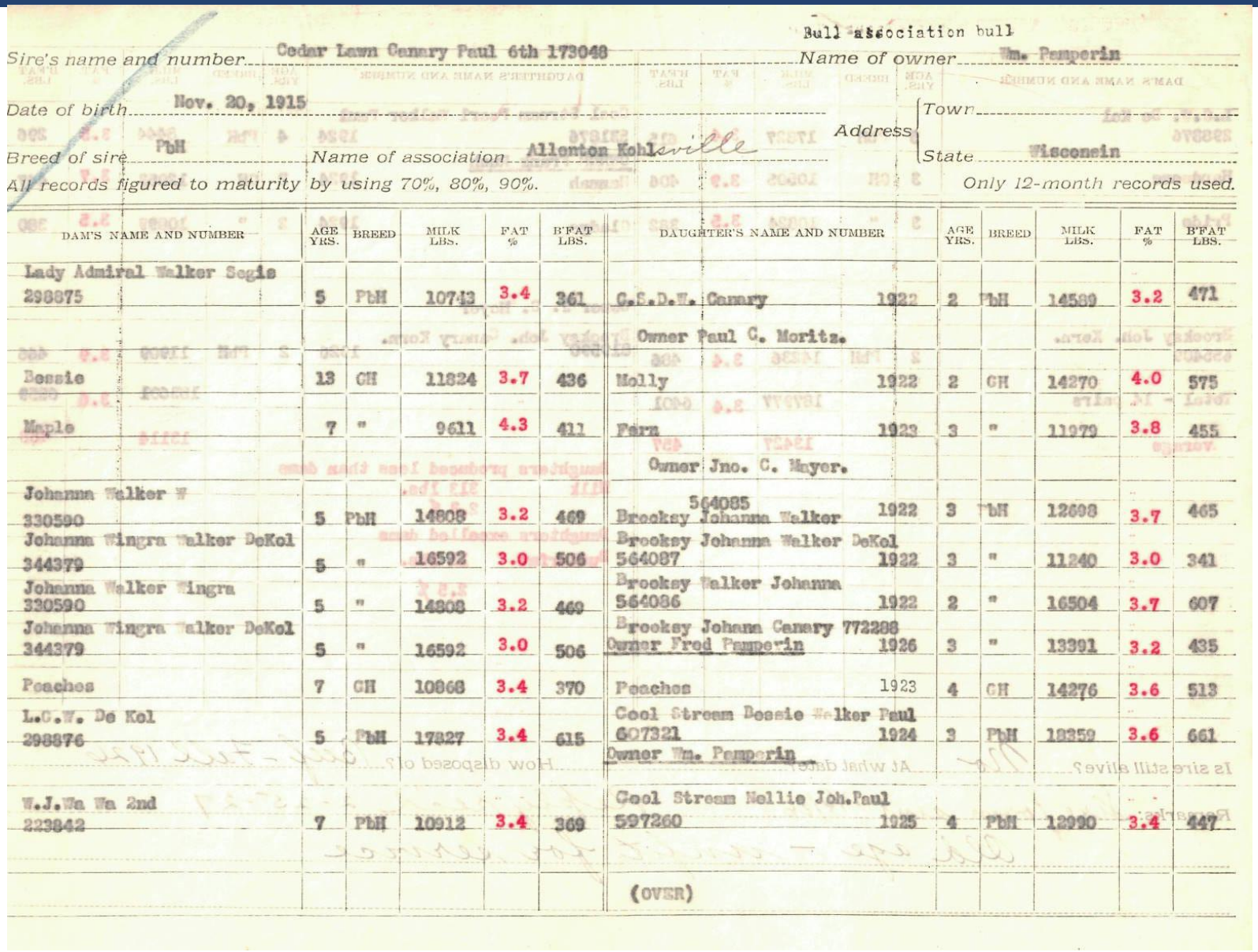
By R. R. GRAVES, *Specialist in Dairy Cattle Breeding, Bureau of Dairy Industry*

CONTENTS

BREEDING STUDIES ON DAIRY CATTLE

Dairy-cattle breeding experiments were initiated by the Bureau of Dairy Industry in 1918 with the object of determining what method of mating would give the most uniformly good results in the transmission of large milk and butterfat producing ability. These experi-

First USDA PTA sent to WI in 1926 (1 of the 23 bulls)



Sire's name and number Cedar Lawn Canary Paul 6th 173048 — Bull association bull — Name of owner Wm. Pemperin

Date of birth Nov. 20, 1915 — Address Town

Breed of sire PbH — Name of association Allenton Kohlsville — State Wisconsin

All records figured to maturity by using 70%, 80%, 90%. — Only 12-month records used.

DAM'S NAME AND NUMBER	AGE YRS.	BREED	MILK LBS.	FAT %	B'FAT LBS.	DAUGHTER'S NAME AND NUMBER	AGE YRS.	BREED	MILK LBS.	FAT %	B'FAT LBS.
Lady Admiral Walker Segis 298875	5	PbH	10743	3.4	361	C.L.D.W. Canary 1922	2	PbH	14589	3.2	471
						Owner Paul C. Moritz.					
Bessie	13	GH	11824	3.7	436	Molly 1922	2	GH	14270	4.0	575
Maple	7	"	9611	4.3	411	Fern 1923	3	"	11979	3.8	455
						Owner Jno. C. Mayer.					
Johanna Walker W 330590	5	PbH	14808	3.2	469	564085 Brooksy Johanna Walker 1922	3	PbH	12698	3.7	465
Johanna Wingra Walker DeKol 344379	5	"	16592	3.0	506	Brooksy Johanna Walker DeKol 564087 1922	3	"	11240	3.0	341
Johanna Walker Wingra 330590	5	"	14808	3.2	469	Brooksy Walker Johanna 564086 1922	2	"	16504	3.7	607
Johanna Wingra Walker DeKol 344379	5	"	16592	3.0	506	Brooksy Johann Canary 772288 Owner Fred Pemperin 1926	3	"	13391	3.2	435
Peaches	7	GH	10868	3.4	370	Peaches 1923	4	GH	14276	3.6	513
L.C.W. De Kol 298876	5	PbH	17827	3.4	615	Cool Stream Bessie Walker Paul 607321 1924 Owner Wm. Pemperin	3	PbH	18359	3.6	661
W.J.Wa Wa 2nd 223842	7	PbH	10912	3.4	369	Cool Stream Nellie Joh.Paul 597260 1925	4	PbH	12990	3.4	447
						(OVER)					

Cedar Lawn Canary Paul 6th

Born: Nov 20, 1915
Sold for beef in Fall, 1926

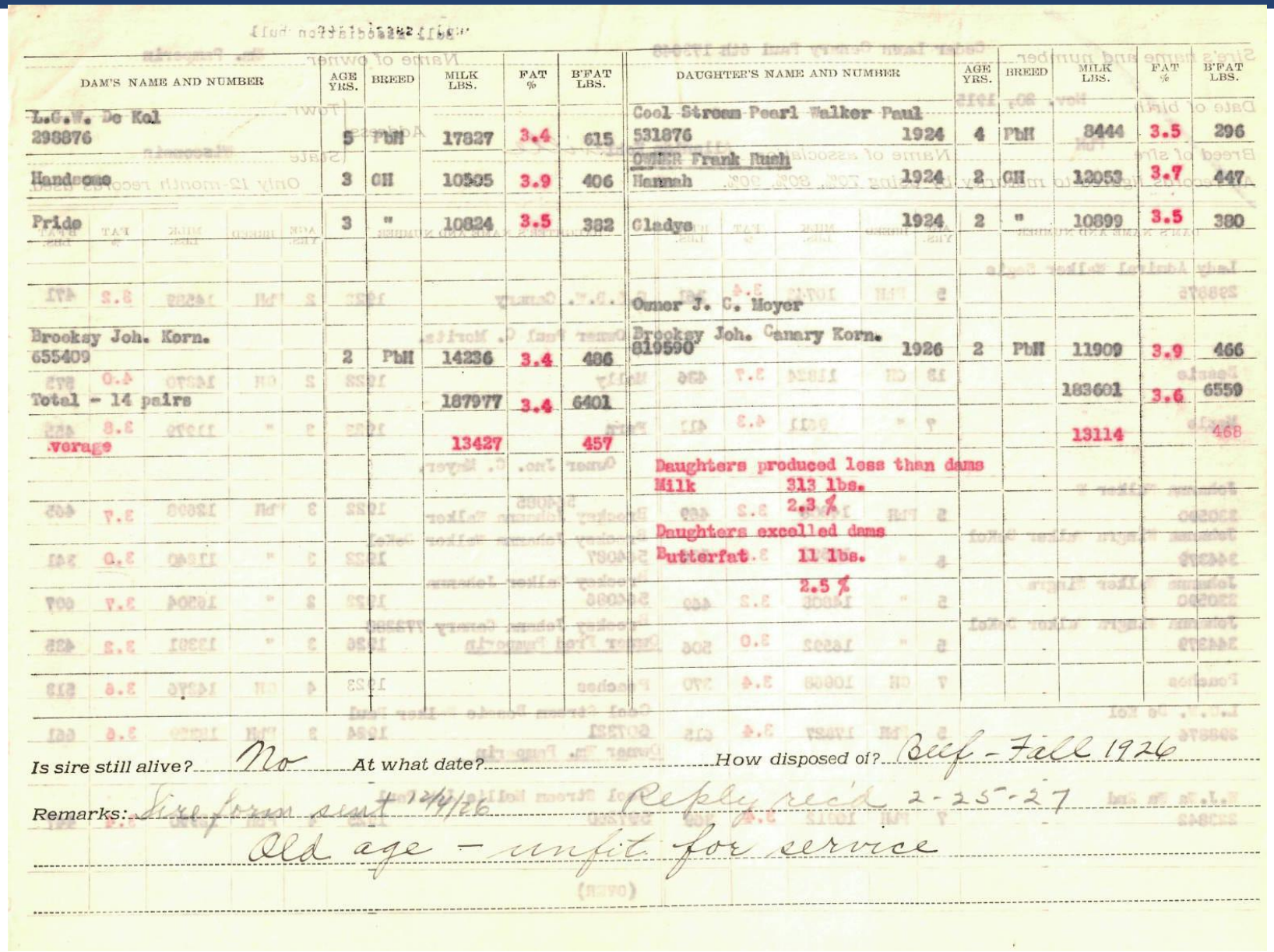
HO registration: 173048

Owned by: Wm. Pemperin
Bull Association: Allenton Kohlsville

The farm began in 1847 in the West Bend area. In 1992 they moved to Farmington at the crossroads of Cheeseville. They are a family dairy farm with 230 dairy cows.

George and Kathy Muth, current owners

Page 2 of the first USDA PTA sent to WI in 1926



DAM'S NAME AND NUMBER	AGE YRS.	BREED	MILK LBS.	FAT %	B'FAT LBS.	DAUGHTER'S NAME AND NUMBER	AGE YRS.	BREED	MILK LBS.	FAT %	B'FAT LBS.
L.G.W. De Kol 298876	5	PbH	17827	3.4	615	Cool Stream Pearl Walker Paul 531876 1924	4	PbH	8444	3.5	296
Handsome	3	GH	10505	3.9	406	OWNER Frank Rush Hannah 1924	2	GH	12053	3.7	447
Pride	3	"	10824	3.5	382	Cladys 1924	2	"	10899	3.5	380
						Owner J. C. Hoyer					
Brooksy Joh. Korn. 655409	2	PbH	14236	3.4	486	Brooksy Joh. Canary Korn. 819590 1926	2	PbH	11909	3.9	466
Total - 14 pairs			187977	3.4	6401				183601	3.6	6559
Average			13427		457				13114		468
						Daughters produced less than dams Milk 313 lbs. 2.3 %					
						Daughters excelled dams Butterfat 11 lbs. 2.5 %					

Is sire still alive? No At what date? How disposed of? Beef - Fall 1926

Remarks: Sire form sent 12/4/26 Reply rec'd 2-25-27

Old age - unfit for service

14 daughter - dam pairs

Daughters produced less than dams for Milk but excelled dams for Butterfat.

Bull PTA:

-313 lbs. Milk

+11 lbs. Butterfat

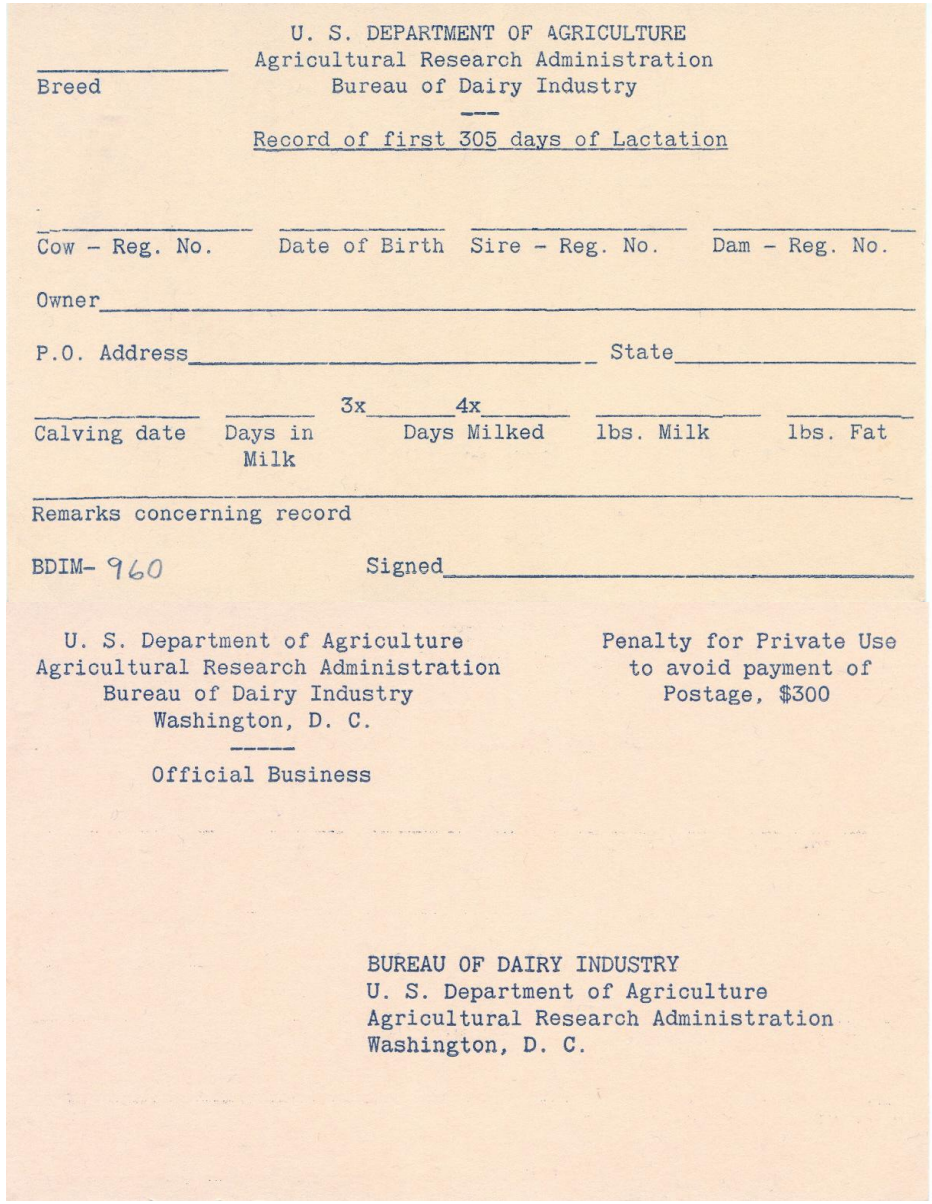
Sire form sent 12-4-1926

Owner replied 2-25-1927

Is sire still alive? No

Sire sold for: Old age – unfit for service

Postcards used to send the lactation records



U. S. DEPARTMENT OF AGRICULTURE
Agricultural Research Administration
Bureau of Dairy Industry

Breed

Record of first 305 days of Lactation

Cow – Reg. No. | Date of Birth | Sire – Reg. No. | Dam – Reg. No.

Owner

P.O. Address _____ State

Calving date | Days in Milk | 3x | 4x Days Milked | lbs. Milk | lbs. Fat

Remarks concerning record

BDIM- 960 Signed

U. S. Department of Agriculture
Agricultural Research Administration
Bureau of Dairy Industry
Washington, D. C.

Official Business

Penalty for Private Use
to avoid payment of
Postage, $300

BUREAU OF DAIRY INDUSTRY
U. S. Department of Agriculture
Agricultural Research Administration
Washington, D. C.

Format 4 Postcard
(305-d lactation record)

BDIM – 960

Used in 1930s

Postage paid by USDA

Calculated by hand in USDA South Bldg, 1926-1959

1922 pedigree relationship matrix – USDA, Beltsville

Wright, S. 1922. Coefficients of inbreeding and relationship. The American Naturalist 56:330-338.

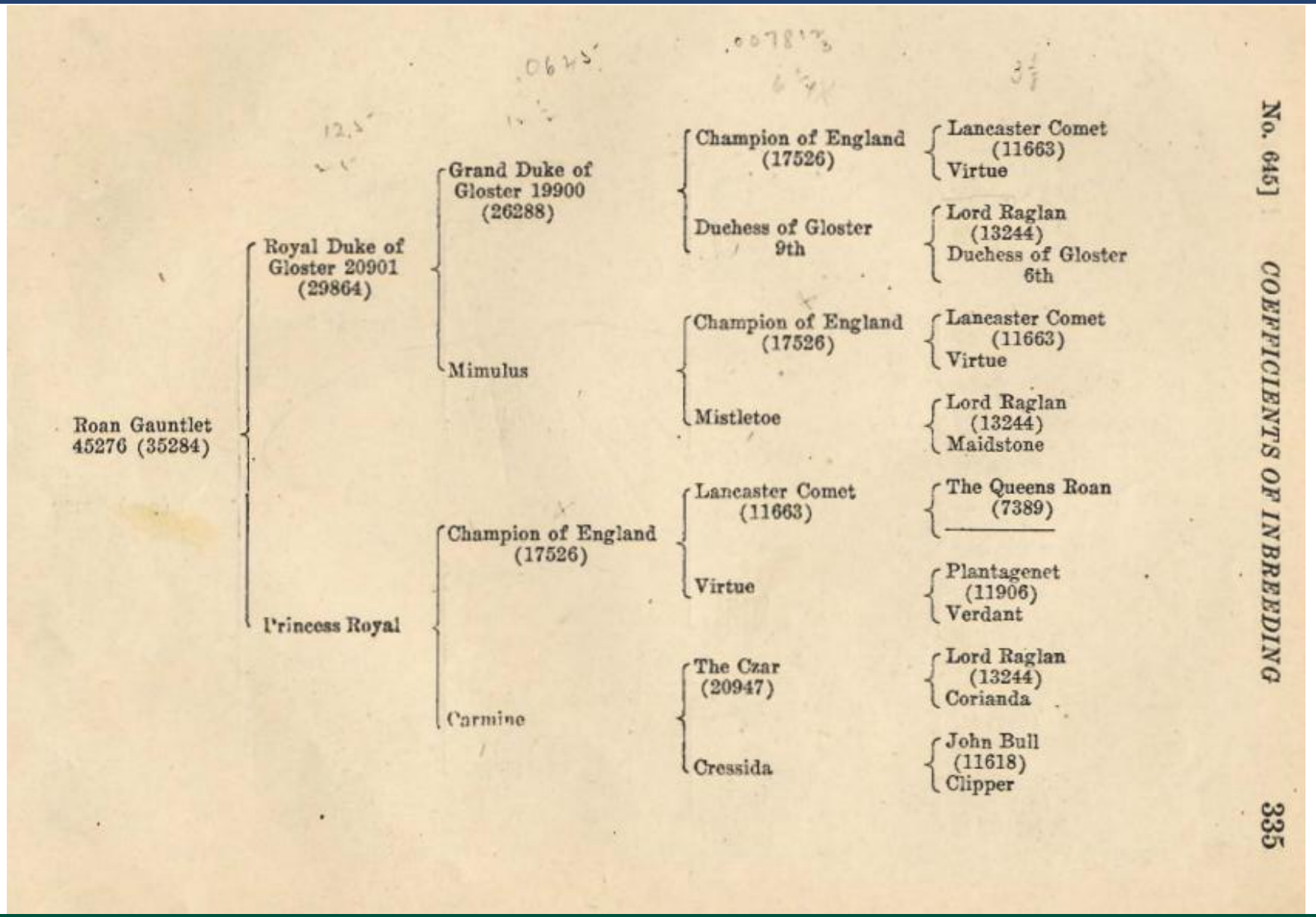
COEFFICIENTS OF INBREEDING AND RELATIONSHIP

DR. SEWALL WRIGHT

BUREAU OF ANIMAL INDUSTRY, UNITED STATES DEPARTMENT OF AGRICULTURE

IN the breeding of domestic animals consanguineous matings are frequently made. Occasionally matings are made between very close relatives—sire and daughter, brother and sister, etc.—but as a rule such close inbreeding is avoided and there is instead an attempt to concentrate the blood of some noteworthy individual by what is known as line breeding. No regular system of mating such as might be followed with laboratory animals is practicable as a rule.

1922 Example pedigree: Shorthorn cattle



No. 645] COEFFICIENTS OF INBREEDING 335

Roan Gauntlet 45276 (35284)

- Royal Duke of Gloster 20901 (29864)
  - Grand Duke of Gloster 19900 (26288)
    - Champion of England (17526)
      - Lancaster Comet (11663)
      - Virtue
    - Duchess of Gloster 9th
      - Lord Raglan (13244)
      - Duchess of Gloster 6th
  - Mimulus
    - Champion of England (17526)
      - Lancaster Comet (11663)
      - Virtue
    - Mistletoe
      - Lord Raglan (13244)
      - Maidstone
- Princess Royal
  - Champion of England (17526)
    - Lancaster Comet (11663)
      - The Queens Roan (7389)
      - ———
    - Virtue
      - Plantagenet (11906)
      - Verdant
  - Carmine
    - The Czar (20947)
      - Lord Raglan (13244)
      - Corianda
    - Cressida
      - John Bull (11618)
      - Clipper

Pedigrees

- Computer storage since 1960, herdbook pedigrees back to 1900
- Sources from industry:
  - Breed associations, DHI, NAAB (bulls), Interbull (foreign)
  - Genotype suppliers
  - DNA discovery of ancestors
  - Report conflicts and errors back to pedigree suppliers
- Conflicts between sources and with DNA must be resolved!

1989 USDA animal model

- Programs by George Wiggans with help from Ignacy
  - Main limitation was memory (500 MB)
  - Data sorted by superherd, herd, sire, cow, calving date
  - Compact data files read and written in each iteration
  - 10 million cows with data
- Single-trait milk and fat, only protein could have missing data
- Single-breed model revised to multi-breed + crossbreds in 2006
- Programs used until 2014

2014 revised pedigree model code (used currently)

- New model options included:
  - Multi-trait models
  - Multiple class and regress variables
  - Suppress some factors for each trait
  - Random regressions
  - Foreign data
  - Parallel processing
- Compute Reliability, YD, DYD, and renumber factors in same program

2014 software revision test (VanRaden et al, 2014)

- Data in 7-trait MT animal model included:
  - Traits: Milk, Fat, Protein, SCS, PL, DPR
  - 76 million lactation phenotypes / trait
  - 63 million animals in pedigree
  - 30 million permanent environment effects
  - 7 million herd management groups
  - 11 million herd by sire interactions
- Genotypes processed separately

2014 Compare ST and MT to previous software

	Correlations within breed		
Trait	Prev, ST	Prev, MT	ST, MT
Milk	.986	.986	.999
Fat	.985	.985	.999
Protein	.986	.988	.999
Somatic Cell Score	.998	.994	.995
Productive Life	.975	.945	.972
Dtr Pregnancy Rate	.987	.958	.970

Domestic bulls with ≥ 10 USA daughters

2025 Data and models for 48 traits in 12 groups

Pedigree trait group	Traits	Model	All-breed	Data Since
Yield (milk, fat, protein)	3	MT	Yes	1960
Type	21 x 6 breeds	MT	No	1976
Productive life	1	ST	Yes	1960
Somatic cell score	1	ST	Yes	1982
Female fertility	4	MT	Yes	1960
Calving ease and stillbirth	4	ST	No	1980
Cow livability	2	MT	Yes	1960
Heifer livability and health	3	ST	Yes	2009
Cow health	6	ST	Yes	1985
Gestation length	1	ST sire	Yes	1970
Residual feed intake	1	ST	HO only	2000
Milking speed	1	ST	Yes	2024

Genotypes: How many loci?

- "SNP chips" available since 2008 for cattle genotyping
  - 50,000 SNPs chosen by Matukumalli, Van Tassell, Sonstegard
  - Higher and lower density chips were mostly nested:
    - 777,000 SNP (HD chip) includes ~90% of the 50,000
    - 50,000 SNP includes nearly all the 3,000 chip from 2011
    - 6,909 SNP chosen by USA, FRA, AUS with custom add-on
- Nesting allows parent confirmation and accurate imputation
- Now include QTLs, gene tests, and sequence data

Imputation of missing genotypes

- Many genotypes are missing and must be estimated
  - With just 1 chip < 10% of genotyped loci are unknown
  - With more chips and high density >99% of genotypes may be unknown and need to be imputed for a given locus
  - Different densities save $, avoid regenotyping all animals
- Pedigree imputation uses SNPs of closest relatives
- Population imputation chooses matching haplotypes from list
- Accuracy high with large reference population of same breed

Example Bull: O-Style

Pedigree Relationship Matrix (1922 math)

	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 (2008 math)

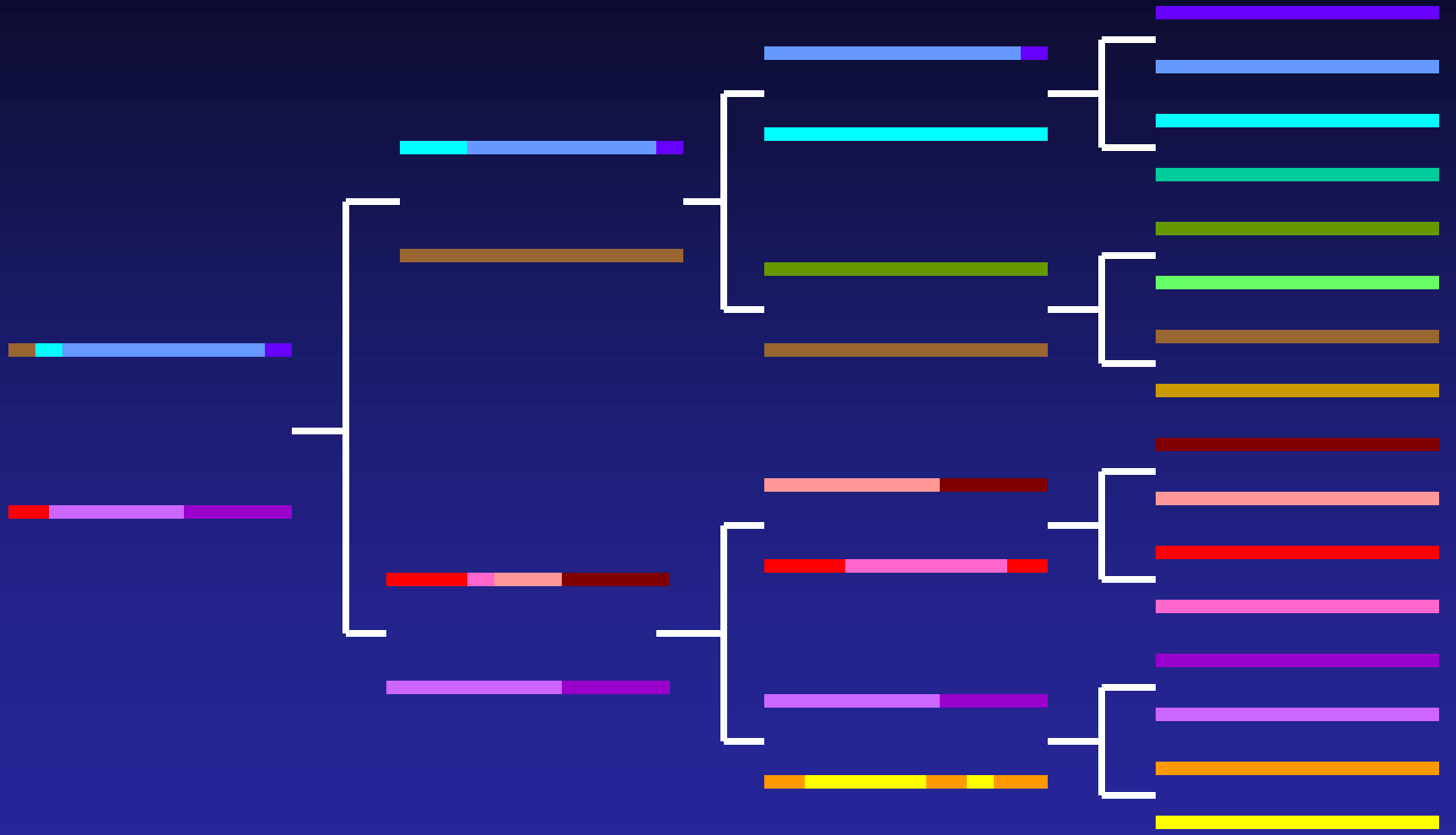
	PGS	PGD	MGS	MGD	Sire	Dam	Bull
Manfred	1.201	.058	.050	.093	.609	.054	.344
Jezebel	.058	1.131	.008	.135	.618	.079	.357
Teamster	.050	.008	1.110	.100	.014	.613	.292
Dima	.093	.135	.100	1.139	.131	.610	.401
O-Man	.609	.618	.014	.131	1.166	.080	.626
Deva	.054	.079	.613	.610	.080	1.148	.613
O-Style	.344	.357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

Difference (Genomic – Pedigree)

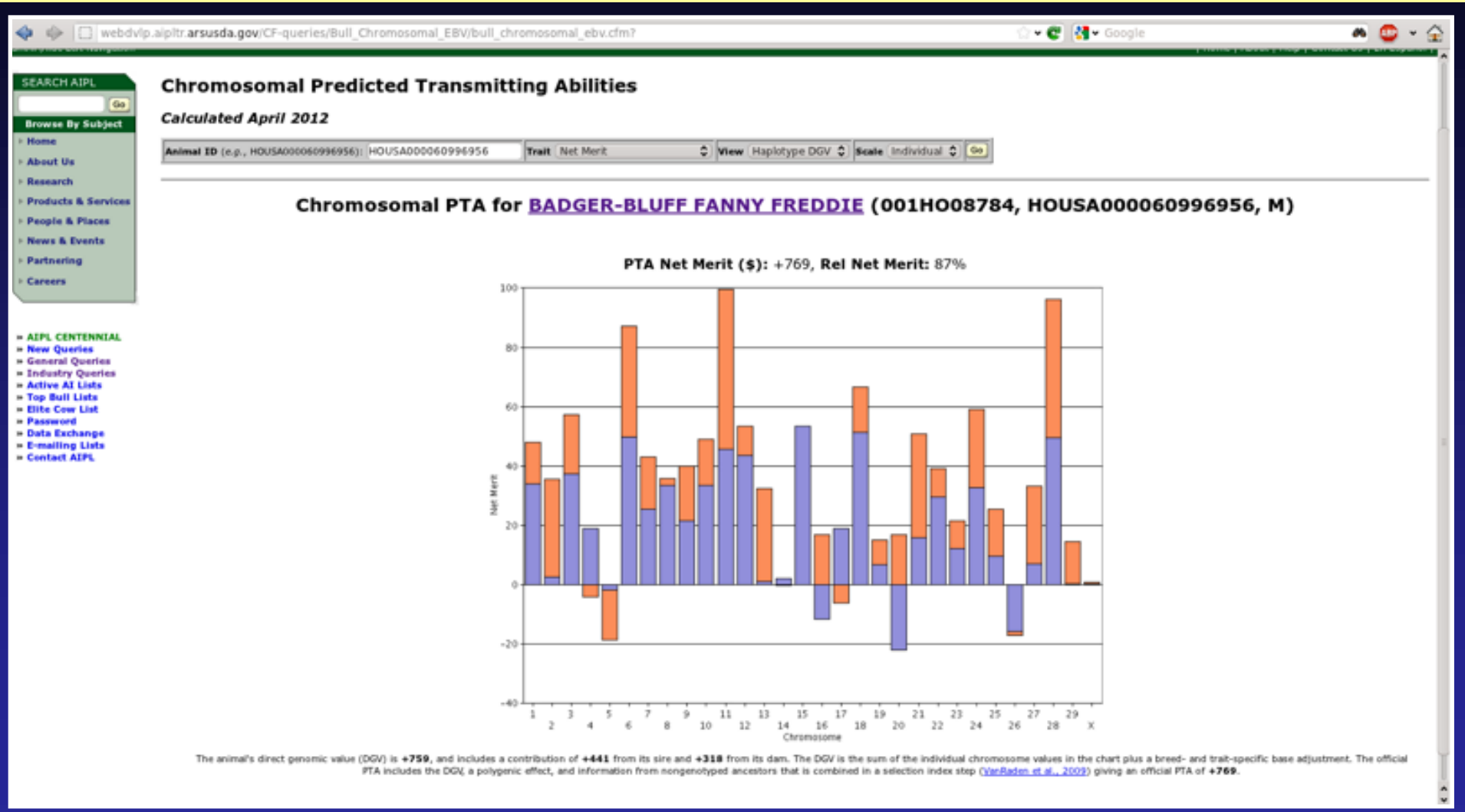
	PGS	PGD	MGS	MGD	Sire	Dam	Bull
Manfred	.149	-.032	-.040	-.012	.038	-.043	.010
Jezebel	-.032	.095	-.043	.036	.055	.004	.038
Teamster	-.040	-.043	.075	-.021	-.057	.035	-.032
Dima	-.012	.036	-.021	.097	.029	.029	.059
O-Man	.038	.055	-.057	.029	.121	-.006	.060
Deva	-.043	.004	.035	.029	-.006	.087	.040
O-Style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

O-Style Haplotypes

chromosome 15

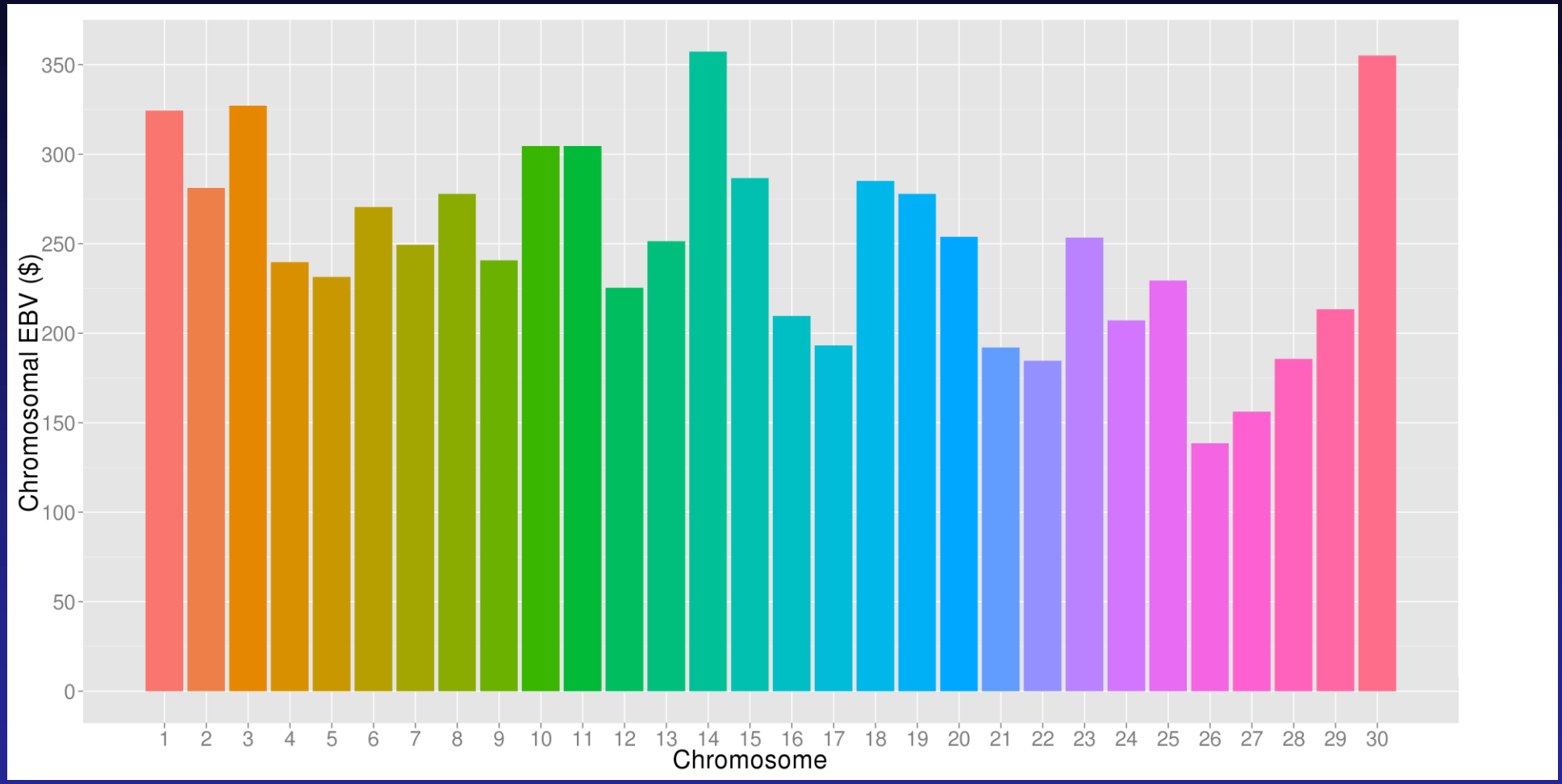


2012 haplotype merit display



Cole, J.B., and Null, D.J. 2012. AIPL Research Report GENOMIC2: Use of chromosomal predicted transmitting abilities. Available: http://aipl.arsusda.gov/reference/chromosomal_pta_query.html.

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

2008 Derivation of genomic relationship matrix

- Pages 1-3 of 1980 population genetics notes (U IL, Dairy Sci 316)
  - Let gene content be coded as 0, 1, or 2 copies of an allele
  - Mean of gene content is twice allele frequency = 2p
  - Variance of gene content = 2p(1-p)
- Genomic relationship matrix algebra 2008
  - Define Z_{ij} for locus i, animal j as the gene content – $2p_i$
  - Subtract mean, divide by the sum of variances = $\sum 2p_i(1-p_i)$
  - Matrix G = Z Z' / $\sum 2p_i(1-p_i)$

Genomic relationship matrices: bovine vs. human

- Simple quadratic :
  - $G = Z Z' / \Sigma 2p_j(1-p_j)$
  - Same math for each G_{ik}
  - Clones related by 1.0
  - Positive semi-definite
- VanRaden 2008
  - 5300 citations

- Different math for G_{ii} vs. G_{ij}:
  - Divide Z_{ij} by sqrt $[2p_j(1-p_j)]$
  - Diagonals G_{ii} get adjusted
  - Clones not related by 1.0
  - Negative eigenvalues
- Yang et al 2010
  - 4700 citations
  - https://www.nature.com/articles/ng.608

Using genotypes to fill in missing pedigrees

- Millions of discovered grandsires are now included in pedigree file
- They make matrix A match G better
- If no dam ID available, construct missing dam IDs to link to grandsires
  - HOUSADAM000000001, for example
  - Discovered pedigrees are used in predictions and made public
- First check if true dam can be discovered in same herd (33,810 found)
  - Match birth and fresh dates (only 1 dam's pedigree matches calf's)
- Postdoc Juan Nani documented benefits from pedigree discovery

How to model the 7,000 clones or identical twins?

- Clones have the largest G - A matrix differences
- Improve genetic evaluations for clones and progeny
- Genotypes are same but polygenic effects treated as full sibs
- Pedigree inbreeding coefficients underestimated for descendants of clones
- Combine progeny counts for cloned bulls instead of reporting (since 2008) the daughter count of the clone with the most
- Improve ancestor discovery (clones are tied)

Conclusions about cloning

- Genetic evaluations can account for identical animals:
  - Link progeny of clones to the source sire or dam
  - Remove clone copies from the pedigree before analysis
  - Restore clone IDs and copy EBVs from the source animal
- Milk production was as expected but some other traits lower.
- Many Holsteins may soon have clones in their pedigrees.
- Cloning techniques also enable gene editing.
- Models may need revision again for that technology.

Reliability estimate for each EBV

- Establishing bounds on the accuracies of predictions of breeding value. Master's Thesis, ISU, 1984
- Similar projects:
  - Daughter equivalents in "Derivation, calculation, and use of national animal model information", 1991
  - "Reliability of genomic predictions for North American Holstein bulls", 2009
- Interbull Genomic Reliability Working Group, member 2013-2024

REML variance estimation programs

- Computational strategies for estimation of variance components. PhD thesis, ISU, 1986
- Fortran programs used for:
  - Angus carcass trait EPDs (1986-1997)
  - Holstein productive life heritability (1994)
  - Fertility and type heritabilities and correlations (2004, 2016)
  - Heifer livability and abortion h^2 (Neupane et al, 2021, 2023)
- Sire model still useful for many millions of records

Software sharing history

- Animal breeders invented mixed models and BLUP
  - C.R. Henderson (Cornell)
- Animal breeders coded mixed models and variance estimation
  - W.R. Harvey (USDA, Ohio State)
  - Many groups coded and shared software and subroutines
  - Government support for free software declining
  - Human geneticists share free code but less efficient
  - Future: Better fitting models or artificial intelligence?

Walter R. Harvey

- **Dr. Harvey wrote general software for USDA Biometrics 1953-62**
- **"What a wonderful man! When I came to Beltsville I was so grateful to find him there. Software was basically nonexistent. He is an example where the number of papers does not begin to disclose his contributions." (Dr. Bob Miller)**
- **3,351 citations for his 1960 software package LSML (a little like BLUPF90)**

Year	Where	Event
1919	New Mexico	Born
1952	Iowa State	PhD, Animal Breeding (with Lush)
1955	USDA-ARS Beltsville	Electrical computing in agriculture
1960	USDA-ARS Beltsville	Least Squares Analysis of Data
1963	Ohio State	Professor
2015	Columbus	Died, age 95

My USDA genomics project staffing 2010-2026

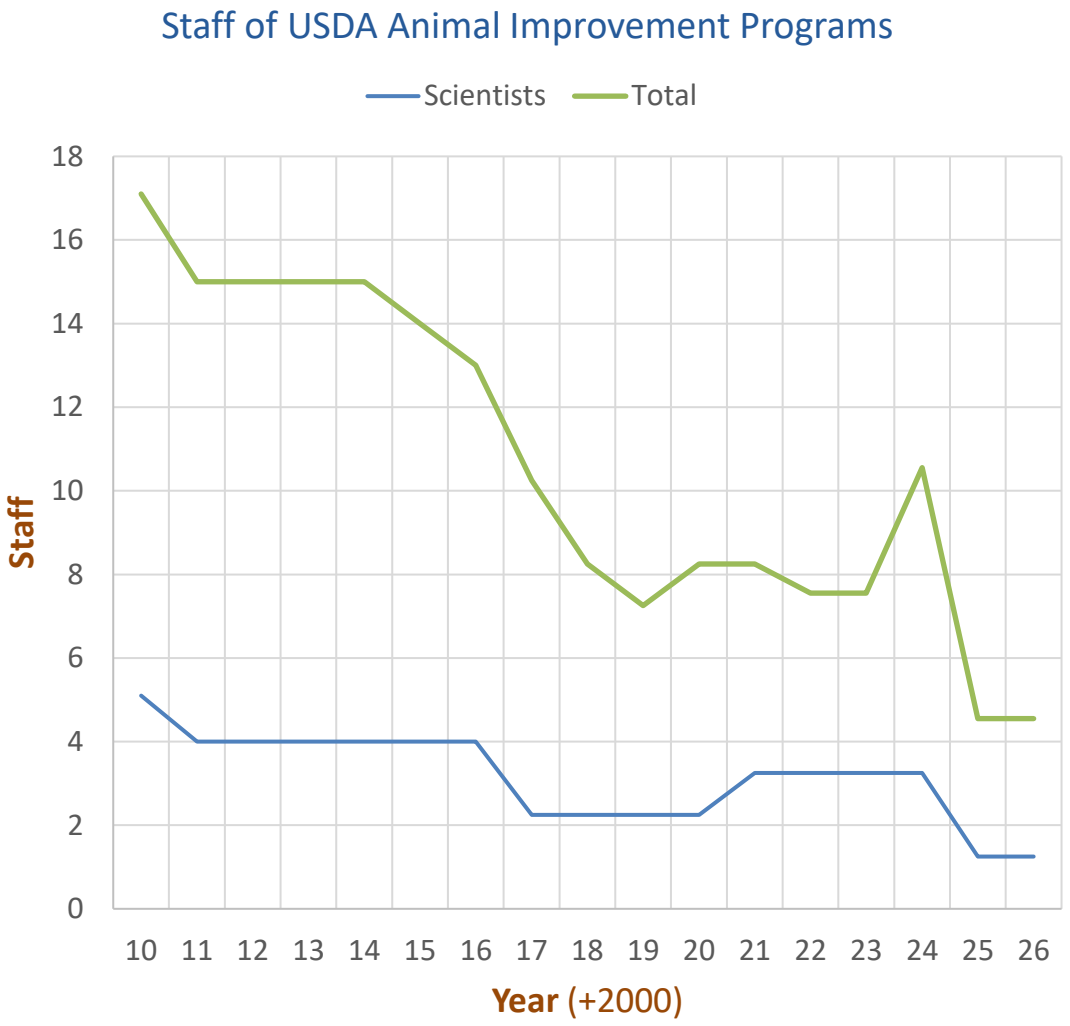


Photo at Bldg 306, Beltsville, April 2025

Not pictured: Jana Hutchison

Extra help: Jeff O'Connell, U MD - Baltimore

USDA computer site history: Remote or local?

- Early computers were expensive and often shared
- Data transfer always by computer tapes until late 1990s, then internet
- CDCB owns computers at a shared, local facility
- Remote (cloud) computing becoming popular again
- More parallel processing

Year	Computer owned by:	Location
1959	Commodity Credit branch, USDA	New Orleans, LA
1965	Navy Bureau of Personnel	Arlington, VA
1968	Computer Center, Statistical Reporting Service, USDA	Washington, DC
1970	Sears and Roebuck, Inc.	Rockville, MD
1973	National Agricultural Library, USDA	Beltsville, MD
1975	Washington Computer Center, USDA	Washington, DC
1985	Cornell University (evaluation only)	Ithaca, NY
1987	USDA Computing Center (database only)	Kansas City, MO
1993-2013	Animal Improvement Programs Lab, USDA	Beltsville, MD

“Cloud” computing, 1988

- Holstein staff flew data to Madison, WI twice per year
- “On-time delivery of PTAs is more important than better accuracy” (Holstein CEO)
- MT sire model tested at IBM Research Center, Los Angeles
- Research on memory sharing
- Replaced by animal model of Ignacy Misztal in 1992 (U GA)

Holstein type trait evaluations: 1988-present

- VanRaden, Jensen, Lawlor, and Funk. 1990. Prediction of transmitting abilities for Holstein type traits. J. Dairy Sci.
- MT sire model
- REML used 779,391 daughters of 871 sires
- Used by HO USA for 4 years

- Misztal, Lawlor, Short, and VanRaden. 1992. Multiple-trait estimation of variance components of yield and type traits using an animal model. J. Dairy Sci.
- MT animal model
- REML used 20,836 cows
- Used by HO USA for 34 years

Email exchanges – 1989

- *To* **Paul**, *From* **Ignacy**
- **I prepared an animal relationship for your REMLD program. The convergence was not reached until round 185 compared to 25 for your sire model example. My program showed full convergence but was 3-5 times slower. So the poor convergence was a result of Animal Model, it will perhaps be even worse with groups for unknown parents.**
- **Following your general advice I added a "relaxation factor." For some values of relax the convergence rate increases drastically (2-20 times), but too high relax and more traits can cause divergence.**

Conclusions

- Transmitting abilities of dairy bulls computed nationally for a century (1926-2026)
- Pedigree models were easy to compute
  - Many researchers developed their own code
- Genomic selection is very successful
  - For previously recorded traits with large databases
  - Enables faster selection for new or less heritable traits
  - Software must keep up with growth of data

Acknowledgments

- Taxpayers for funding USDA-ARS-AGIL project 8042-31000-002-00, “Improving dairy animals by increasing accuracy of genomic prediction, evaluating new traits, and redefining selection goals”

- AGIL staff for doing this research

- Council on Dairy Cattle Breeding (CDCB) and its industry suppliers for data