Prediction of complex traits

Daniel Gianola

Sewall Wright Emeritus Professor of Animal Breeding and Genetics

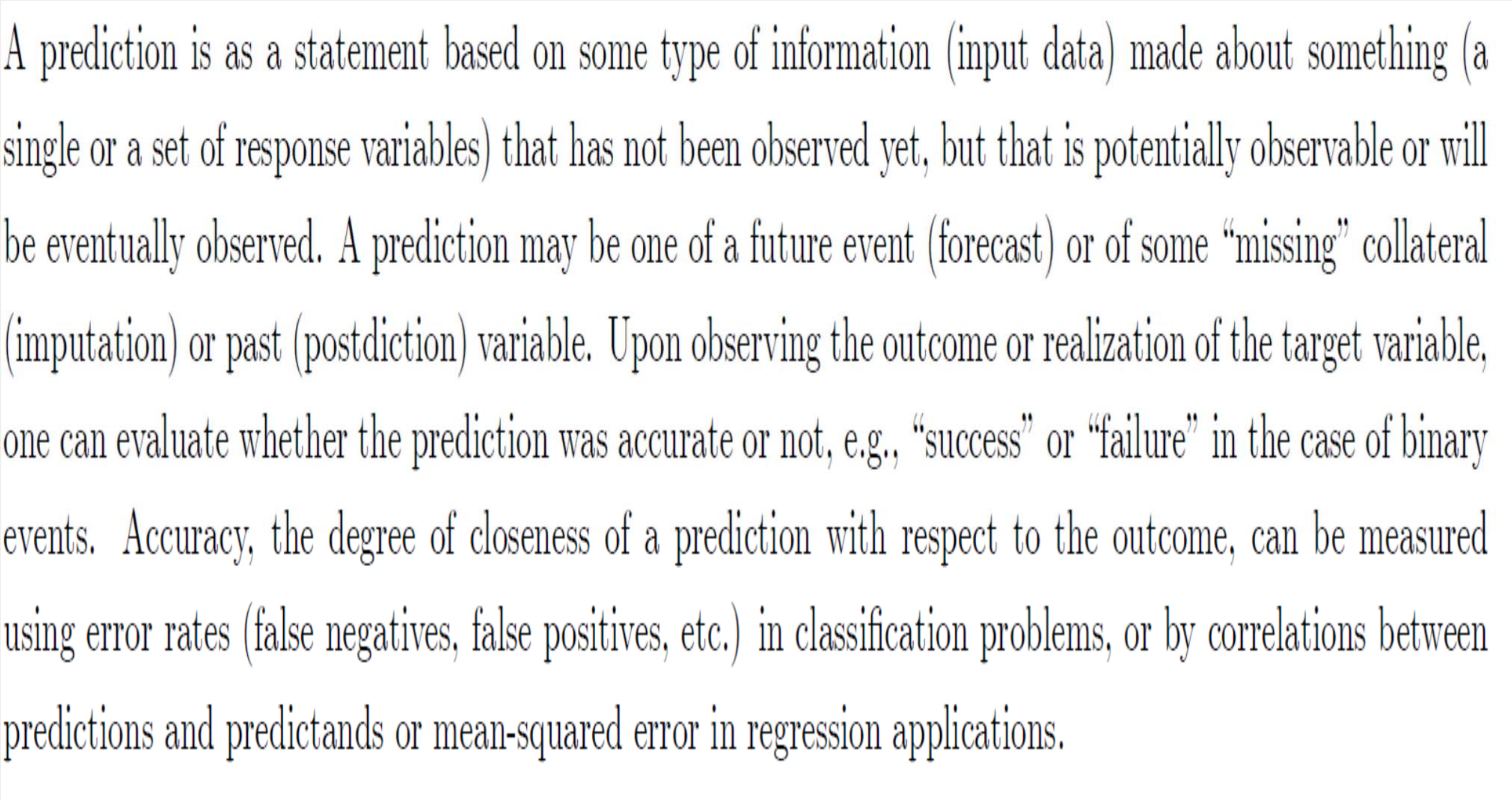
University of Wisconsin-Madison
USA

UW–MADISON
ANIMAL SCIENCES

Dairy Science

Universidad Politecnica de Valencia, Spain
May 22-26, 2017

0. DEFINITION OF PREDICTION



A prediction is as a statement based on some type of information (input data) made about something (a single or a set of response variables) that has not been observed yet, but that is potentially observable or will be eventually observed. A prediction may be one of a future event (forecast) or of some "missing" collateral (imputation) or past (postdiction) variable. Upon observing the outcome or realization of the target variable, one can evaluate whether the prediction was accurate or not, e.g., "success" or "failure" in the case of binary events. Accuracy, the degree of closeness of a prediction with respect to the outcome, can be measured using error rates (false negatives, false positives, etc.) in classification problems, or by correlations between predictions and predictands or mean-squared error in regression applications.

1. UNDERLYING PHILOSOPHY OF THE COURSE

Predictive inference (Geisser 1993)

"Clearly hypothesis testing and estimation as stressed in almost all statistics books involve parameters...this presumes the truth of the model and imparts an inappropriate existential meaning to an index or parameter...inferring about observables is more pertinent since they can occur and be validated to a degree that is not possible for parameters".

-Bayesian methods: important role in machine learning.

-Bayes theorem: provides predictive distribution automatically.

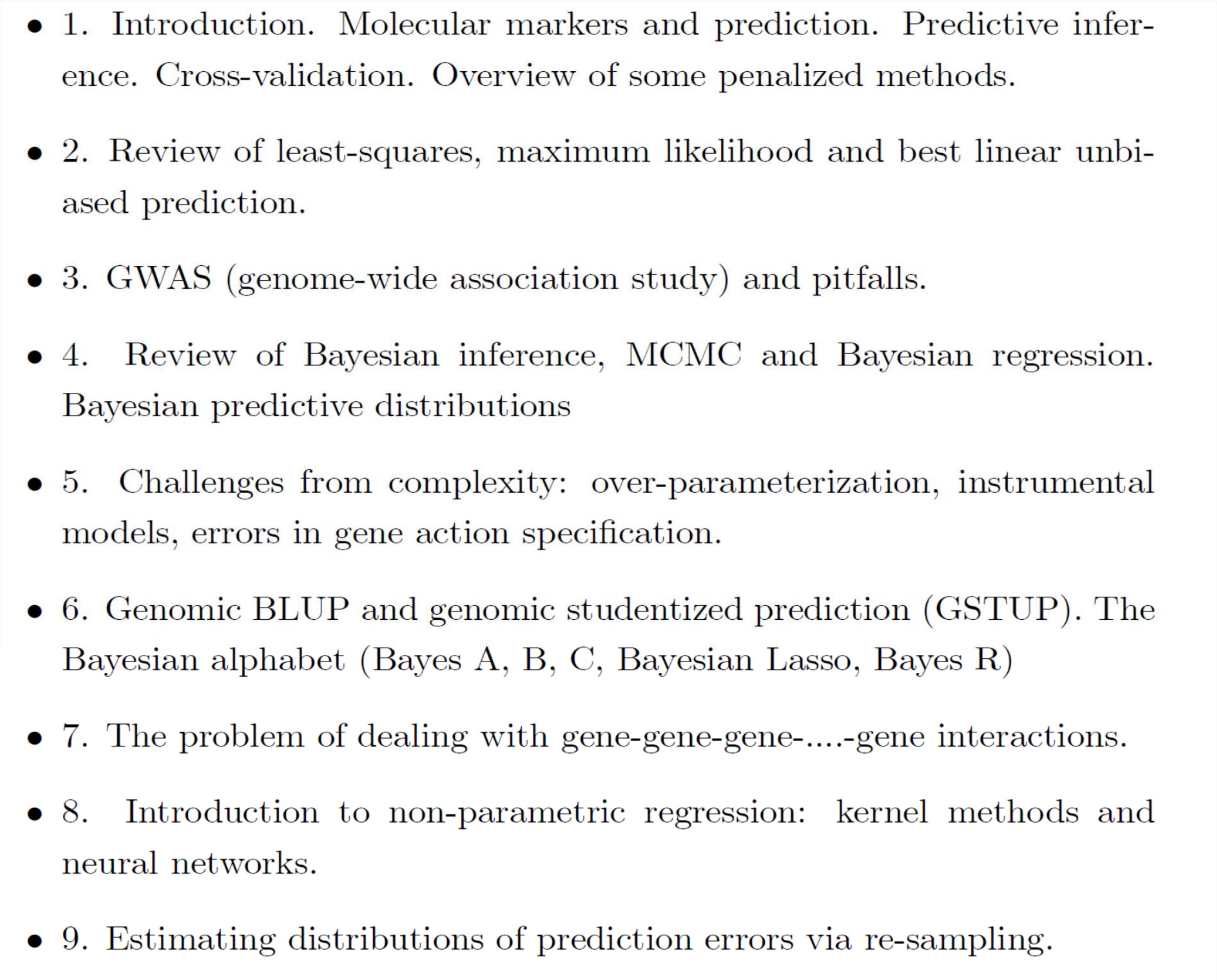
-Has not been appreciated in full yet in whole-genome prediction literature.

-This distribution, however, is based on a model. There is model uncertainty

-Impossible to arrive at a model that can be taken seriously, at least mechanistically, for any complex trait…

-We will argue later why this is the case, at least partially!

2. TOPICS COVERED (ORDER MAY VARY)

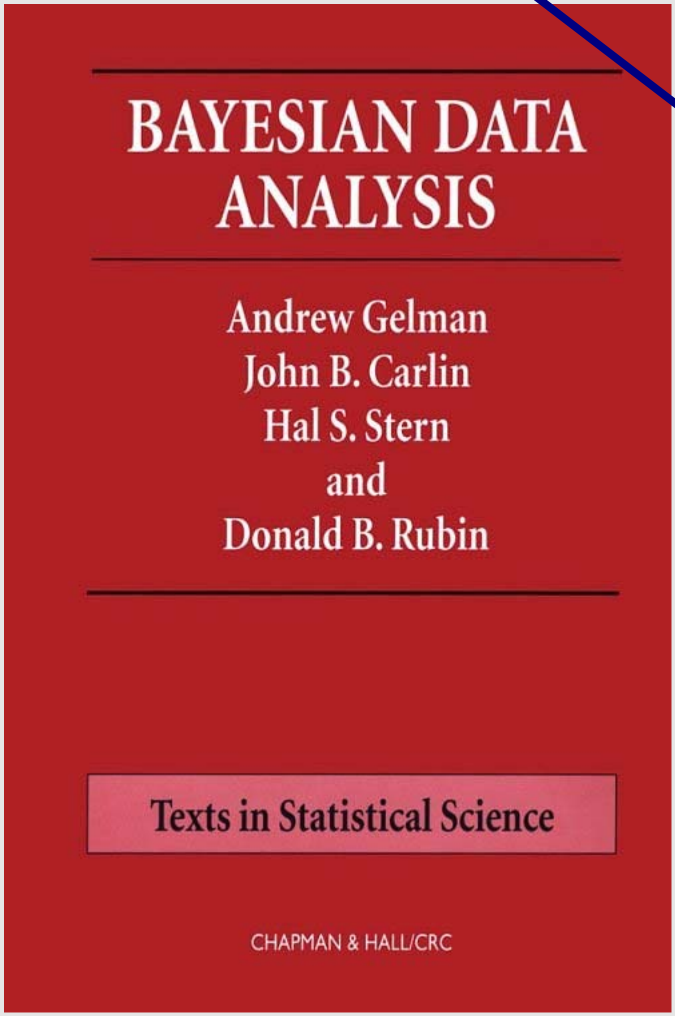


- 1. Introduction. Molecular markers and prediction. Predictive inference. Cross-validation. Overview of some penalized methods.
- 2. Review of least-squares, maximum likelihood and best linear unbiased prediction.
- 3. GWAS (genome-wide association study) and pitfalls.
- 4. Review of Bayesian inference, MCMC and Bayesian regression. Bayesian predictive distributions
- 5. Challenges from complexity: over-parameterization, instrumental models, errors in gene action specification.
- 6. Genomic BLUP and genomic studentized prediction (GSTUP). The Bayesian alphabet (Bayes A, B, C, Bayesian Lasso, Bayes R)
- 7. The problem of dealing with gene-gene-gene-....-gene interactions.
- 8. Introduction to non-parametric regression: kernel methods and neural networks.
- 9. Estimating distributions of prediction errors via re-sampling.

TEXTS THAT MAY BE USEFUL AS SUPPORTING MATERIALS FOR THIS COURSE

Gelman A., Carlin, J. B., Stern, H. and Rubin, D. B. 1995. Chapman&Hall/CRC

Gelman A., Carlin, J. B., Stern, H., Dunson D. V.; Vethari A; Rubin, D. B. 2013. Chapman&Hall/CRC

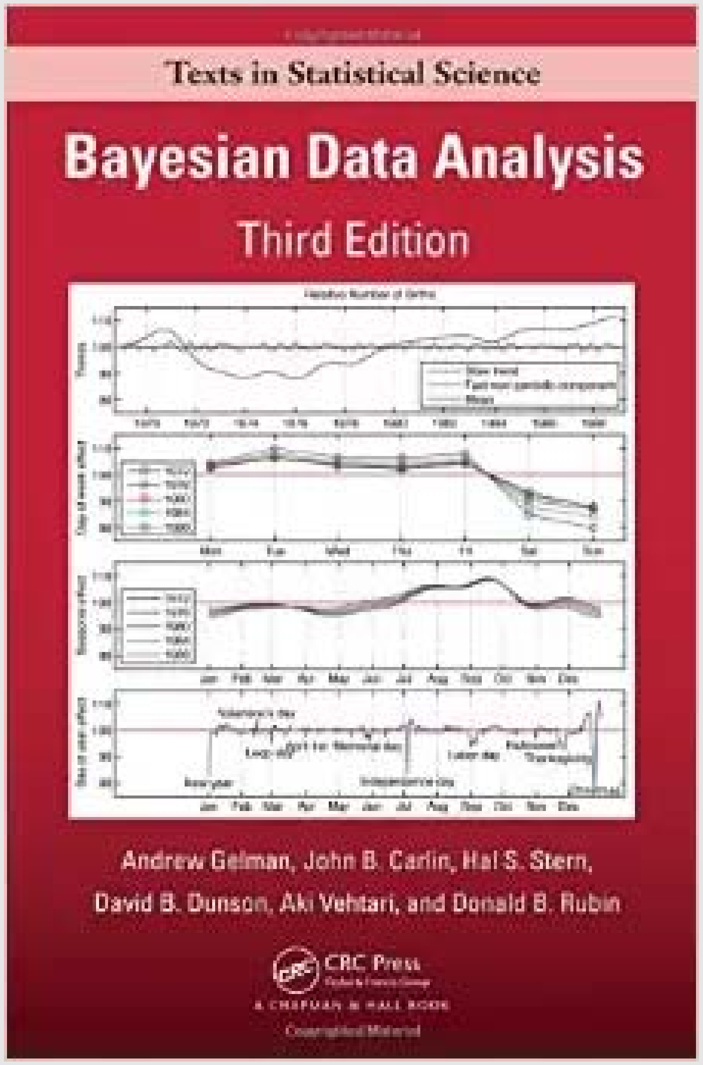


BAYESIAN DATA ANALYSIS

Andrew Gelman
John B. Carlin
Hal S. Stern
and
Donald B. Rubin

Texts in Statistical Science

CHAPMAN & HALL/CRC



Copyrighted Material

Texts in Statistical Science

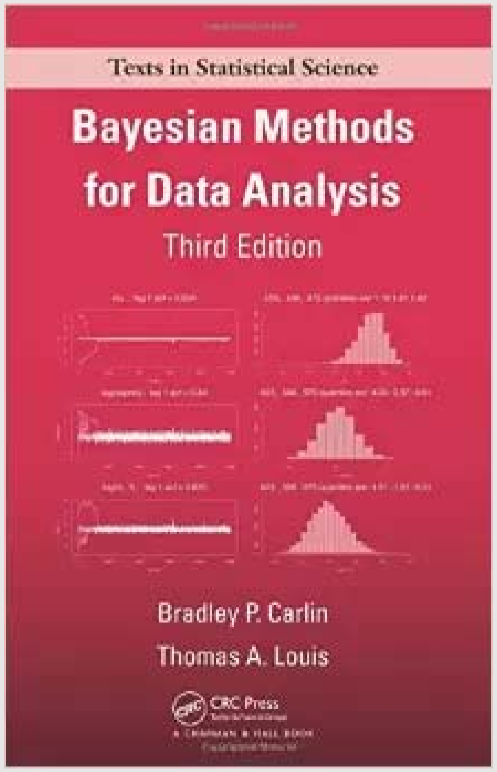
Bayesian Data Analysis

Third Edition

Andrew Gelman, John B. Carlin, Hal S. Stern,
David B. Dunson, Aki Vehtari, and Donald B. Rubin

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Copyrighted Material



Texts in Statistical Science

Bayesian Methods for Data Analysis

Third Edition

Bradley P. Carlin

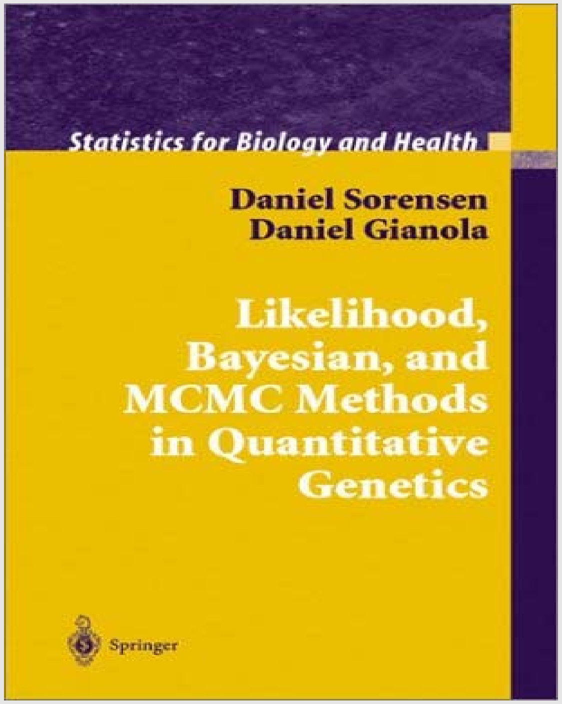
Thomas A. Louis

CRC Press

A CHAPMAN & HALL BOOK

Carlin, B. P. and Louis, T. A. 2008.
Bayesian Methods for Data Analysis,
Third Edition Chapman & Hall/CRC

Sorensen, D. and Gianola, D. 2002. Springer



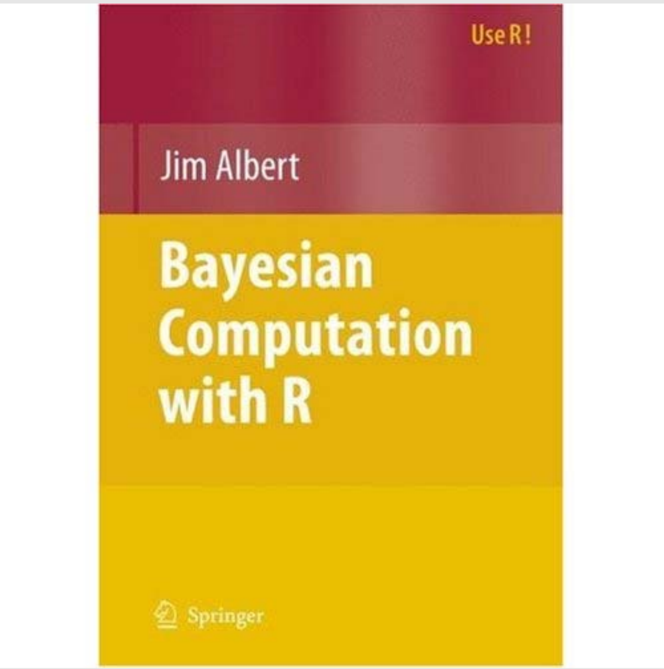
Statistics for Biology and Health

Daniel Sorensen
Daniel Gianola

Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics

Springer

Albert, J. 2009. Bayesian Computation with R. Second Edition



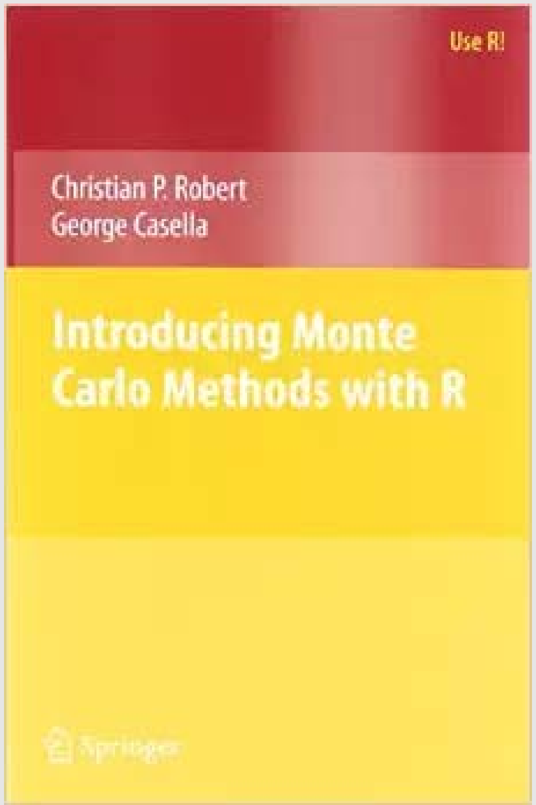
Use R!

Jim Albert

Bayesian Computation with R

Springer

Robert C. P; Casella G. 2010.
Introducing Monte Carlo Methods with R.
Springer.



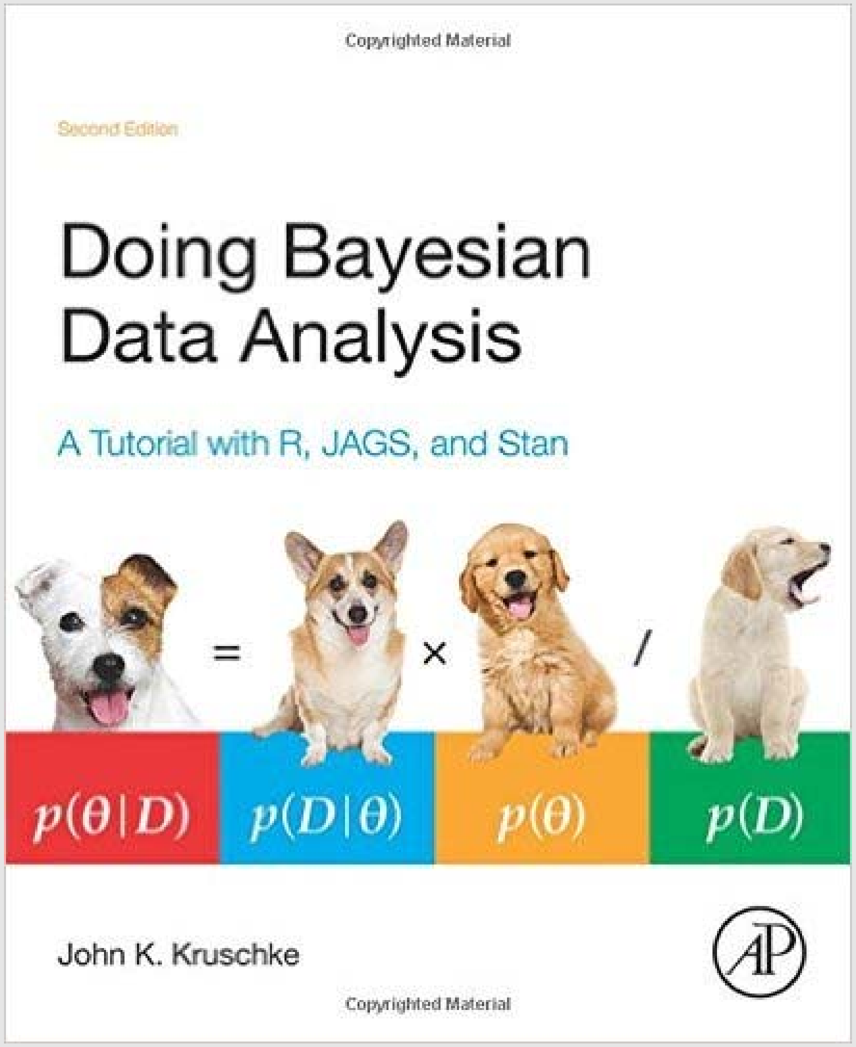
Use R!

Christian P. Robert
George Casella

Introducing Monte Carlo Methods with R

Springer

SOME NEW DEVELOPMENTS IN TEXTBOOKS WITH APPLICATIONS AND SOFTWARE



Copyrighted Material

Second Edition

Doing Bayesian Data Analysis

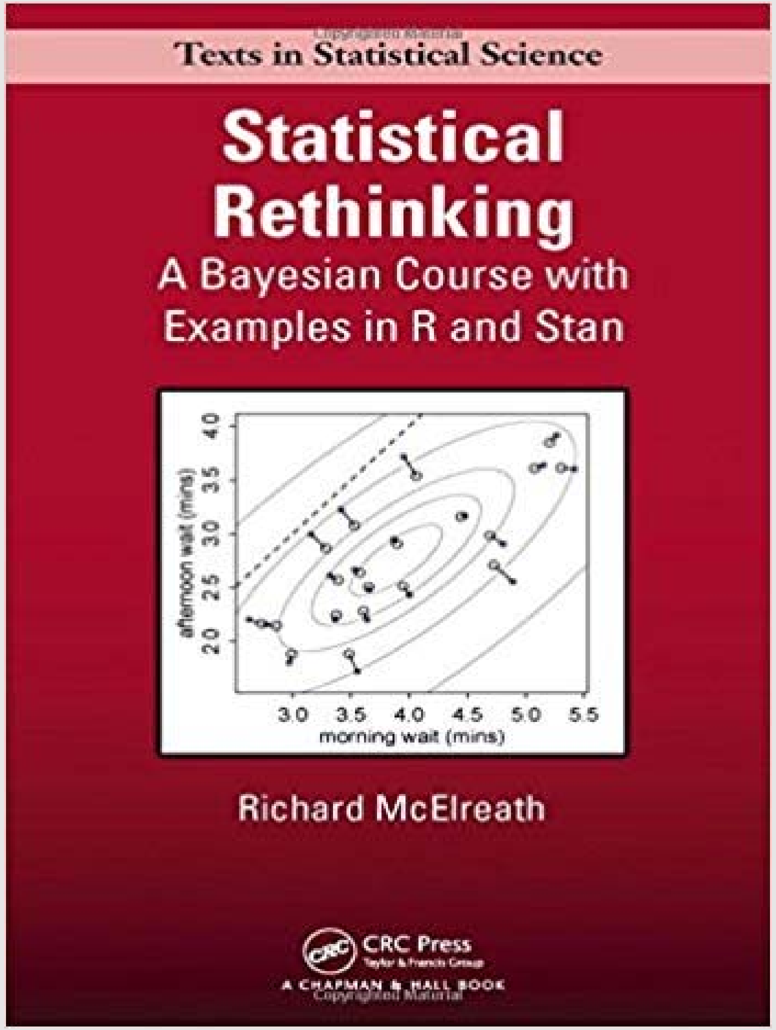
A Tutorial with R, JAGS, and Stan

$p(\theta|D) = p(D|\theta) \times p(\theta) / p(D)$

John K. Kruschke

Copyrighted Material

2014



Texts in Statistical Science

Statistical Rethinking

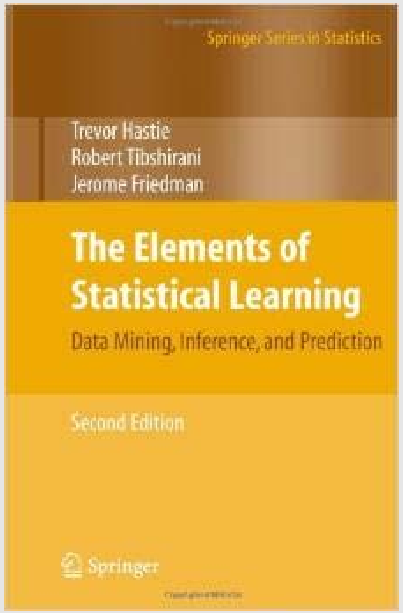
A Bayesian Course with Examples in R and Stan

Richard McElreath

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

2015

Statistical and machine learning


Springer Series in Statistics

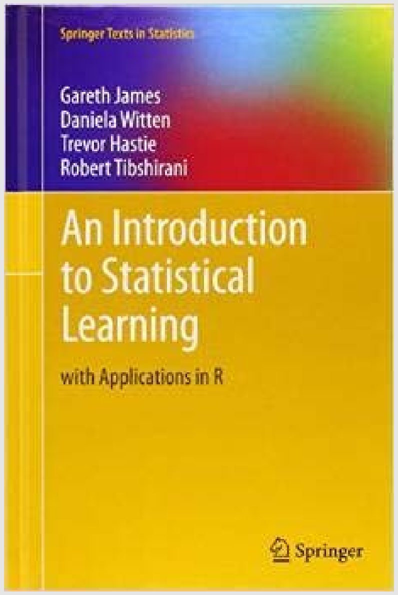
Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer

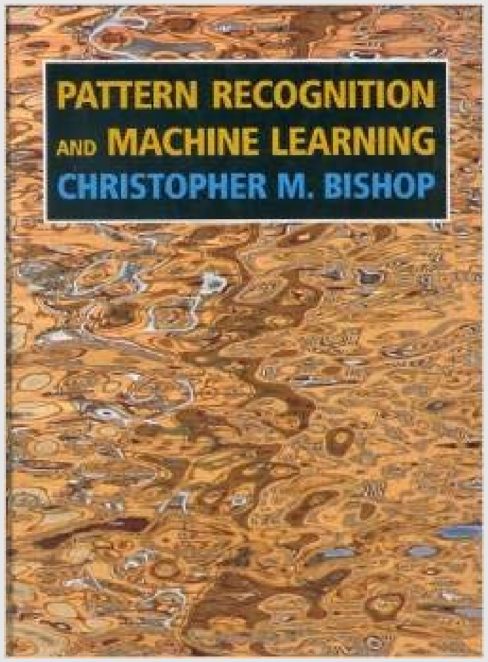

Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

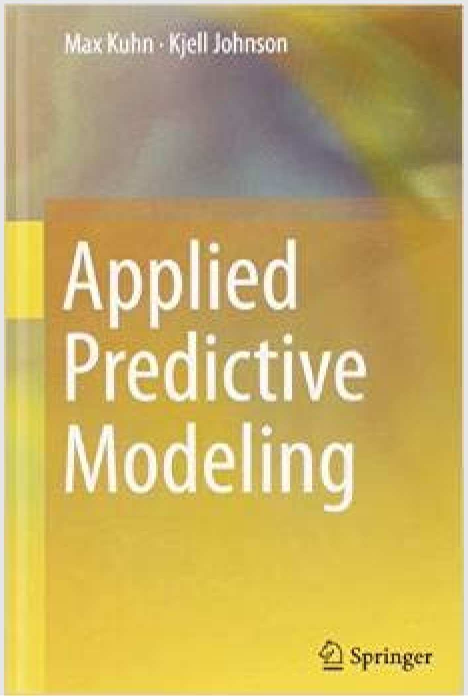
An Introduction to Statistical Learning

with Applications in R

Springer

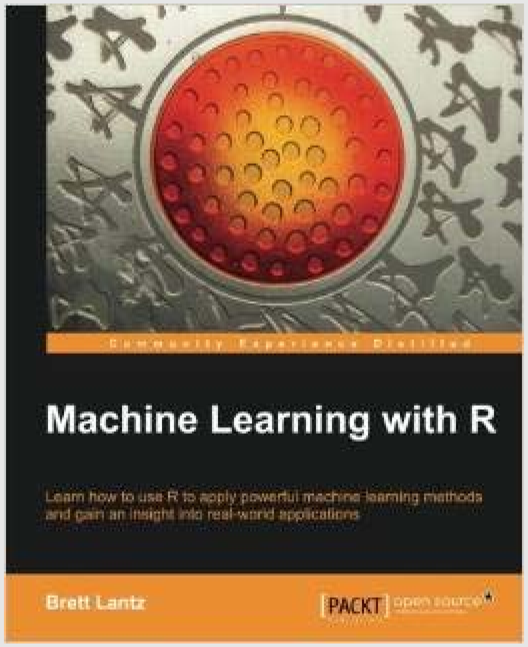

PATTERN RECOGNITION AND MACHINE LEARNING

CHRISTOPHER M. BISHOP


Max Kuhn · Kjell Johnson

Applied Predictive Modeling

Springer

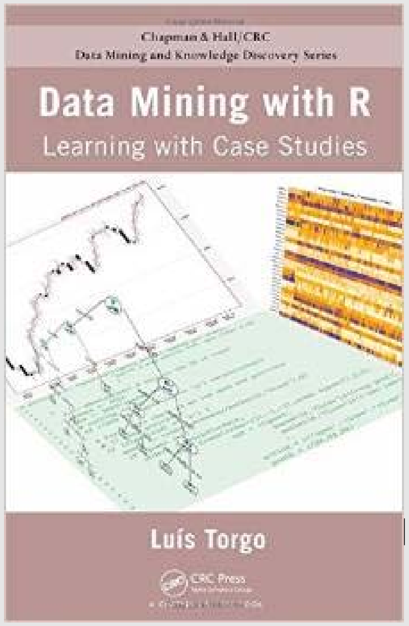

Community Experience Distilled

Machine Learning with R

Learn how to use R to apply powerful machine-learning methods and gain an insight into real-world applications

Brett Lantz

PACKT open source


Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

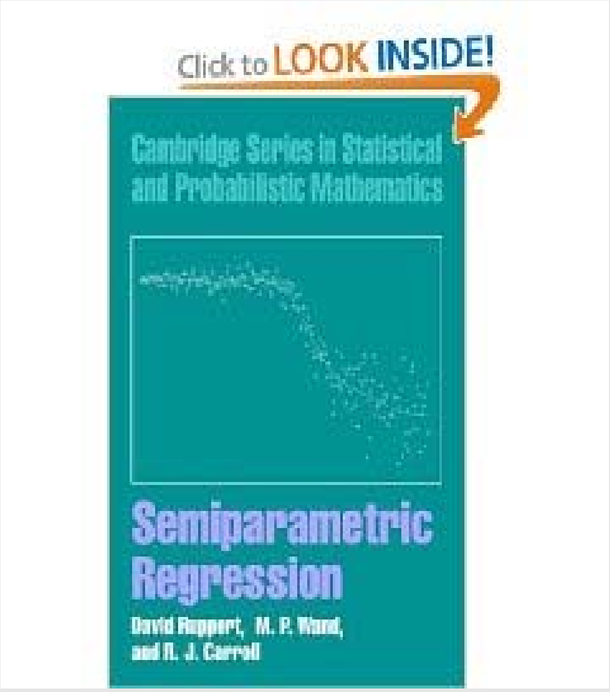
Data Mining with R

Learning with Case Studies

Luís Torgo

CRC Press

Non-parametric regression

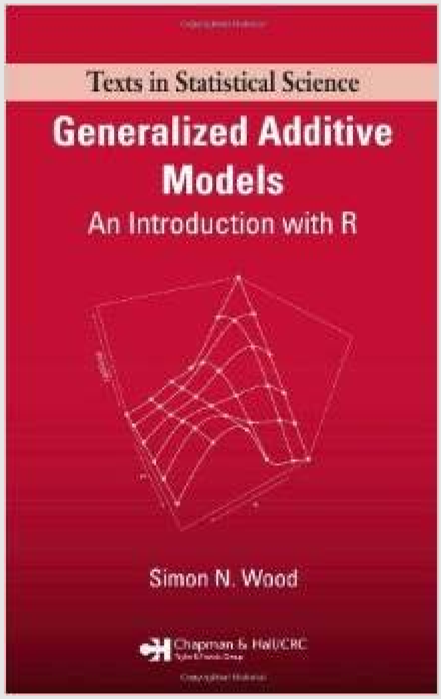


Click to LOOK INSIDE!

Cambridge Series in Statistical and Probabilistic Mathematics

Semiparametric Regression

David Ruppert, M. P. Wand, and R. J. Carroll



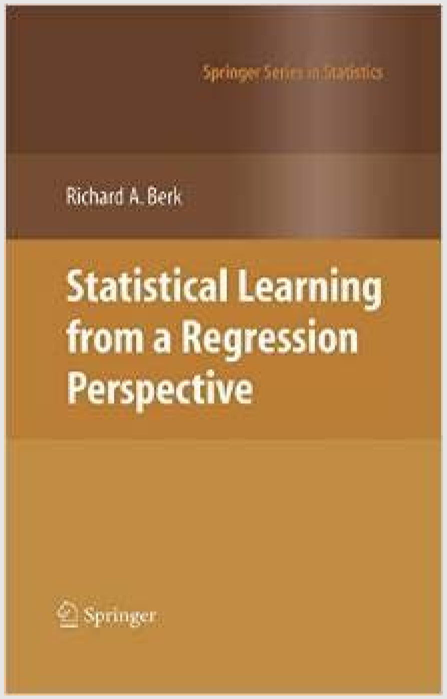
Texts in Statistical Science

Generalized Additive Models

An Introduction with R

Simon N. Wood

Chapman & Hall/CRC



Springer Series in Statistics

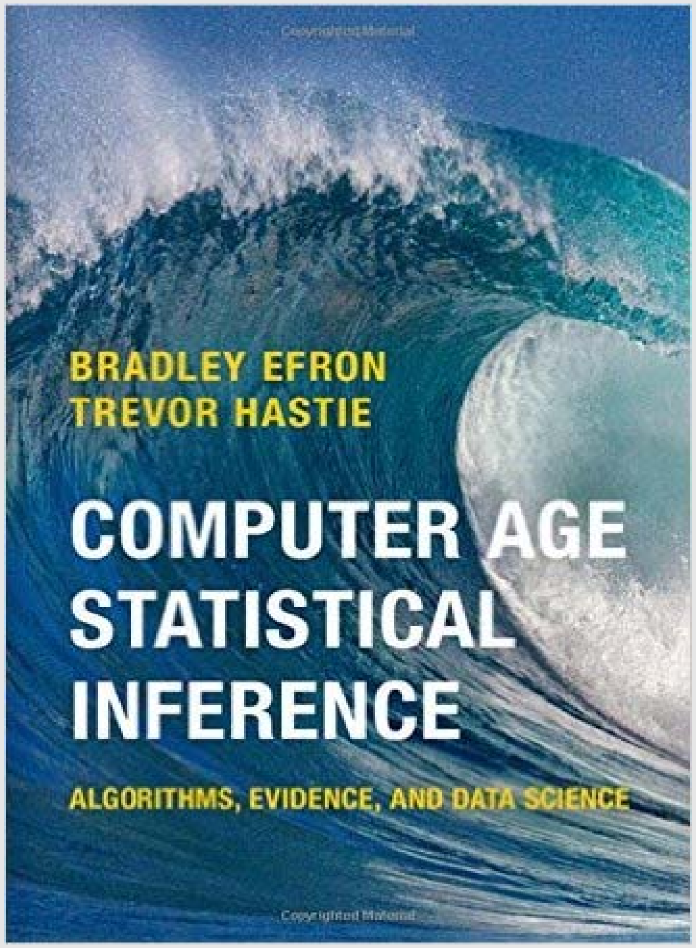
Richard A. Berk

Statistical Learning from a Regression Perspective

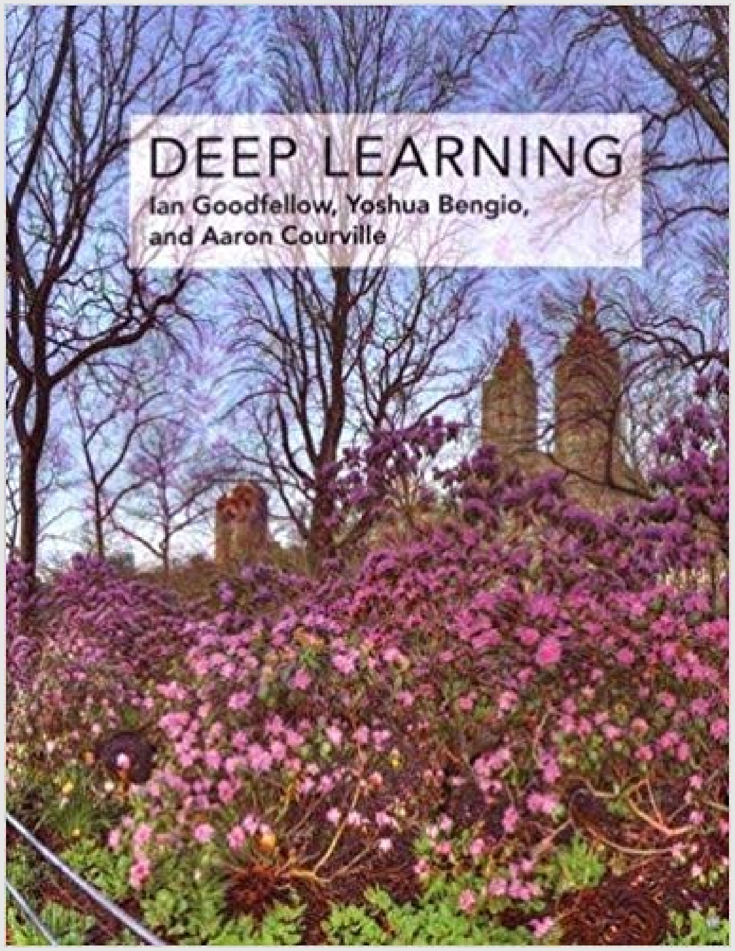
Springer

THE END OF HISTORY?

2016


Copyrighted Material
BRADLEY EFRON
TREVOR HASTIE
COMPUTER AGE STATISTICAL INFERENCE
ALGORITHMS, EVIDENCE, AND DATA SCIENCE
Copyrighted Material

2016



I am not a robot…or am I?

3. EVOLUTION OF STATISTICAL METHODS IN ANIMAL BREEDING AND QUANTITATIVE GENETICS

Balding et al. (2007)
“Handbook of Statistical Genetics”. Wiley

Chapter 20

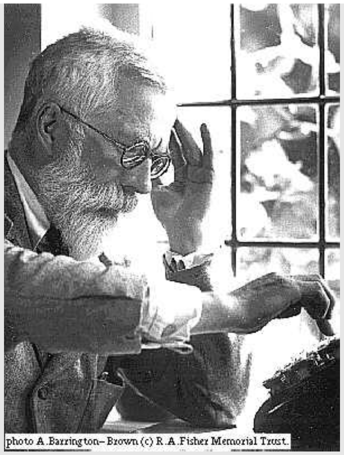
D. Gianola

“Inferences from Mixed Models in Quantitative Genetics”

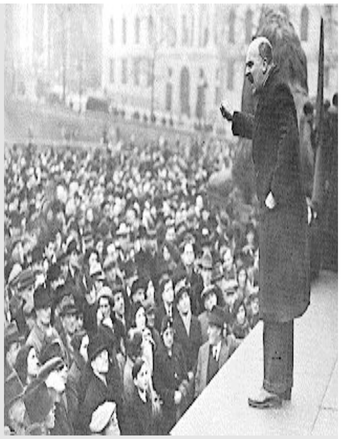
Gianola and Rosa (2015)
“One hundred years of statistical developments in animal breeding”

Annu. Rev. Anim. Biosci. 2015. 3:19–56

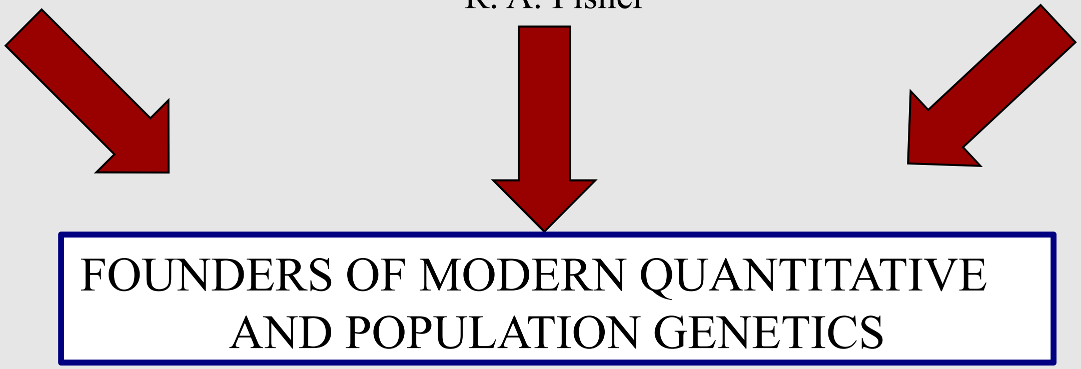
Sewall Wright



R. A. Fisher



J. B. S. Haldane

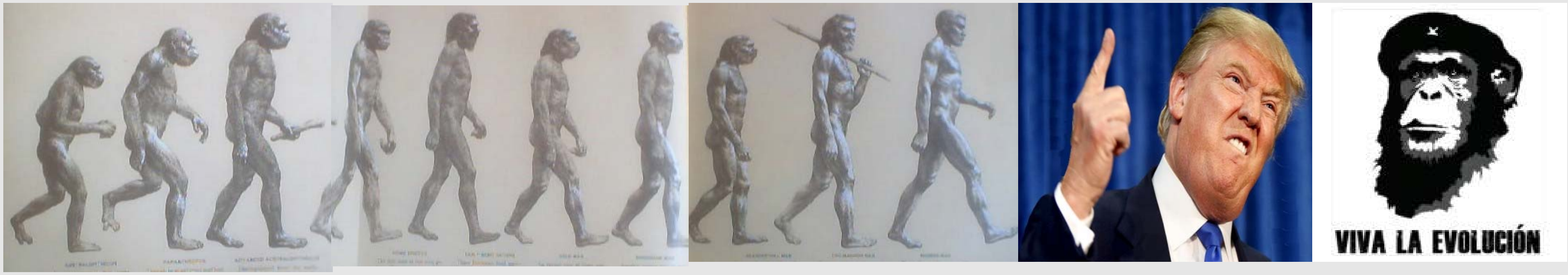


FOUNDERS OF MODERN QUANTITATIVE AND POPULATION GENETICS

SCIENTIFIC FOUNDATIONS
OF
ANIMAL (PLANT) BREEDING

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

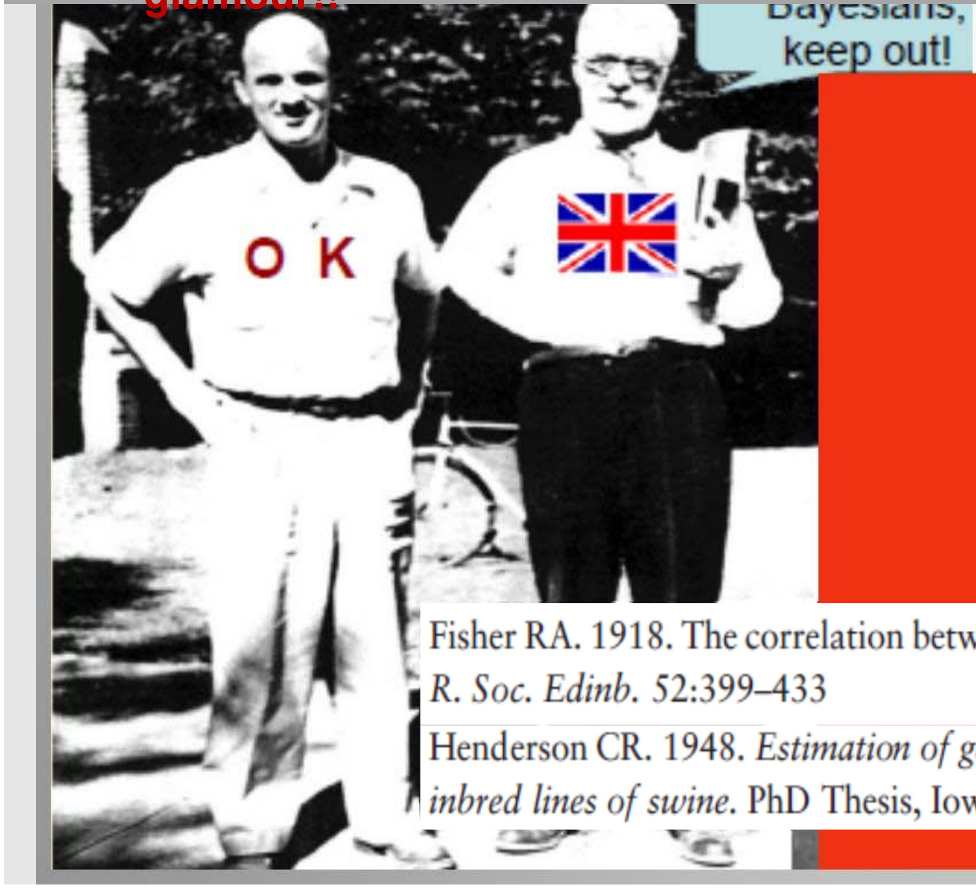
HISTORICAL PROGRESSION



Archaen	Galton’s regression; Pearson: density estimation	[Early 20th century]
Pathozoic	Fisher’s 1918, Path analysis, “Animal Breeding Plans”	[1918-1945]
Anovian	Least-squares, (CO) Variance components: Henderson’s Method 1, Selection index	[1936-1943]
Post-anovian	Henderson’s 2+3, Rao’s MINQUE and MIVQUE	[1953-1973]
Blupassic	Mixed models, BLUP, animal model, multi-traits, random regression	[1948-2009]
Remlian	Maximum likelihood: VCE, ASREML, DMU, WOMBAT	[1971-2009]
Posteriozoic	Threshold models, Survival, MCMC	[1982-2008]
Genomacic	QTLs, GWAS, whole-genome prediction, machine learning, networks, “causal variants”	[2001-present]

TOME 1: FROM FISHER TO HENDERSON

- Fisher’s mean became a mean vector
- Fisher’s additive variance became a covariance matrix
- BLUP and the mixed model equations were developed
- Wright’s NRM matrix found to have an easy inverse
- BLUP extended to cross-sectional, longitudinal and multivariate data
- BLUP even for non-existing individuals…
- Good estimates of dispersion parameters needed! ML and REML (Thompson)
- More efficient production of milk, meat, eggs and fiber!
- No genes, no MAF, no LD, no “causal variants”. No Nature Genetics or NIH, no glamour!!



Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* 52:399–433

Henderson CR. 1948. *Estimation of general, specific and maternal combining ability in crosses among inbred lines of swine.* PhD Thesis, Iowa State Univ.

THE n<<p ERA

(in animal breeding, began in 1948-1973: C. R. HENDERSON)

$$\mathbf{y} = \underset{\overset{\uparrow}{\text{Fixed}}}{\mathbf{X}\boldsymbol{\beta}} + \underset{\overset{\nearrow}{\text{Random}}}{\mathbf{Z}\mathbf{u}} + \mathbf{e}$$

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \mathbf{R} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R})$$

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$$

BLUP=Best linear unbiased predictor

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

BLUP= Conditional posterior mean in Bayes Gaussian linear hierarchical model

BLUP=penalized (L2) maximum likelihood

BLUP=Similar to kriging in geostatistics

BLUP=special case of RKHS regression

BLUP=Exactly like the Holy Roman Empire (Gelman)

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

Rev. Thomas Bayes

1702 London, England
1761 Tunbridge Wells, Kent, England

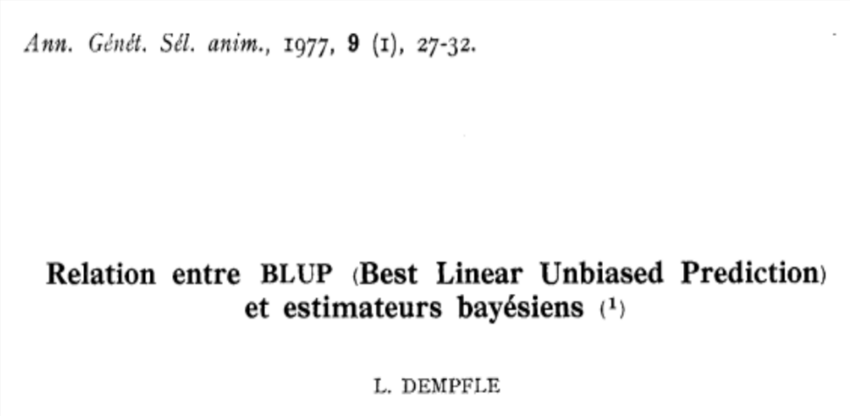
1763. “An essay towards solving a problem in the doctrine of chances”. *Philosophical Transactions of the Royal Society of London* **53**, 370-418.

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

TOME 2: FROM HENDERSON TO BAYES

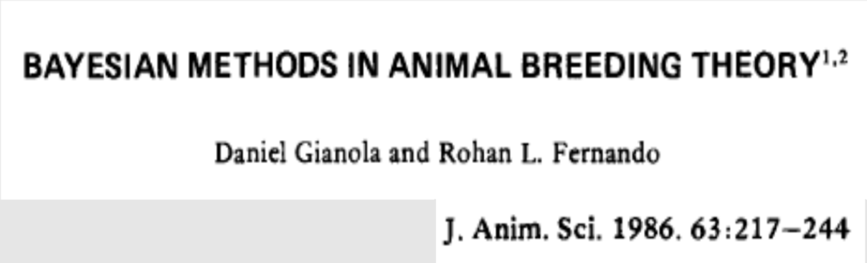


Ann. Génét. Sél. anim., 1977, **9** (1), 27-32.

Relation entre BLUP (Best Linear Unbiased Prediction) et estimateurs bayésiens (1)

L. DEMPFLE

PRE-MCMC

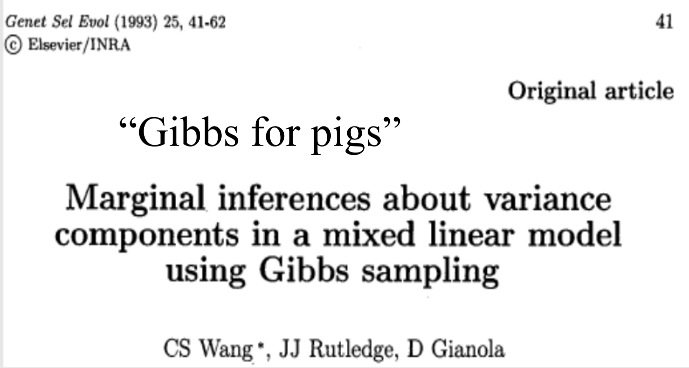


BAYESIAN METHODS IN ANIMAL BREEDING THEORY[1,2]

Daniel Gianola and Rohan L. Fernando

J. Anim. Sci. 1986. 63:217–244

Gianola, D., Foulley, J.L. Non-linear prediction of latent genetic liability with binary expression: an empirical Bayes approach. in: Proc. 2nd World Congr. Genet. Appl. Livest. Prod. VII. ; 1982: 293–303.



Genet Sel Evol (1993) 25, 41-62 41
© Elsevier/INRA

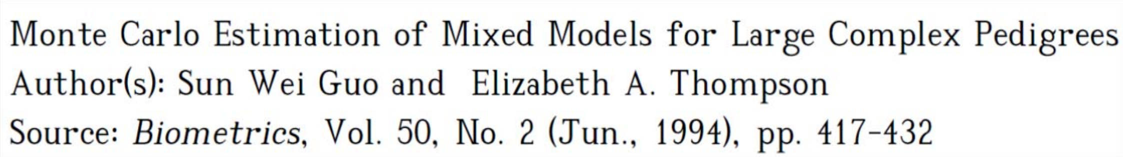
Original article

"Gibbs for pigs"

Marginal inferences about variance components in a mixed linear model using Gibbs sampling

CS Wang*, JJ Rutledge, D Gianola

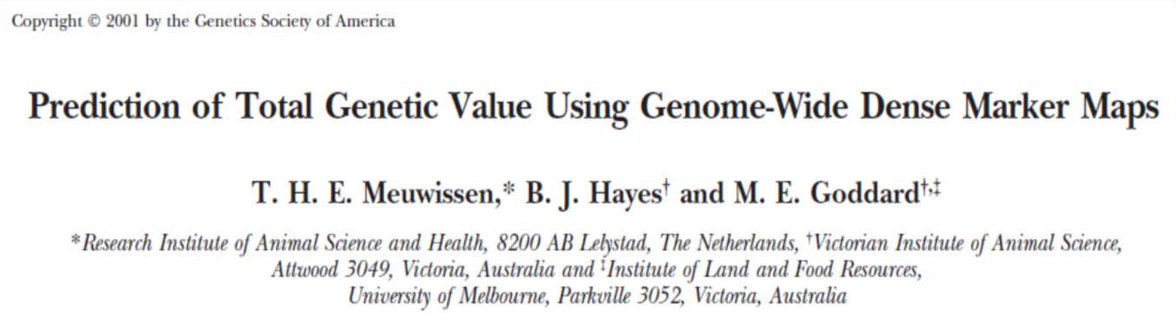
MCMC



Monte Carlo Estimation of Mixed Models for Large Complex Pedigrees
Author(s): Sun Wei Guo and Elizabeth A. Thompson
Source: *Biometrics*, Vol. 50, No. 2 (Jun., 1994), pp. 417-432

MCMC WITH MARKER DATA

GENOMIC SELECTION



Copyright © 2001 by the Genetics Society of America

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*

Bayesian methods in Genetics: today

-Classification of genotypes
-Molecular evolution
-Linkage mapping
-QTL cartography
-Genetic risk analysis
-Gaussian linear and non-linear models
: cross-sectional+ longitudinal univariate+ multivariate
-Generalized linear models
-Survival analysis
-Thick-tailed processes
-Mixtures
-Semi-parametrics
-Transcriptional analysis
-ABC in population genetics
-Structural equation modeling
-Bayesian proteomics with wavelets
-Bayesian non-parametrics (Dirichlet process priors, RKHS)
-EPITOME: Methods for genomic selection
(the Bayesian Alphabet—A, B, C,C-pi, L,R… and more)
-Bayesian multi-omics (DNA+methylation, gene expression, environmentomics)

RED: animal breeders made strong contributions

4. BASIC QUANTITATIVE GENETICS:

the additive genetic model

$$u_i = W_{i1}a_1 + W_{i2}a_2 + \ldots + W_{iK}a_K$$

$$W_{ij}a_j = \begin{cases} -a_j & \text{if } W_{ij} = -1\,(aa);\ \Pr(W_{ij} = -1) = (1-p_j)^2 \\ 0 & \text{if } W_{ij} = 0\,(Aa);\ \Pr(W_{ij} = 0) = 2p_j(1-p_j) \\ a_j & \text{if } W_{ij} = 1\,(AA);\ \Pr(W_{ij} = 1) = p_j^2 \end{cases}$$

Random genotypes (*W's) ;* fixed effects of QTL (*a'*s) ➔ *u* is a random "genetic signal"

+

+…

+

= **'additive genetic value'**

$$\begin{aligned} E(u) &= E(W_1 a_1 + ... + W_K a_K) \\ &= E(W_1) a_1 + ... + E(W_K) a_K \end{aligned}$$

$$\begin{aligned} Var(u) &= a_1^2 Var(W_1) + ... + a_K^2 Var(W_K) && \text{EQUILIBRIUM} \\ &\quad + 2a_1 a_2 Cov(W_1, W_2) + ... + 2a_{K-1} a_K Cov(W_{K-1}, W_K) && \text{DISEQUILIBRIUM} \end{aligned}$$

$Cov(W_1, W_2) = 2D_{12}$

***D is the* linkage disequilibrium statistic**

***If* $\Pr(W_1, W_2) = \Pr(W_1)\Pr(W_2)$: “linkage equilibrium**

THERE IS ALWAYS LINKAGE DISEQUILIBRIUM!

IN AN EQUILIBRIUM POPULATION

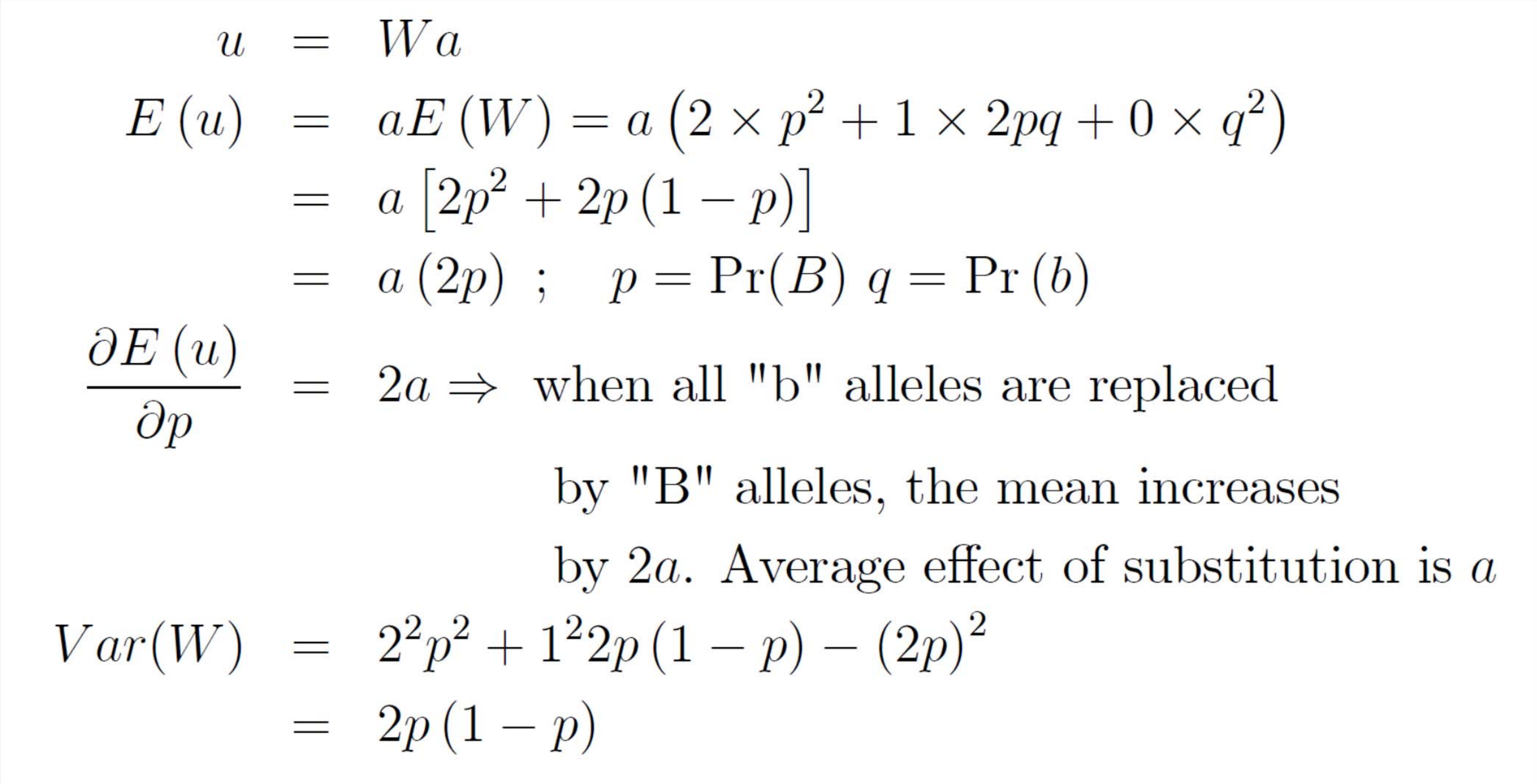
$$Var(u) = \sum_{k=1}^{K} a_k^2 Var(W_k) = \sum_{k=1}^{K} \tau_k^2$$

$$\begin{aligned} \text{IF } K &\rightarrow \infty \\ u &\rightarrow N(E(u), Var(u)) \end{aligned}$$

"INFINITESIMAL MODEL"

Effect of an allelic substitution

W(AA)=2
W(Aa)=1
W(aa)=0



$$\begin{aligned}
u &= Wa \\
E\left(u\right) &= aE\left(W\right) = a\left(2 \times p^2 + 1 \times 2pq + 0 \times q^2\right) \\
&= a\left[2p^2 + 2p\left(1-p\right)\right] \\
&= a\left(2p\right)\ ; \quad p = \Pr(B)\ q = \Pr\left(b\right) \\
\frac{\partial E\left(u\right)}{\partial p} &= 2a \Rightarrow \text{ when all "b" alleles are replaced} \\
&\qquad \text{by "B" alleles, the mean increases} \\
&\qquad \text{by } 2a\text{. Average effect of substitution is } a \\
Var(W) &= 2^2p^2 + 1^2 2p\left(1-p\right) - \left(2p\right)^2 \\
&= 2p\left(1-p\right)
\end{aligned}$$

Epistasis: interaction between genotypes at different loci

$$\begin{aligned}
u &= W_1a_1 + W_2a_2 + W_1W_2a_{12} \Longrightarrow \text{additive x additive epistasis}\\
E\left(u\right) &= a_1E\left(W_1\right) + a_2E\left(W_2\right) + a_{12}E\left(W_1W_2\right)\\
&= a_1E\left(W_1\right) + a_2E\left(W_2\right) + a_{12}E\left(W_1W_2\right)\left\{\begin{array}{c} E\left(W_1\right)E\left(W_2\right) \text{ if equilibrium}\\ E\left(W_1\right)E\left(W_2\right) + Cov\left(W_1W_2\right) \text{ if not}\end{array}\right.\\
Cov\left(W_1W_2\right) &= 2D_{12} \left[\text{D here is the LD statistic (covariance between allelic codes)}\right]
\end{aligned}$$

Epistasis: substitution effect depends on allelic frequency distribution at other loci

$$\frac{\partial E\left(u\right)}{\partial p_1} = a_1\frac{\partial E\left(W_1\right)}{\partial p_1} + a_{12}\frac{\partial E\left(W_1W_2\right)}{\partial p_1}\left\{\begin{array}{c} \frac{\partial E(W_1)}{\partial p_1}E\left(W_2\right) \text{ if equilibrium}\\ \frac{\partial E(W_1)}{\partial p_1}E\left(W_2\right) + \frac{\partial Cov(W_1W_2)}{\partial p_1} \text{ if not}\end{array}\right.$$

Genetic variance in two-locus model

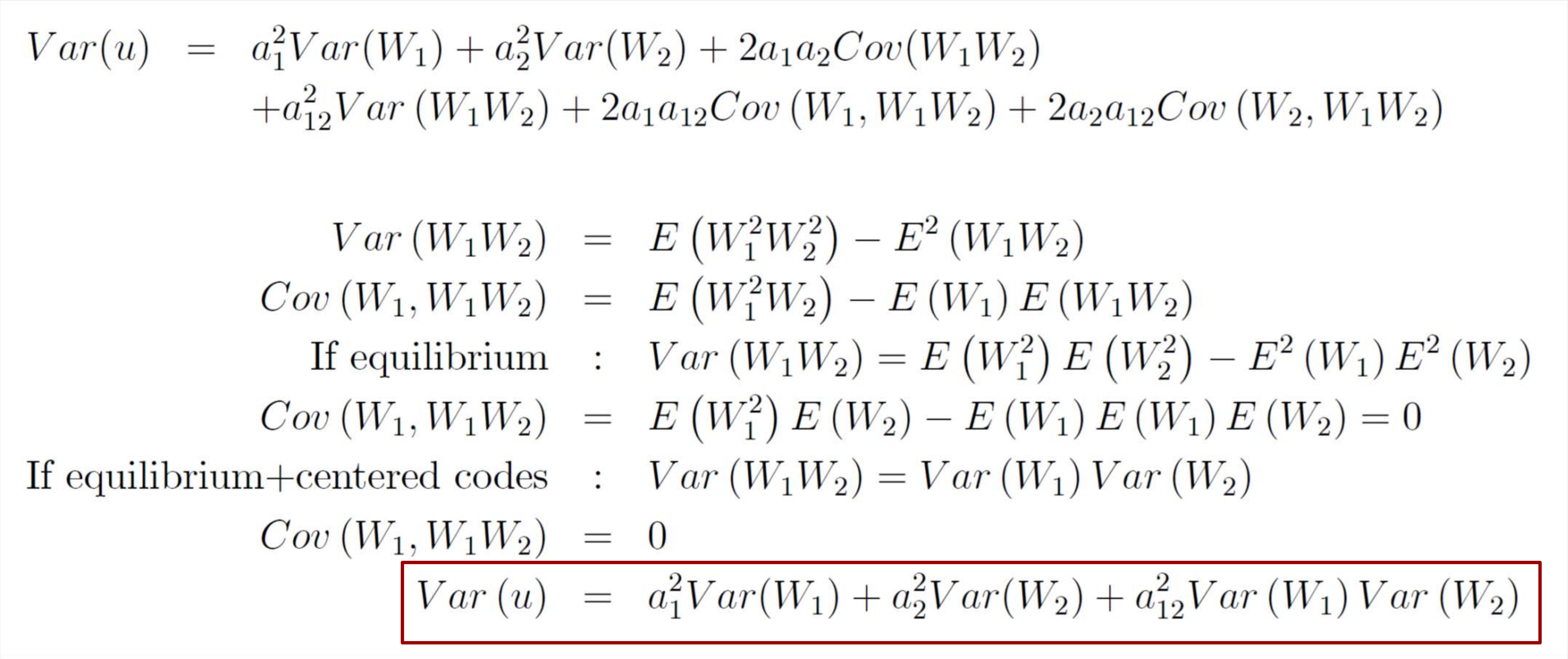
$$u = W_1 a_1 + W_2 a_2 + W_1 W_2 a_{12}$$

$$\begin{aligned} Var(u) &= a_1^2 Var(W_1) + a_2^2 Var(W_2) + 2a_1 a_2 Cov(W_1 W_2) \\ &+ a_{12}^2 Var\left(W_1 W_2\right) + 2a_1 a_{12} Cov\left(W_1, W_1 W_2\right) + 2a_2 a_{12} Cov\left(W_2, W_1 W_2\right) \end{aligned}$$

$$\begin{aligned} Var\left(W_1 W_2\right) &= E\left(W_1^2 W_2^2\right) - E^2\left(W_1 W_2\right) \\ Cov\left(W_1, W_1 W_2\right) &= E\left(W_1^2 W_2\right) - E\left(W_1\right) E\left(W_1 W_2\right) \end{aligned}$$

$$Cov(W_2, W_1 W_2) = E\left(W_1 W_2{}^2\right) - \mathrm{E}(W_2) E(W_1 W_2)$$

Disequilibrium variance



$$\begin{aligned}
Var(u) &= a_1^2 Var(W_1) + a_2^2 Var(W_2) + 2a_1 a_2 Cov(W_1 W_2) \\
&\quad + a_{12}^2 Var\left(W_1 W_2\right) + 2a_1 a_{12} Cov\left(W_1, W_1 W_2\right) + 2a_2 a_{12} Cov\left(W_2, W_1 W_2\right)
\end{aligned}$$

$$\begin{aligned}
Var\left(W_1 W_2\right) &= E\left(W_1^2 W_2^2\right) - E^2\left(W_1 W_2\right) \\
Cov\left(W_1, W_1 W_2\right) &= E\left(W_1^2 W_2\right) - E\left(W_1\right) E\left(W_1 W_2\right) \\
\text{If equilibrium} &: Var\left(W_1 W_2\right) = E\left(W_1^2\right) E\left(W_2^2\right) - E^2\left(W_1\right) E^2\left(W_2\right) \\
Cov\left(W_1, W_1 W_2\right) &= E\left(W_1^2\right) E\left(W_2\right) - E\left(W_1\right) E\left(W_1\right) E\left(W_2\right) = 0 \\
\text{If equilibrium+centered codes} &: Var\left(W_1 W_2\right) = Var\left(W_1\right) Var\left(W_2\right) \\
Cov\left(W_1, W_1 W_2\right) &= 0 \\
\boxed{Var\left(u\right) = a_1^2 Var(W_1) + a_2^2 Var(W_2) + a_{12}^2 Var\left(W_1\right) Var\left(W_2\right)}
\end{aligned}$$

Equilibrium variance

Genetic variance in three-locus model (assuming HW and linkage equilibrium)

$$u = W_1a_1 + W_2a_2 + W_3a_3 + W_1W_2a_{12} + W_1W_2a_{13} + W_2W_3a_{13} + W_1W_2W_3a_{123}$$

$$\begin{aligned}
\text{If equilibrium+centered} &: \\
Var\left(W_iW_j\right) &= Var\left(W_i\right)Var\left(W_j\right) \\
Var\left(W_iW_jW_k\right) &= Var\left(W_i\right)Var\left(W_j\right)Var\left(W_{jk}\right) \\
Cov\left(W_i, W_j\right) &= 0 \\
Cov\left(W_i, W_iW_j\right) &= 0 \\
Cov\left(W_i, W_jW_k\right) &= 0 \\
Cov\left(W_iW_j, W_kW_l\right) &= 0 \\
Cov\left(W_i, W_jW_kW_l\right) &= 0 \\
Var(u) &= \sum_{k=1}^{3} 2p_kq_ka_k^2 + \sum_{k<m} a_{kj}^2 4p_kq_kp_mq_m + a_{123}^2 8p_1q_1p_2q_2p_3q_3
\end{aligned}$$

TO FULLY CHARACTERIZE STATISTICAL GENETIC ARCHITECTURE OF COMPLEX TRAITS, NEED KNOWING MORE THAN NUMBER OF LOCI, SUBSTITUTION EFFECTS AND ALLELIC FREQUENCIES .

NEED TO KNOW HIGHLY-DIMENSIONAL GENOTYPE DISTRIBUTIONS!

FOR EXAMPLE

$$E\left(W_1W_2W_3\right)=\sum\sum\sum w_1w_2w_3\Pr\left(W_1=w_j,W_2=w_k,W_3=w_l\right)$$

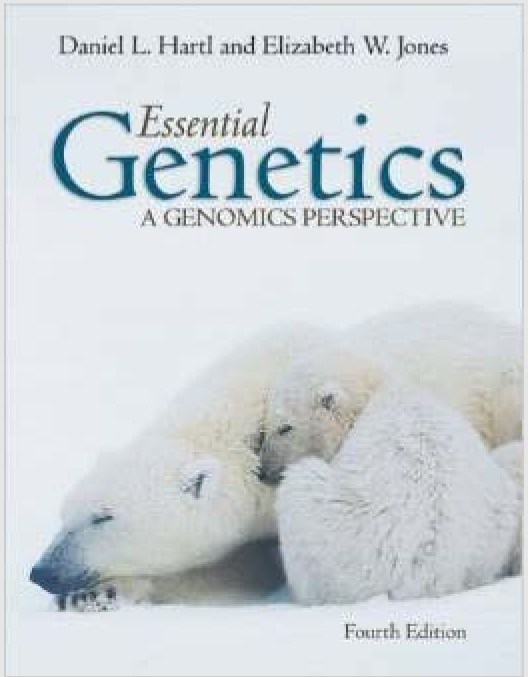
$$\begin{aligned} Var\left(W_1W_2W_3\right) &= \sum\sum\sum\left(w_1w_2w_3\right)^2\Pr\left(W_1=w_j,W_2=w_k,W_3=w_l\right) \\ &\quad -\left[\sum\sum\sum w_1w_2w_3\Pr\left(W_1=w_j,W_2=w_k,W_3=w_l\right)\right]^2 \end{aligned}$$

$$Cov(W_l^n,W_j^{n'}W_k^{n''}W_l^{n'''}W_m^{n''''})=E\left(W_j^{n'}W_k^{n''}W_l^{n+n'''}W_m^{n''''}\right)-E\left(W_l^n\right)E\left(W_j^{n'}W_k^{n''}W_l^{n'''}W_m^{n''''}\right)$$

ARGUABLY HIGH LEVEL DISEQUILIBRIUM MAY MATTER

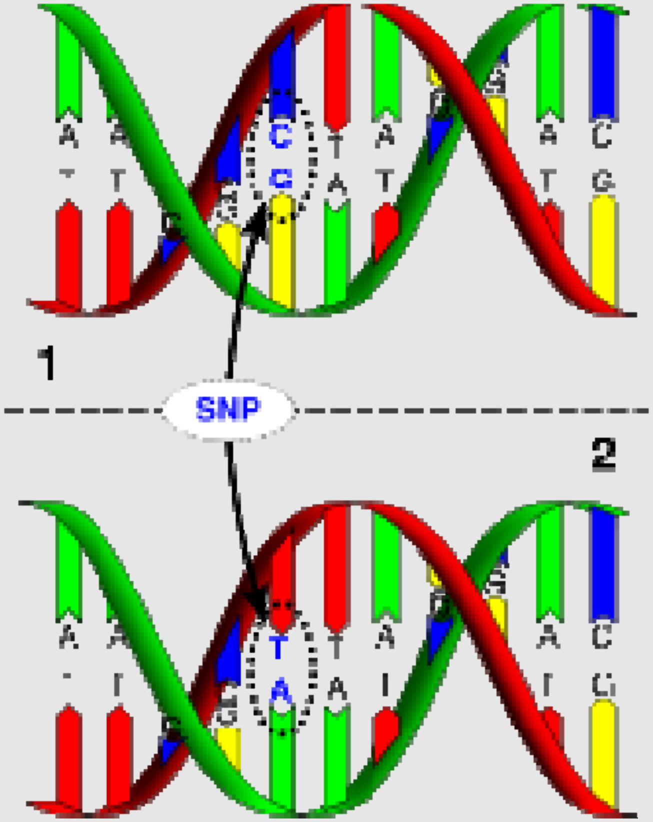
5. DATA TYPICALLY USED AS INPUTS (COVARIATES) IN PREDICTION MODELS

Primer on genomic data



GENOMIC DATA: Single nucleotide polymorphisms

All you wanted to know about SNPs
but were afraid to ask…



SNP= DNA sequence variation occurring when a single nucleotide - A, T, C, or G in the genome differs between members of a species (or between paired chromosomes)

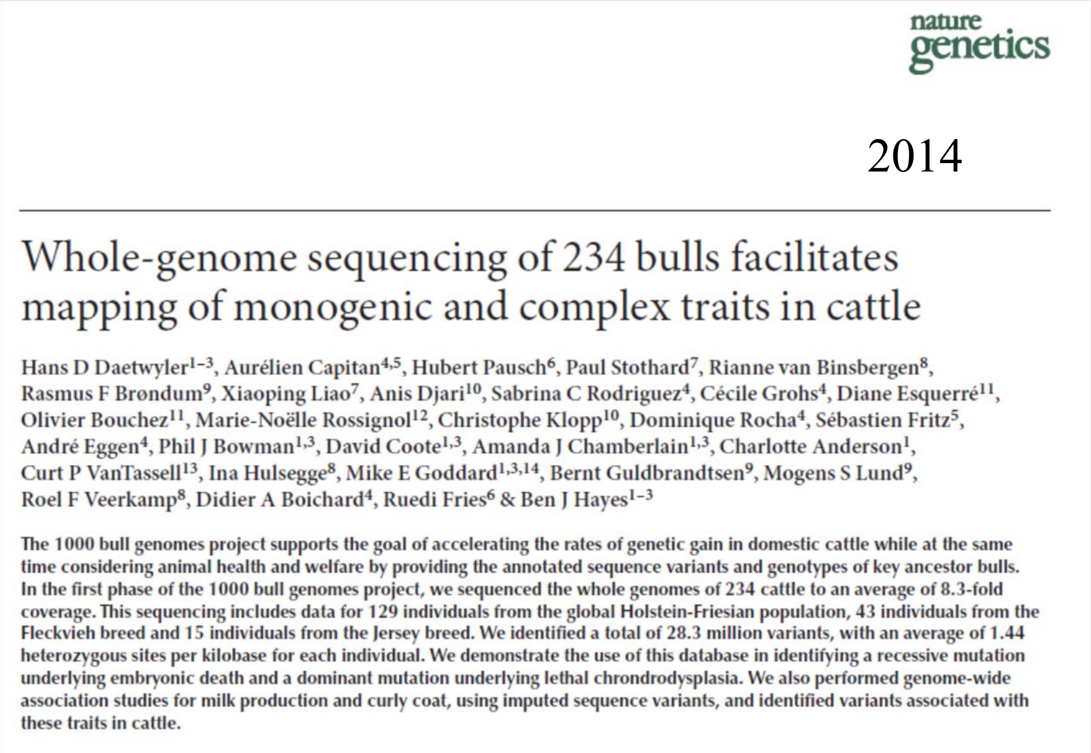
ABOVE: two sequenced DNA fragments
AAGC**C**TA to AAGC**T**TA, contain a difference in a single nucleotide.

we say that there are two *alleles* : C and T

SEQUENCES FOR THOUSANDS OF ANIMALS AND PLANTS (WITHIN SPECIES) AVAILABLE OR COMING SOON!

Schnable P, Ware D, Fulton RS, et al. (22 November 2009). "The B73 Maize Genome: Complexity, Diversity, and Dynamics". *Science*. **326** (5956): 1112–1115.

2,300Mbp	39,656 genes predicted



nature genetics

2014

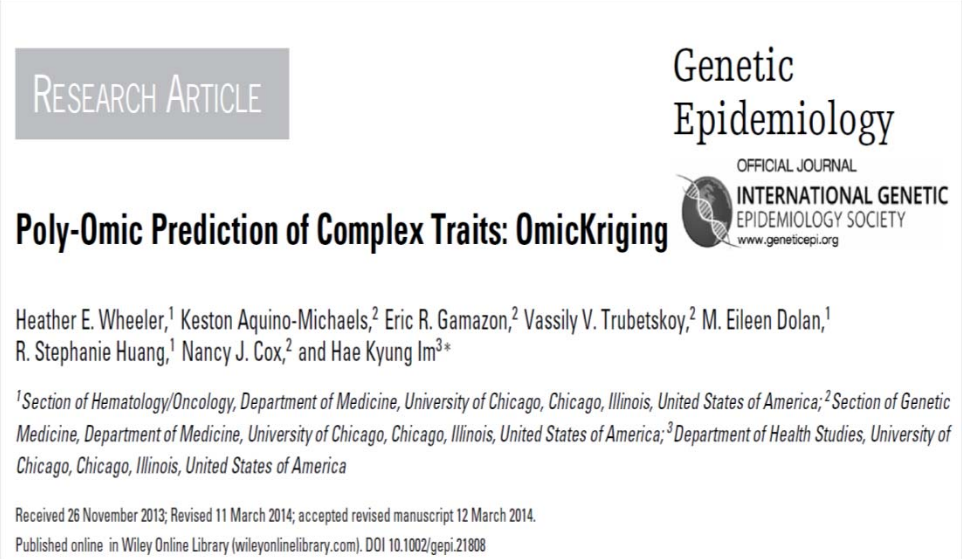
Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle

Hans D Daetwyler[1–3], Aurélien Capitan[4,5], Hubert Pausch[6], Paul Stothard[7], Rianne van Binsbergen[8], Rasmus F Brøndum[9], Xiaoping Liao[7], Anis Djari[10], Sabrina C Rodriguez[4], Cécile Grohs[4], Diane Esquerré[11], Olivier Bouchez[11], Marie-Noëlle Rossignol[12], Christophe Klopp[10], Dominique Rocha[4], Sébastien Fritz[5], André Eggen[4], Phil J Bowman[1,3], David Coote[1,3], Amanda J Chamberlain[1,3], Charlotte Anderson[1], Curt P VanTassell[13], Ina Hulsegge[8], Mike E Goddard[1,3,14], Bernt Guldbrandtsen[9], Mogens S Lund[9], Roel F Veerkamp[8], Didier A Boichard[4], Ruedi Fries[6] & Ben J Hayes[1–3]

The 1000 bull genomes project supports the goal of accelerating the rates of genetic gain in domestic cattle while at the same time considering animal health and welfare by providing the annotated sequence variants and genotypes of key ancestor bulls. In the first phase of the 1000 bull genomes project, we sequenced the whole genomes of 234 cattle to an average of 8.3-fold coverage. This sequencing includes data for 129 individuals from the global Holstein-Friesian population, 43 individuals from the Fleckvieh breed and 15 individuals from the Jersey breed. We identified a total of 28.3 million variants, with an average of 1.44 heterozygous sites per kilobase for each individual. We demonstrate the use of this database in identifying a recessive mutation underlying embryonic death and a dominant mutation underlying lethal chrondrodysplasia. We also performed genome-wide association studies for milk production and curly coat, using imputed sequence variants, and identified variants associated with these traits in cattle.

8.3-fold ave., 28.3 million variants, 1.44 heterozygous sites/kilobase

MULTI-OMICS OR "OTHER" OMICS



RESEARCH ARTICLE

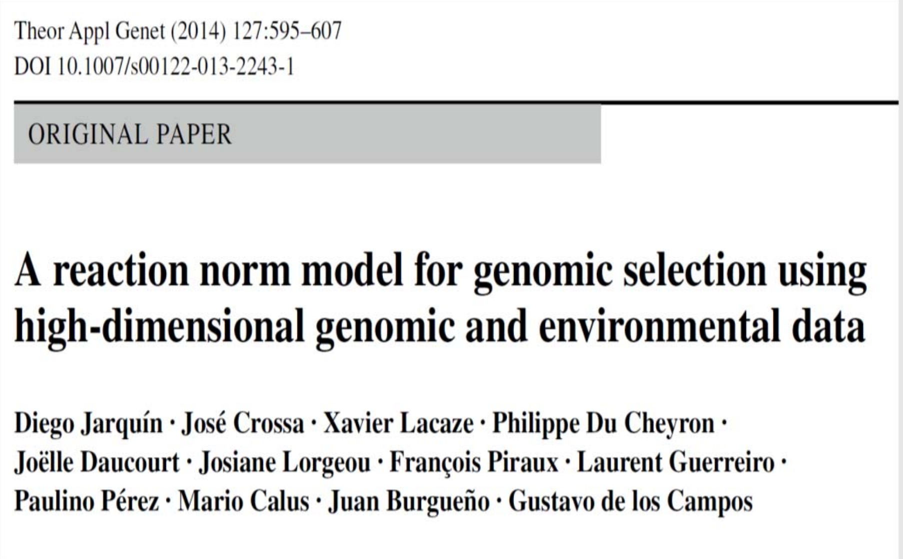
Genetic
Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC
EPIDEMIOLOGY SOCIETY
www.geneticepi.org

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*

Received 26 November 2013; Revised 11 March 2014; accepted revised manuscript 12 March 2014.
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21808

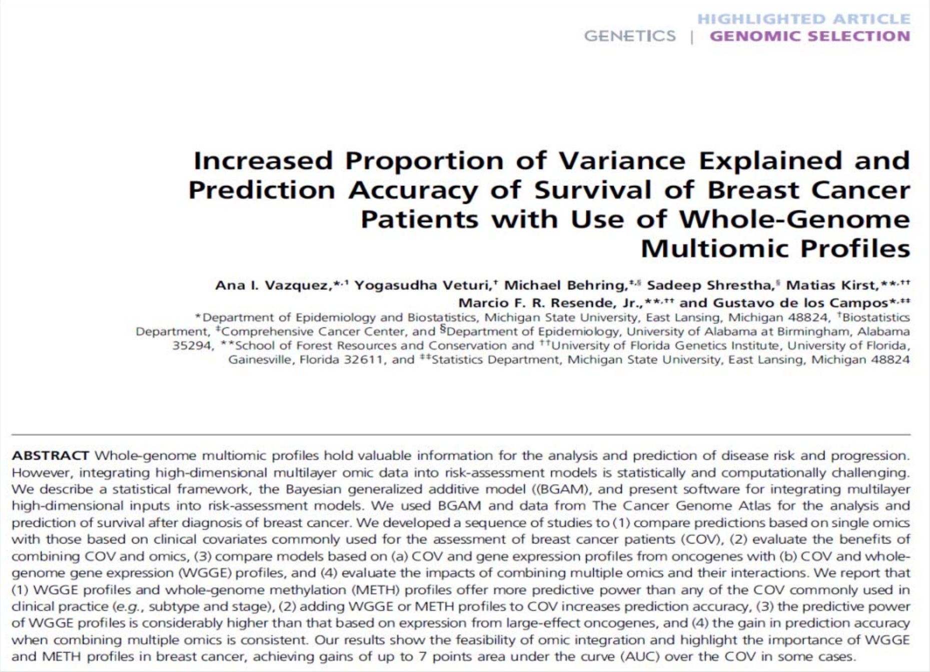


Theor Appl Genet (2014) 127:595–607
DOI 10.1007/s00122-013-2243-1

ORIGINAL PAPER

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



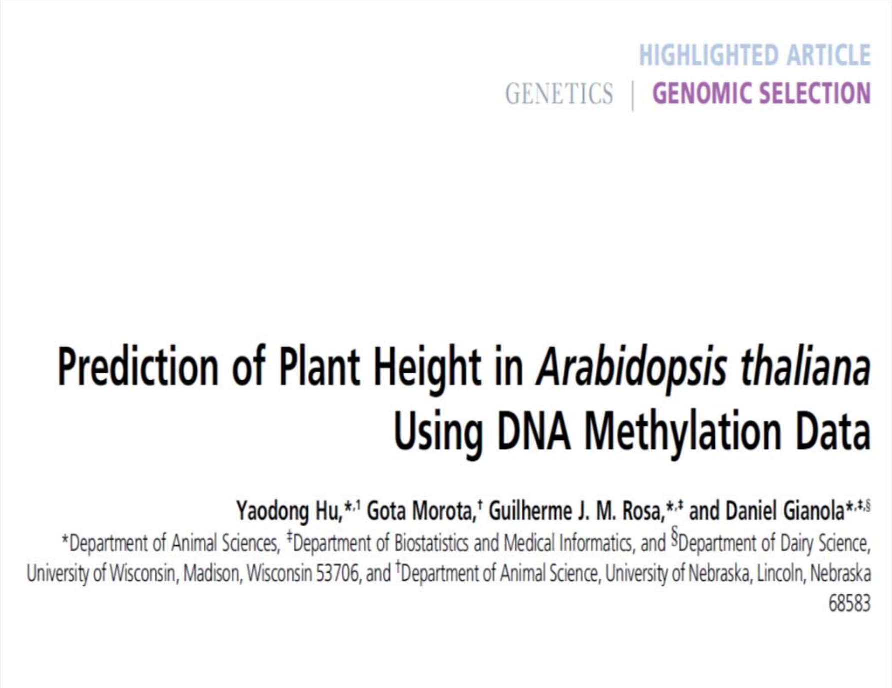
HIGHLIGHTED ARTICLE
GENETICS | GENOMIC SELECTION

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.



HIGHLIGHTED ARTICLE
GENETICS | GENOMIC SELECTION

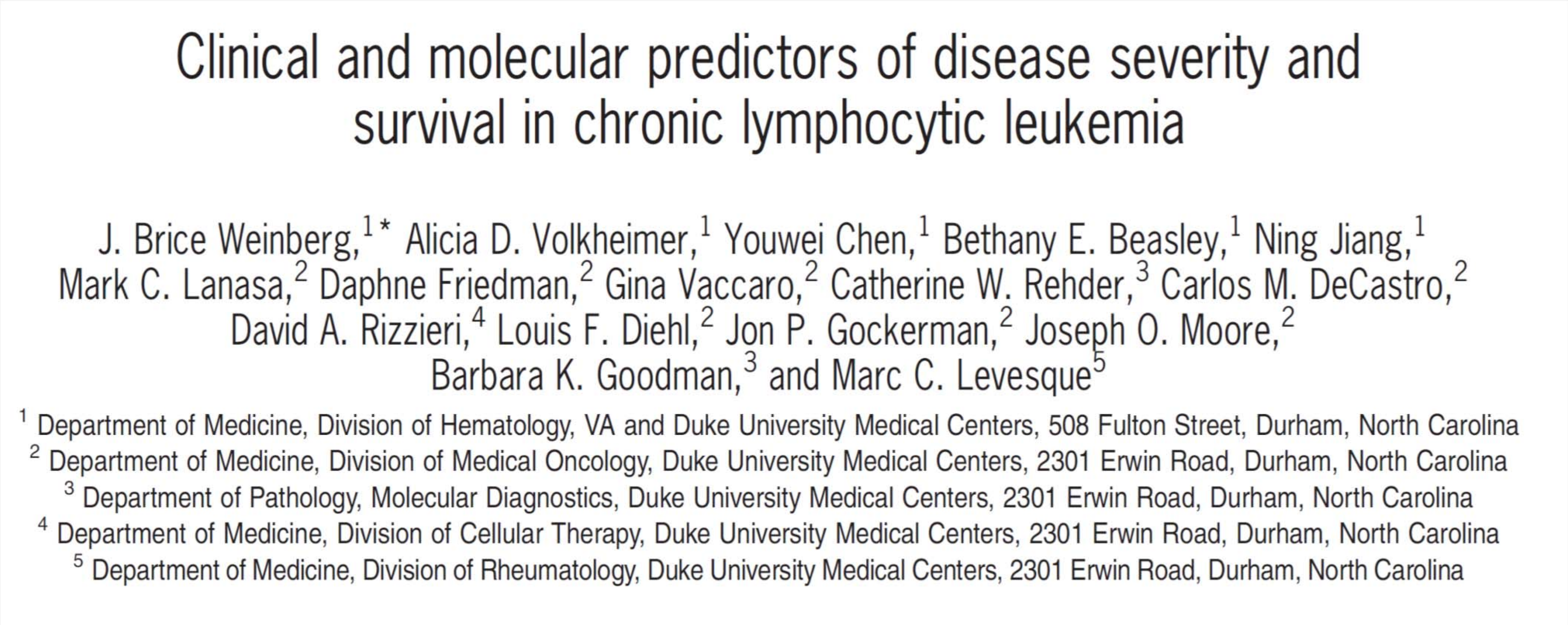
Prediction of Plant Height in *Arabidopsis thaliana* Using DNA Methylation Data

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

6. GENOME-ENABLED PREDICTION

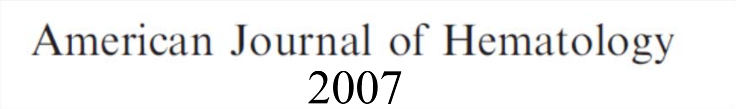
EXAMPLE 1: MEDICINE
Prediction of clinical outcomes



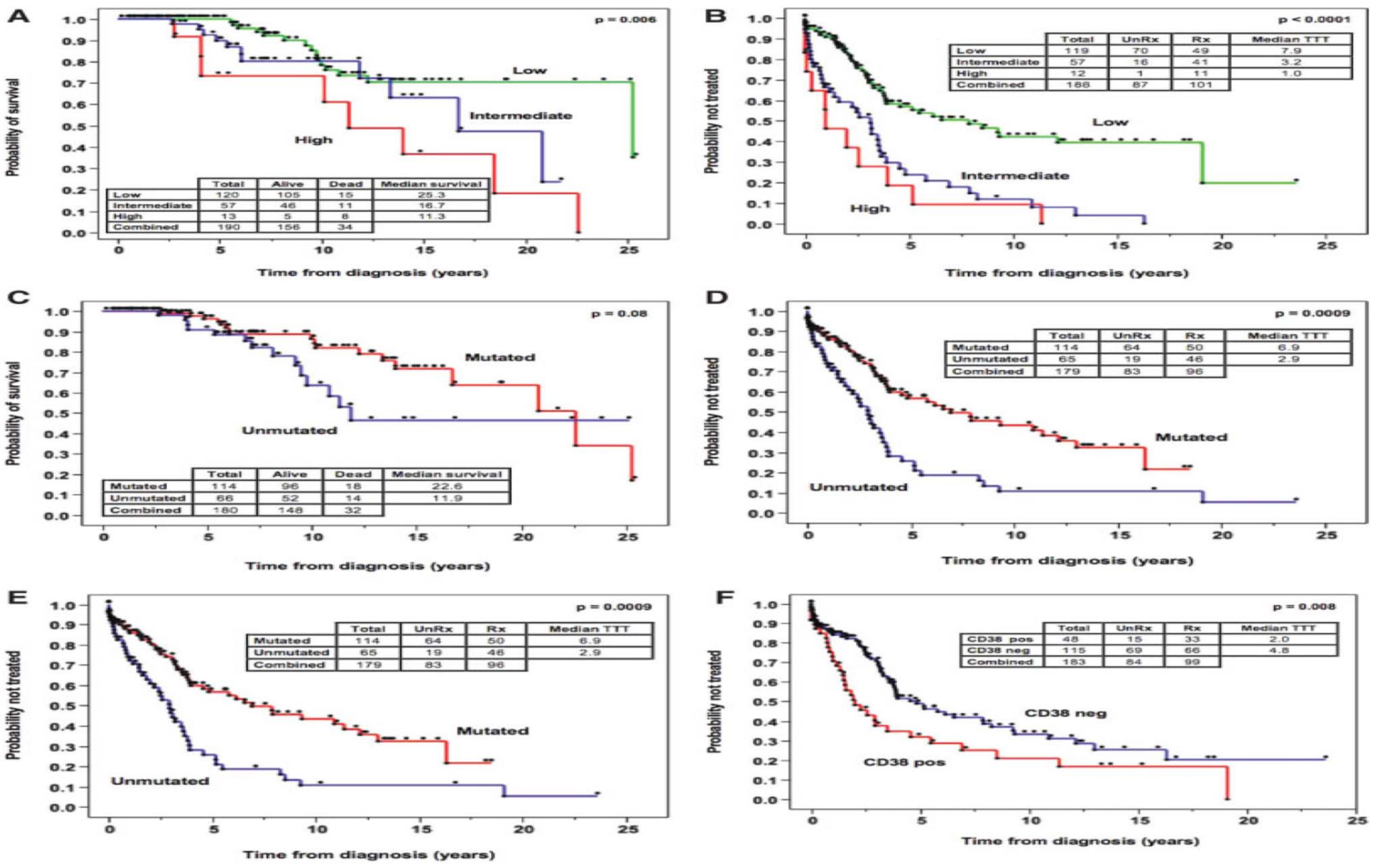
Clinical and molecular predictors of disease severity and survival in chronic lymphocytic leukemia

J. Brice Weinberg,[1]* Alicia D. Volkheimer,[1] Youwei Chen,[1] Bethany E. Beasley,[1] Ning Jiang,[1]
Mark C. Lanasa,[2] Daphne Friedman,[2] Gina Vaccaro,[2] Catherine W. Rehder,[3] Carlos M. DeCastro,[2]
David A. Rizzieri,[4] Louis F. Diehl,[2] Jon P. Gockerman,[2] Joseph O. Moore,[2]
Barbara K. Goodman,[3] and Marc C. Levesque[5]

[1] Department of Medicine, Division of Hematology, VA and Duke University Medical Centers, 508 Fulton Street, Durham, North Carolina
[2] Department of Medicine, Division of Medical Oncology, Duke University Medical Centers, 2301 Erwin Road, Durham, North Carolina
[3] Department of Pathology, Molecular Diagnostics, Duke University Medical Centers, 2301 Erwin Road, Durham, North Carolina
[4] Department of Medicine, Division of Cellular Therapy, Duke University Medical Centers, 2301 Erwin Road, Durham, North Carolina
[5] Department of Medicine, Division of Rheumatology, Duke University Medical Centers, 2301 Erwin Road, Durham, North Carolina



American Journal of Hematology
2007



	Total	Alive	Dead	Median survival
Low	120	105	15	25.3
Intermediate	57	46	11	16.7
High	13	5	8	11.3
Combined	190	156	34	

	Total	UnRx	Rx	Median TTT
Low	119	70	49	7.9
Intermediate	57	16	41	3.2
High	12	1	11	1.0
Combined	188	87	101	

	Total	Alive	Dead	Median survival
Mutated	114	96	18	22.6
Unmutated	66	52	14	11.9
Combined	180	148	32	

	Total	UnRx	Rx	Median TTT
Mutated	114	64	50	6.9
Unmutated	65	19	46	2.9
Combined	179	83	96	

	Total	UnRx	Rx	Median TTT
Mutated	114	64	50	6.9
Unmutated	65	19	46	2.9
Combined	179	83	96	

	Total	UnRx	Rx	Median TTT
CD38 pos	48	15	33	2.0
CD38 neg	115	69	66	4.8
Combined	183	84	99	

Figure 2. Survival and time-to-treatment (TTT) according to modified Rai stage (A,B), IgV$_H$ gene mutation status (C,D), and CD38 positivity (E,F). Kaplan-Meier plots of survival (A,C,E) and TTT (B,D,F) by modified Rai stage. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

PROBLEMS?

APART FROM THE ERROR IN GRAPH E,
MODEL WAS NOT CROSS-VALIDATED

WHOLE-GENOME ASSISTED PREDICTION STARTED IN ANIMAL BREEDING.

USE ALL SNP MARKERS IN MODELS PREDICTION OF “BREEDING VALUE”

Meuwissen, Hayes and Goddard (Genetics,2001),
“Genomic selection”
Better terms:
“Genome-enabled selection”
“Genome-assisted selection”

Prototypical linear regression model (no nuisance parameters)

SNP effects combined additively

Effect of chromosomal segment alleles, haplotypes

$$\mathbf{y} = \mu \mathbf{1}_n + \Sigma_i \mathbf{X}_i \mathbf{g}_i + \mathbf{e},$$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{g}_i} = \boldsymbol{X}'_i$$

Linear model in parameters

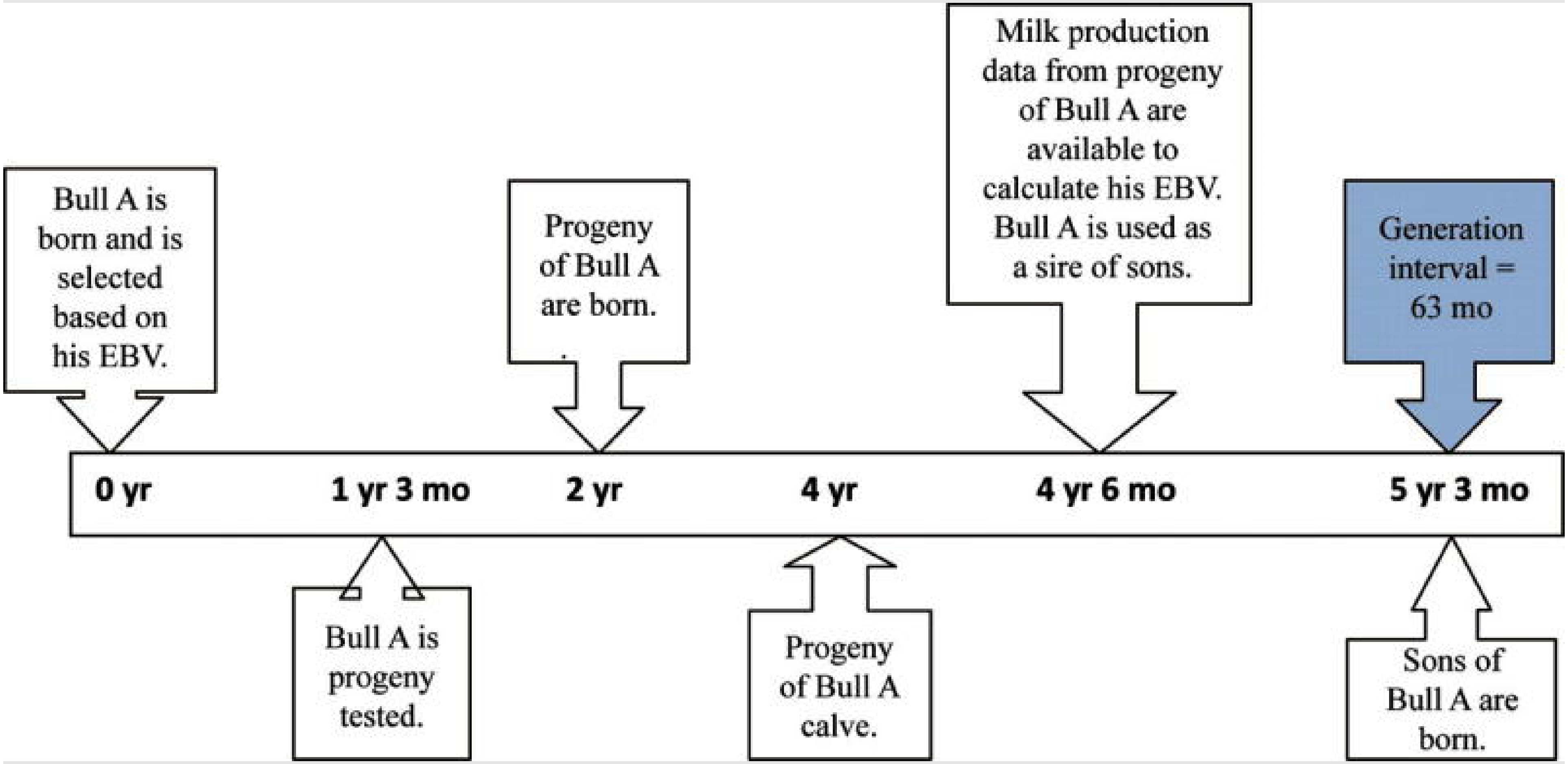
QUESTIONS:
ABANDON QTLS, PEDIGREES, KNOWN GENES?
“Whole-genome prediction” (whole or part? Variable selection?)

Essentials of genome-enabled prediction (and selection in breeding)

- Fit (train) some regression model to data set with markers and phenotypes
- Estimate marker substitution effects or marked genetic signal, or omic-captured signal
- Predict signal or phenotype in a new sample (testing or validation sample) for which input information is available
- Once phenotype (or pseudo-phenotype) is observed, asses quality of prediction. For example, calculate predictive correlation or mean squared error of prediction **(choice of metric?)**
- Objective: gain reliability over pedigree or covariate-based prediction. If new sample is of juveniles, plan medical strategy or reduce generation interval in breeding. Dispense with progeny testing? Reduce frequency of phenotyping in some programs?

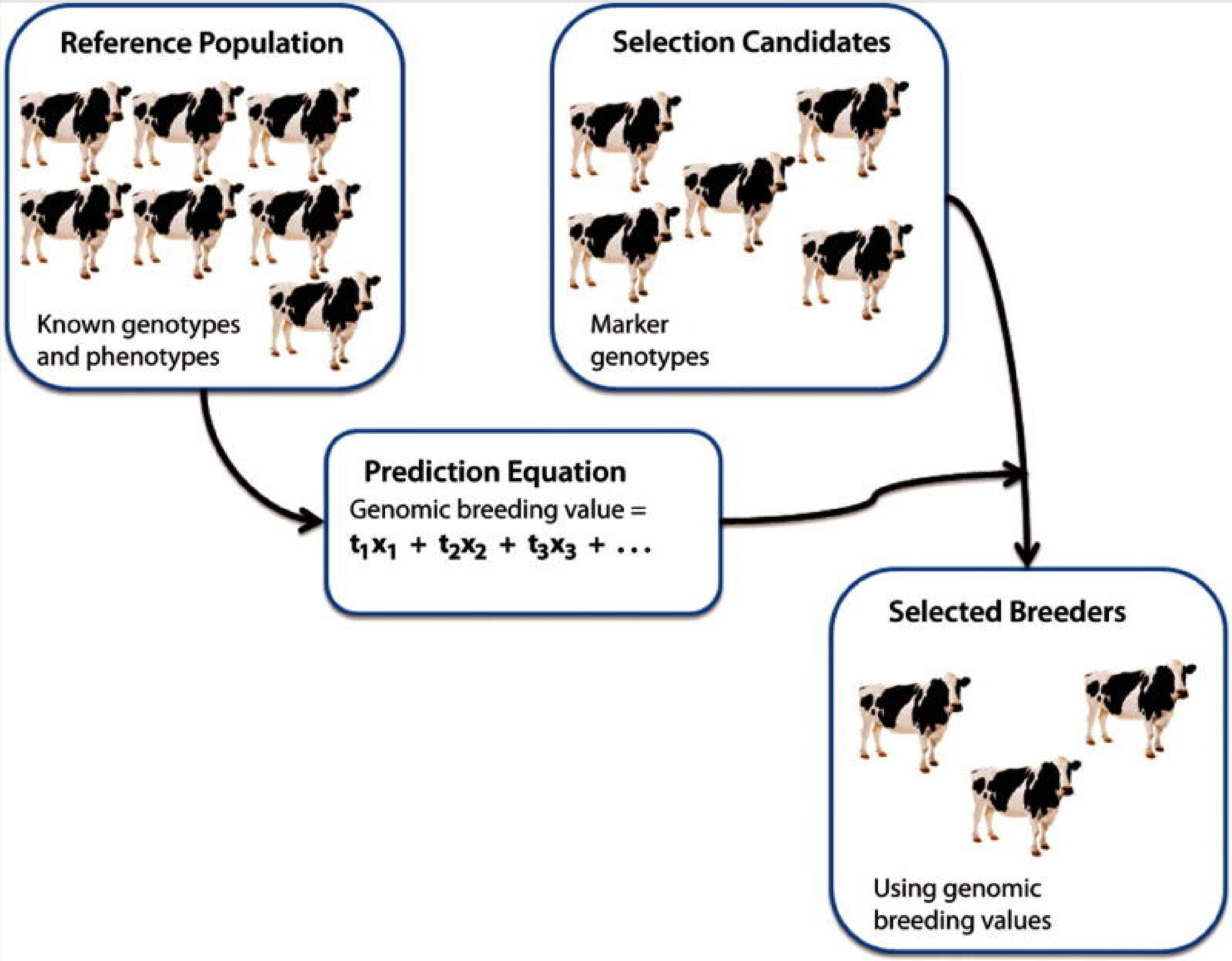
EXAMPLE 2: DAIRY CATTLE BREEDING

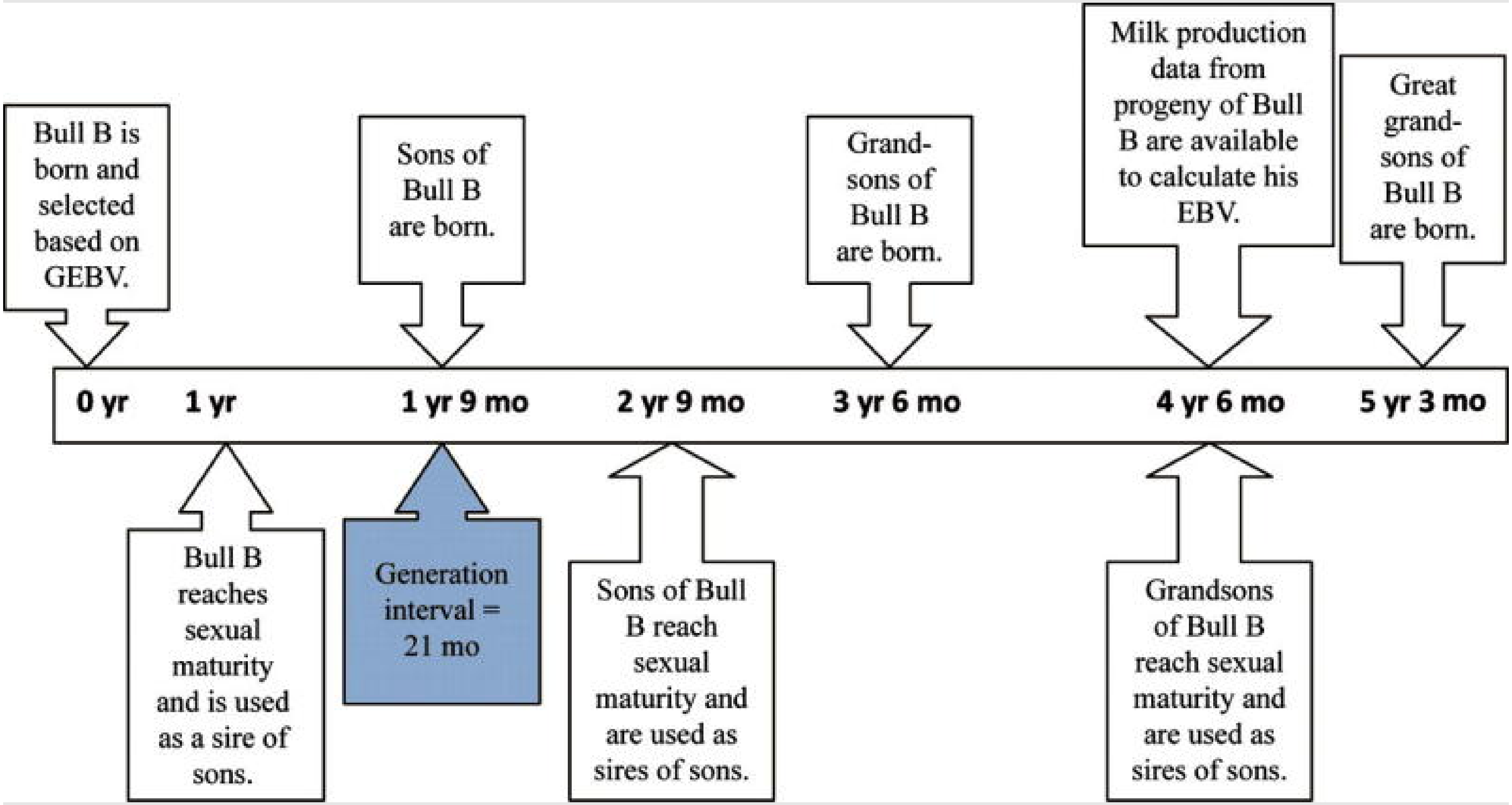
Prediction of progeny performance


Bull A is born and is selected based on his EBV.
Progeny of Bull A are born.
Milk production data from progeny of Bull A are available to calculate his EBV. Bull A is used as a sire of sons.
Generation interval = 63 mo
0 yr
1 yr 3 mo
2 yr
4 yr
4 yr 6 mo
5 yr 3 mo
Bull A is progeny tested.
Progeny of Bull A calve.
Sons of Bull A are born.

Classical progeny testing scheme

Genome-enabled selection




Bull B is born and selected based on GEBV.

Sons of Bull B are born.

Grand-sons of Bull B are born.

Milk production data from progeny of Bull B are available to calculate his EBV.

Great grand-sons of Bull B are born.

0 yr 1 yr 1 yr 9 mo 2 yr 9 mo 3 yr 6 mo 4 yr 6 mo 5 yr 3 mo

Bull B reaches sexual maturity and is used as a sire of sons.

Generation interval = 21 mo

Sons of Bull B reach sexual maturity and are used as sires of sons.

Grandsons of Bull B reach sexual maturity and are used as sires of sons.

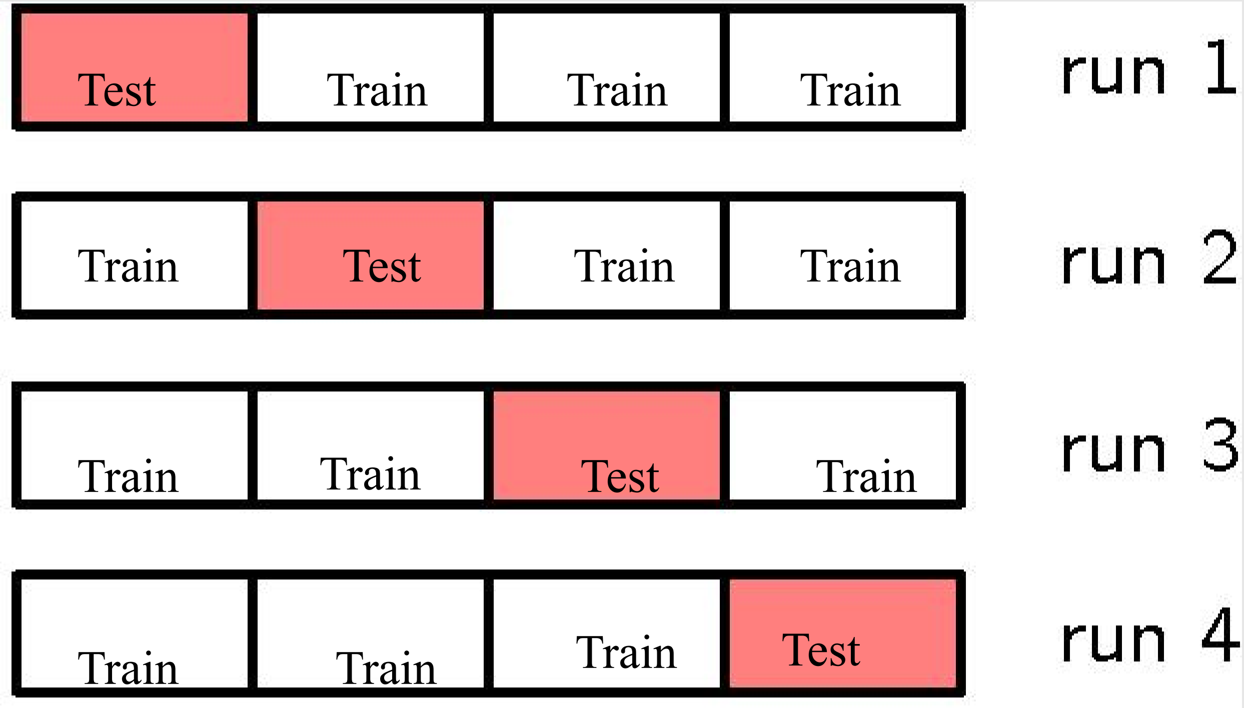
TWO ISSUES:

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

7. CROSS-VALIDATION (CV)

- Data available (genomic, multiomic, phenotyes)
- Data generated by unknown process
- Split into training (fitting)- testing (predictand) sets
- How to split? Random not always possible…
- Fitting process describes current data (model is always wrong). Sample may be idiosyncratic
- Use training process to make statement about yet-to-be observed data (testing set)
- Prediction error (conditional and unconditional): point estimate is obtained
- Distribution of prediction errors (conditional or unconditional): interval estimate. For this, CV must be replicated or pseudo-replicated

ILLUSTRATION OF A 4-FOLD CROSS-VALIDATION (red: testing set; white: training set)



Test	Train	Train	Train	run 1
Train	Test	Train	Train	run 2
Train	Train	Test	Train	run 3
Train	Train	Train	Test	run 4

ALGORITHM:

```
Choose a loss function L (e.g., mean squared error between predicted and observed outcome
Choose a set of training-testing splits (K=4)
Choose a set of regularization parameter values $\tau_1, \tau_2, \ldots, \tau_A$
for a=1 to A do
    for k=1 to K do
        train model and find "best" parameter estimates corresponding to $\tau_a$
    end for
    $L(\tau_a) = \frac{\sum_1^K L_k(\tau_a)}{K}$
end for
$\tau_{opt} = \arg min_{\tau_A}(L(\tau_a))$
```

Important

This is a single-realization from the CV distribution. Repeat many times! (data structure issues here)

CROSS-VALIDATION

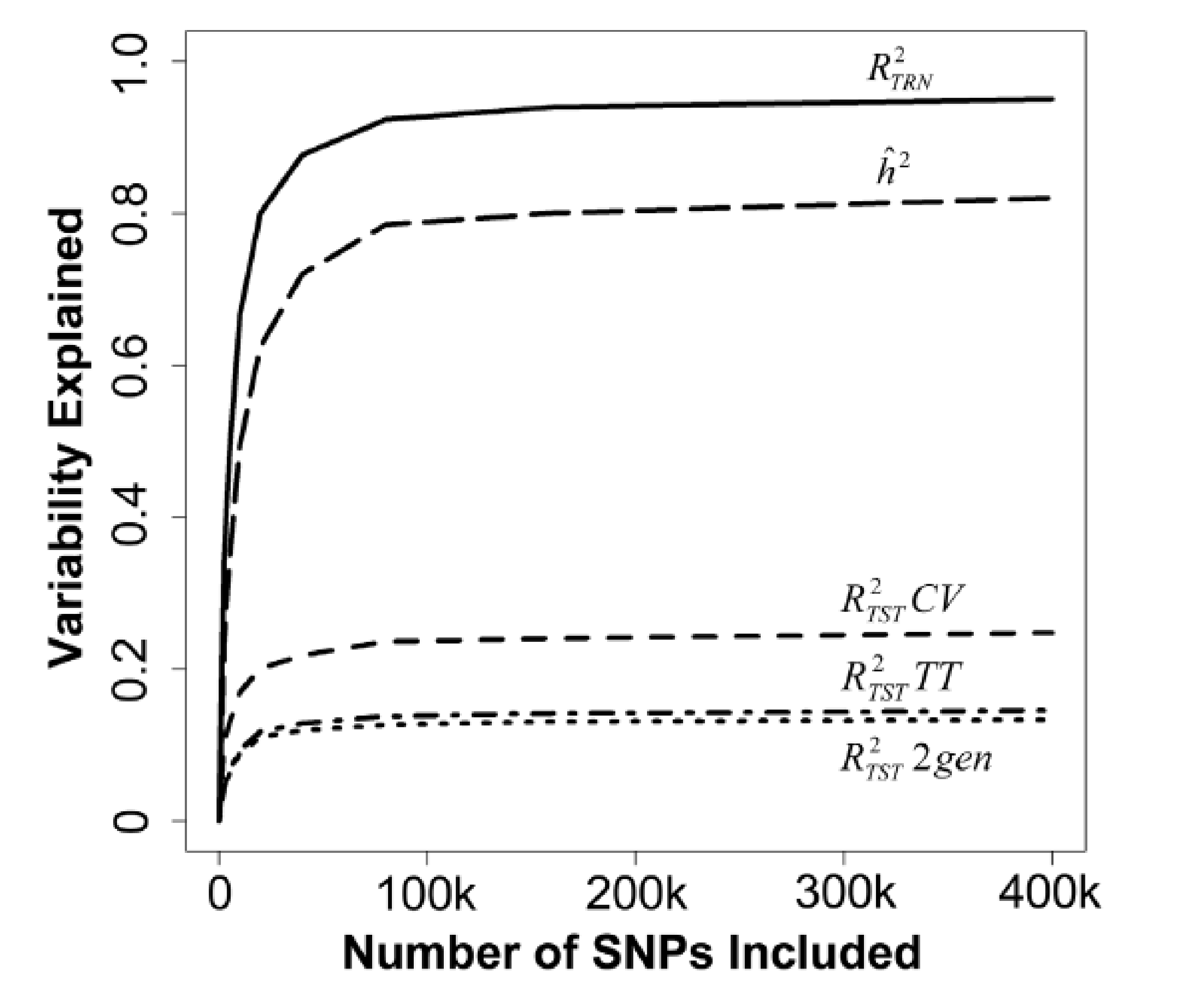
seldom done in animal breeding in the pre-genomic era. Often absent in GWAS and medical studies)

➔A. Prediction and goodness of fit are different: a model that fits well to training data may predict badly. A mechanistically poor model can give better predictions that “fancier” models

➔B. Any cross-validation scheme (e.g., k-folds) has a cross-validation distribution

THIS IS THE DISTRIBUTION THAT MATTERS AND NOT THAT BASED ON THEORETICAL CONSIDERATIONS FROM SOME MODEL

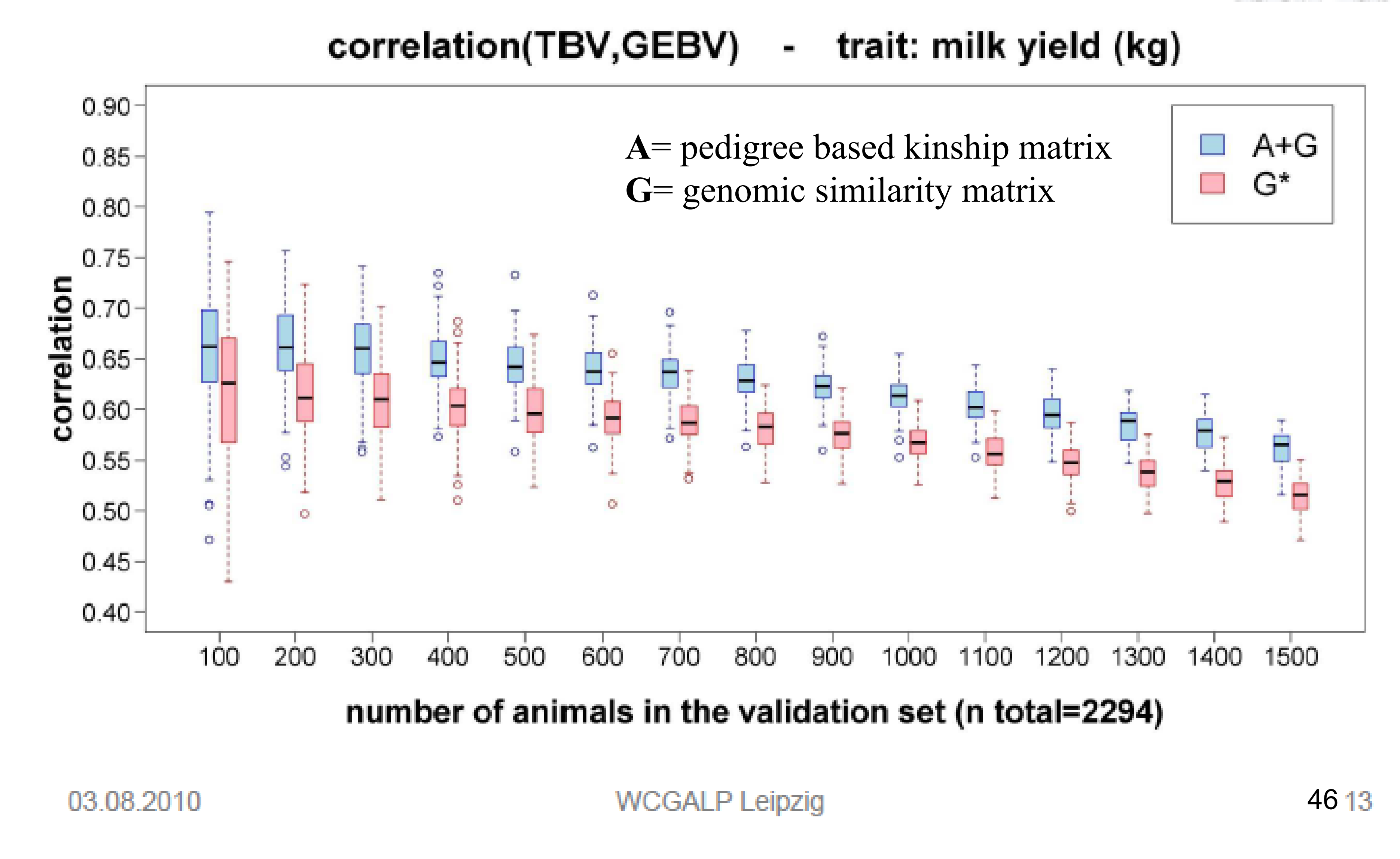
GOODNESS OF FIT (TRAINING= TRN) vs. **PREDICTIVE ABILITY** (TESTING= TST)



HUMAN STATURE: MAKOWSKY et al. , Plos Genetics 2011

CROSS-VALIDATION UNCERTAINTY AND IMPACT OF LAYOUT:
2294 dairy bulls with progeny tests (“TBV”)

(Erbe et al. 2010)



correlation(TBV,GEBV) - trait: milk yield (kg)

A= pedigree based kinship matrix
G= genomic similarity matrix

03.08.2010 WCGALP Leipzig 13

7. METHODS FOR GENOME-ENABLED PREDICTION ARE SOMEWHAT “DIFFERENT”

Maximum Likelihood (parameters are fixed): General Linear Model with Known Dispersion Structure and rank[X (n x p)]=p

A matrix with marker genotype codes

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \qquad \boldsymbol{\beta} \text{ is unknown}$$

$$L(\boldsymbol{\beta}|\mathbf{V},\mathbf{y}) \propto \exp\left[-\tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] \qquad \text{(likelihood)}$$

$$S(\boldsymbol{\beta}|\mathbf{y}) = \frac{\partial\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]}{\partial\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \qquad \text{(score)}$$

Equations satisfying first-order condition

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad\Longrightarrow\quad \boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}}$$

ML estimator

Expected information matrix

$$\mathbf{I}(\boldsymbol{\theta}) = E_{\mathbf{y}}\left[\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)'\right] = -E_{\mathbf{y}}\left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}\right)$$

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})$$

In the linear regression model

Variance-covariance matrix of the estimates = inverse of the **information** matrix:

$$\Rightarrow (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

WHAT HAPPENS IF n<p (AS WITH WHOLE-GENOME MARKER MODELS)

“generalized inverse”: one of an infinite number

$$\beta^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}; \quad E(\beta^0|\beta) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta$$

-rank(X)< min(n,p), there is an infinite number of solutions to the ml equations, all giving the same likelihood. This is called UNIDENTIFIABILITY

-Technically speaking, the likelihood contains information about at most rank(X) linear combinations of the regression vector.

→

MAXIMUM LIKELIHOOD DOES NOT PROVIDE AN ANSWER.
MUST DO SOMETHING ELSE!!!

ANSWER:

- Must reduce size of regression coefficients
- Such that the “effective” number of parameters< n
- It is equivalent to rationing…
- Pizza analogy: there are 10 portions (n=10), but more and more persons show up. Portion size is gradually reduced.
- Suppose n=1000 Holstein sequences and fit sequence-model with 28.4 million variants. Not much pizza per variant!

PENALIZED and BAYESIAN METHODS DO THIS!

- The idea of “penalty: typically ad-hoc
- It does not arise “naturally” in classical inference
- It appears “naturally” in Bayesian inference

➔ L_2 penalty: equivalent to Gaussian prior

➔ L_1 penalty: equivalent to double exponential prior

➔ Penalties on covariance matrices equivalent to priors (e.g., inverse Wishart)

➡ Bayesian methods lend themselves for predictive inference.
The prior becomes part of a prediction machine which can
ALWAYS be calibrated in some manner [contrary to inference]

The concept of penalized likelihood (example: ridge regression viewed from this perspective)

$$
\begin{aligned}
y &= X\beta + e;\ e \sim N(0, I\sigma_e^2) \\
SSR &= (y - X\beta)'(y - X\beta) \\
L(\beta|y) &\sim \exp\left[-\frac{(y - X\beta)'(y - X\beta)}{2\sigma_e^2}\right] \\
\text{Penalty} &\sim \exp\left[-\frac{\beta'\beta}{2\sigma_\beta^2}\right] \\
\text{Penalized likelihood} &\sim \exp\left[-\frac{(y - X\beta)'(y - X\beta)}{2\sigma_e^2}\right]\exp\left[-\frac{\beta'\beta}{2\sigma_\beta^2}\right] \\
\text{Penalized sum of squares} &= -2\log[\text{Penalized likelihood}] \\
&= \frac{(y - X\beta)'(y - X\beta)}{2\sigma_e^2} + \frac{\beta'\beta}{2\sigma_\beta^2}
\end{aligned}
$$

Ridge regression estimator obtained by minimizing penalized SS over β

$$\frac{\partial\left(\text{Penalized sum of squares}\right)}{\partial\beta} = -X'\frac{(y - X\beta)}{\sigma_e^2} + \frac{\beta}{\sigma_\beta^2}$$

$$\Rightarrow \text{Set to 0}$$

$$\left(X'X + I\frac{\sigma_e^2}{\sigma_\beta^2}\right)\hat{\beta} = X'y$$

$$\boxed{\hat{\beta} = (X'X + I\lambda)^{-1}X'y;} \quad \lambda = \frac{\sigma_e^2}{\sigma_\beta^2}$$

Looks like BLUP(β)

Verify minimum:

$$\frac{\partial^2\left(\text{Penalized sum of squares}\right)}{\partial\beta\partial\beta'} = \left(\frac{X'X}{\sigma_e^2} + \frac{I}{\sigma_\beta^2}\right) = \left(X'X + I\frac{\sigma_e^2}{\sigma_\beta^2}\right)\sigma_e^2$$

Positive-definite➔ minimum

The concept of penalized likelihood (example in the mixed linear model)

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$$

$$\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\mathbf{R} \sim N(\mathbf{X\beta}+\mathbf{Zu},\mathbf{R})$$

$$\mathbf{u} \sim N(\mathbf{0},\mathbf{G})$$

$$p(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\mathbf{R}) = \frac{1}{(2\pi)^{\frac{N}{2}}|\mathbf{R}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu})\right]$$

$$p(\mathbf{u}|\mathbf{G}) = \frac{1}{(2\pi)^{\frac{q}{2}}|\mathbf{G}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}\right]$$

Assuming known variance components, the log of the joint density of the data and random effects is termed “penalized likelihood

$$l(\boldsymbol{\beta},\mathbf{u}|\mathbf{y},\mathbf{R},\mathbf{G}) = K - \tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) - \tfrac{1}{2}\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}$$

$$-2l(\beta,u|y,R,G) = K + (y-X\beta-Zu)'(y-X\beta-Zu) + u'G^{-1}u$$

Penalized SS

$$\frac{\partial l(\boldsymbol{\beta},\mathbf{u}|\mathbf{y},\mathbf{R},\mathbf{G})}{\partial\boldsymbol{\beta}} = \mathbf{X}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})$$

$$\frac{\partial l(\boldsymbol{\beta},\mathbf{u}|\mathbf{y},\mathbf{R},\mathbf{G})}{\partial\mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) - \mathbf{G}^{-1}\mathbf{u}$$

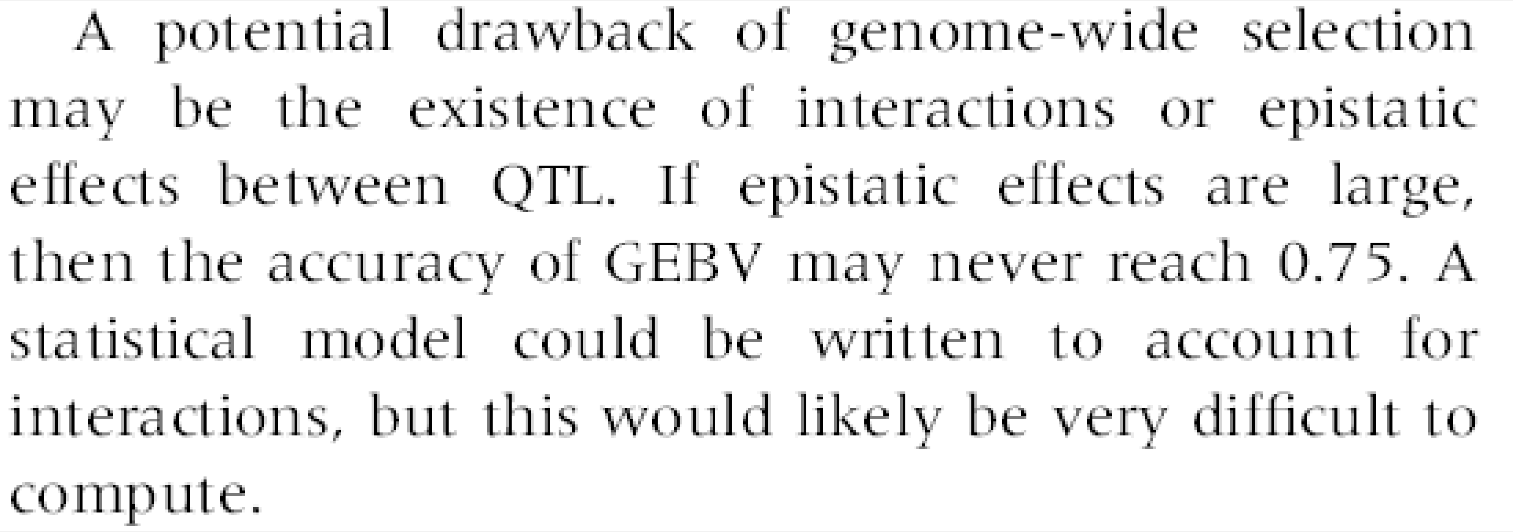
Setting the derivatives to 0 yields

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

- The solution to these equations produces the “maximum penalized likelihood” estimates of **β** and **u**
- These solutions are also the BLUE(**β**) and BLUP(**u**)

8. COMPLICATIONS: EPISTASIS

Schaeffer (2006, Journal of Animal Breeding and Genetics), Wrote:



A potential drawback of genome-wide selection may be the existence of interactions or epistatic effects between QTL. If epistatic effects are large, then the accuracy of GEBV may never reach 0.75. A statistical model could be written to account for interactions, but this would likely be very difficult to compute.

YES, IT WOULD BE DIFFICULT!
SEE NEXT…

Dealing with epistatic interactions and non-linearities

gene x gene

gene x gene x gene

gene x gene x gene x gene

…………..

IS EPISTASIS AN ANOMALY?

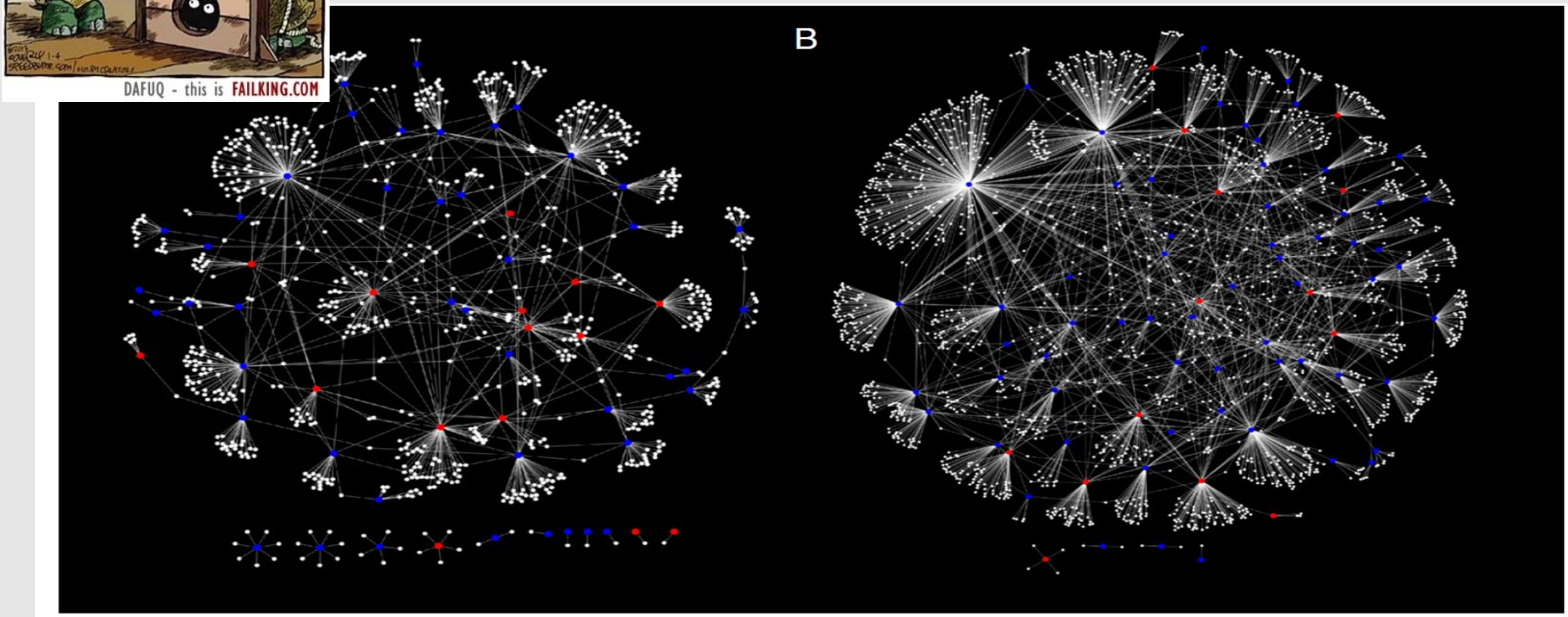
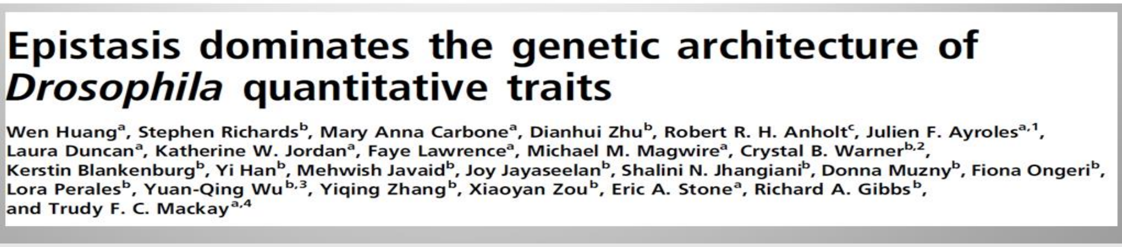


Fig. 5. Networks of epistatic interactions. Interaction networks are depicted for (*A*) starvation resistance and (*B*) chill coma recovery. Nodes depict genes, and edges significant interactions. Red nodes are genes containing significant SNPs from the Flyland analysis. Blue nodes are genes containing significant SNPs from DGRP analysis.

PNAS, 2012



Epistasis dominates the genetic architecture of *Drosophila* quantitative traits

Wen Huang[a], Stephen Richards[b], Mary Anna Carbone[a], Dianhui Zhu[b], Robert R. H. Anholt[c], Julien F. Ayroles[a,1], Laura Duncan[a], Katherine W. Jordan[a], Faye Lawrence[a], Michael M. Magwire[a], Crystal B. Warner[b,2], Kerstin Blankenburg[b], Yi Han[b], Mehwish Javaid[b], Joy Jayaseelan[b], Shalini N. Jhangiani[b], Donna Muzny[b], Fiona Ongeri[b], Lora Perales[b], Yuan-Qing Wu[b,3], Yiqing Zhang[b], Xiaoyan Zou[b], Eric A. Stone[a], Richard A. Gibbs[b], and Trudy F. C. Mackay[a,4]

RANDOM EFFECTS MODELS FOR ASSESSING EPISTASIS REST ON:

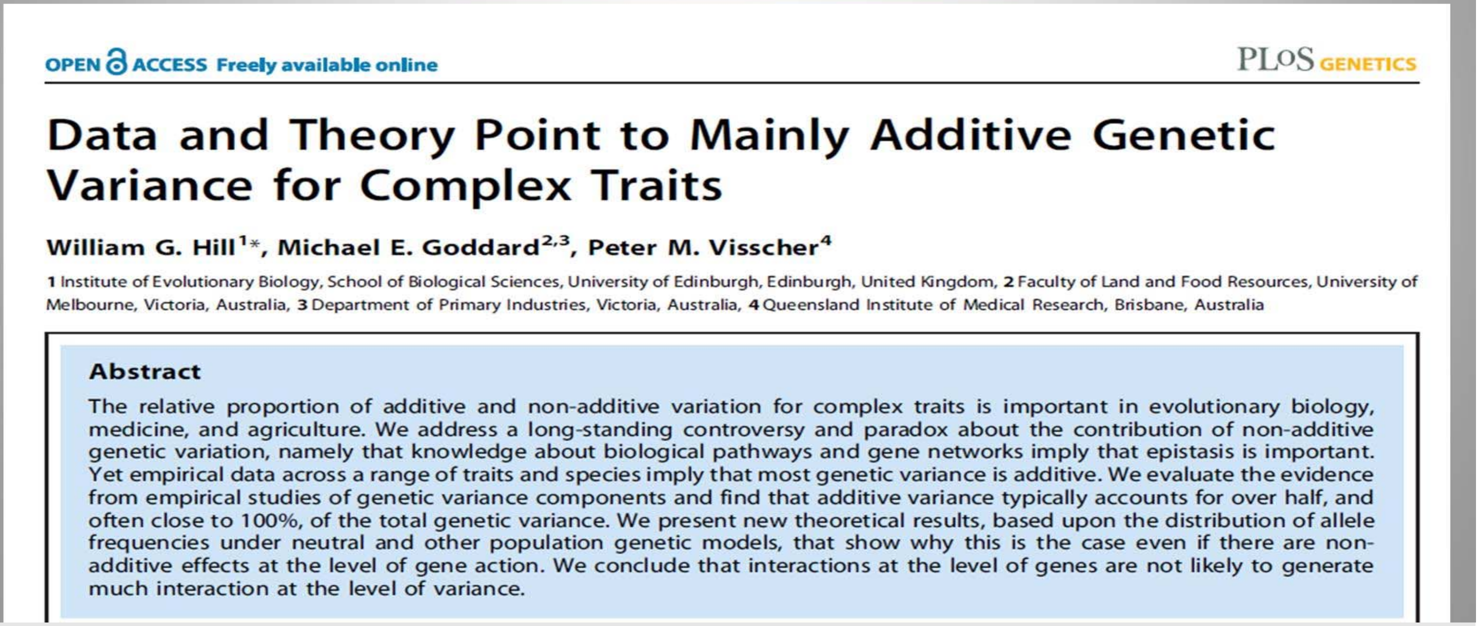
Cockerham (1954) and Kempthorne (1954)

--Orthogonal partition of genetic variance into additive, dominance
additive x additive, etc. **ONLY** if

- □No selection
- □No inbreeding
- □No assortative mating
- □No mutation
- □No migration
- □No linkage, Linkage equilibrium

Just consider
Linkage disequilibrium

ALL
ASSUMPTIONS
VIOLATED!



OPEN ACCESS Freely available online

PLoS GENETICS

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.

Influence of Gene Interaction on Complex Trait Variation with Multilocus Models

Genetics, Vol. 198, 355–367

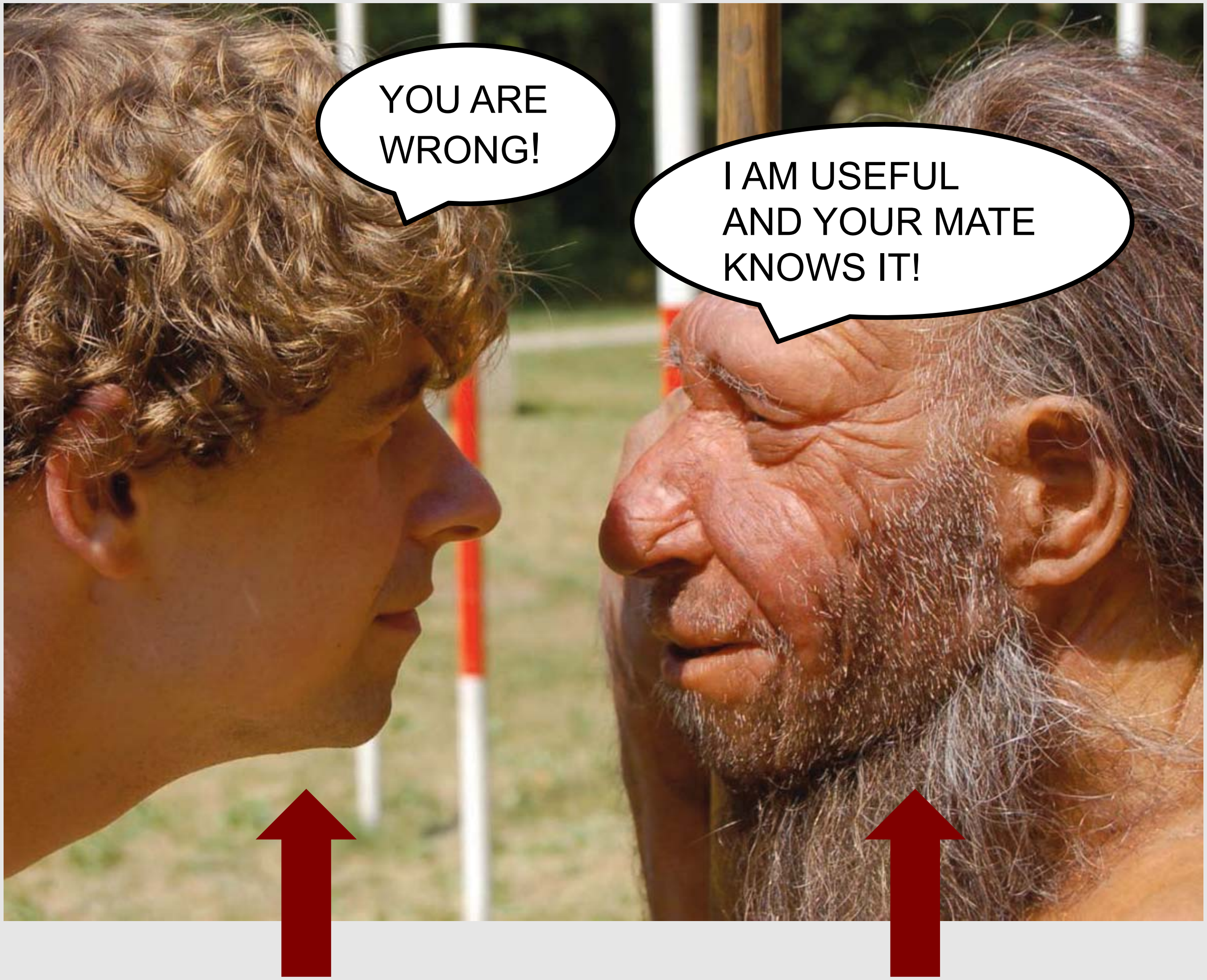
Asko Mäki-Tanila*,[1] and William G. Hill[†]

A prevailing view, and for good reasons (Hill et al., 2008; Crow, 2010; Hill, 2010)

- Fisher’s theorem of natural selection
- Interactions are second-order effects; likely tiny and hard to detect
- Epistasis probably arises with genes of large effects, unlikely to be observed in outbred populations
- Epistatic systems generate additive variance and “release” it, so why worry?

CLOSE ENCOUNTERS OF THE PREHISTORIC KIND

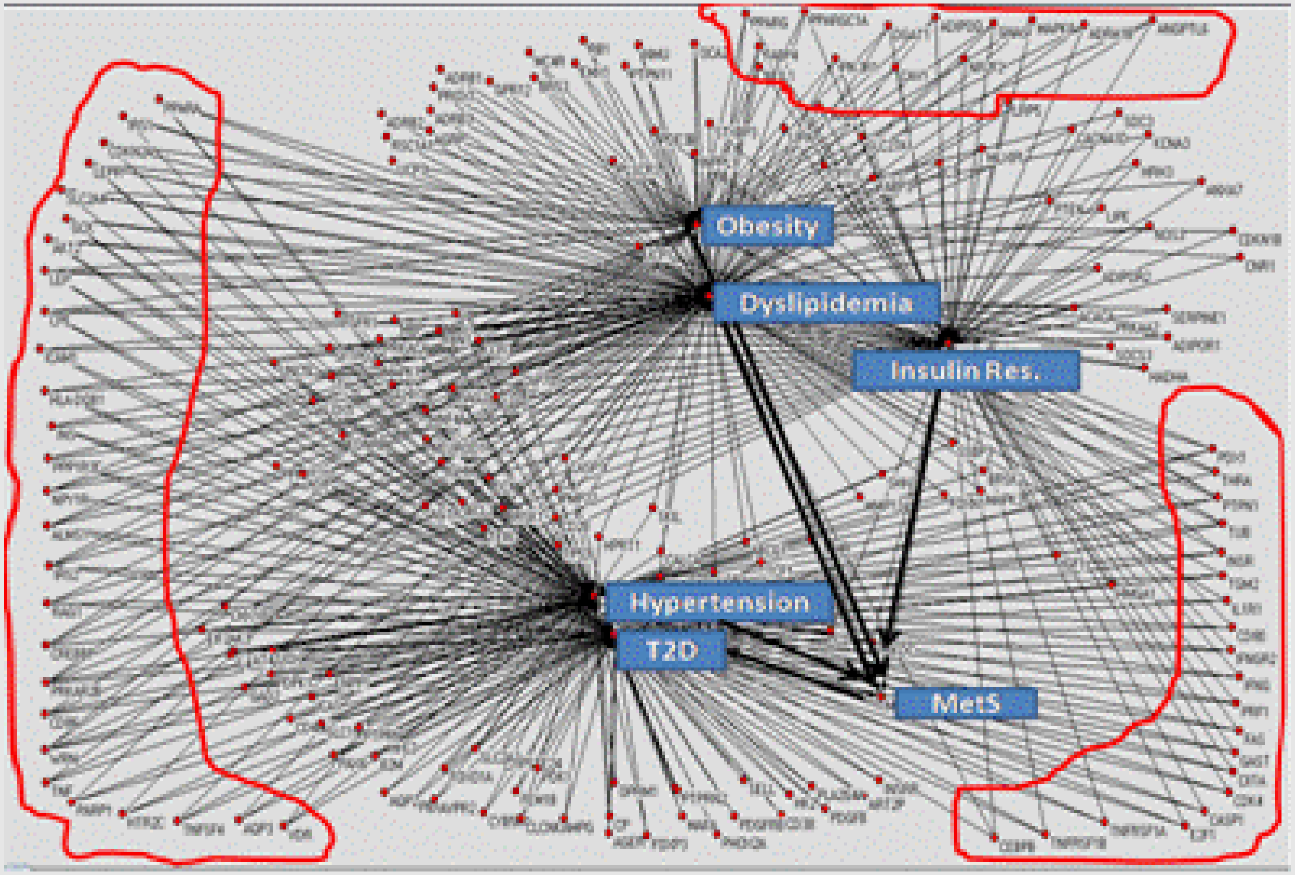
Homo sapiens



Neanderthal

GENOMICS AND COMPLEX BIOLOGY (networks)

THE ADDITIVE GENETIC MODEL (linear regressions)



”We may have to confront the complexity of our networks and systems head on, from the beginning, or we may fail at both goals of discovery and characterization. One thing is clear. The current paradigm, as successful as it now is, will reach its limit all too quickly. We need to begin thinking about the problems of the future, not the ones of the present.”

https://dsgweb.wustl.edu/Mission.html

A less popular view (Gianola and a few others)

- If everything behaves as additive, can additive models allow us to learn about “genetic architecture”?
- In areas where phenotypic prediction is crucial (medicine, precision mating) can exploting interaction have added value?
- Is so, should we consider enriching our battery of tricks?

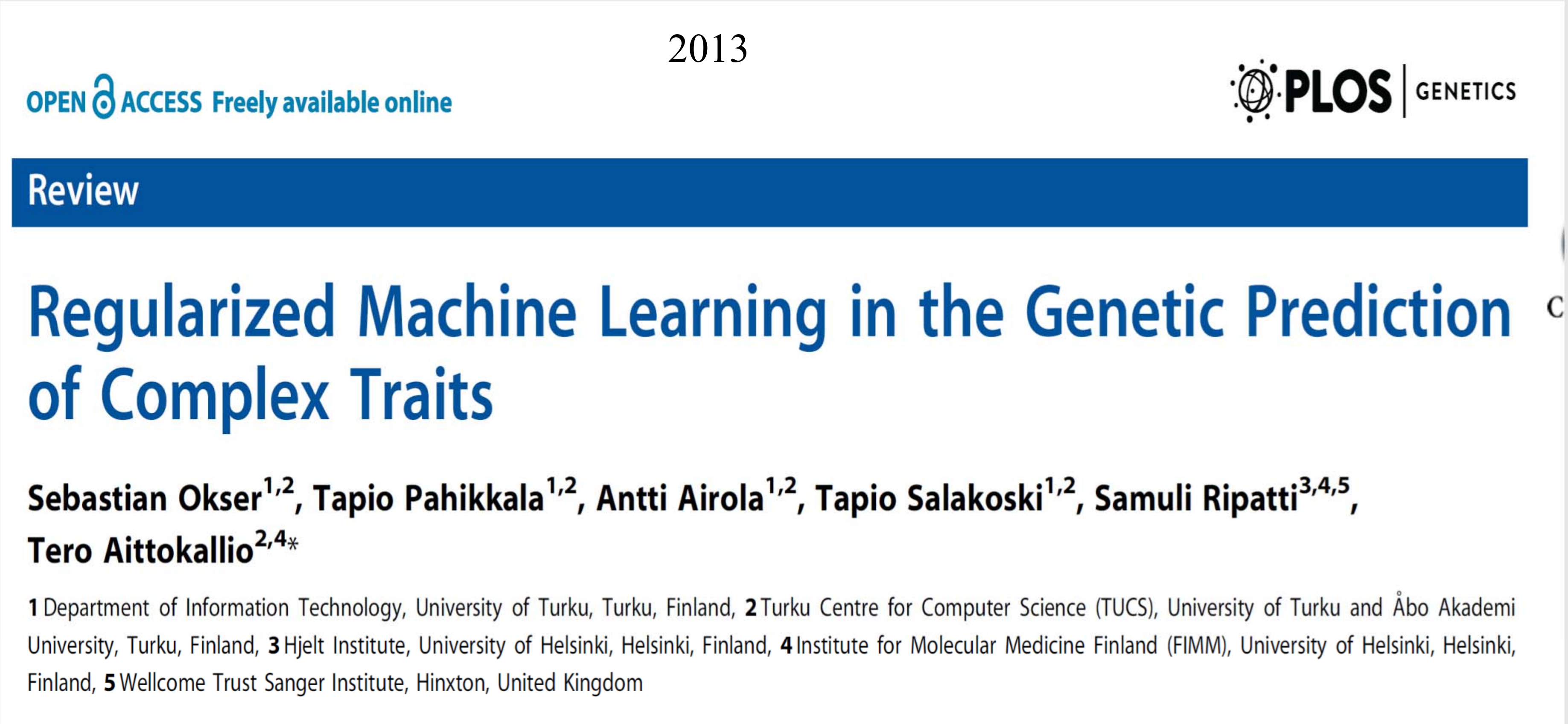
The era of machine learning and artificial intelligence

(largely non-parametric)

Distinctive aspects of non-parametric fitting

- **I**nvestigate patterns free of strictures imposed by parametric models
- **R**egression coefficients appear but (typically) do not have an obvious interpretation
- **O**ften: very good predictive performance in cross-validation
- **T**uning methods and algorithms (maximization, MCMC) similar to those of parametric methods
- **O**ften produce surprising results

2013



OPEN ACCESS Freely available online

PLOS | GENETICS

Review

Regularized Machine Learning in the Genetic Prediction of Complex Traits

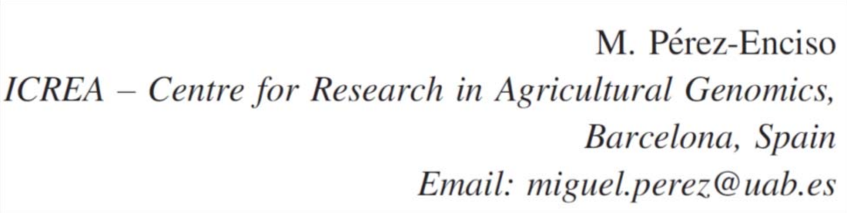
Sebastian Okser[1,2], Tapio Pahikkala[1,2], Antti Airola[1,2], Tapio Salakoski[1,2], Samuli Ripatti[3,4,5], Tero Aittokallio[2,4]*

1 Department of Information Technology, University of Turku, Turku, Finland, 2 Turku Centre for Computer Science (TUCS), University of Turku and Åbo Akademi University, Turku, Finland, 3 Hjelt Institute, University of Helsinki, Helsinki, Finland, 4 Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland, 5 Wellcome Trust Sanger Institute, Hinxton, United Kingdom

EDITORIAL

WILEY Journal of Animal Breeding and Genetics

Animal Breeding learning from machine learning



M. Pérez-Enciso
ICREA – Centre for Research in Agricultural Genomics,
Barcelona, Spain
Email: miguel.perez@uab.es

GUYS?

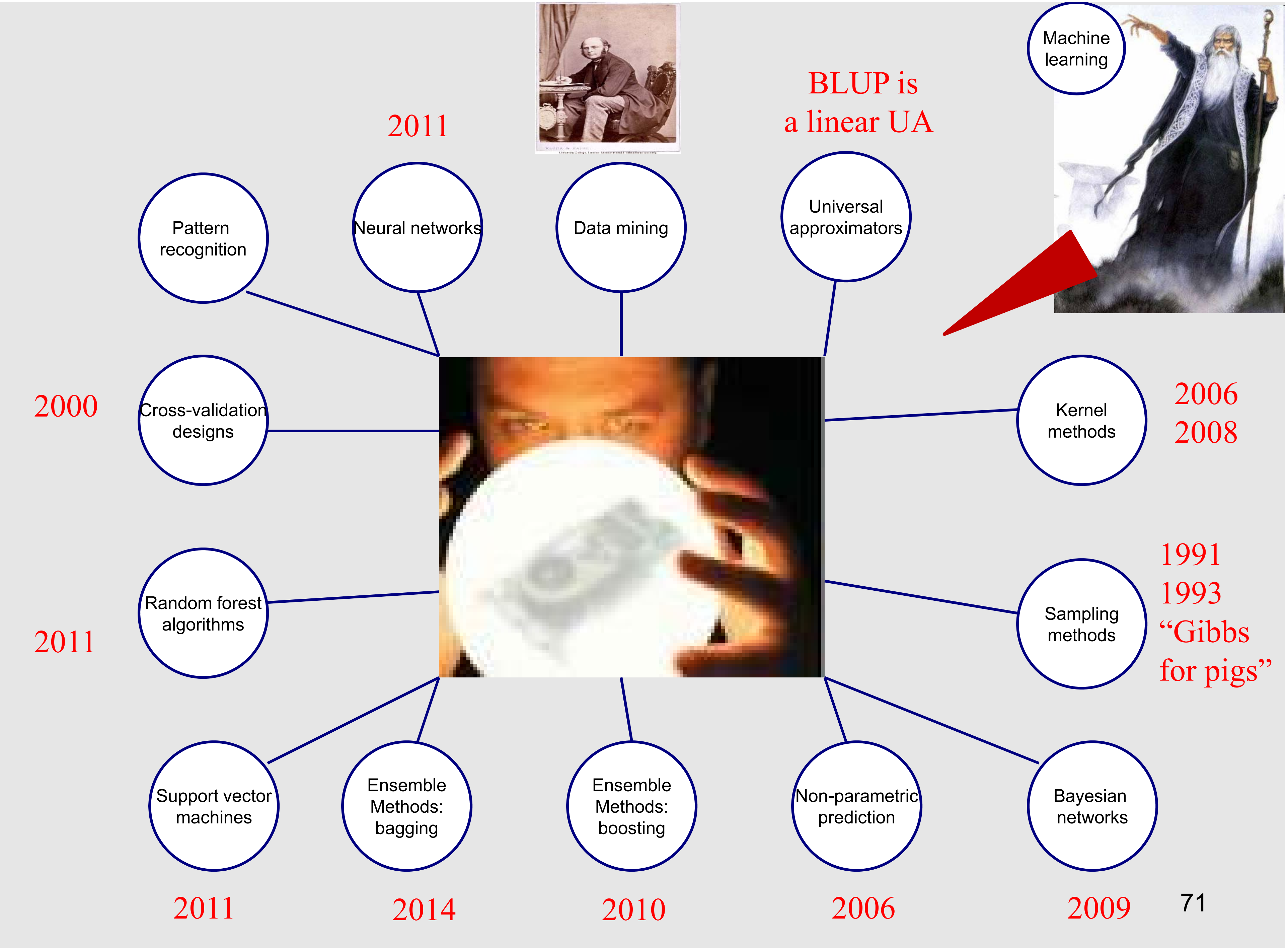
READ THE ANIMAL+ PLANT BREEDING LITERATURE!

A VIEW OF LINEAR MODELS (as employed in q. genetics)

Mathematically, can be viewed as a “local” approximation of a complex process

$$f(x) = \underbrace{\underbrace{\underbrace{f(a) + f'(a)(x-a)}_{\text{Linear approximation}} + \frac{f''(a)}{2!}(x-a)^2}_{\text{Quadratic approximation}} + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n}_{n^{th}\text{ order approximation}} + \ldots$$

FELDMAN and LEWONTIN (1975)
CHEVALET (1994)



Machine learning

BLUP is a linear UA

2011

Pattern recognition

Neural networks

Data mining

Universal approximators

2000

Cross-validation designs

Kernel methods

2006
2008

Random forest algorithms

2011

Sampling methods

1991
1993
“Gibbs for pigs”

Support vector machines

2011

Ensemble Methods: bagging

2014

Ensemble Methods: boosting

2010

Non-parametric prediction

2006

Bayesian networks

2009

WILL WE EVER GET THE ARCHITECTURE OF COMPLEX TRAITS RIGHT?

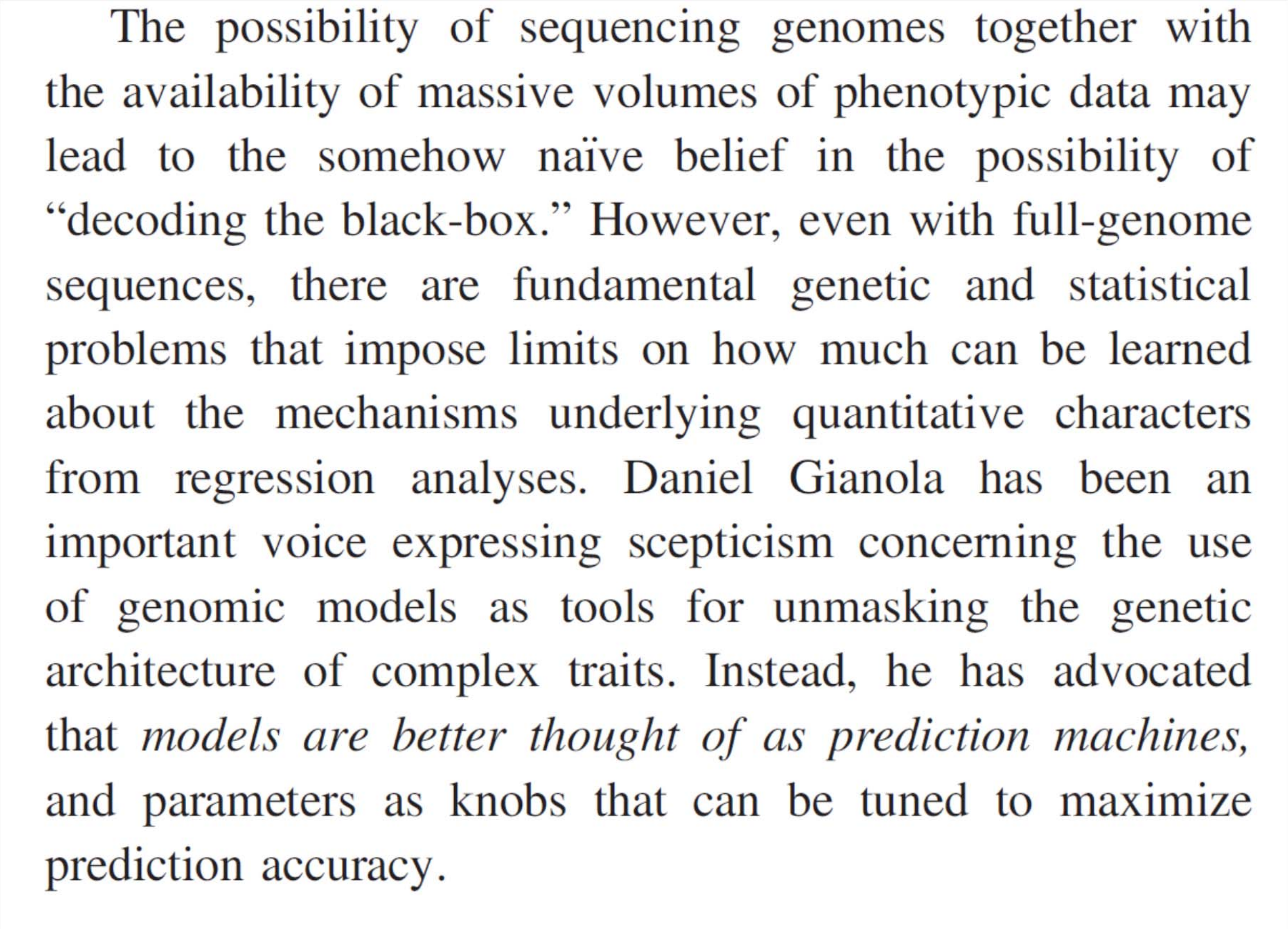
ZAHA HADID

RAFAEL VIÑOLY

FRANK GHERY

SANTIAGO CALATRAVA

SORENSEN AND de los CAMPOS (2011) WROTE THE FOLLOWING (J. Animal Breeding and Genetics, 2017)



The possibility of sequencing genomes together with the availability of massive volumes of phenotypic data may lead to the somehow naïve belief in the possibility of "decoding the black-box." However, even with full-genome sequences, there are fundamental genetic and statistical problems that impose limits on how much can be learned about the mechanisms underlying quantitative characters from regression analyses. Daniel Gianola has been an important voice expressing scepticism concerning the use of genomic models as tools for unmasking the genetic architecture of complex traits. Instead, he has advocated that *models are better thought of as prediction machines,* and parameters as knobs that can be tuned to maximize prediction accuracy.

THIS WILL BE THE VIEW ADVANCED IN THIS COURSE!!