From Galton to Machine Learning: a brief account of quantitative methods in animal genetics

Daniel Gianola

Sewall Wright Professor Emeritus
of Animal Breeding and Genetics

University of Wisconsin
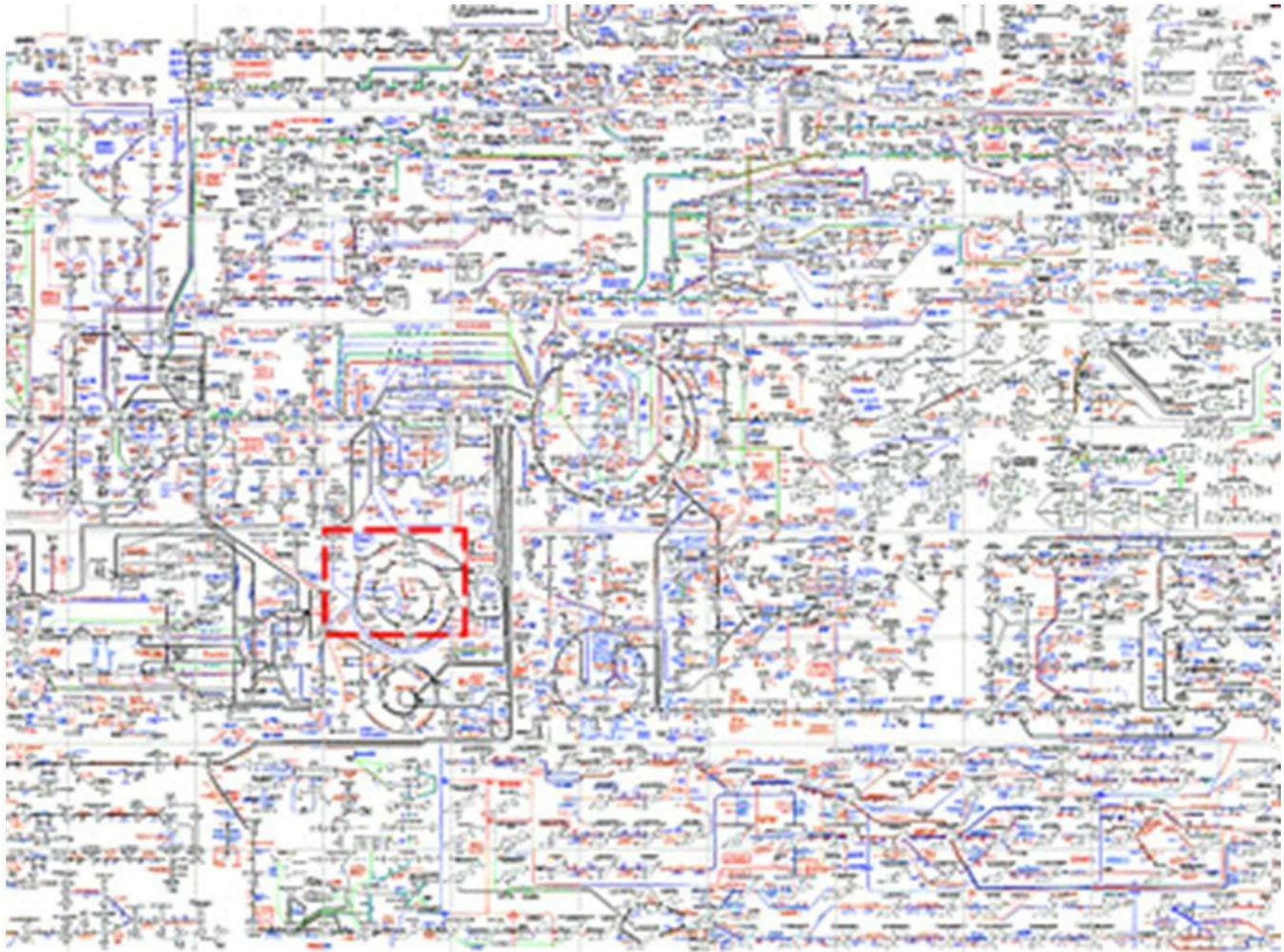**Madison, USA**

UW–MADISON ANIMAL SCIENCES

Proposition 1

It must be true that quantitative traits are "complex", in any sense of the word.
Why?

A "complex" trait involves many metabolic pathways: Roche's Chart

Coping with complexity

(WELCOME TO THE WORLD OF ABSTRACTIONS)

First assumption: there is a genetic signal and an environmental signal
Second assumption: the joint effect translates into a phenotye **y**

$$Y = f(G,E)$$

For some **UNKNOWN** function f

Choices?

$$Y = G^E?$$
$$Y = E^G?$$
$$Y = G + E + GE?$$ → Is an assumption
$$Y = (G + E)^{GE}?$$
$$Y = G + E?$$ → Is an even a stronger assumption

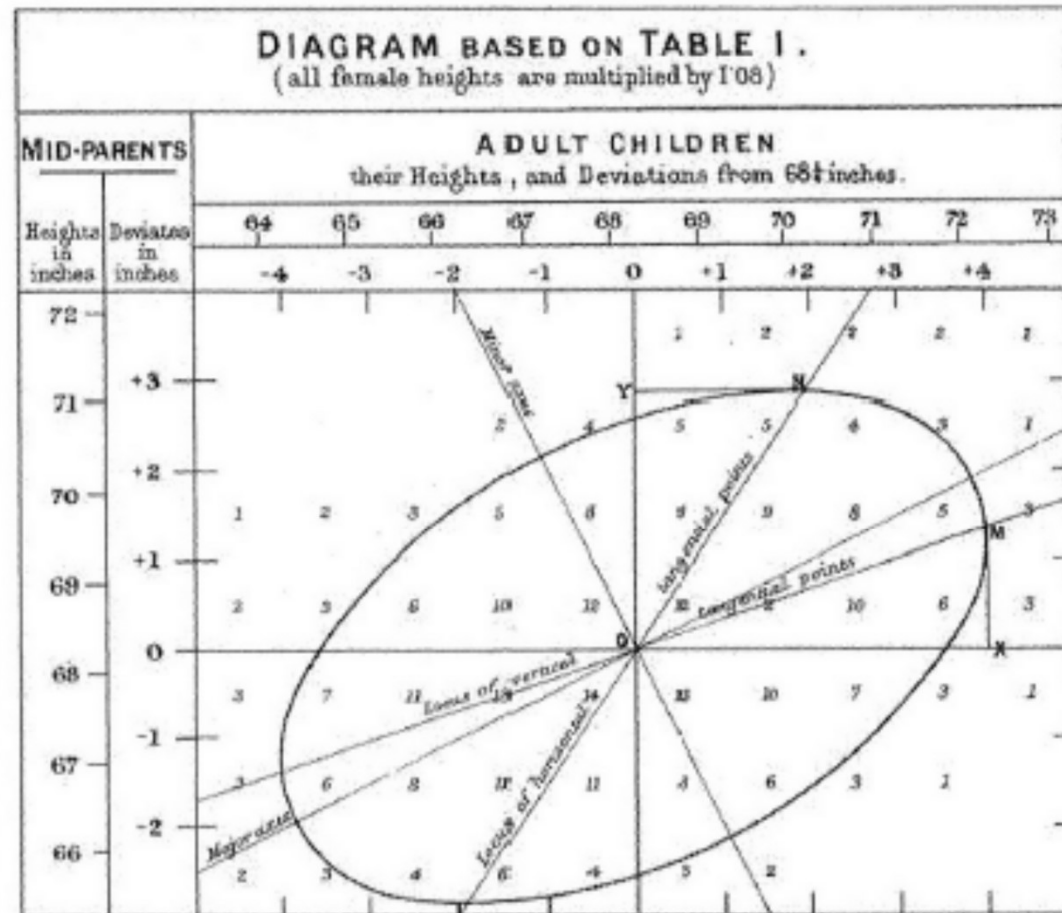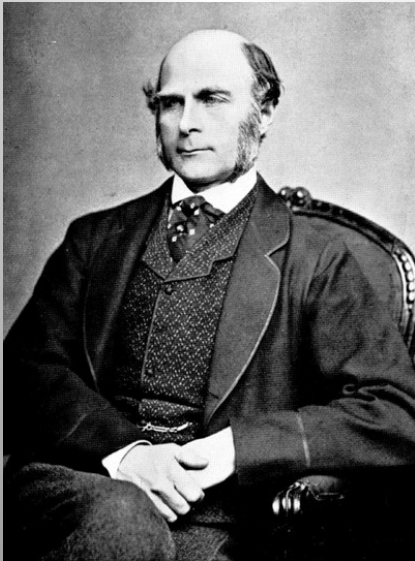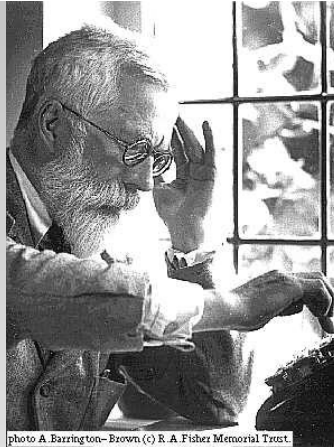# GALTON'S (1822-1911) REGRESSION OF OFFSPRING ON PARENT:
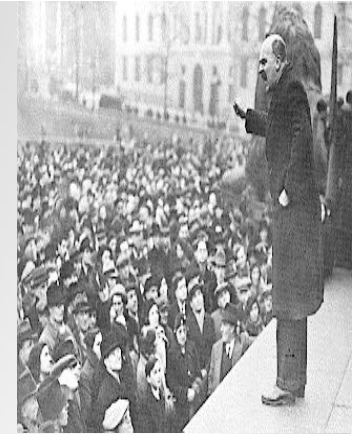## impetus for linear models

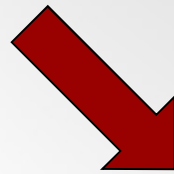Figure 1. Galton's fitted regression model.
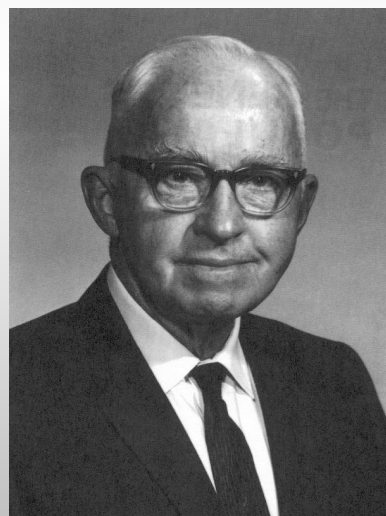
Sewall Wright

R. A. Fisher

J. B. S. Haldane

FOUNDERS OF MODERN QUANTITATIVE
AND POPULATION GENETICS

SCIENTIFIC FOUNDATIONS
OF
ANIMAL (PLANT) BREEDING

Jay L. Lush, Iowa State University
(animal breeding)

6

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52:399-433.

XV.—**The Correlation between Relatives on the Supposition of Mendelian Inherit-ance**. By **R. A. Fisher**, B.A. *Communicated by* Professor J. ARTHUR THOMSON. (With Four Figures in Text.)

CONTENTS.

DECOMPOSITION OF VARIANCE (under some assumptions)

TOTAL VARIANCE=

ADDITIVE+DOMINANCE+EPISTATIC+ENVIRONMENTAL

TYPICALLY ADDITIVE VARIANCE IS 1-40% OF TOTAL

SELECTION (main tool) EXPLOITS ADDITIVE VARIANCE

Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits

William G. Hill[1]*, Michael E. Goddard[2,3], Peter M. Visscher[4]

1 Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, 2 Faculty of Land and Food Resources, University of Melbourne, Victoria, Australia, 3 Department of Primary Industries, Victoria, Australia, 4 Queensland Institute of Medical Research, Brisbane, Australia

Abstract

The relative proportion of additive and non-additive variation for complex traits is important in evolutionary biology, medicine, and agriculture. We address a long-standing controversy and paradox about the contribution of non-additive genetic variation, namely that knowledge about biological pathways and gene networks imply that epistasis is important. Yet empirical data across a range of traits and species imply that most genetic variance is additive. We evaluate the evidence from empirical studies of genetic variance components and find that additive variance typically accounts for over half, and often close to 100%, of the total genetic variance. We present new theoretical results, based upon the distribution of allele frequencies under neutral and other population genetic models, that show why this is the case even if there are non-additive effects at the level of gene action. We conclude that interactions at the level of genes are not likely to generate much interaction at the level of variance.

THE GENETIC CORRELATION (Hazel, 1943)
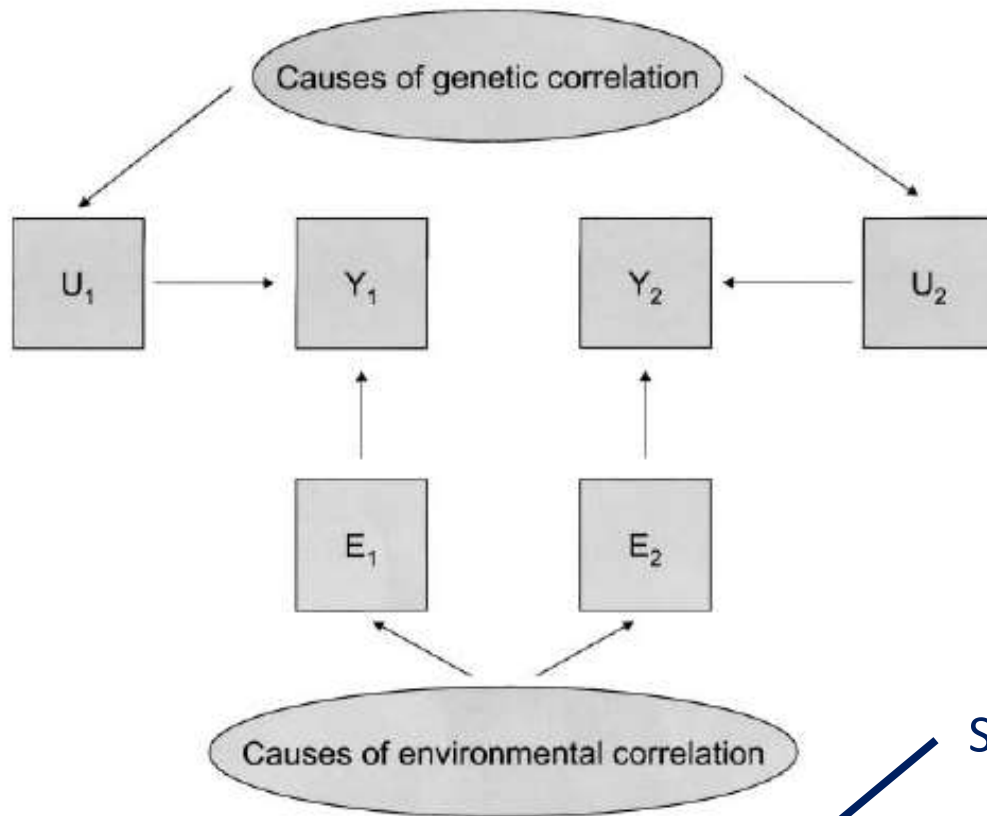## ➜ INPUT IN MULTIPLE-TRAIT IMPROVEMENT

FIGURE 1.—Standard bivariate model used in quantitative genetics: Y_1 and Y_2 are the phenotypic values; U_1 and U_2 are additive genetic effects acting on the traits; E_1 and E_2 are residual effects. A single-headed arrow (*e.g.*, $A \rightarrow B$) indicates that variable A affects variable B.

Square root of heritability of trait Y

$$r_{XY} = r_G h_X h_Y + r_E \sqrt{1 - h_X^2} \sqrt{1 - h_Y^2}$$

Genetic correlation   Environmental correlation

CONCEPTO CENTRAL DEL MEJORAMIENTO GENETICO

$$Merito\,(M) \quad : \quad caracteres,\ valor\ economico,$$

$$Direccion \quad : \quad extremos,\ intermedios,\ valores\ optimos$$

$$E\,(\Delta_G) \quad = \quad \frac{intensity \times correlation\,(EM,M) \times var.\ genetica}{generation\ interval}$$

$$\max[E\,(\Delta_G)]\ ST \quad : \quad inbreeding,\ conservation,\ e-impact$$
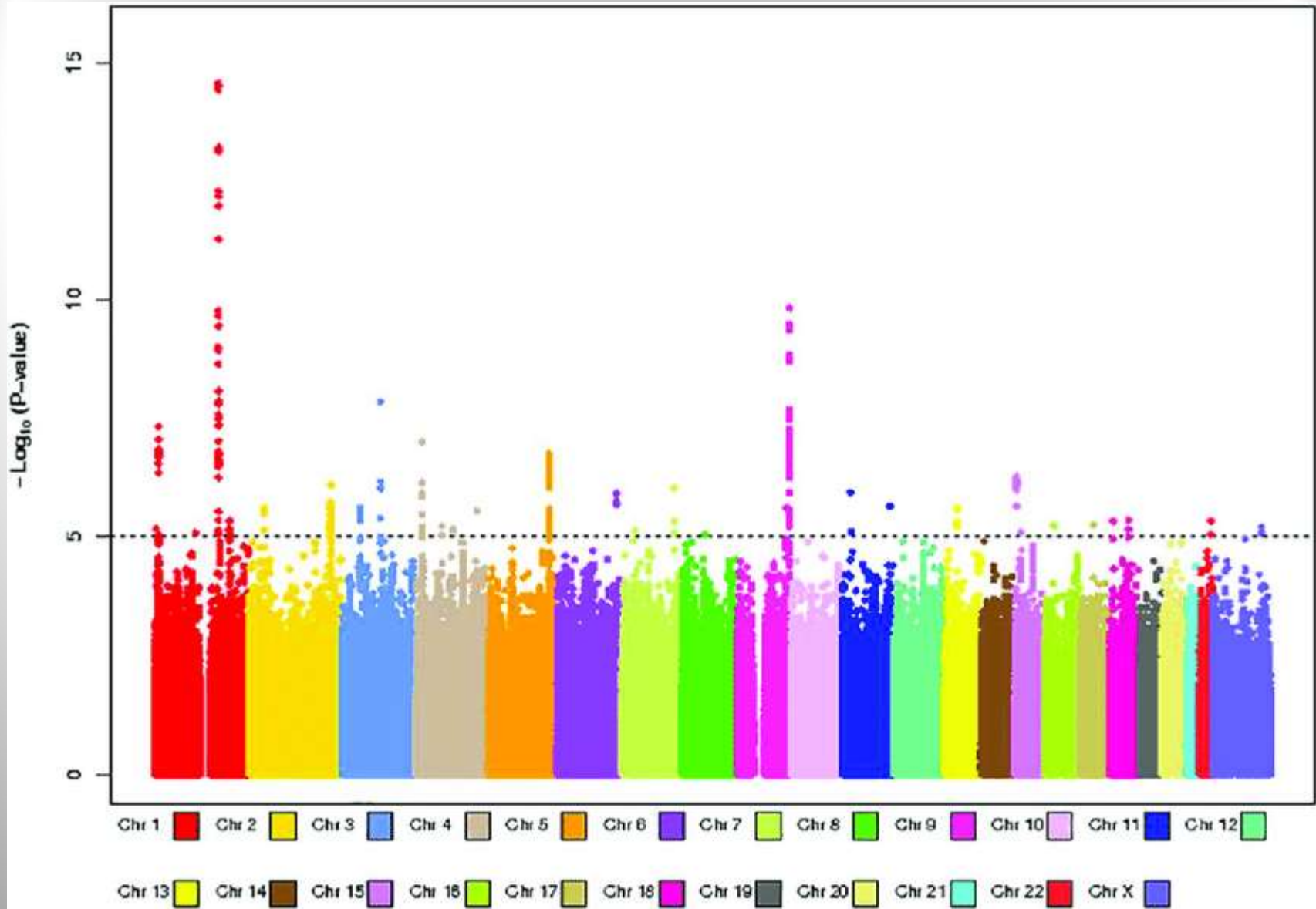
Evolution of four breeding programs
BLACK= 2% per year BLUE= 1.5% per year
GREEN= 1% per year RED= 0.5% per year
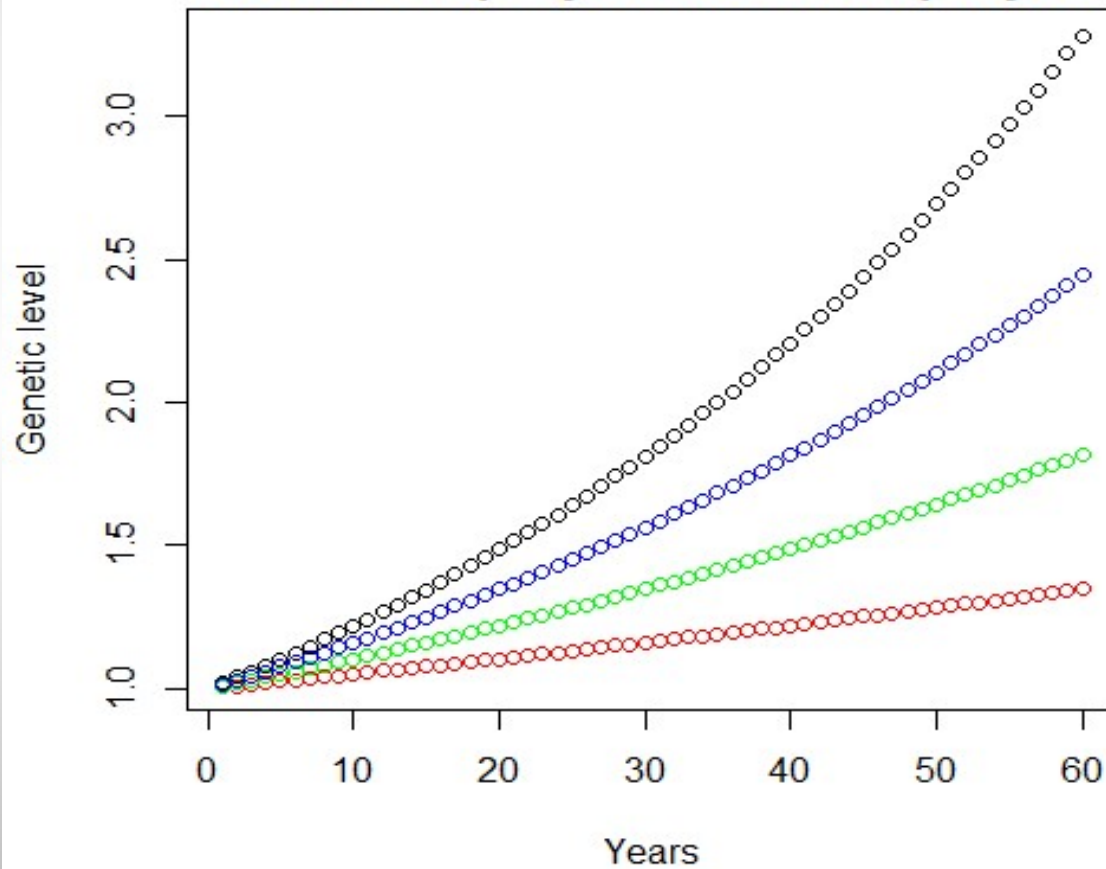
THREE PARADIGMS FOR GENETIC ANALYSIS IN ANIMAL BREEDING

PARADIGM 1
(QTL discovery)

GWAS: search for association between some marker or genomic region, and a target phenotype.

THROMBO-EMBOLISM

PARADIGM 2
(variance components, indexing, BLUP)
Fisher's infinitesimal model
(extended vectorially by C. R. Henderson)

1987

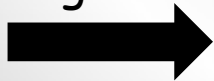# THE n<<p ERA
(In animal breeding ≈1948-1973: C. R. HENDERSON)

$$y = X\beta + Zu + e$$

$$y|\beta, u, R \sim N(X\beta + Zu, R)$$

Fixed Random $$u \sim N(0, G)$$

BLUP=Best linear unbiased predictor

MME: an algorithm

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

BLUP= Conditional posterior mean in Bayesian Gaussian linear hierarchical model

BLUP=penalized (L2) maximum likelihood

BLUP=Similar to kriging in geostatistics

BLUP=special case of RKHS regression
BLUP=single layer NN (input= **A**), linear activation function

BAYESIAN INFERENCE AND THE NEO-BAYES-LAPLACE REVOLUTION
(James-Stein, Lindley, Box, Zellner…)

Rev. Thomas Bayes

1702 London, England
1761 Tunbridge Wells, Kent, England

1763. "An essay towards solving a problem in the doctrine of chances".
Philosophical Transactions of the Royal Society of London **53**, 370-418.
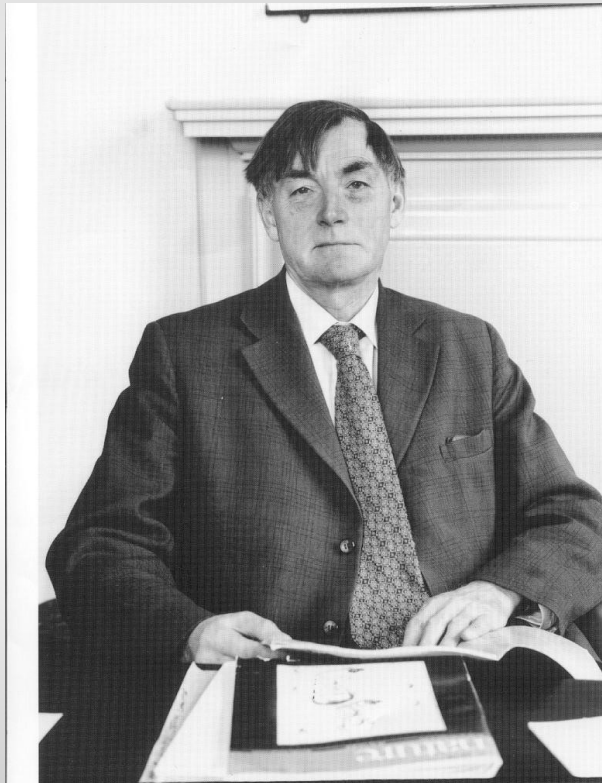

Pierre-Simon Laplace

1749 Beaumont-en-Auge, France
1827 Paris, France

1774. "Mémoire sur la probabilité des causes par les événements".
Savants étranges **6**, 621-656. *Oeuvres* **8**, 27-65

THE EDINBURGH SCHOOL KEPT REMINDING US OF THE GENES, SETTING THE STAGE FOR THE QUANTITATIVE GENOMICS ERA: THREE GIANTS!

1. D. S. Falconer

2. Alan Robertson

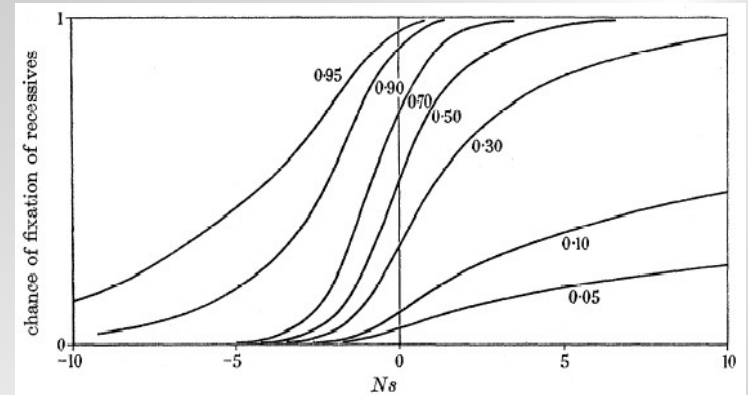$$u(q) = \frac{\displaystyle\int_0^q e^{-2Nsq^2}\,dq}{\displaystyle\int_0^1 e^{-2Nsq^2}\,dq}.$$

FIGURE 2. The chance of fixation of a recessive gene. The curves are drawn for different initial recessive frequencies.

Disequilibrium Among Several Linked Neutral Genes in Finite Population

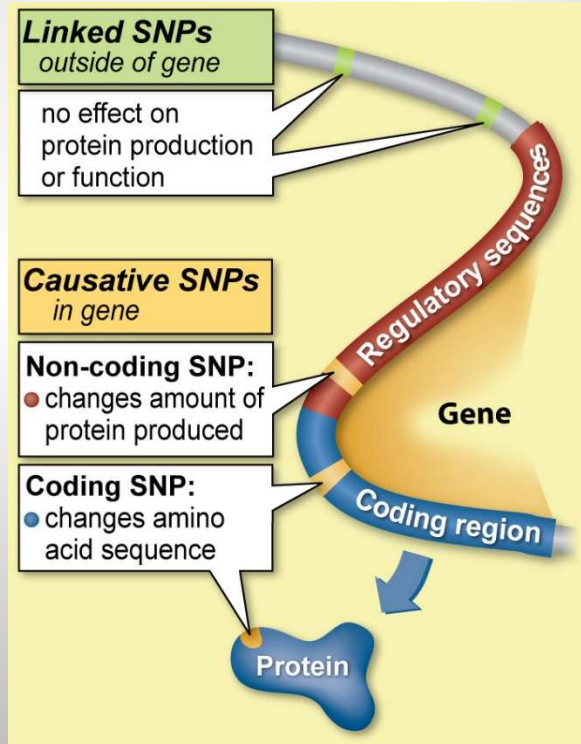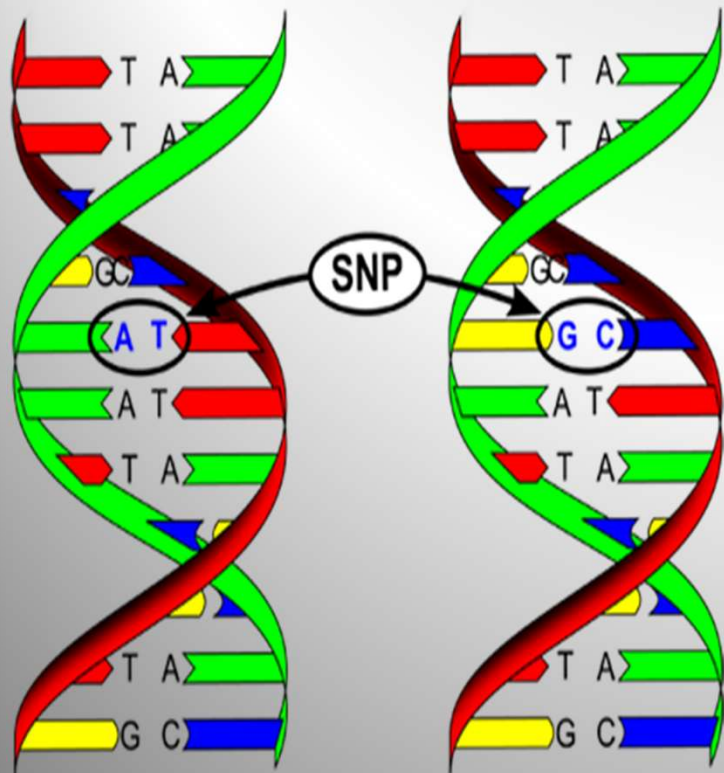### I. Mean Changes in Disequilibrium*

WILLIAM G. HILL

3. William G. Hill

THE GENOMIC ERA

MASSIVE NUMBERS OF MOLECULAR MARKERS AVAILABLE: DNA sequences
(cattle: 3 Gb; maize: 2500 Mb)

Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes[†] and M. E. Goddard[†,‡]

*Research Institute of Animal Science and Health, 8200 AB Lelystad, The Netherlands, [†]Victorian Institute of Animal Science, Attwood 3049, Victoria, Australia and [‡]Institute of Land and Food Resources, University of Melbourne, Parkville 3052, Victoria, Australia

On Marker-Assisted Prediction of Genetic Value: Beyond the Ridge

Daniel Gianola,*,[1] Miguel Perez-Enciso[†] and Miguel A. Toro[‡]

*Department of Animal Sciences, University of Wisconsin, Madison, Wisconsin 53706, [†]Station d'Amelioration Génétique des Animaux, Institut National de la Recherche Agronomique, 31326 Castanet-Tolosan, France and [‡]Departamento de Mejora Genética Animal, Instituto Nacional de Investigaciones Agrarias, 28040-Madrid, Spain

Genome-enabled selection

Reference Population

Known genotypes and phenotypes

Selection Candidates

Marker genotypes

Prediction Equation

Genomic breeding value =
$t_1x_1 + t_2x_2 + t_3x_3 + \ldots$

Selected Breeders

Using genomic breeding values
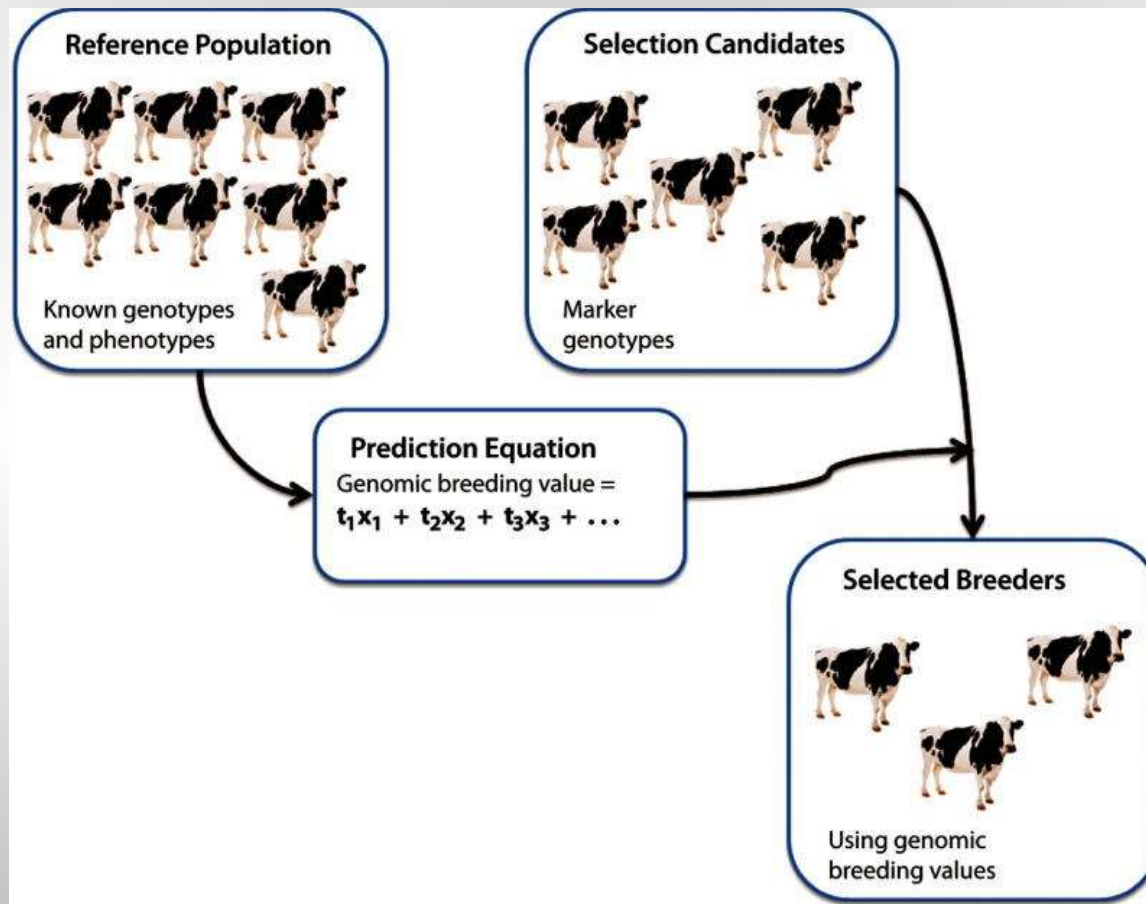
CLASSICAL DAIRY CATTLE BREEDING
Prediction of progeny performance
(progeny testing)
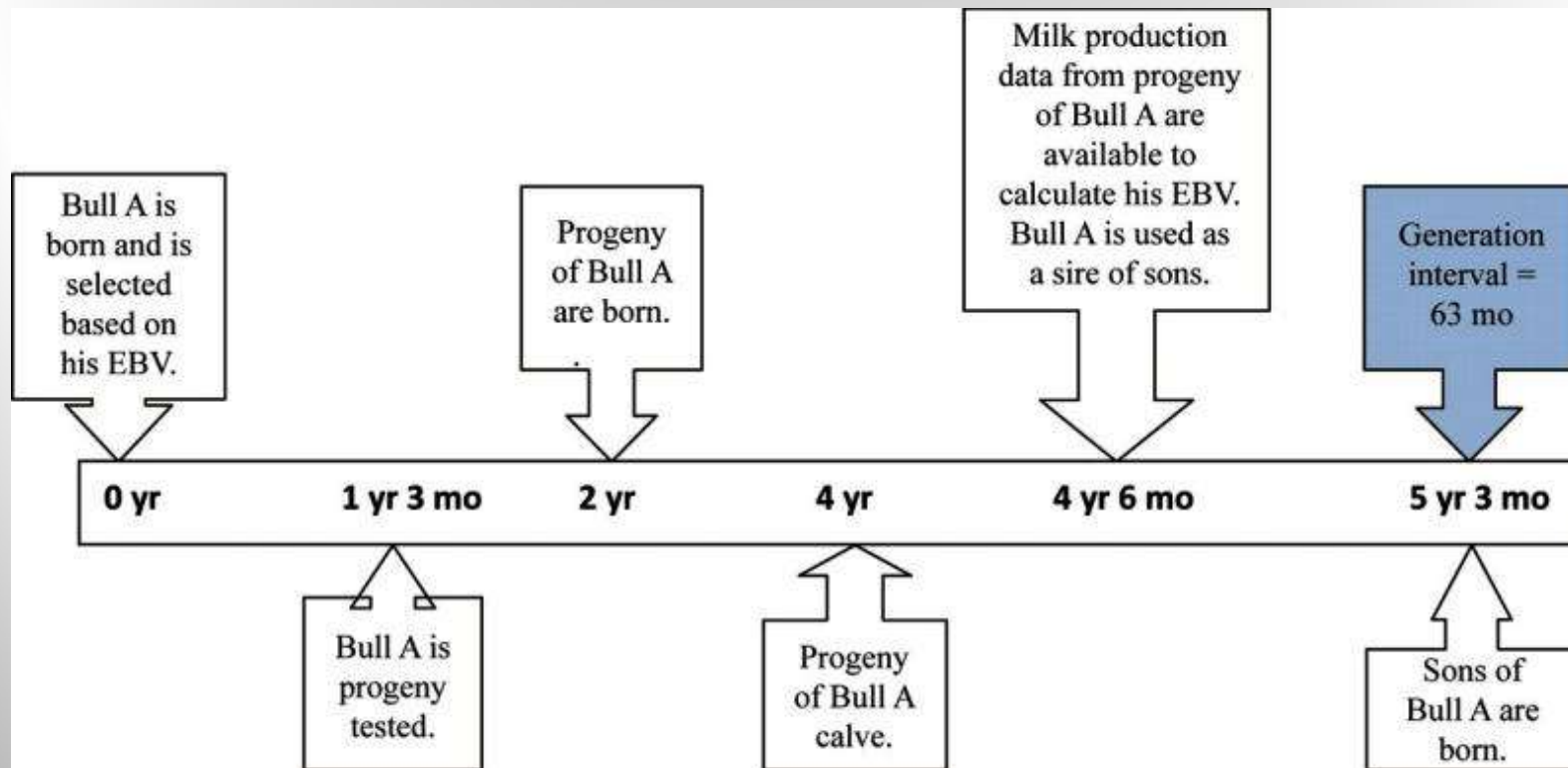
Classical progeny testing scheme

TWO IMPACTS OF GENOMIC SELECTION:

1) Generation interval drastically reduced.
2) Genome-enabled predictions (GEBV) may be more accurate than EBV

GENOMIC SELECTION IN DAIRY CATTLE (USA)
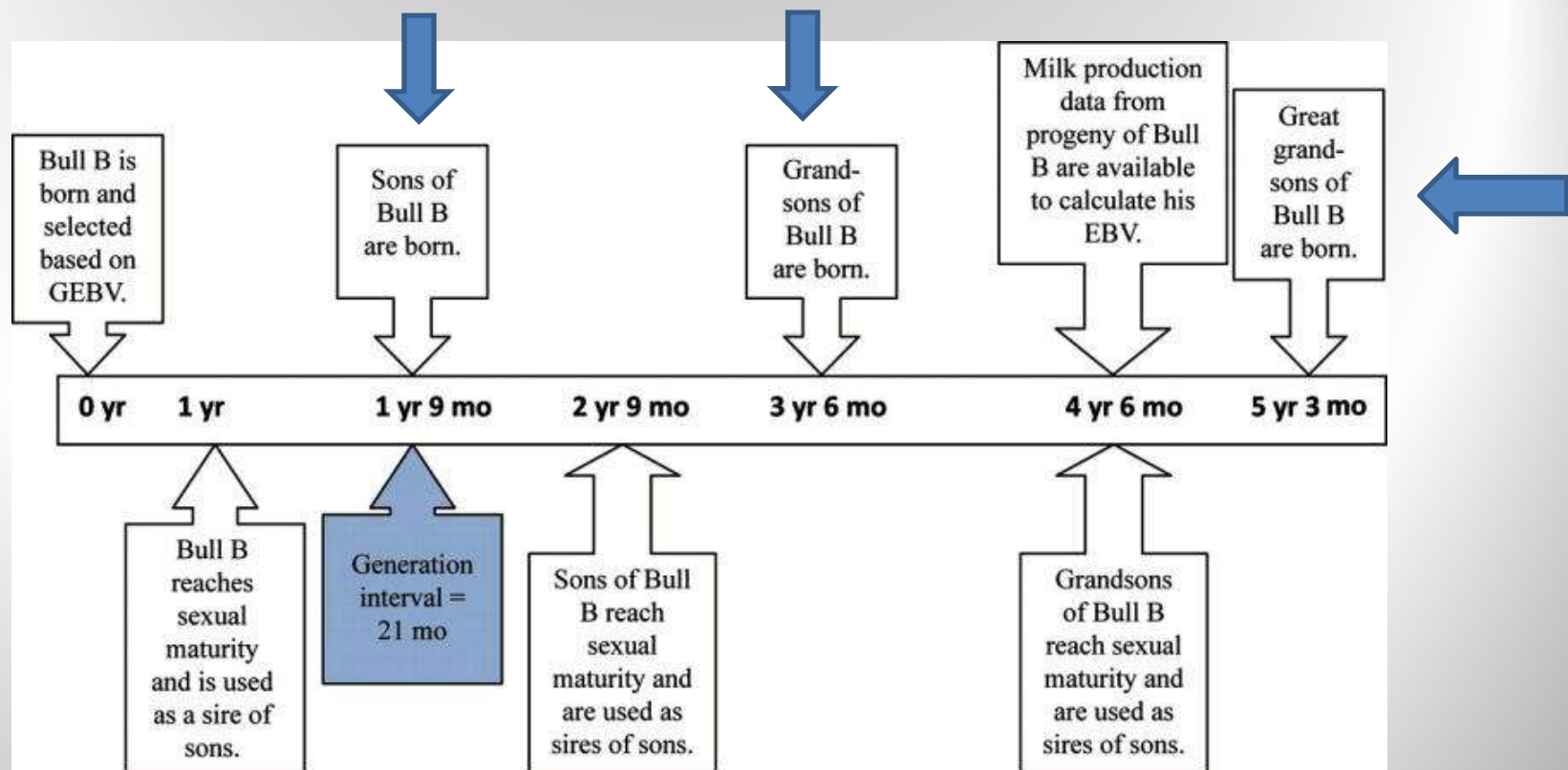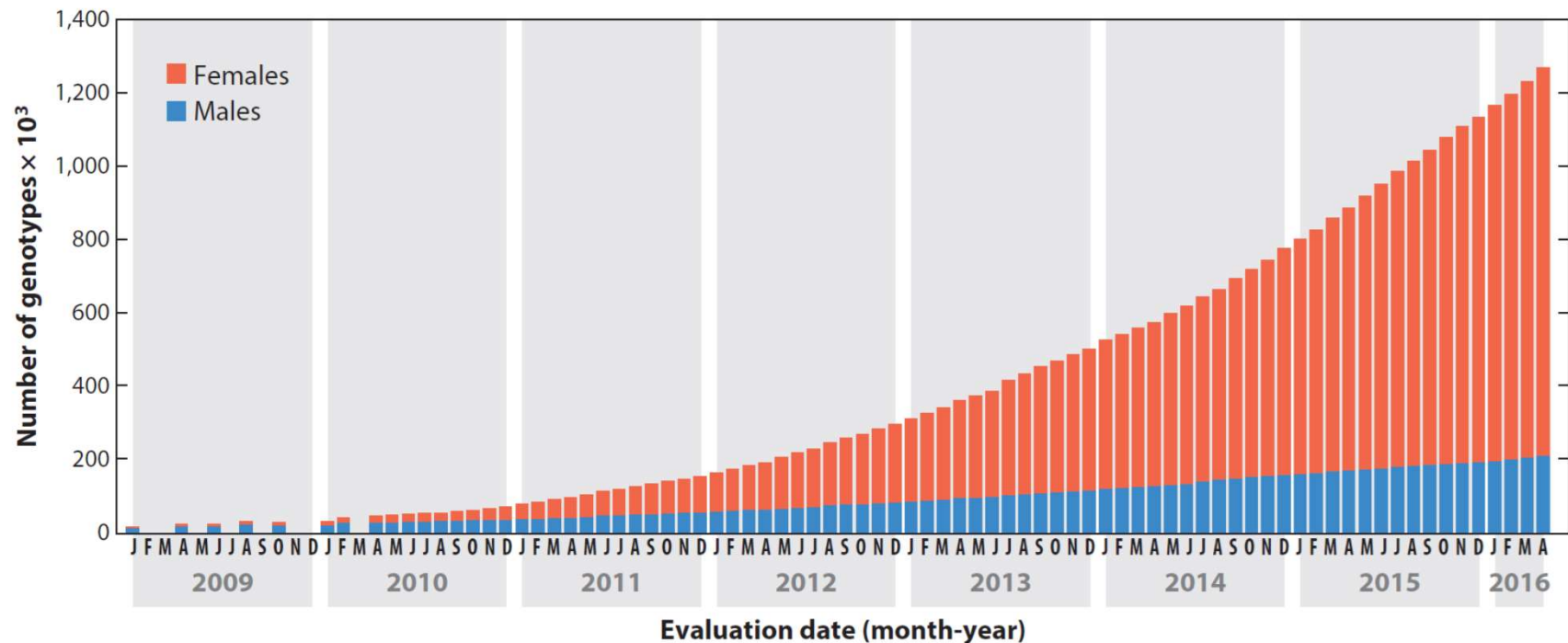

Figure 1

Number of genotyped animals included in US genomic evaluations for dairy cattle since January 2009. Official US genomic evaluations were first released to the dairy industry in January 2009 for Holsteins and Jerseys, in August 2009 for Brown Swiss, in April 2013 for Ayrshires, and in April 2016 for Guernseys. Data for figure generation were reported by the Council on Dairy Cattle Breeding (27). Months without data represent months in which official evaluations were not released.

WIGGANS (2017), Ann. Reviews of Animal Biosciences

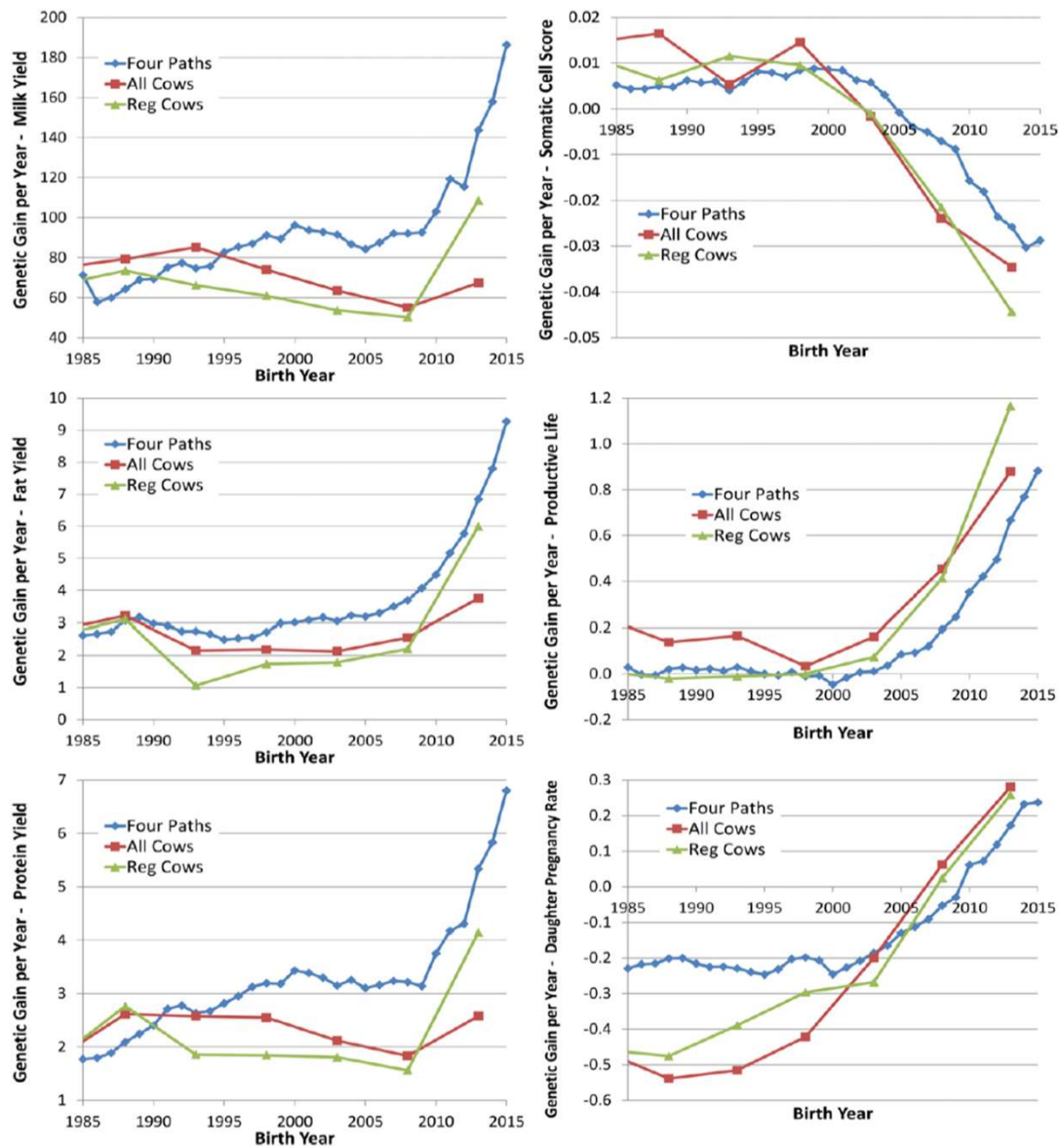# RECENT ESTIMATES OF GENOMIC BREEDING VALUE TRENDS

Fig. 3. Genetic gain per year estimates from four paths of selection (Four Paths) and segmented regressions of trait PBV on birth year for all cows (All Cows) or the subset of cows registered in the national herdbook (Reg Cows) for six traits (milk, fat, and protein yields; SCS; PL; and DPR).

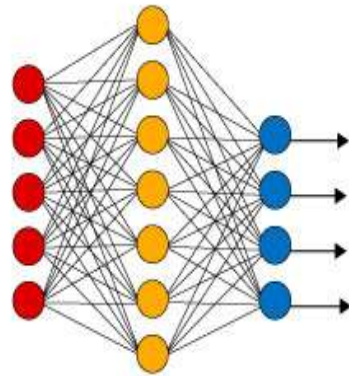**GARCIA-RUIZ, 2016. PNAS**

PARADIGM 3:

**MINE AND INTERROGATEDATA
"HYPOTHESIS-FREE DISCOVERY",
CLASSIFY, PREDICT!**

⬇   ⬇

(machine learning: largely non-parametric)
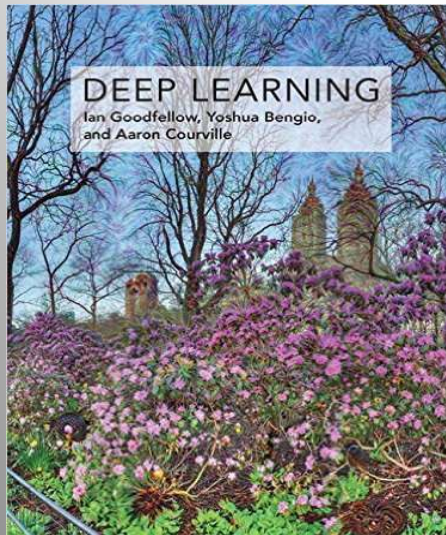
The return of the multi-layer neural networks…
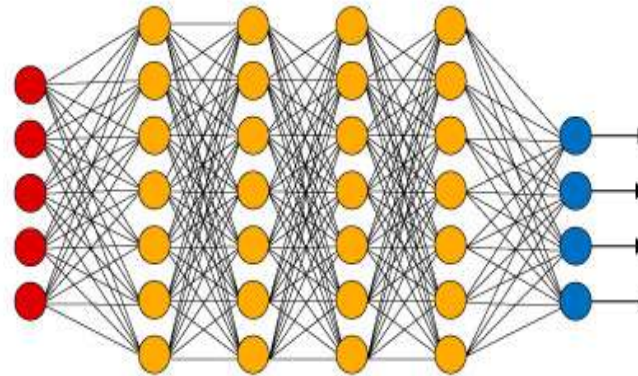
RKHS:
Reproducing Kernel Hilbert Spaces Regression

Mathematics knows no races or geographic boundaries; for mathematics, the cultural world is one country.

— *David Hilbert* —

AZ QUOTES
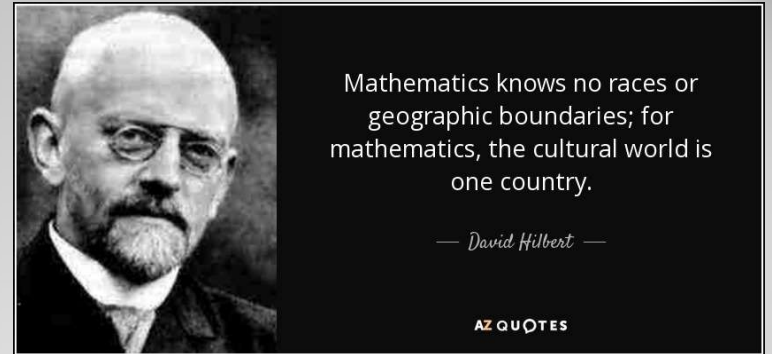
SPECIAL CASES OF RKHS➔SIMILARITY MATRICES

- BLUP using pedigrees
- BLUP using markers (GBLUP)
- Kriging in geostatistics
- Linear combinations of kernels+ Hadamard-Product kernels
- Genomic, enviromentomic, epigenomic, metagenomic kernels
- G x E KERNELS
- Support vector machines in regression or classification

MULTI-OMICS OR "OTHER" OMICS

Genetic Epidemiology

Poly-Omic Prediction of Complex Traits: OmicKriging

Heather E. Wheeler,[1] Keston Aquino-Michaels,[2] Eric R. Gamazon,[2] Vassily V. Trubetskoy,[2] M. Eileen Dolan,[1] R. Stephanie Huang,[1] Nancy J. Cox,[2] and Hae Kyung Im[3]*

[1]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, Illinois, United States of America; [2]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, Illinois, United States of America; [3]Department of Health Studies, University of Chicago, Chicago, Illinois, United States of America

A reaction norm model for genomic selection using high-dimensional genomic and environmental data

Diego Jarquín · José Crossa · Xavier Lacaze · Philippe Du Cheyron · Joëlle Daucourt · Josiane Lorgeou · François Piraux · Laurent Guerreiro · Paulino Pérez · Mario Calus · Juan Burgueño · Gustavo de los Campos

Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with Use of Whole-Genome Multiomic Profiles

Ana I. Vazquez,*,[1] Yogasudha Veturi,[†] Michael Behring,[‡,§] Sadeep Shrestha,[§] Matias Kirst,**,[††] Marcio F. R. Resende, Jr.,**,[††] and Gustavo de los Campos*,[‡‡]

*Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan 48824, [†]Biostatistics Department, [‡]Comprehensive Cancer Center, and [§]Department of Epidemiology, University of Alabama at Birmingham, Alabama 35294, **School of Forest Resources and Conservation and [††]University of Florida Genetics Institute, University of Florida, Gainesville, Florida 32611, and [‡‡]Statistics Department, Michigan State University, East Lansing, Michigan 48824

ABSTRACT Whole-genome multiomic profiles hold valuable information for the analysis and prediction of disease risk and progression. However, integrating high-dimensional multilayer omic data into risk-assessment models is statistically and computationally challenging. We describe a statistical framework, the Bayesian generalized additive model ((BGAM), and present software for integrating multilayer high-dimensional inputs into risk-assessment models. We used BGAM and data from The Cancer Genome Atlas for the analysis and prediction of survival after diagnosis of breast cancer. We developed a sequence of studies to (1) compare predictions based on single omics with those based on clinical covariates commonly used for the assessment of breast cancer patients (COV), (2) evaluate the benefits of combining COV and omics, (3) compare models based on (a) COV and gene expression profiles from oncogenes with (b) COV and whole-genome gene expression (WGGE) profiles, and (4) evaluate the impacts of combining multiple omics and their interactions. We report that (1) WGGE profiles and whole-genome methylation (METH) profiles offer more predictive power than any of the COV commonly used in clinical practice (e.g., subtype and stage), (2) adding WGGE or METH profiles to COV increases prediction accuracy, (3) the predictive power of WGGE profiles is considerably higher than that based on expression from large-effect oncogenes, and (4) the gain in prediction accuracy when combining multiple omics is consistent. Our results show the feasibility of omic integration and highlight the importance of WGGE and METH profiles in breast cancer, achieving gains of up to 7 points area under the curve (AUC) over the COV in some cases.

Prediction of Plant Height in *Arabidopsis thaliana* Using DNA Methylation Data

Yaodong Hu,*,[1] Gota Morota,[†] Guilherme J. M. Rosa,*,[‡] and Daniel Gianola*,[‡,§]

*Department of Animal Sciences, [‡]Department of Biostatistics and Medical Informatics, and [§]Department of Dairy Science, University of Wisconsin, Madison, Wisconsin 53706, and [†]Department of Animal Science, University of Nebraska, Lincoln, Nebraska 68583

STRETCHING THE BORDERS!

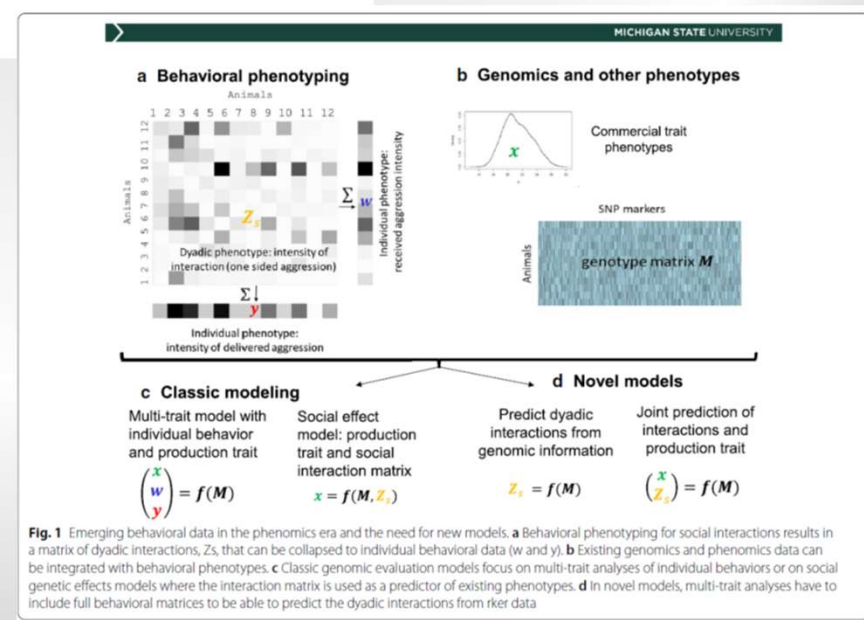GSE — Genetics Selection Evolution

OPINION

Open Access

Check for updates

Phenomes: the current frontier in animal breeding

Miguel Pérez-Enciso[1,2]* and Juan P. Steibel[3,4]

Behavioral traits measured with wearable sensors and computer vision

Fig. 1 Emerging behavioral data in the phenomics era and the need for new models. **a** Behavioral phenotyping for social interactions results in a matrix of dyadic interactions, Zs, that can be collapsed to individual behavioral data (w and y). **b** Existing genomics and phenomics data can be integrated with behavioral phenotypes. **c** Classic genomic evaluation models focus on multi-trait analyses of individual behaviors or on social genetic effects models where the interaction matrix is used as a predictor of existing phenotypes. **d** In novel models, multi-trait analyses have to include full behavioral matrices to be able to predict the dyadic interactions from rker data

GENOMIC EPIDEMIOLOGY: PANDEMICS AND ZOONOSIS
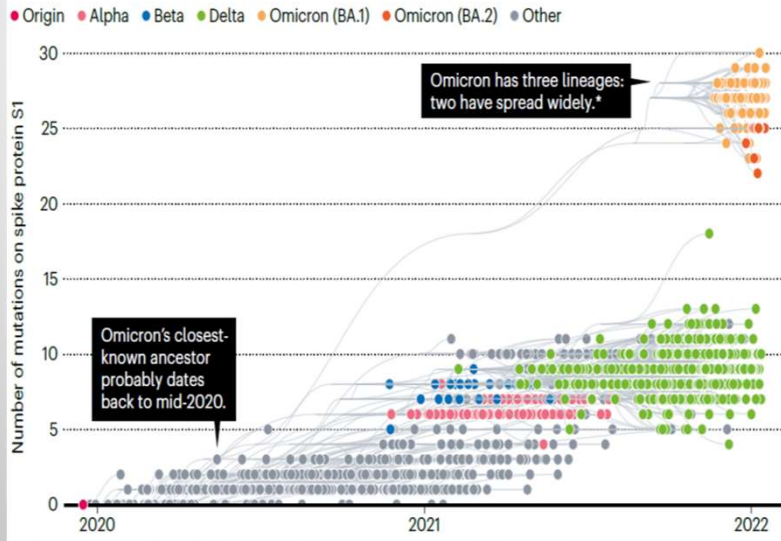
SARS-COV-2

No. mutaciones proteína espícula . Mayormente neutrales (sinventajas adaptativas); algunas incrementan en frecuencia, otras son desplazadas. De Nature (2022)
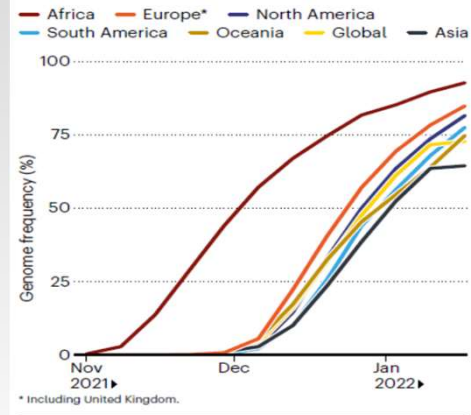
Omicron: porcentaje en muestras
secuenciadas en diferentes continentes.
Nature (2022)

Casos SARS-COV-2 examinados a nivel molecular
para mutaciones y seguimiento epidemiológico. Nature (2022).

AGRO-ECOGENOMICS

METAGENOMICS

ORIGINAL ARTICLE

Journal of Animal Breeding and Genetics | **WILEY**

Structural equation models to disentangle the biological relationship between microbiota and complex traits: Methane production in dairy cattle as a case of study

Alejandro Saborío-Montero[1,2] | Mónica Gutiérrez-Rivas[1] | Aser García-Rodríguez[3] | Raquel Atxaerandio[3] | Idoia Goiri[3] | Evangelina López de Maturana[4] | José Antonio Jiménez-Montero[5] | Rafael Alenda[6] | Oscar González-Recio[1,6]

Yinglin Xia · Jun Sun · Ding-Geng Chen

Statistical Analysis of Microbiome Data with R

Springer
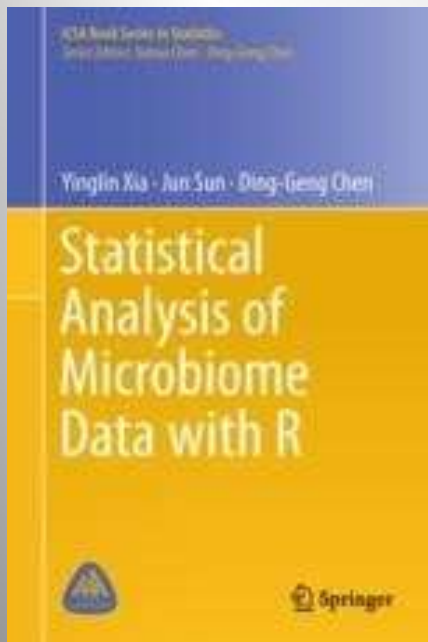
BREEDING 1.0

Incidental selection
by farmers

BREEDING 2.0

Statistical and
experimental design
to improve selection
effort

BREEDING 3.0

Integration of
genetic and
genomic data;
current state of
the art

BREEDING 4.0

Ability to combine
any known alleles
into optimal
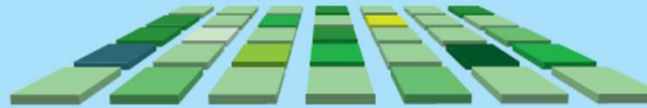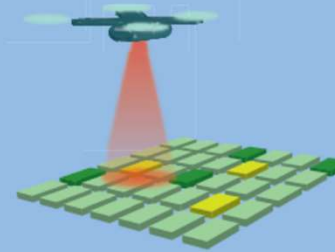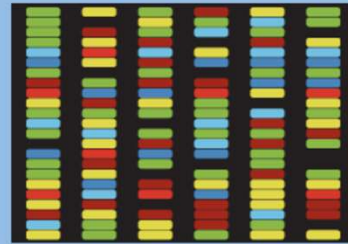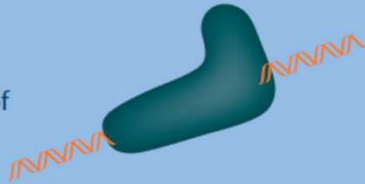combinations; will
be reached soon for
some crops
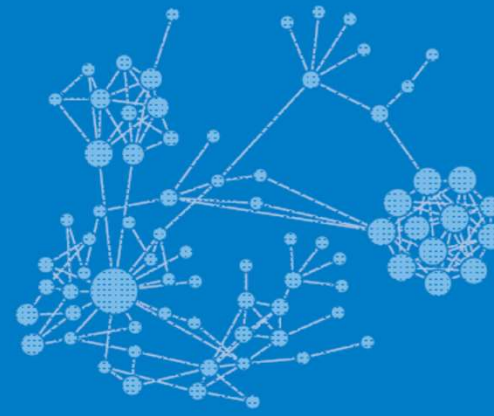
ANNUAL REVIEWS

Annual Review of Genetics

On the Road to Breeding
4.0: Unraveling the Good,
the Bad, and the Boring of
Crop Quantitative Genomics

Jason G. Wallace,[1] Eli Rodgers-Melnick,[2]
and Edward S. Buckler[3,4]

[1]Department of Crop and Soil Sciences, The University of Georgia, Athens,
Georgia 30602, USA; email: jason.wallace@uga.edu
[2]Corteva Agriscience, DowDuPont, Johnston, Iowa 50111, USA
[3]United States Department of Agriculture, Agricultural Research Service,
Ithaca, New York 14853, USA
[4]Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853, USA

THE
CODE
BREAKER

Jennifer Doudna, Gene Editing,
and the Future of the Human Race

WALTER
ISAACSON

BESTSELLING AUTHOR OF *Leonardo da Vinci* and *Steve Jobs*
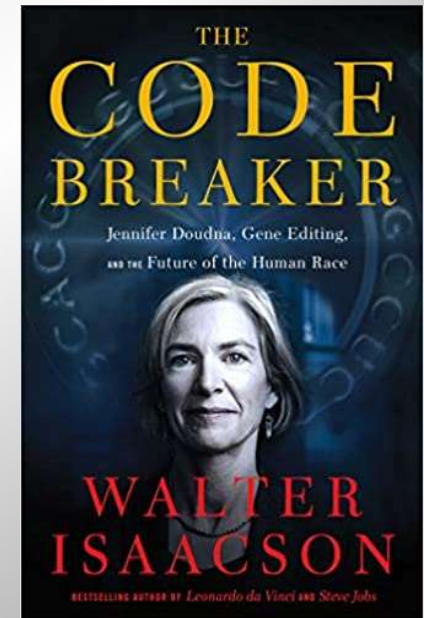
AGRI PHARMACOGENOMICS

Potentials and Challenges of Genomics for Breeding Cannabis Cultivars

Gianni Barcaccia[*], Fabio Palumbo, Francesco Scariolo, Alessandro Vannozzi, Marcello Borin and Stefano Bona

Pomegranate Breeding: Utilization of Molecular and Genetic Data for Improvement of Fruit Quality and Adaptation to Different Climatic Conditions

R. Harel-Beja[1], I. Bar-Ya'akov[1], K. Hatib[1], T. Trainin[1], Z. Ben-Simhon[1], D. Holland[1], R. Eshed[2], M. Sharabi[2], M. Rubinstein[2], R. Ophir[2] and A. Sherman[2]
[1] Unit of Deciduous Fruit Tree Sciences, Newe Ya'ar Research Center, Agricultural Research Organization, PO Box 1021, Ramat Yishay 30095, Israel
[2] Genomic Unit, Plant Sciences Institute, Volcani Center, Agricultural Research Organization, PO Box 6, Bet Dagan 50250, Israel

Keywords: *Punica granatum* L., germplasm collection, segregating populations, transcriptome, SNP markers, genetic map

LOOKING BACKWARDLY AND FORWARDLY...
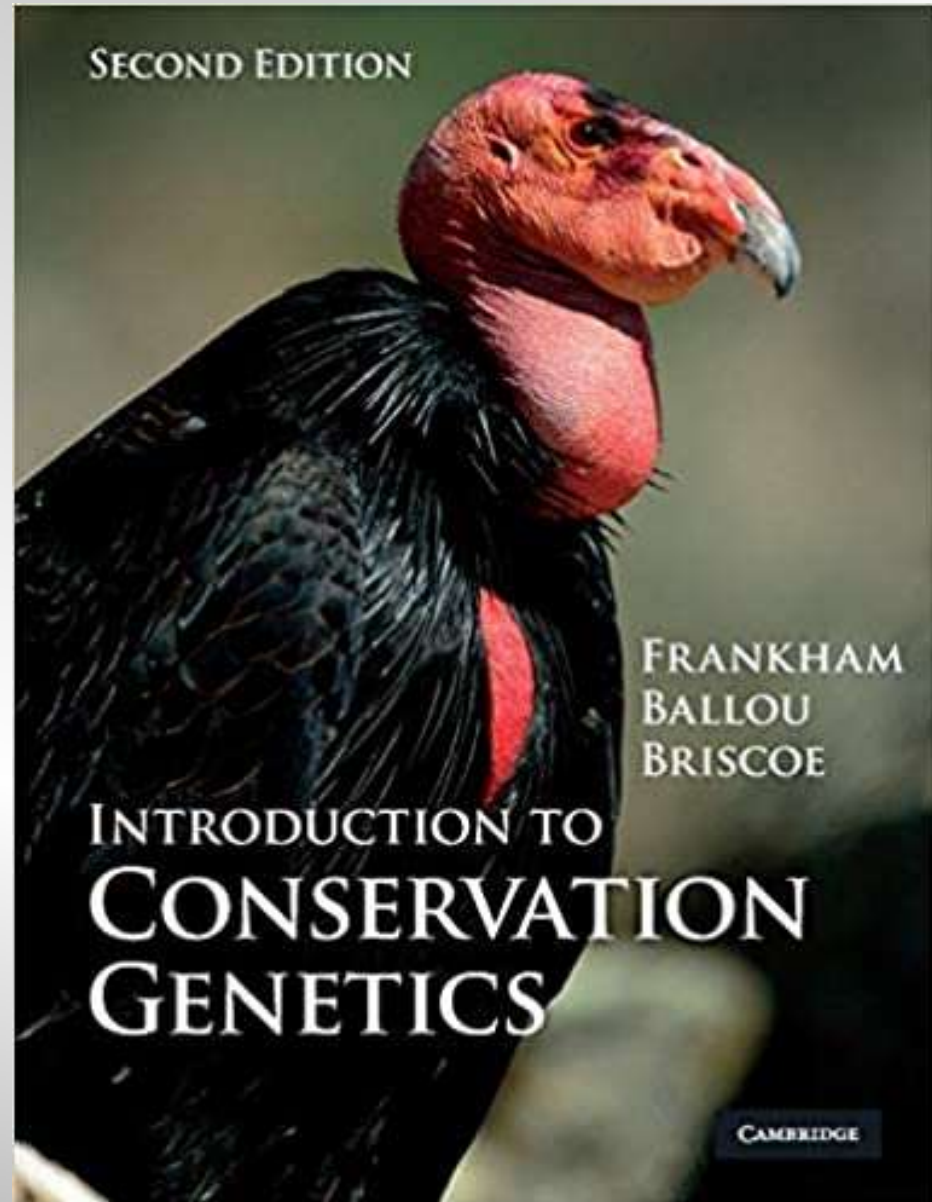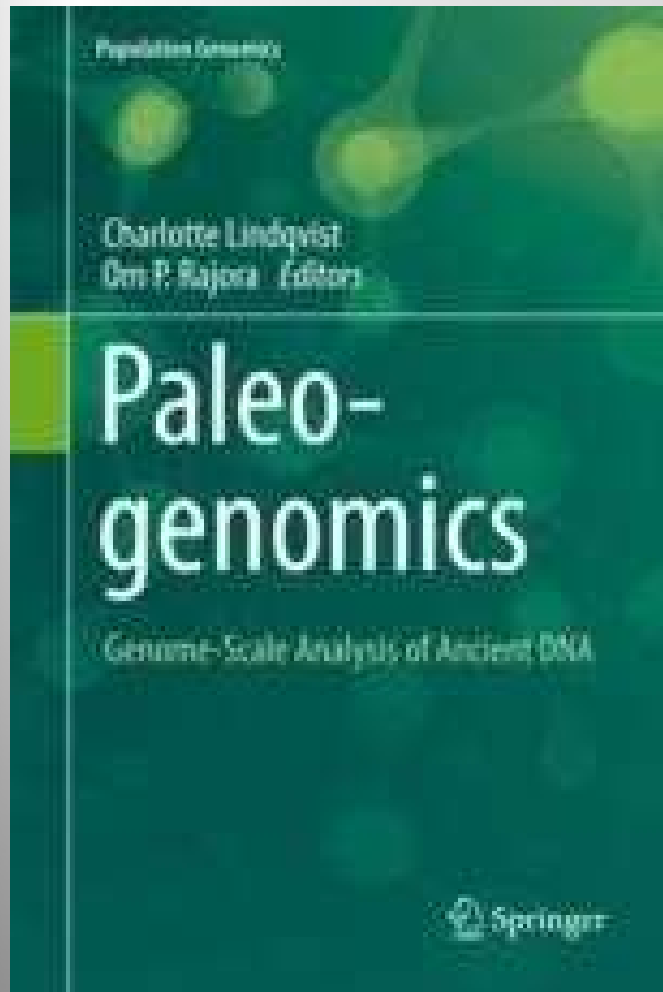
REMARK 1. Breeding objectives (1936: Smith--1943: Hazel) BUT WHAT DO WE BREED FOR TODAY?

$$H \quad = \quad \text{aggregate genetic value. Includes trait genotypes to improve THAT MATTER in some merit function (linear or non-linear) and their (socioeconomic) values}$$

$$\mathbf{y}|\mathbf{X} \quad = \quad \text{vector of RELEVANT MEASURES and explanatory variables}$$

$$H, \mathbf{y} \quad = \quad \text{Must have a statistical non-trivial joint distribution such that}$$

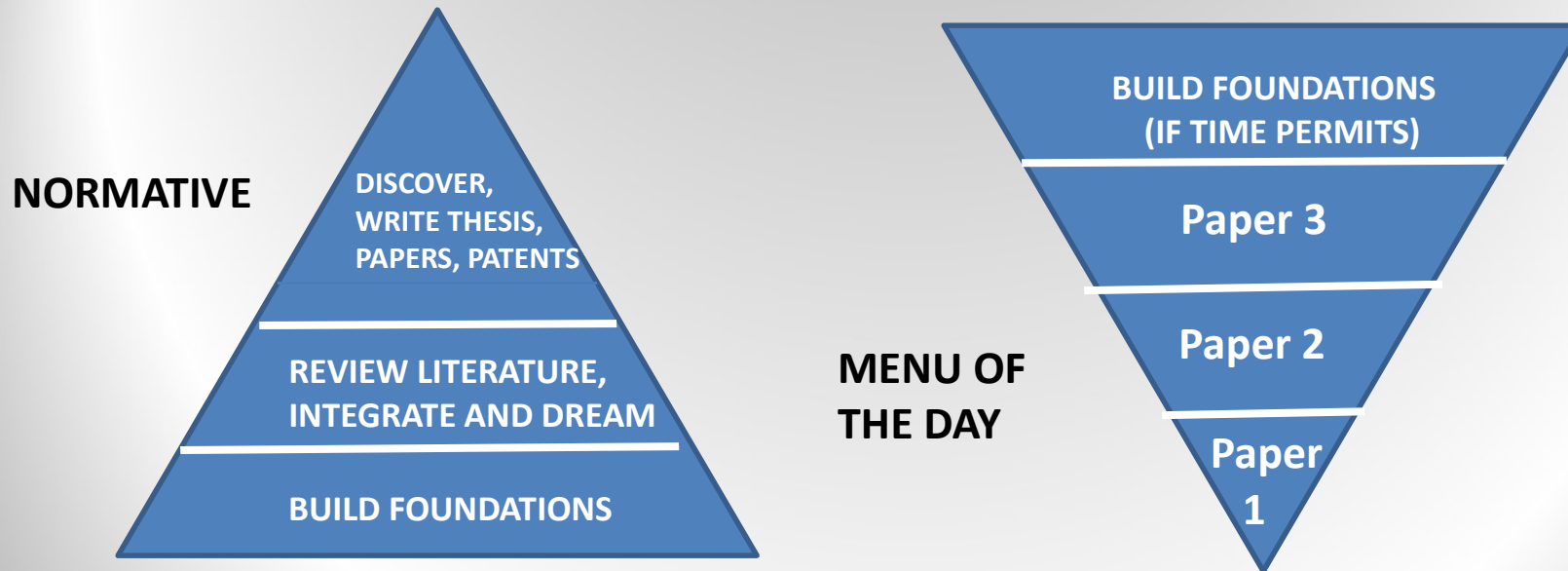$$Entropy(H) \quad > \quad Entropy(H|\mathbf{y})$$

[animal breeders live in a flat earth and use correlation]

TODAY'S WAPATULI:

Brittleness, fragility, robustness, resilience, sustainability, animal welfare (activity, sociability, inclusiveness, dignity, flatulence, biometrics), epigenome, interactome, metagenome, environmental frailtome, noiseome, eco-friendliness, drones, sensors, images, smart phones, infra-red measures, spectrometry, metabolome, "chipomics", crops and livestock models

plus **STILL** relevant production, reproduction and health, pedigree

REMARK 2. Diatribes: the learning triangle

NORMATIVE

DISCOVER,
WRITE THESIS,
PAPERS, PATENTS

REVIEW LITERATURE,
INTEGRATE AND DREAM

BUILD FOUNDATIONS

MENU OF THE DAY

BUILD FOUNDATIONS
(IF TIME PERMITS)

Paper 3

Paper 2

Paper 1

- Algorithms, software, visualization used by students with deficiencies in biology, experimental design, causality, logic, and statistical science. Basic science fundamental.

- Breeding objectives cannot be delineated by specialists.

- AI will not generate fertile interactions if compartments do not intersect