PREDICTION OF COMPLEX TRAITS WITH KERNEL METHODS

RKHS
(largely non-parametric)

PARAMETRIC APPROACHES

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? $\left\{ \begin{array}{l} Y = G^{E}? \\ Y = E^{G}? \\ Y = G + E + GE? \\ Y = (G+E)^{GE}? \\ Y = G + E? \end{array} \right.$

$Y = G + E + GE?$ → Is an assumption

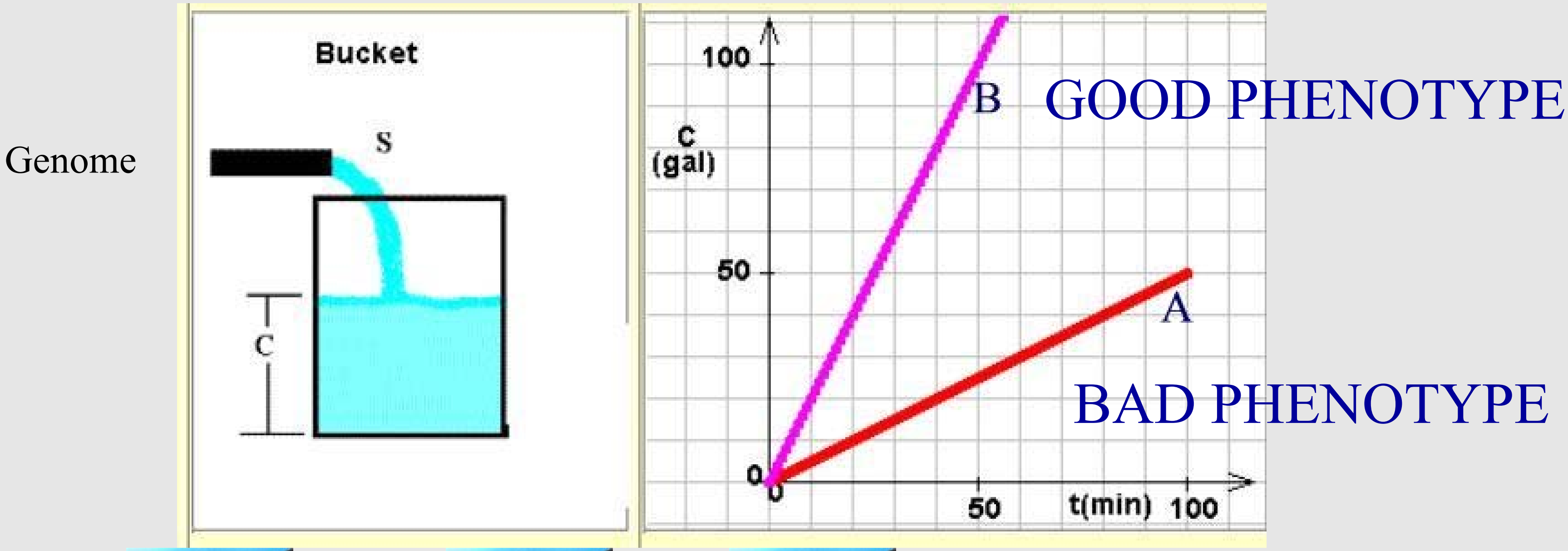
$Y = G + E?$ → Is an even a stronger assumption

THE CENTRAL DOGMA OF QUANTITATIVE GENETICS:

the additive genetic model

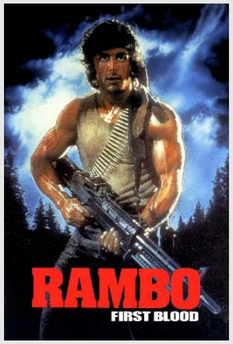
$$u_i = W_{i1}a_1 + W_{i2}a_2 + \ldots + W_{iK}a_K$$

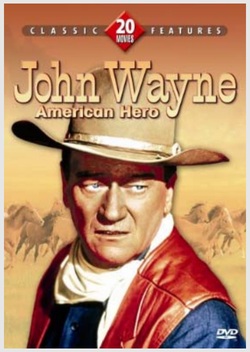
$$W_{ij}a_j = \begin{cases} -a_j & \text{if } W_{ij} = -1(aa);\ \Pr\left(W_{ij} = -1\right) = \left(1-p_j\right)^2 \\ 0 & \text{if } W_{ij} = 0(Aa);\ \Pr\left(W_{ij} = 0\right) = 2p_j\left(1-p_j\right) \\ a_j & \text{if } W_{ij} = 1(AA);\ \Pr\left(W_{ij} = 1\right) = p_j^2 \end{cases}$$

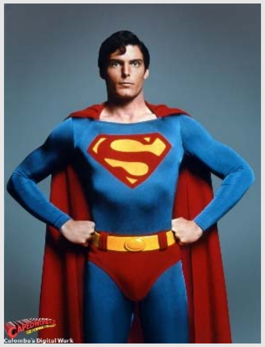


+ + + = **‘additive genetic value’**

A few complications….





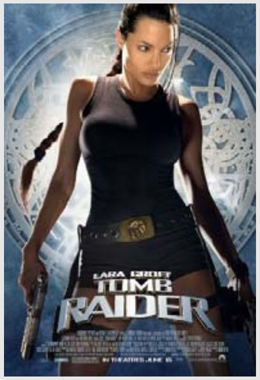


Dealing with epistatic interactions and non-linearities

gene x gene

gene x gene x gene

gene x gene x gene x gene



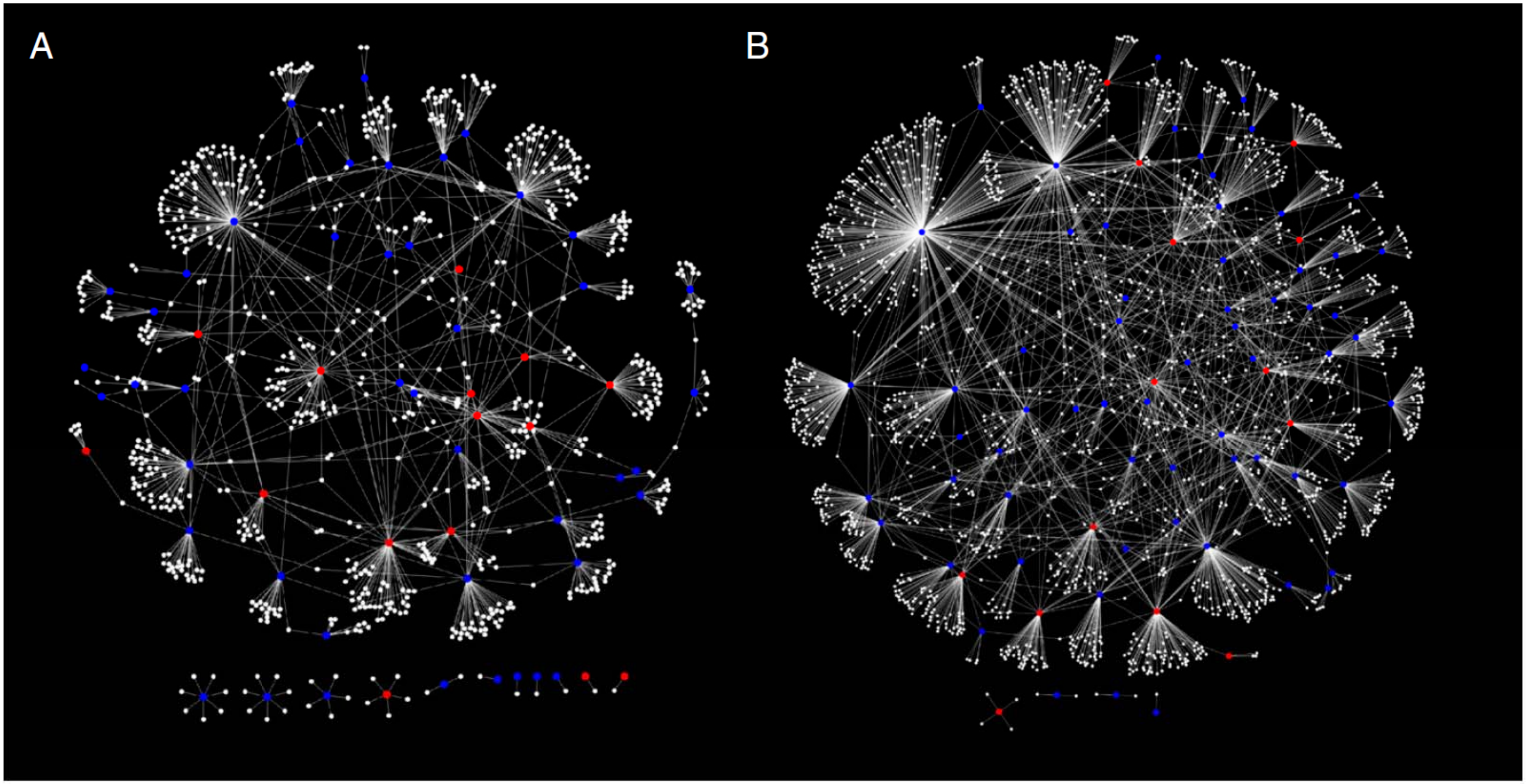
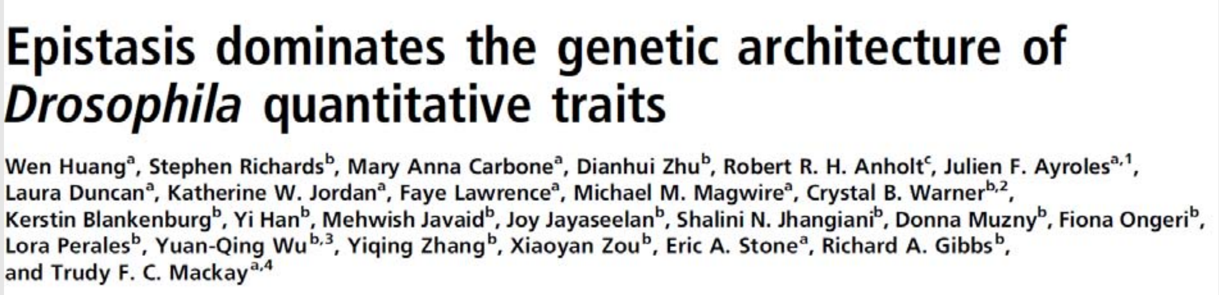


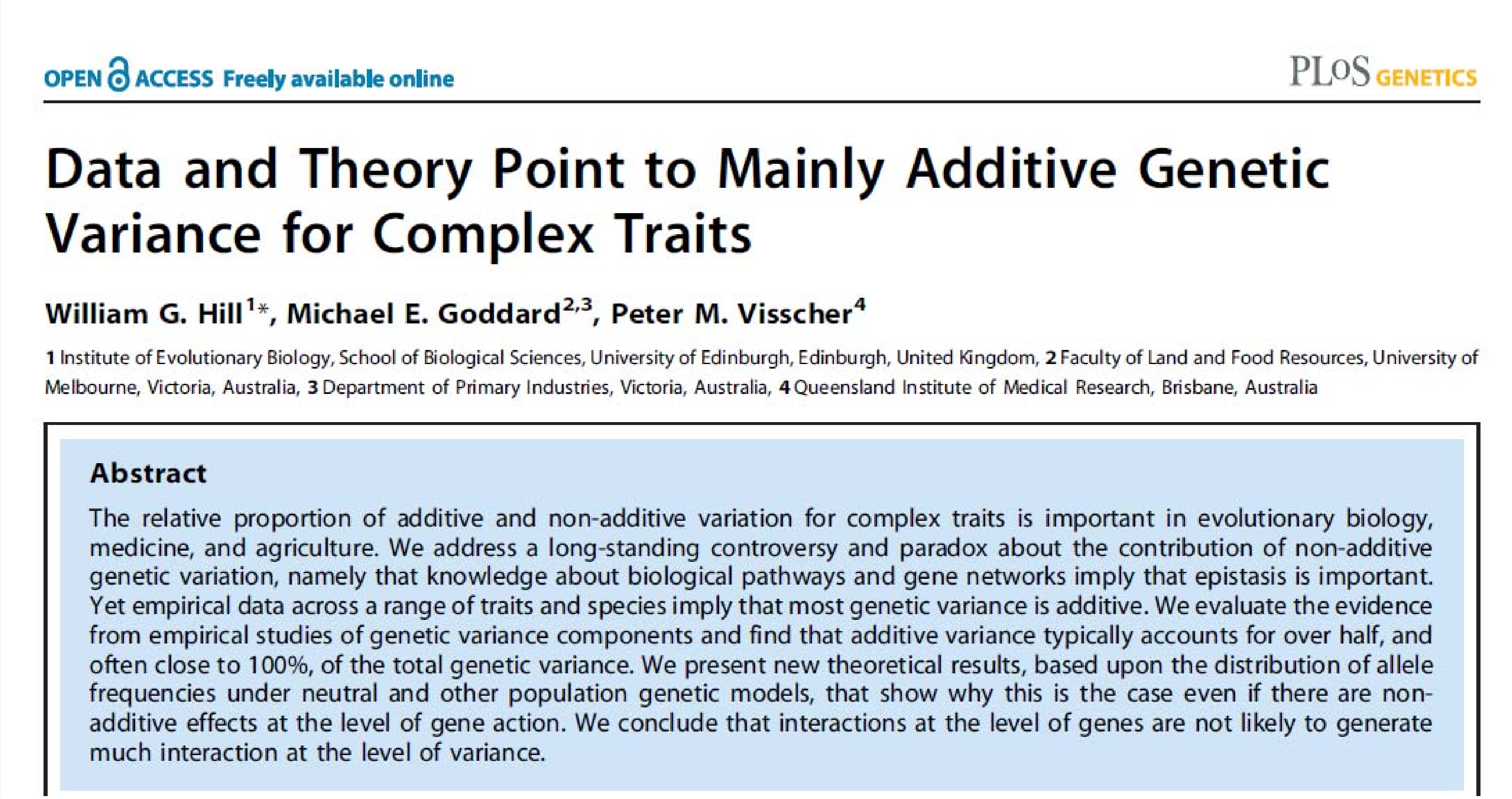
Fig. 5. Networks of epistatic interactions. Interaction networks are depicted for (*A*) starvation resistance and (*B*) chill coma recovery. Nodes depict genes, and edges significant interactions. Red nodes are genes containing significant SNPs from the Flyland analysis. Blue nodes are genes containing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



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.

- ***UPPER LIMIT PLACED ON VARIANCE COMPONENTS FOR DISCOVERY PURPOSES: EVERYTHING TURNS ADDITIVE EVEN IF NOT SO…***

- ***ADDITIVE MODEL DENIES WHAT IT IS AND EXPLAINS WHAT IT IS NOT!***

"Good news. The test results show it's a metaphor."

HMM, NICE TO SEE SOME GOOD NEWS FOR A CHANGE.
NEWS
GLOBAL WARMING ON THE RISE

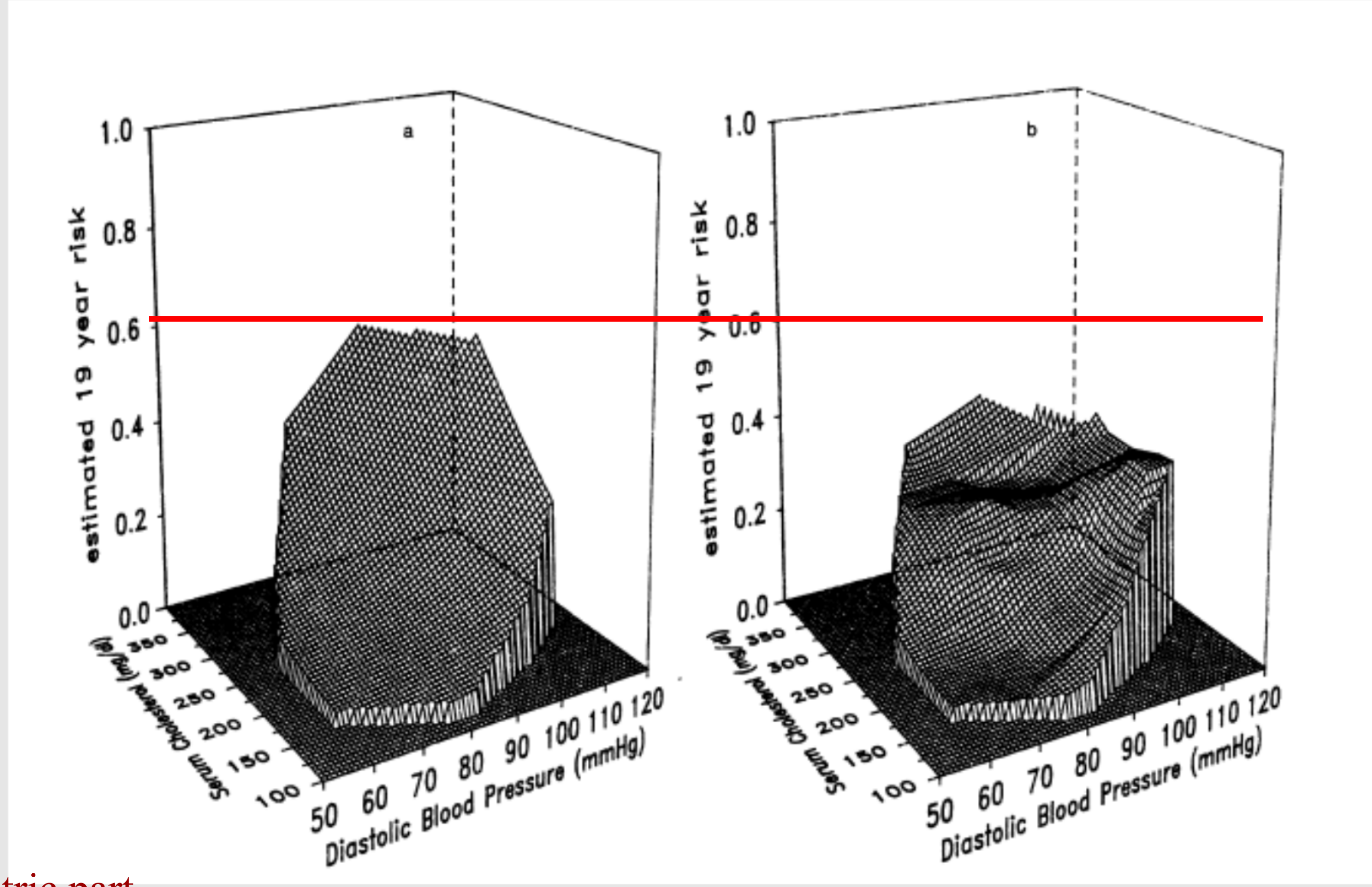
Rubes By Leigh Rubin

Creators Syndicate, Inc. rubes2@earthlink.net
© 2006 Leigh Rubin! www.creators.com

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

Logistic regression with thin-plate splines



parametric part

$$f(\mathbf{x}_i) = \boxed{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}} + \sum_{j=1}^{N} \alpha_j \left[(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 \right] \log \left[(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 \right]$$

Risk of heart attack after 19 years as a function of cholesterol level and blood pressure. Left: logistic regression model. Right: thin plate spline fit. Wahba (2007)

"Would you refuse your dinner because you do not understand the digestive system?"

quote by British mathematician in "*The emperor of the maladies: a biography of cancer*",2010, by *Siddhartha Mujkherjee*

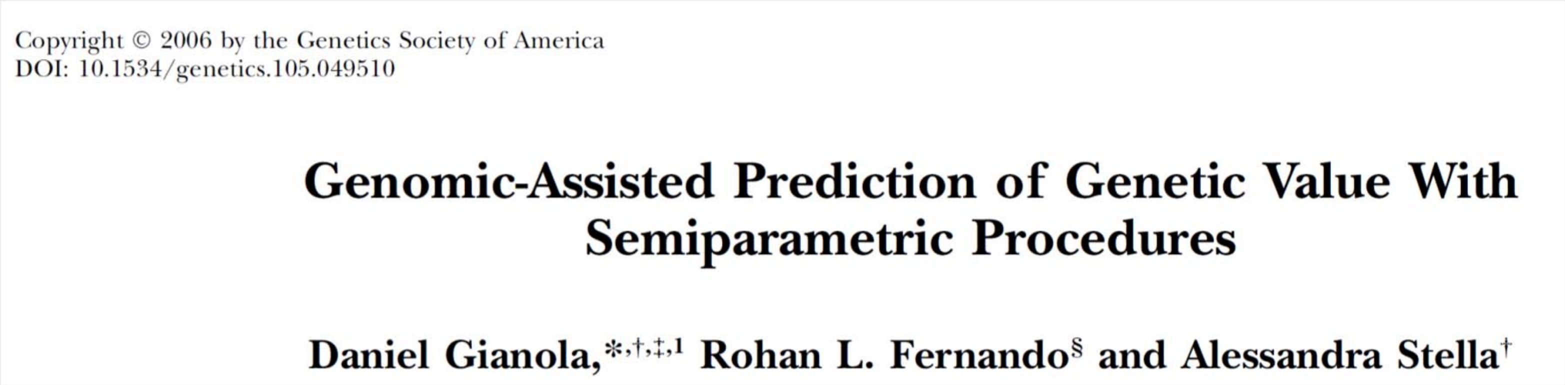
SOME PREDICTION MACHINES. YOU HAVE HEARD OF:

- BLUP using pedigrees
- BLUP using markers (GBLUP)
- Support vector machines in regression or classification?
- Kriging in geostatistics
- Kernel machines in computer science

THESE METHODS ARE SPECIAL CASES OF A GENERAL FRAMEWORK: RKHS

(Reproducing kernel Hilbert spaces methodology)
Sounds scary…

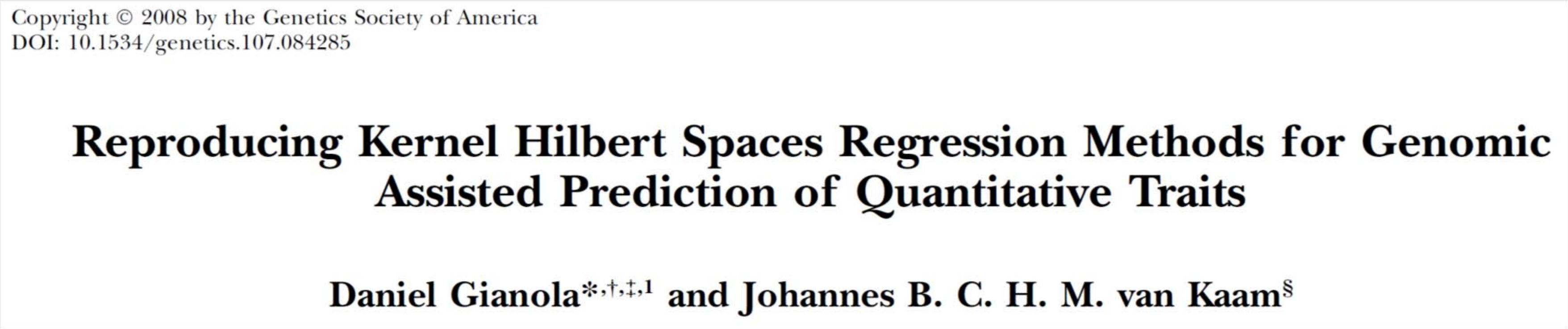
“And at the beginning there was light…”



DOI: 10.1534/genetics.105.049510

Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures

Daniel Gianola,[*,†,‡,1] Rohan L. Fernando[§] and Alessandra Stella[†]



DOI: 10.1534/genetics.107.084285

Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits

Daniel Gianola[*,†,‡,1] and Johannes B. C. H. M. van Kaam[§]

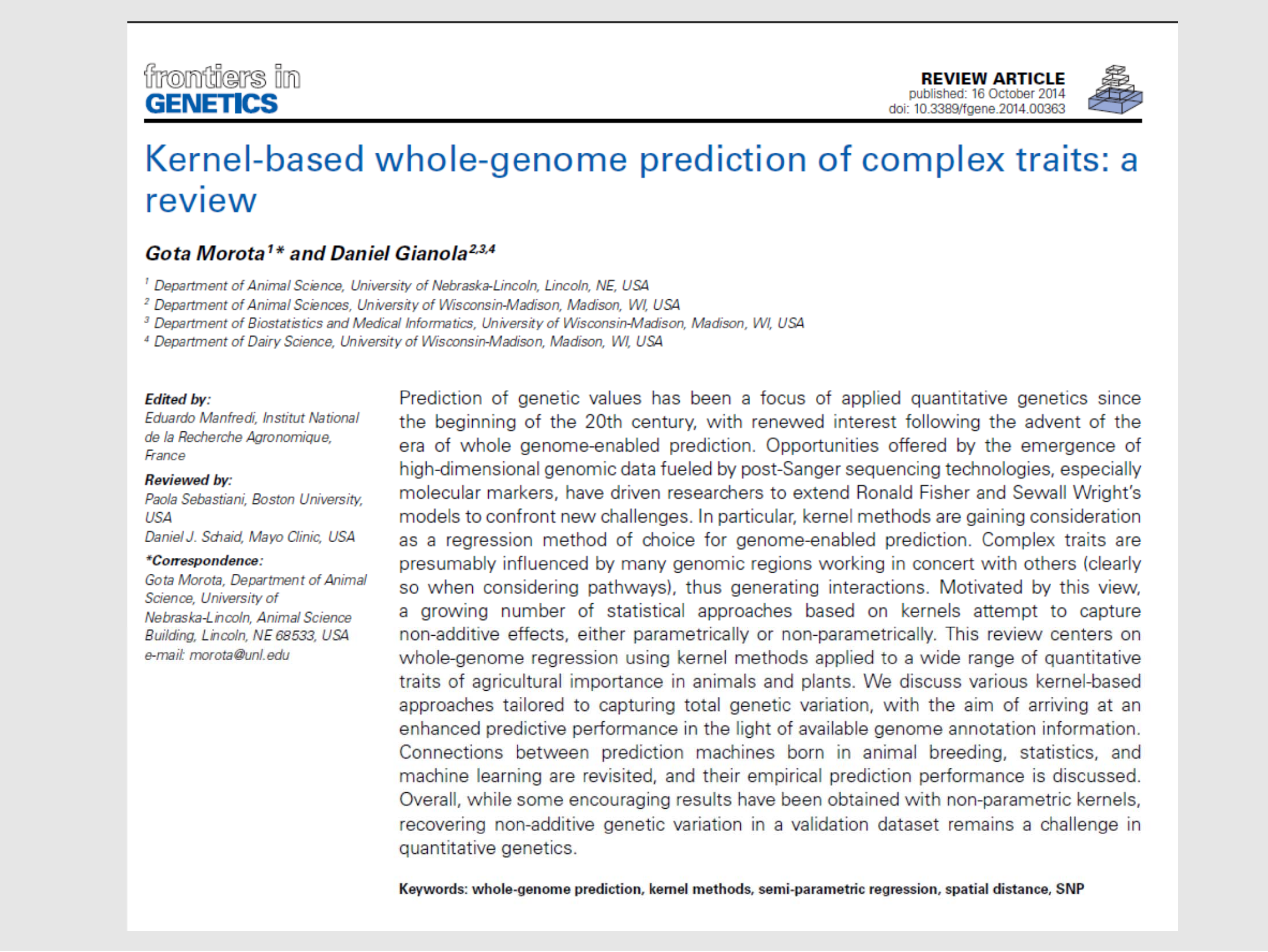
DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1136

February 21, 2007

Statistical Learning in Medical Data Analysis

Grace Wahba[1]



frontiers in
GENETICS

REVIEW ARTICLE
published: 16 October 2014
doi: 10.3389/fgene.2014.00363

Kernel-based whole-genome prediction of complex traits: a review

Gota Morota[1]* and Daniel Gianola[2,3,4]

[1] Department of Animal Science, University of Nebraska-Lincoln, Lincoln, NE, USA
[2] Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI, USA
[3] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA
[4] Department of Dairy Science, University of Wisconsin-Madison, Madison, WI, USA

Edited by:
Eduardo Manfredi, Institut National de la Recherche Agronomique, France

Reviewed by:
Paola Sebastiani, Boston University, USA
Daniel J. Schaid, Mayo Clinic, USA

***Correspondence:**
Gota Morota, Department of Animal Science, University of Nebraska-Lincoln, Animal Science Building, Lincoln, NE 68533, USA
e-mail: morota@unl.edu

Prediction of genetic values has been a focus of applied quantitative genetics since the beginning of the 20th century, with renewed interest following the advent of the era of whole genome-enabled prediction. Opportunities offered by the emergence of high-dimensional genomic data fueled by post-Sanger sequencing technologies, especially molecular markers, have driven researchers to extend Ronald Fisher and Sewall Wright's models to confront new challenges. In particular, kernel methods are gaining consideration as a regression method of choice for genome-enabled prediction. Complex traits are presumably influenced by many genomic regions working in concert with others (clearly so when considering pathways), thus generating interactions. Motivated by this view, a growing number of statistical approaches based on kernels attempt to capture non-additive effects, either parametrically or non-parametrically. This review centers on whole-genome regression using kernel methods applied to a wide range of quantitative traits of agricultural importance in animals and plants. We discuss various kernel-based approaches tailored to capturing total genetic variation, with the aim of arriving at an enhanced predictive performance in the light of available genome annotation information. Connections between prediction machines born in animal breeding, statistics, and machine learning are revisited, and their empirical prediction performance is discussed. Overall, while some encouraging results have been obtained with non-parametric kernels, recovering non-additive genetic variation in a validation dataset remains a challenge in quantitative genetics.

Keywords: whole-genome prediction, kernel methods, semi-parametric regression, spatial distance, SNP

Reproducing Kernel Hilbert spaces mixed model regression

Function of molecular information x (e.g., vector of SNPs)

non-parametric

$$SS[g(\mathbf{x}),\lambda] = \sum_{i=1}^{n}[y_i - \overbrace{\mathbf{w}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{u}}^{\text{parametric}} - g(x_i)]^2 + \lambda||g(\mathbf{x})||_H^2$$

Smoothing parameter ($\lambda=\frac{1}{\theta}$)

"Penalized sum of squares"

Norm under
Hilbert space (*H*) of
functions, a huge class

Variational problem: find *g(**x**)* over entire space of functions minimizing SS(.)

1) Solution to variational problem: linear function (Kimeldorf & Wahba, 1971)

No. individuals with molecular data

$$g(.) \quad = \alpha_0 + \sum_{j=1}^{n} \alpha_j K(.\,, \mathbf{x}_j)$$

reduction of dimension p (# SNPs)➔ # indiv.

Regression coefficient

Reproducing kernel (may contain bandwidth parameters)

2) Model becomes

$$\begin{aligned} y &= X\beta + Zu + g\left(\text{markers}\right) + e \\ &= X\beta + Zu + K\alpha + e \\ \alpha &\sim N\left(0, K^{-1}\sigma_\alpha^2\right) \end{aligned}$$

KERNEL CONSTRUCTION: MAIN ISUES

-KERNEL MUST BE AN n x n SYMMETRIC PSD MATRIX

-NOTION OF DISTANCE (“similarity”) BETWEEN GENOTYPES OF PAIRS OF INDIVIDUAL

-MATHEMATICAL FORM (LINEAR OR NON-LINEAR TRANSFORMATION OF INPUTS: THE MARKER CODES)

EXAMPLES OF MEASURES OF DISTANCE (marker genotypes in pairs of individuals THAT CAN BE USED IN KERNELS

Euclidean

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

$$\|\mathbf{q}-\mathbf{p}\| = \sqrt{\|\mathbf{p}\|^2 + \|\mathbf{q}\|^2 - 2\mathbf{p}\cdot\mathbf{q}}.$$

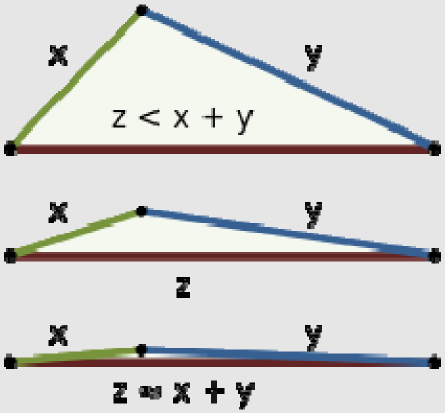
Manhattan

$$d(x,y)=\sum_{k=1}^{p}|x_k-y_k|,$$

Bray-Curtis

$$d_{ij} = (\Sigma_k \mid x_{ik} - x_{jk} \mid)/(\Sigma_k x_{ik} + x_{jk})$$

Distances must meet triangle inequality



EXAMPLES OF MATHEMATICAL FORMS OF KERNELS

Standard (no bandwidth)

A : numerator relationship matrix

G : genomic relationship matrices

Gaussian

$$K_h(\mathbf{x}, \mathbf{x}_j) = \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_j)'(\mathbf{x}-\mathbf{x}_j)}{h}\right]$$

“bandwidth”

$h = \frac{1}{\theta}$

Gaussian on relationships

$$K_\theta = \left\{\exp\left(-\theta \frac{a_{ij}{}^2}{\max\left(a_{ij}{}^2\right)}\right)\right\}$$

t-kernel

$$K(\mathrm{x}_i, \mathrm{x}_j) = [1 + \frac{(\mathrm{x}_i - \mathrm{x}_j)'\Sigma^{-1}(\mathrm{x}_i - \mathrm{x}_j)}{m\nu}]^{-\frac{(\nu+m)}{2}}$$

Bandwidth parameter

m= #markers
v =bandwidth

Possible alternatives for *t-kernel* for an S x 1 vector of markers

Bandwidth parameter

$$k_{\nu,\Sigma}(x_i,x_j) = \left[1 + \frac{(\mathbf{x}_i-\mathbf{x}_j)' \sum^{-1} (\mathbf{x}_i-\mathbf{x}_j)}{S\nu} \right]^{-\left(\frac{1+\nu}{2}\right)}$$

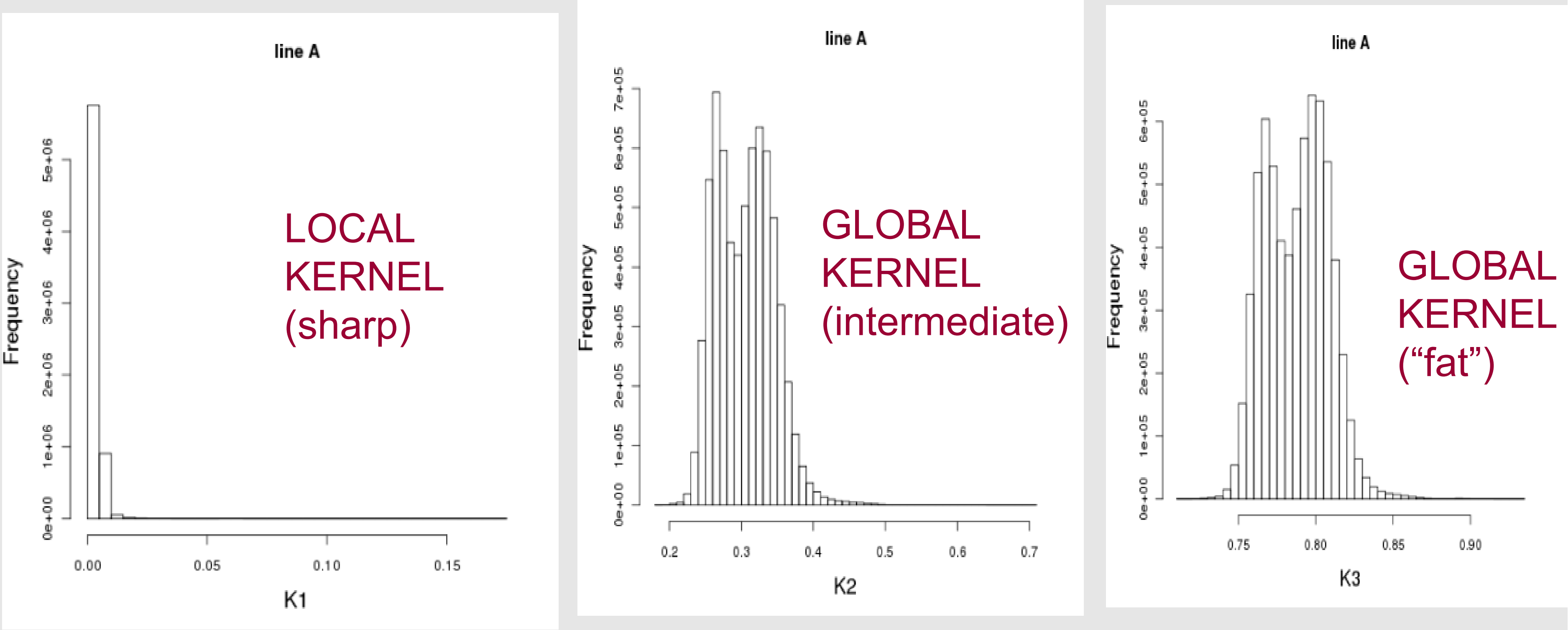
$\Sigma^{-1} = Diag(2p_k q_k),$

$\Sigma = Diag(2p_k q_k)$

$\Sigma^{-1} = R$ where R is a matrix containing r^2 from LD

$\Sigma = R$

GAUSSIAN KERNEL WITH 3 BANDWIDTHS

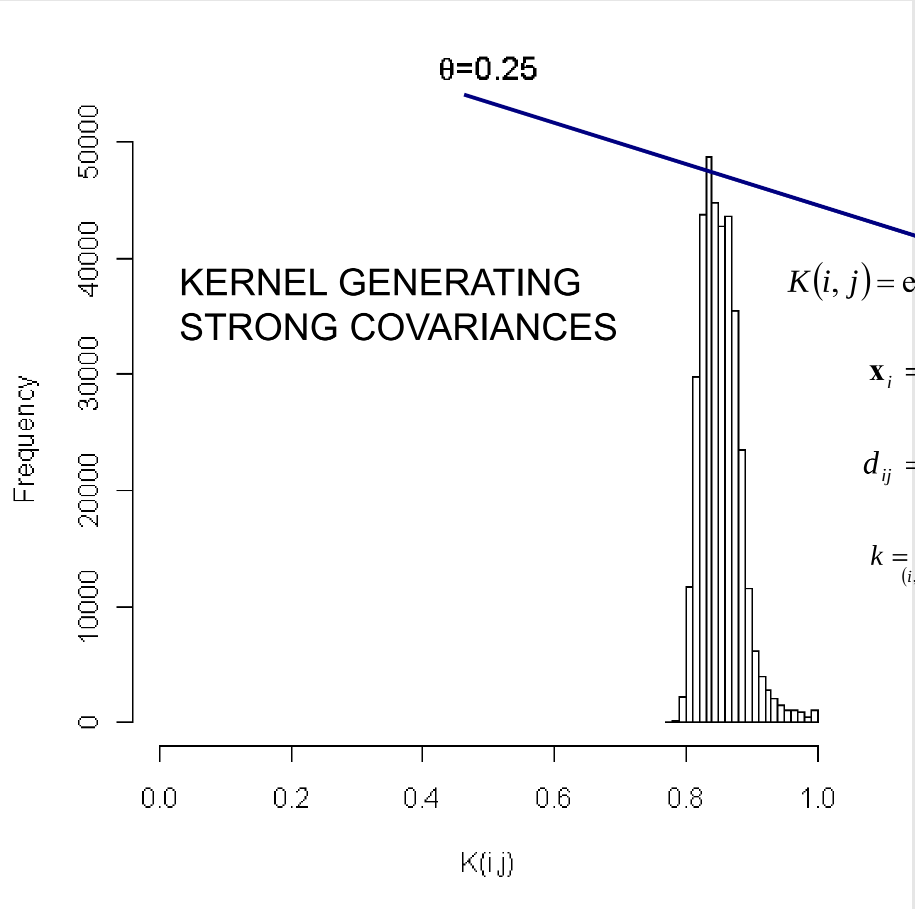


LOCAL KERNEL (sharp)

GLOBAL KERNEL (intermediate)

GLOBAL KERNEL (“fat”)

Histograms of the entries of $K=\{K(\boldsymbol{x}_i, \boldsymbol{x}_{i'})\}$, .



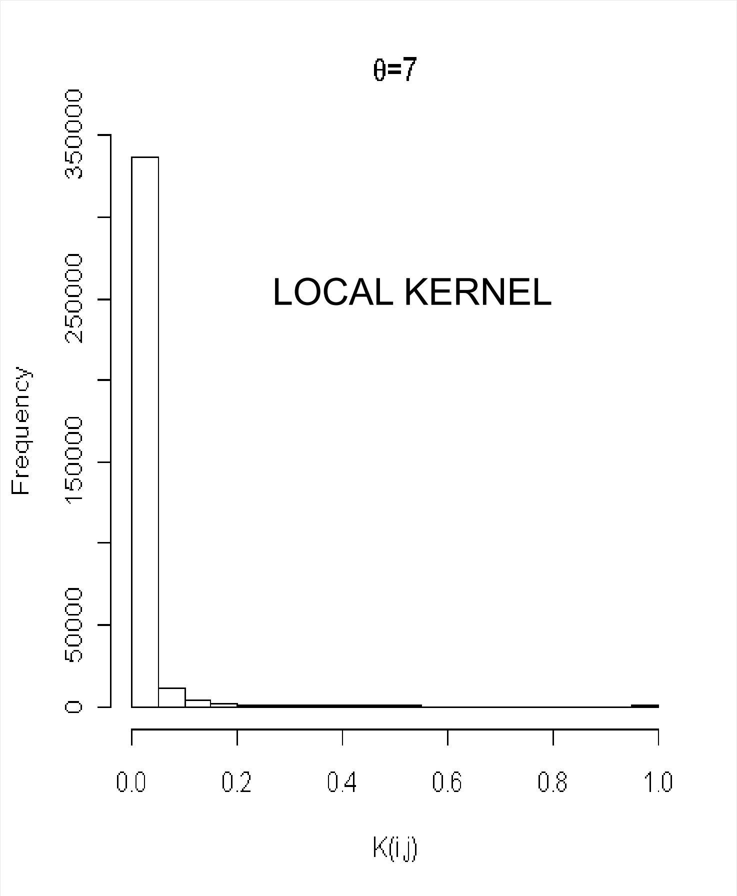
KERNEL GENERATING STRONG COVARIANCES

$$K(i,j) = \exp\left\{-\theta\, k^{-1} d_{ij}\right\}$$

$$\mathbf{x}_i = \left(x_{i1},\dots,x_{ip}\right)'$$

$$d_{ij} = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2$$

$$k = \max_{(i,j)}\left\{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right\}$$



LOCAL KERNEL

Histogram of evaluations of Gaussian kernel by value of bandwidth parameter

"Excessively" sharp kernels approach ***I*** *(n x n)* and may copy noise, plus pose identification problem (variance component issue)

Mixed model representation of a semi-parametric regression

nuisances

$$y_i \;=\; \mathbf{w}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u} + \sum_{j=1}^{n} \exp\left[-\frac{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}{h}\right]\alpha_j + e_i$$

Infinitesimal additive effect

Gaussian kernel on markers used

Type equation here.

Define row vector

$$\boldsymbol{k}'_i(h) \;=\; \left\{\exp\left[-\frac{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}{h}\right]\right\}$$

$$\mathbf{K}(h) = \begin{bmatrix} \mathbf{k}_1'(h) \\ \mathbf{k}_2'(h) \\ . \\ . \\ . \\ \mathbf{k}_n'(h) \end{bmatrix}$$

*(**K** is n x n and symmetric so **K=K’**)*

Then:

Bandwidth parameter

$$\boxed{\mathbf{y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{K}(h)\boldsymbol{\alpha} + \mathbf{e}}$$

$$\sigma_{\alpha}^{2} = \frac{1}{\lambda}$$

Do:

$$\boxed{\boldsymbol{\alpha} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{K}^{-1}(h)\sigma_{\alpha}^{2})}$$

Smoothing parameter

T = K below (sorry, I discovered that I had used ***T*** instead of ***K***)

$$\begin{bmatrix} \mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{T}(h) \\ \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\dfrac{\sigma_e^2}{\sigma_u^2} & \mathbf{Z}'\mathbf{T}(h) \\ \mathbf{T}'(h)\mathbf{W} & \mathbf{T}'(h)\mathbf{Z} & \mathbf{T}'(h)\mathbf{T}(h) + \mathbf{T}(h)\dfrac{\sigma_e^2}{\sigma_\alpha^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \mathbf{W}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{T}'(h)\mathbf{y} \end{bmatrix}$$

h assumed known here

THE "ANIMAL MODEL" IS A PARTICULAR CASE OF RKHS

$$
\begin{aligned}
y &= A\alpha + e \\
\alpha &\sim N(0, A^{-1}\sigma_a^2) \\
e &\sim N(0, I\sigma_e^2)
\end{aligned}
$$

*Use **A** as kernel matrix*

$$\Rightarrow u = A\alpha \sim N(0, A\sigma_a^2)$$

$$\left(A'A + A\frac{\sigma_e^2}{\sigma_a^2}\right)\hat{\alpha} = A'y$$

$$A\left(A + I\frac{\sigma_e^2}{\sigma_a^2}\right)\hat{\alpha} = Ay$$

$$\hat{\alpha} = \left(A + I\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1} y$$

Predicted Genetic signal

$$A\hat{\alpha} = \left(I + A^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1} y = \text{BLUP(additive effects)}$$

GENOMIC BLUP IS A PARTICULAR CASE OF RKHS

$$y = XX'\alpha + e$$

$$\alpha \sim N\left(0, (XX')^{-1}\sigma_\beta^2\right)$$

$$e \sim N(0, I\sigma_e^2)$$

$$\Rightarrow u = XX'\alpha \sim N(0, XX'\sigma_\beta^2)$$

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

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

$$\hat{\alpha} = \left(XX' + I\frac{\sigma_e^2}{\sigma_\beta^2}\right)^{-1} y$$

Predicted Genetic signal

$$XX'\hat{\alpha} = XX'\left(XX' + I\frac{\sigma_e^2}{\sigma_\beta^2}\right)^{-1} y$$

$$\hat{u} = \left(I + (XX')^{-1}\frac{\sigma_e^2}{\sigma_\beta^2}\right)^{-1} y = \text{"GENOMIC BLUP"}$$

Penalized estimation

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\arg\min} \left\{ \left(\mathbf{y}-\mathbf{K}\boldsymbol{\alpha}\right)'\left(\mathbf{y}-\mathbf{K}\boldsymbol{\alpha}\right) + \lambda\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} \right\}$$

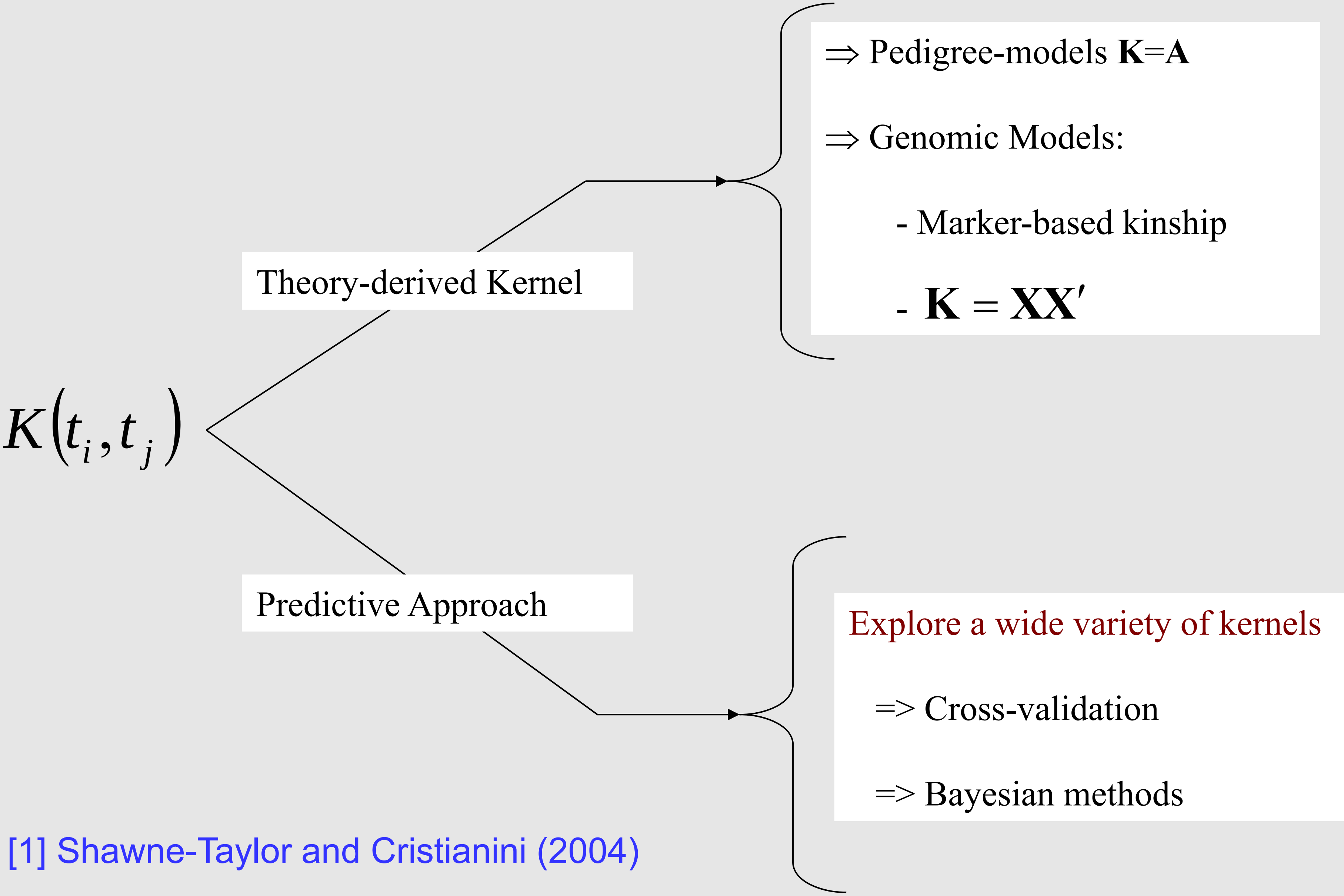
Bayesian or REML View

$$\begin{cases} \mathbf{y} = \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \\ p\left(\boldsymbol{\varepsilon},\boldsymbol{\alpha}\right) = N\left(\boldsymbol{\varepsilon}\middle|\mathbf{0},\mathbf{I}\sigma_{\varepsilon}^{2}\right)N\left(\boldsymbol{\alpha}\middle|\mathbf{0},\mathbf{K}^{-1}\sigma_{\alpha}^{2}\right) \end{cases}$$

Place priors, if Bayesian

[1] Kimeldorf, G.S. & Wahba, G. (1970).

How to Choose the Reproducing Kernel? [1]



$K(t_i, t_j)$

Theory-derived Kernel

$\Rightarrow$ Pedigree-models **K**=**A**

$\Rightarrow$ Genomic Models:

- Marker-based kinship

- $\mathbf{K} = \mathbf{XX}'$

Predictive Approach

Explore a wide variety of kernels

=> Cross-validation

=> Bayesian methods

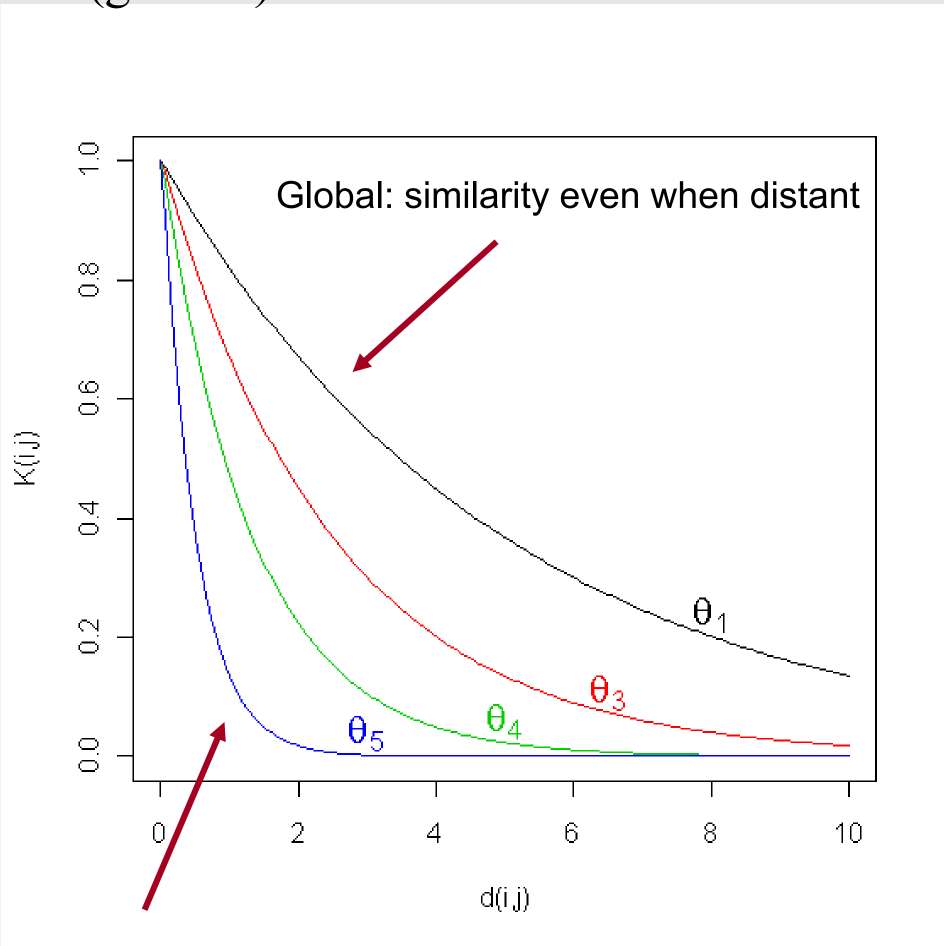
[1] Shawne-Taylor and Cristianini (2004)

Example of multiple-kernel fitting: 4 Gaussians simultaneously

$d\left(\mathbf{x}_i, \mathbf{x}_j\right)$:

(genetic) distance between individuals

$$\Longrightarrow \quad K\left(i, j \middle| \theta\right) = Exp\left\{ -\theta_k \times d\left(\mathbf{x}_i, \mathbf{x}_j\right) \right\}; k = 1,2,3,4$$



Sharp: similarity only if close

Operationally

$$\begin{aligned} y &= X\beta + Zu + \\ &\quad K_1\alpha_1 + K_2\alpha_2 + K_3\alpha_3 + K_4\alpha_4 + e; \\ \alpha_i &\sim N(0, K_i^{-1}\sigma^2_{\alpha_i}) \end{aligned}$$

de los Campos et al. (2010) Genetics Research

FAQ 1

- Why can RKHS capture (potentially) epistasis even when *K encodes* additive marker codes only?

EXAMPLE: 2 LOCUS MODEL WITH EPISTASIS

$$y = x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_{12} + e$$

Effect of allelic substitution at locus 1 depends on locus 2

$$\frac{\delta y}{\delta x_1} = \beta_1 + x_2\beta_{12}$$

EXAMPLE: RKHS WITH GAUSSIAN KERNEL

$$y(\boldsymbol{x}) = \sum_{i=1}^{n} exp\left(\frac{(\boldsymbol{x}-\boldsymbol{x}_i)'(\boldsymbol{x}-\boldsymbol{x}_i)}{h}\right) + \epsilon$$

$$\frac{\delta y}{\delta x_j} = 2\sum_{i=1}^{n} exp\left(\frac{(\boldsymbol{x}-\boldsymbol{x}_i)'(\boldsymbol{x}-\boldsymbol{x}_i)}{h}\right)(\boldsymbol{x}-\boldsymbol{x}_i)'0_j$$

0_j is a p×1 vector of 0's, save for a 1 in position j.

Effect of allelic substitution at locus j depends on ALL other loci
Form of epistasis not represented by linear models

FAQ 2

- Can RKHS produce an “estimated breeding value”?

EXAMPLE:

Consider multi-kernel representation with *c*+2 kernels:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{g} + \boldsymbol{G}\boldsymbol{\alpha}_g + \sum_{i=1}^{c} \boldsymbol{K}_i(\boldsymbol{x}, \boldsymbol{h}_i)\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}.$$

Kernel $\boldsymbol{A}$: infinitesimal effects

Kernel $\boldsymbol{G}$: additive effects of markers

***c* kernels $\boldsymbol{K}_i$** could be Gaussian kernels with varying bandwidth parameters.

RKHS($Ag + G\alpha_g$) would capture additive information from A and G

FAQ 3

- Suppose I construct additive + dominance genomic relationship matrices (kernels) and additive x dominance (another kernel) as:

  - G(add)
  - G(dom)
  - G(add) # G(dom)

Do I obtain meaningful estimates of additive, dominance and additive x dominance variances?

Morota et al. (2014)

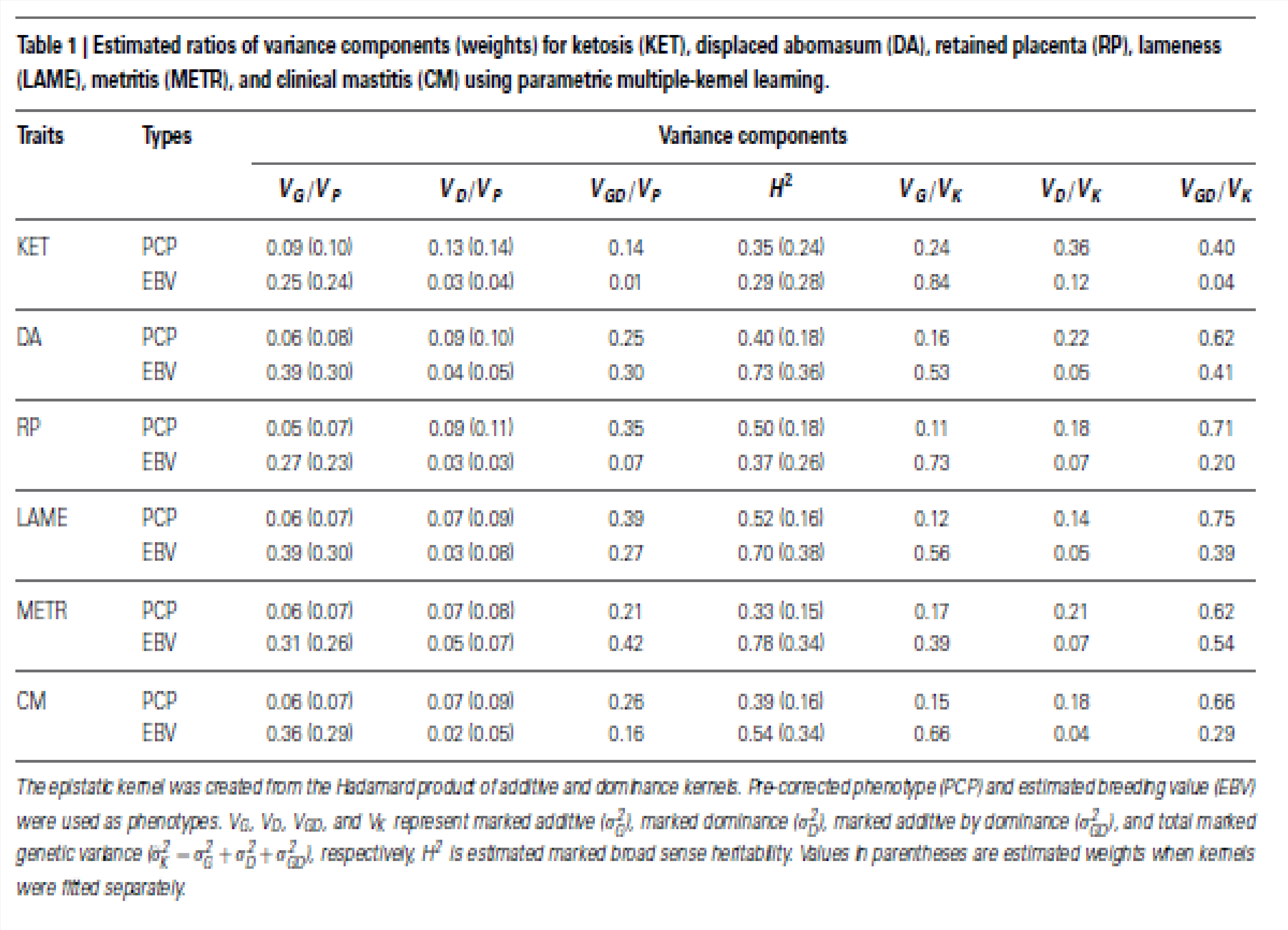


Table 1 | Estimated ratios of variance components (weights) for ketosis (KET), displaced abomasum (DA), retained placenta (RP), lameness (LAME), metritis (METR), and clinical mastitis (CM) using parametric multiple-kernel learning.

Traits	Types	Variance components						
		V_G/V_P	V_D/V_P	V_{GD}/V_P	H^2	V_G/V_K	V_D/V_K	V_{GD}/V_K
KET	PCP	0.09 (0.10)	0.13 (0.14)	0.14	0.35 (0.24)	0.24	0.36	0.40
	EBV	0.25 (0.24)	0.03 (0.04)	0.01	0.29 (0.28)	0.84	0.12	0.04
DA	PCP	0.06 (0.08)	0.09 (0.10)	0.25	0.40 (0.18)	0.16	0.22	0.62
	EBV	0.39 (0.30)	0.04 (0.05)	0.30	0.73 (0.36)	0.53	0.05	0.41
RP	PCP	0.05 (0.07)	0.09 (0.11)	0.35	0.50 (0.18)	0.11	0.18	0.71
	EBV	0.27 (0.23)	0.03 (0.03)	0.07	0.37 (0.26)	0.73	0.07	0.20
LAME	PCP	0.06 (0.07)	0.07 (0.09)	0.39	0.52 (0.16)	0.12	0.14	0.75
	EBV	0.39 (0.30)	0.03 (0.08)	0.27	0.70 (0.38)	0.56	0.05	0.39
METR	PCP	0.06 (0.07)	0.07 (0.08)	0.21	0.33 (0.15)	0.17	0.21	0.62
	EBV	0.31 (0.26)	0.05 (0.07)	0.42	0.78 (0.34)	0.39	0.07	0.54
CM	PCP	0.06 (0.07)	0.07 (0.09)	0.26	0.39 (0.16)	0.15	0.18	0.66
	EBV	0.36 (0.29)	0.02 (0.05)	0.16	0.54 (0.34)	0.66	0.04	0.29

The epistatic kernel was created from the Hadamard product of additive and dominance kernels. Pre-corrected phenotype (PCP) and estimated breeding value (EBV) were used as phenotypes. V_G, V_D, V_{GD}, and V_K represent marked additive (σ_G^2), marked dominance (σ_D^2), marked additive by dominance (σ_{GD}^2), and total marked genetic variance ($\sigma_K^2 - \sigma_G^2 + \sigma_D^2 + \sigma_{GD}^2$), respectively. H^2 is estimated marked broad sense heritability. Values in parentheses are estimated weights when kernels were fitted separately.

ANSWER: NO! KERNELS MUST BE MUTUALLY ORTHOGONAL

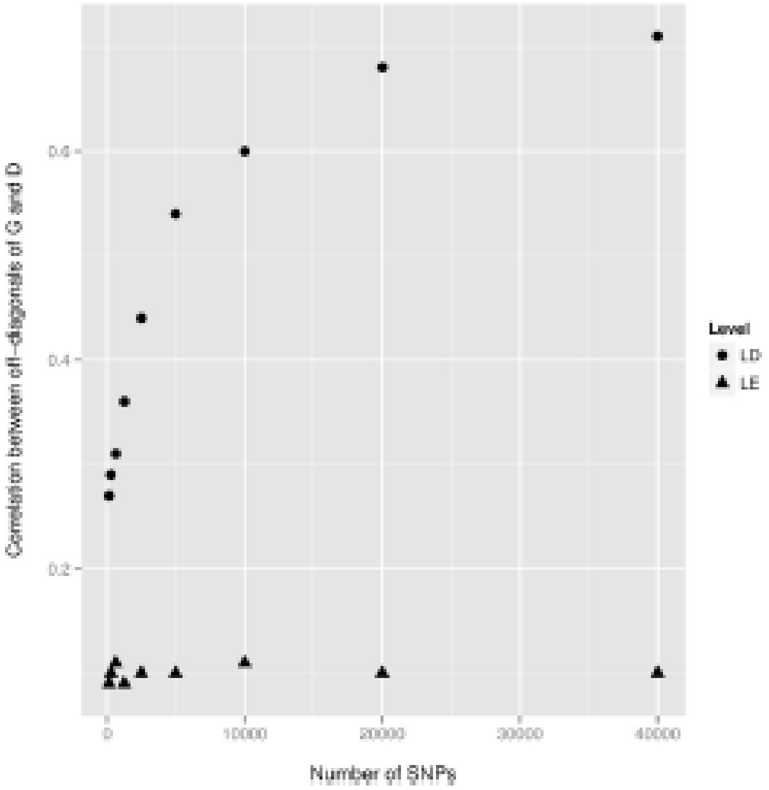


FIGURE 3 | Correlations between off-diagonal elements of the additive genomic relationship matrix G and of the dominance relationship matrix D as a function of the number of SNPs. Genotypes were both randomly sampled from the present study (Level = LD) and via a computer simulation locus by locus (Level = LE) with an average minor allele frequency equal to 0.35. The averages of the r^2 linkage disequilibrium (LD) statistic between adjacent markers were 0.18 and 0.008 for the real and simulated datasets, respectively.

➔Orthogonality destroyed by LD

➔Gets worse with increased # SNP

Corollary: do not take seriously claims of “dominance” and “epistatic” variance from naively constructed genomic relationship matrices.

Example 1 of RKHS

$$\begin{bmatrix} y_2 = 5 \\ y_3 = 3 \\ y_4 = 7 \\ y_5 = 8 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \left(\underset{\substack{\uparrow \\ \text{Additive}}}{\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix}} + \underset{\substack{\uparrow \\ \text{Dominance}}}{\begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{bmatrix}} \right) + \begin{bmatrix} e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$= \mathbf{X\beta} + \mathbf{Z}(\mathbf{a} + \mathbf{d}) + \mathbf{e}.$$

Henderson (1985) assumed $\sigma_a^2 = 5, \sigma_d^2 = 4$ and $\sigma_e^2 = 20$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 1 \end{bmatrix} \text{ and } \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Application of BLUP paradigm leads to

$$\widehat{\boldsymbol{\beta}}' = \begin{bmatrix} 5.145 & 0.241 \end{bmatrix},$$

$$\widehat{\mathbf{a}}' = \begin{bmatrix} 0.045 & -0.192 & -0.343 & 0.096 & 0.242 \end{bmatrix},$$

$$\widehat{\mathbf{d}}' = \begin{bmatrix} 0 & -0.073 & -0.365 & 0.162 & 0.234 \end{bmatrix}.$$

$$\widehat{\mathbf{g}} = \widehat{\mathbf{a}} + \widehat{\mathbf{d}} = \begin{bmatrix} 0.045 & -0.265 & -0.708 & 0.259 & 0.477 \end{bmatrix}$$

Next, do RKHS with K=A+D as positive-definite kernel matrix

$$\mathbf{K} = \mathbf{A} + \mathbf{D} = \begin{bmatrix} 2 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 2 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 2 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & \frac{3}{4} & 2 & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 2 \end{bmatrix}$$

$$\begin{bmatrix} y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 2 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 2 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{3}{4} & 2 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & 2 \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} + \begin{bmatrix} e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$= \mathbf{X\beta} + \mathbf{K\alpha} + \mathbf{e}.$$

$\sigma_{\alpha}^{2} = \sigma_{a}^{2} + \sigma_{d}^{2} = 9$ ➔ This is 1/λ

$$\left[\ \widehat{\beta}_0 = 5.289 \quad \widehat{\beta}_1 = 0.200 \quad \widehat{\alpha}_2 = -0.128 \quad \widehat{\alpha}_3 = -0.781 \quad \widehat{\alpha}_4 = 0.487 \quad \widehat{\alpha}_5 = 0.422\ \right]$$

$$\begin{bmatrix} \widehat{g}_{K,1} \\ \widehat{g}_{K,2} \\ \widehat{g}_{K,3} \\ \widehat{g}_{K,4} \\ \widehat{g}_{K,5} \end{bmatrix} = \begin{bmatrix} 0.036 \\ -0.210 \\ -0.569 \\ 0.206 \\ 0.382 \end{bmatrix}$$

COMPARED WITH

$$\widehat{\mathbf{g}} = \widehat{\mathbf{a}} + \widehat{\mathbf{d}} = \left[\ 0.045 \quad -0.265 \quad -0.708 \quad 0.259 \quad 0.477\ \right]$$

PREDICTING FUTURE RECORDS UNDER THE SAME ENVIRONMENTAL CONDITIONS; PARAMETRICALLY

$$\begin{bmatrix} y_1^f \\ y_2^f \\ y_3^f \\ y_4^f \\ y_5^f \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} + \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{bmatrix} + \begin{bmatrix} e_1^f \\ e_2^f \\ e_3^f \\ e_4^f \\ e_5^f \end{bmatrix}$$

$$= \mathbf{M}_P \boldsymbol{\theta}_P + \mathbf{e}^f,$$

PREDICTION OF FUTURE RECORDS NON-PARAMETRICALLY

$$\begin{bmatrix} y_1^f \\ y_2^f \\ y_3^f \\ y_4^f \\ y_5^f \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 2 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 2 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{3}{4} & 2 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & 2 \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} + \begin{bmatrix} e_1^f \\ e_2^f \\ e_3^f \\ e_4^f \\ e_5^f \end{bmatrix}$$

$$= \mathbf{M}_K \boldsymbol{\theta}_K + \mathbf{e}^f.$$

FOR BOTH APPROACHES THE PREDICTIVE DISTRIBUTION IS

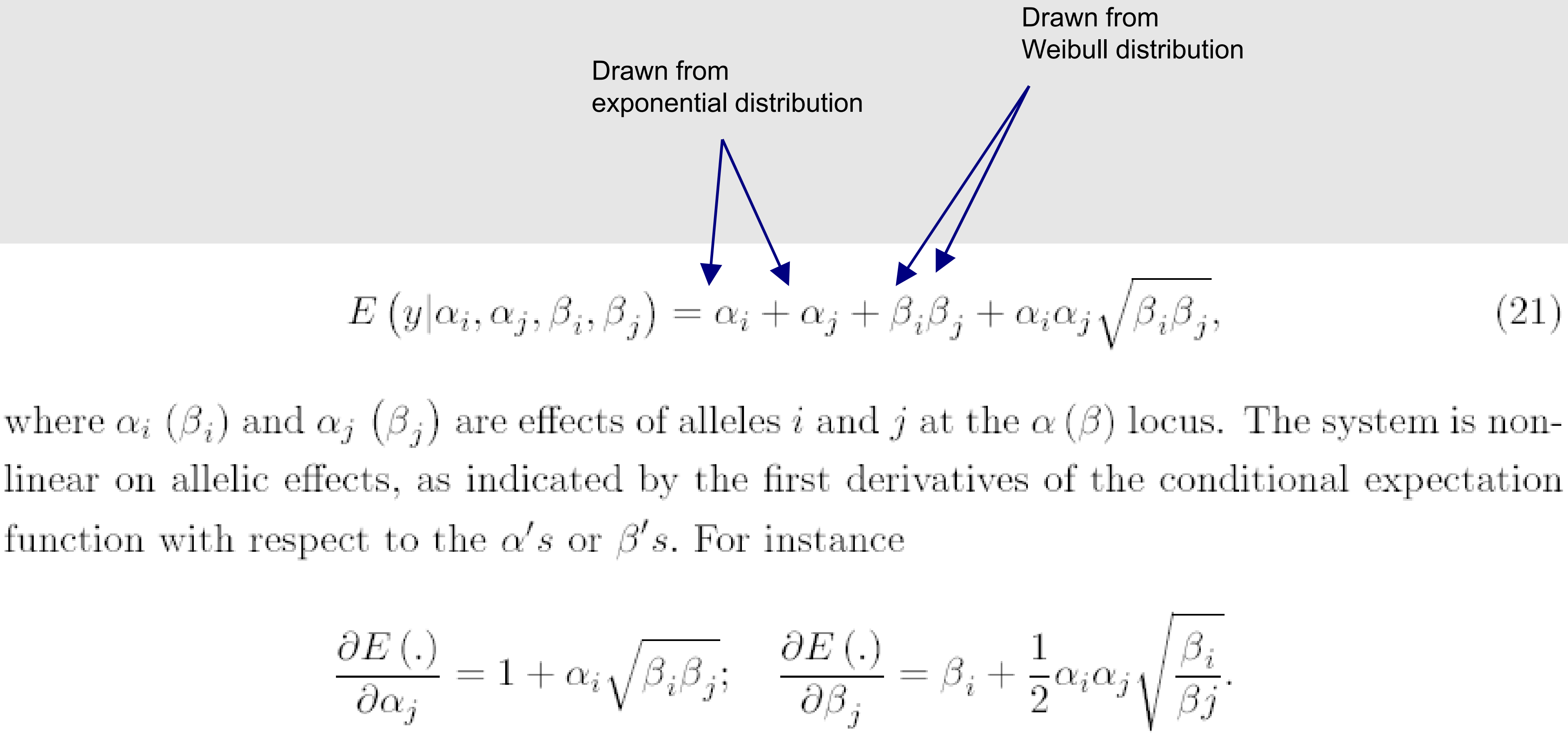
$$\begin{bmatrix} y_1^f \\ y_2^f \\ y_3^f \\ y_4^f \\ y_5^f \end{bmatrix} \Bigg| \begin{bmatrix} y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}, \text{dispersion (smoothing) parameters}$$

$$\sim \left(\mathbf{M}_{\cdot}\widehat{\boldsymbol{\theta}_{\cdot}}, (\mathbf{M}_{\cdot}\mathbf{C}_{\cdot}^{-1}\mathbf{M}_{\cdot}' + \mathbf{I}_f)\sigma_e^2\right),$$

For the two procedures the mean and SD of the predictive distributions are:

$$P = \begin{bmatrix} 5.674 \pm 6.020 \\ 5.364 \pm 5.460 \\ 5.162 \pm 5.353 \\ 5.646 \pm 5.834 \\ 6.828 \pm 6.115 \end{bmatrix}; K = \begin{bmatrix} 5.754 \pm 5.576 \\ 5.286 \pm 5.659 \\ 4.735 \pm 5.561 \\ 5.919 \pm 5.940 \\ 7.061 \pm 6.157 \end{bmatrix}$$

Example 2 of RKHS



Drawn from exponential distribution

Drawn from Weibull distribution

$$E\left(y|\alpha_i,\alpha_j,\beta_i,\beta_j\right)=\alpha_i+\alpha_j+\beta_i\beta_j+\alpha_i\alpha_j\sqrt{\beta_i\beta_j}, \qquad (21)$$

where α_i (β_i) and α_j (β_j) are effects of alleles i and j at the α (β) locus. The system is non-linear on allelic effects, as indicated by the first derivatives of the conditional expectation function with respect to the $\alpha' s$ or $\beta' s$. For instance

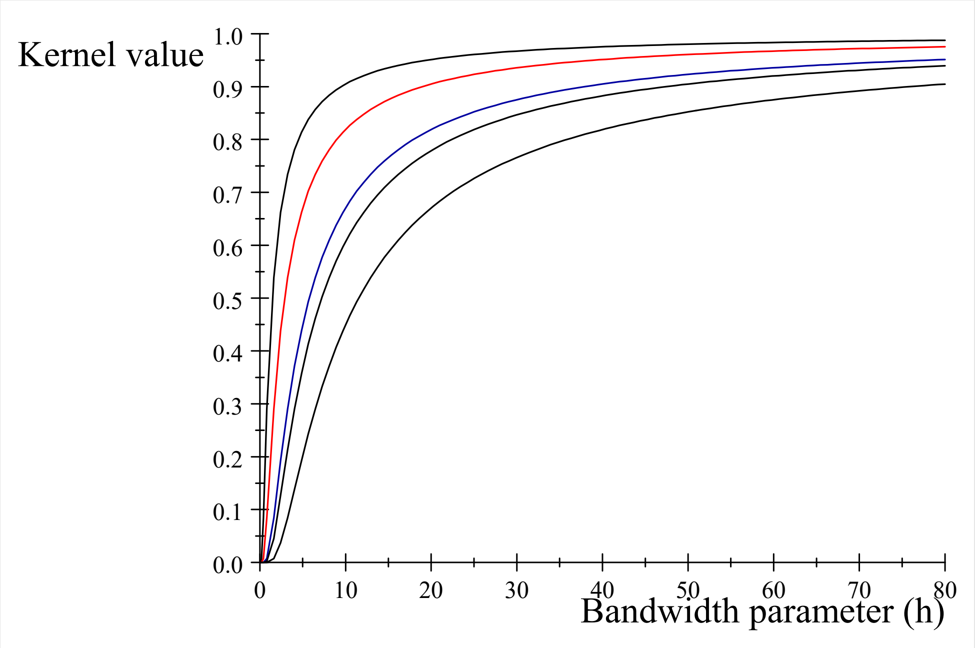
$$\frac{\partial E\left(.\right)}{\partial\alpha_j}=1+\alpha_i\sqrt{\beta_i\beta_j}; \quad \frac{\partial E\left(.\right)}{\partial\beta_j}=\beta_i+\frac{1}{2}\alpha_i\alpha_j\sqrt{\frac{\beta_i}{\beta j}}.$$

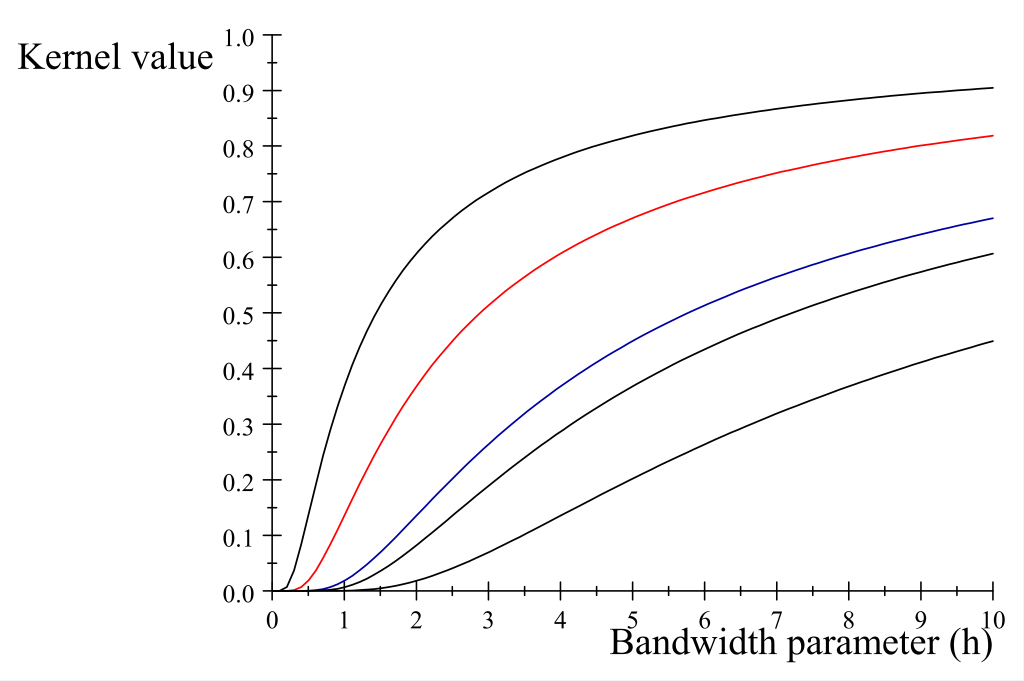
Arbitrary Gaussian kernel adopted for the RKHS regression using as covariate a 2×1 vector: number of alleles at each of the two loci, e.g., $x_{AA} = 2, x_{Aa} = 1$ and $x_{aa} = 0$. For example, the kernel entry $AABB$ and $AAbb$ is

$$k(\mathbf{x}_{AABB}, \mathbf{x}_{AAbb}, h) = \exp\left[-\frac{(2-2)^2 + (2-0)^2}{h}\right] = \exp\left[-\frac{4}{h}\right],$$

$$\mathbf{K}_h = \left[\begin{array}{cccccccccc}
 & AABB & AABb & AAbb & AaBB & AaBb & Aabb & aaBB & aaBb & aabb \\
AABB & 1 & e^{-\frac{1}{h}} & e^{-\frac{4}{h}} & e^{-\frac{1}{h}} & e^{-\frac{2}{h}} & e^{-\frac{5}{h}} & e^{-\frac{4}{h}} & e^{-\frac{5}{h}} & e^{-\frac{8}{h}} \\
AABb & e^{-\frac{1}{h}} & 1 & e^{-\frac{1}{h}} & e^{-\frac{2}{h}} & e^{-\frac{1}{h}} & e^{-\frac{2}{h}} & e^{-\frac{5}{h}} & e^{-\frac{4}{h}} & e^{-\frac{5}{h}} \\
AAbb & e^{-\frac{4}{h}} & e^{-\frac{1}{h}} & 1 & e^{-\frac{5}{h}} & e^{-\frac{2}{h}} & e^{-\frac{1}{h}} & e^{-\frac{8}{h}} & e^{-\frac{5}{h}} & e^{-\frac{4}{h}} \\
AaBB & e^{-\frac{1}{h}} & e^{-\frac{2}{h}} & e^{-\frac{5}{h}} & 1 & e^{-\frac{1}{h}} & e^{-\frac{4}{h}} & e^{-\frac{1}{h}} & e^{-\frac{2}{h}} & e^{-\frac{5}{h}} \\
AaBb & e^{-\frac{2}{h}} & e^{-\frac{1}{h}} & e^{-\frac{2}{h}} & e^{-\frac{1}{h}} & 1 & e^{-\frac{1}{h}} & e^{-\frac{2}{h}} & e^{-\frac{1}{h}} & e^{-\frac{2}{h}} \\
Aabb & e^{-\frac{5}{h}} & e^{-\frac{2}{h}} & e^{-\frac{1}{h}} & e^{-\frac{4}{h}} & e^{-\frac{1}{h}} & 1 & e^{-\frac{5}{h}} & e^{-\frac{2}{h}} & e^{-\frac{1}{h}} \\
aaBB & e^{-\frac{4}{h}} & e^{-\frac{5}{h}} & e^{-\frac{8}{h}} & e^{-\frac{1}{h}} & e^{-\frac{2}{h}} & e^{-\frac{5}{h}} & 1 & e^{-\frac{1}{h}} & e^{-\frac{4}{h}} \\
aaBb & e^{-\frac{5}{h}} & e^{-\frac{4}{h}} & e^{-\frac{5}{h}} & e^{-\frac{2}{h}} & e^{-\frac{1}{h}} & e^{-\frac{2}{h}} & e^{-\frac{1}{h}} & 1 & e^{-\frac{1}{h}} \\
aabb & e^{-\frac{8}{h}} & e^{-\frac{5}{h}} & e^{-\frac{4}{h}} & e^{-\frac{5}{h}} & e^{-\frac{2}{h}} & e^{-\frac{1}{h}} & e^{-\frac{4}{h}} & e^{-\frac{1}{h}} & 1
\end{array}\right]$$

Kernel value $k(.,.;h) = \exp\left(-\frac{S}{h}\right)$ against bandwidth parameter h. Curves, from upper to lower, correspond to $S = 1, 2, 4, 5, 8$





$h = 1.75$ as bandwidth parameter

6 unique entries in the $\mathbf{K}$ matrix:

1.0 (diagonal elements, the two individuals have identical genotypes)

0.565 (3 alleles in common in a pair of individuals)

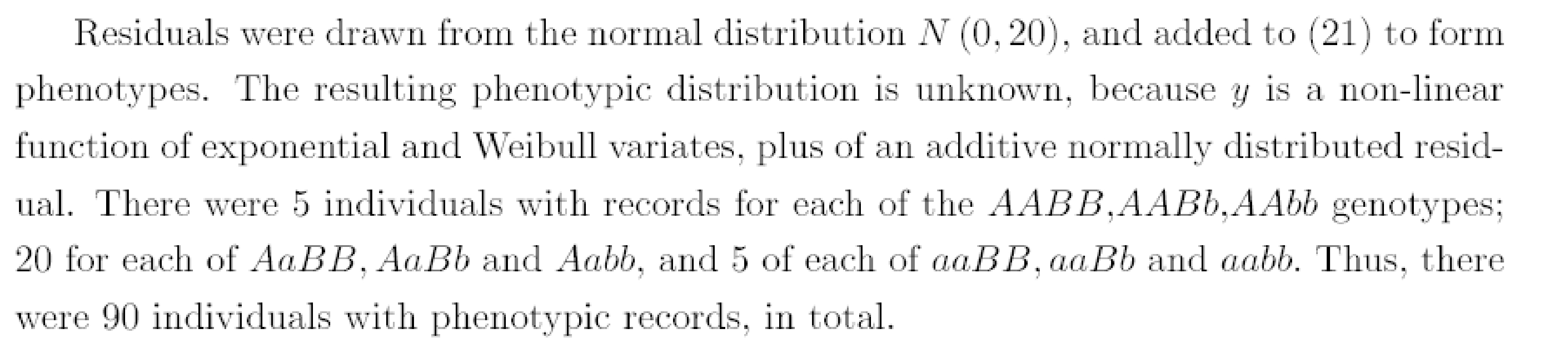
0.319 (2 alleles in common, 1 per locus)

0.102 (2 alleles in common at only one locus)

0.06 (1 allele in common)

0.01 (no alleles shared).

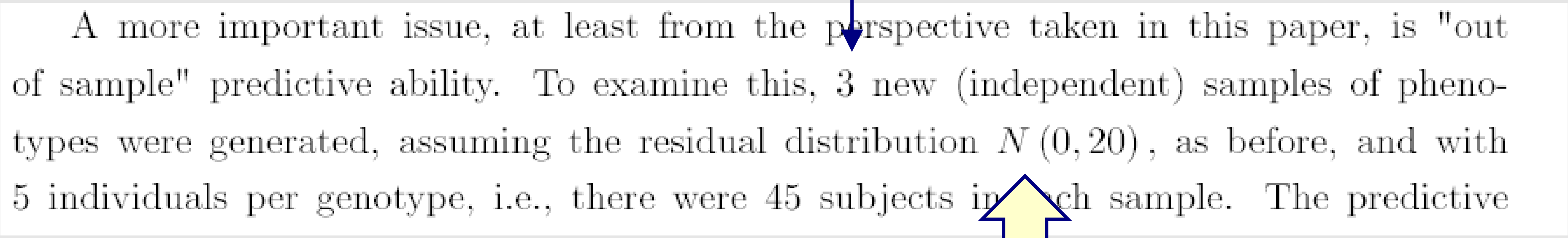
Training set



Residuals were drawn from the normal distribution $N(0,20)$, and added to (21) to form phenotypes. The resulting phenotypic distribution is unknown, because y is a non-linear function of exponential and Weibull variates, plus of an additive normally distributed residual. There were 5 individuals with records for each of the $AABB$,$AABb$,$AAbb$ genotypes; 20 for each of $AaBB$, $AaBb$ and $Aabb$, and 5 of each of $aaBB$, $aaBb$ and $aabb$. Thus, there were 90 individuals with phenotypic records, in total.

Testing set

100



A more important issue, at least from the perspective taken in this paper, is "out of sample" predictive ability. To examine this, 3 new (independent) samples of phenotypes were generated, assuming the residual distribution $N(0,20)$, as before, and with 5 individuals per genotype, i.e., there were 45 subjects in each sample. The predictive

IMPORTANT ISSUE TO DISCUSS HERE

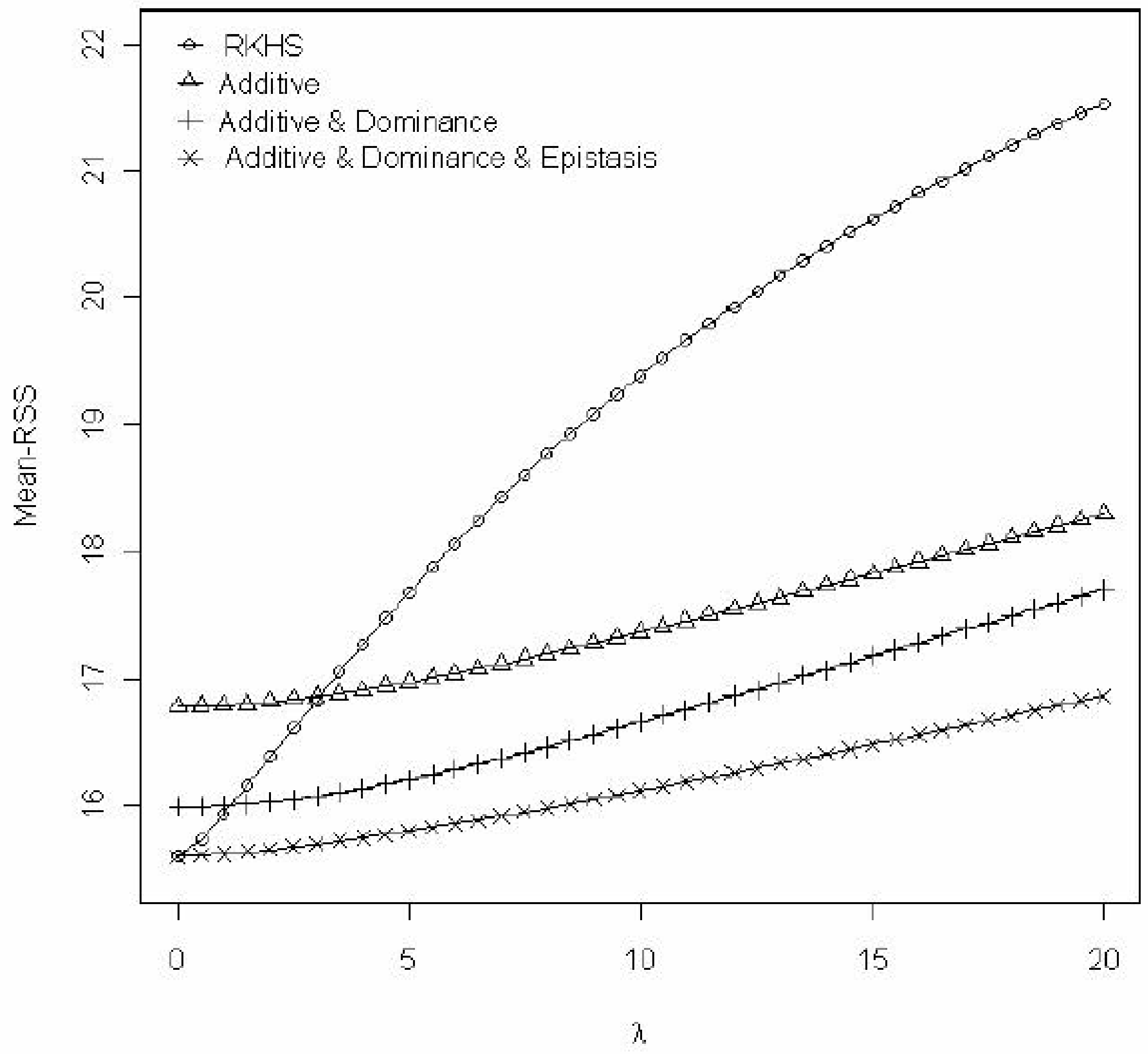


Figure 4. Average (over 90 data points) squared residual for four models fitted to the training sample (RKHS= reproducing kernel Hilbert spaces regression with Gaussian kernel and bandwidth= 1.75) for each value of the smoothing parameter λ.

TRAINING
SET

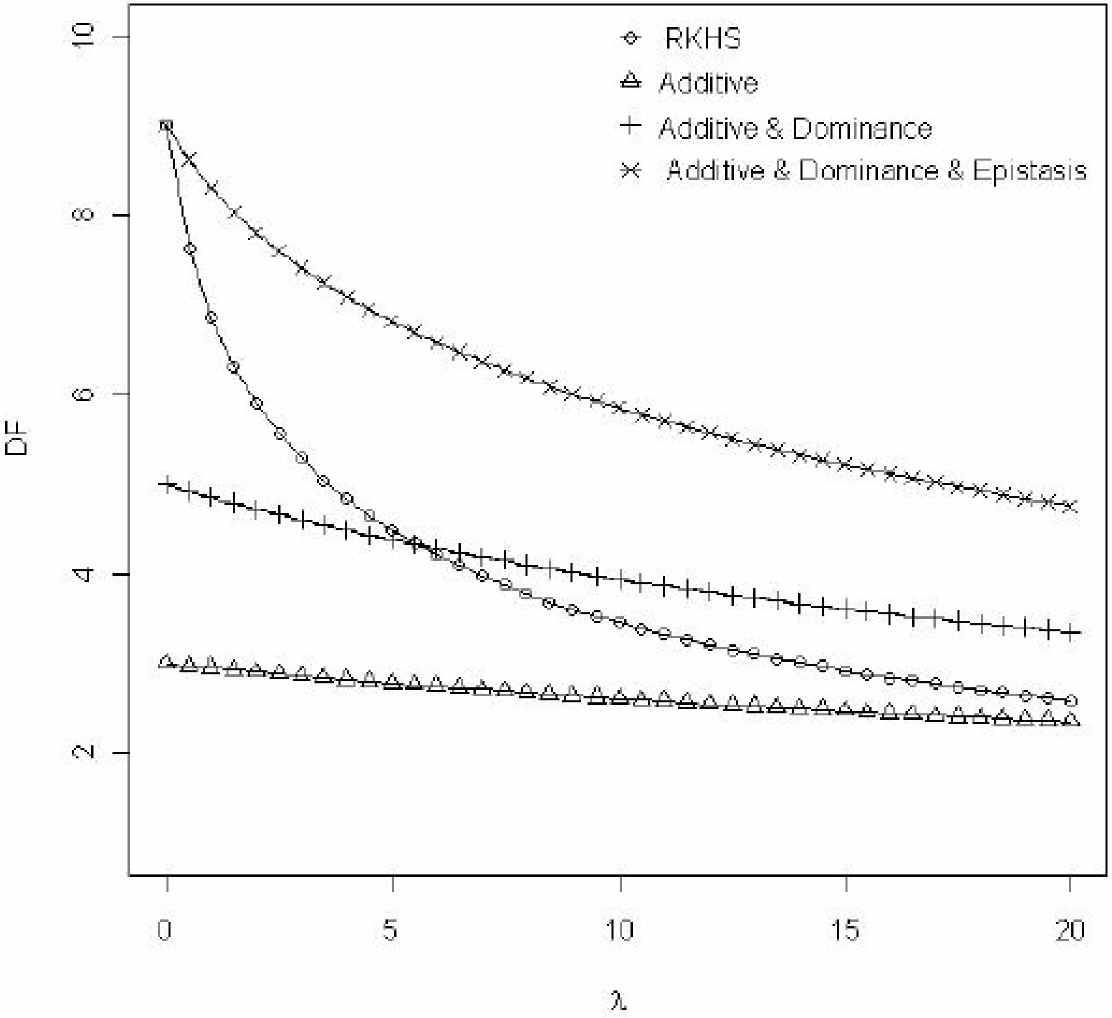


Figure 5. Effective degrees of freedom for four models fitted to the training sample (RKHS= reproducing kernel Hilbert spaces regression with Gaussian kernel and bandwidth= 1.75) at each value of the smoothing parameter λ.

df=tr(X$(X'X)^{-1}$X')
in regression

df=tr(X$(X'X + I\lambda)^{-1}$X')
in ridge regression

df=tr(K$(K'K + K\lambda)^{-1}$K')
=tr($(K + I\lambda)^{-1}$K')
in RKHS regression

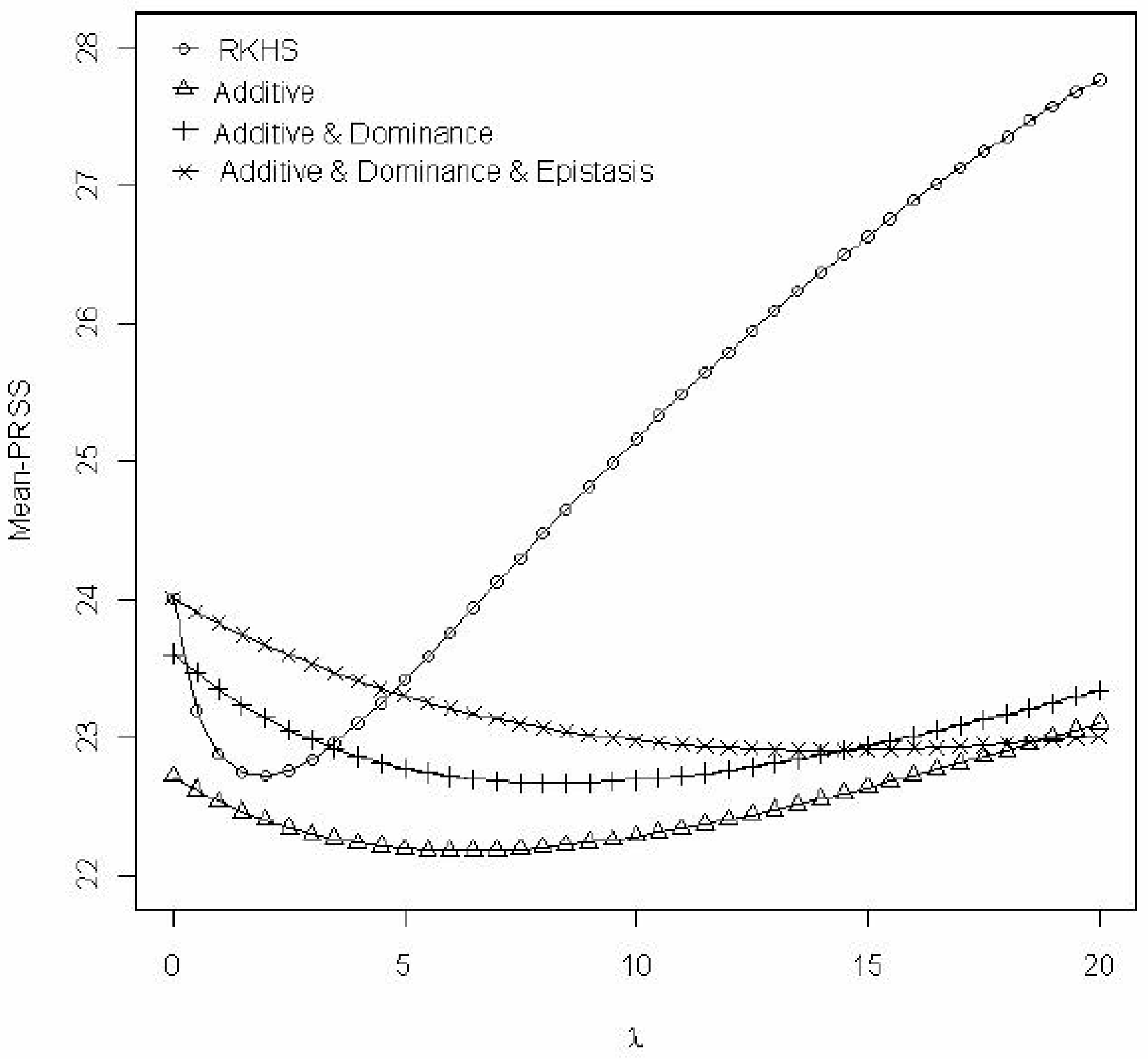
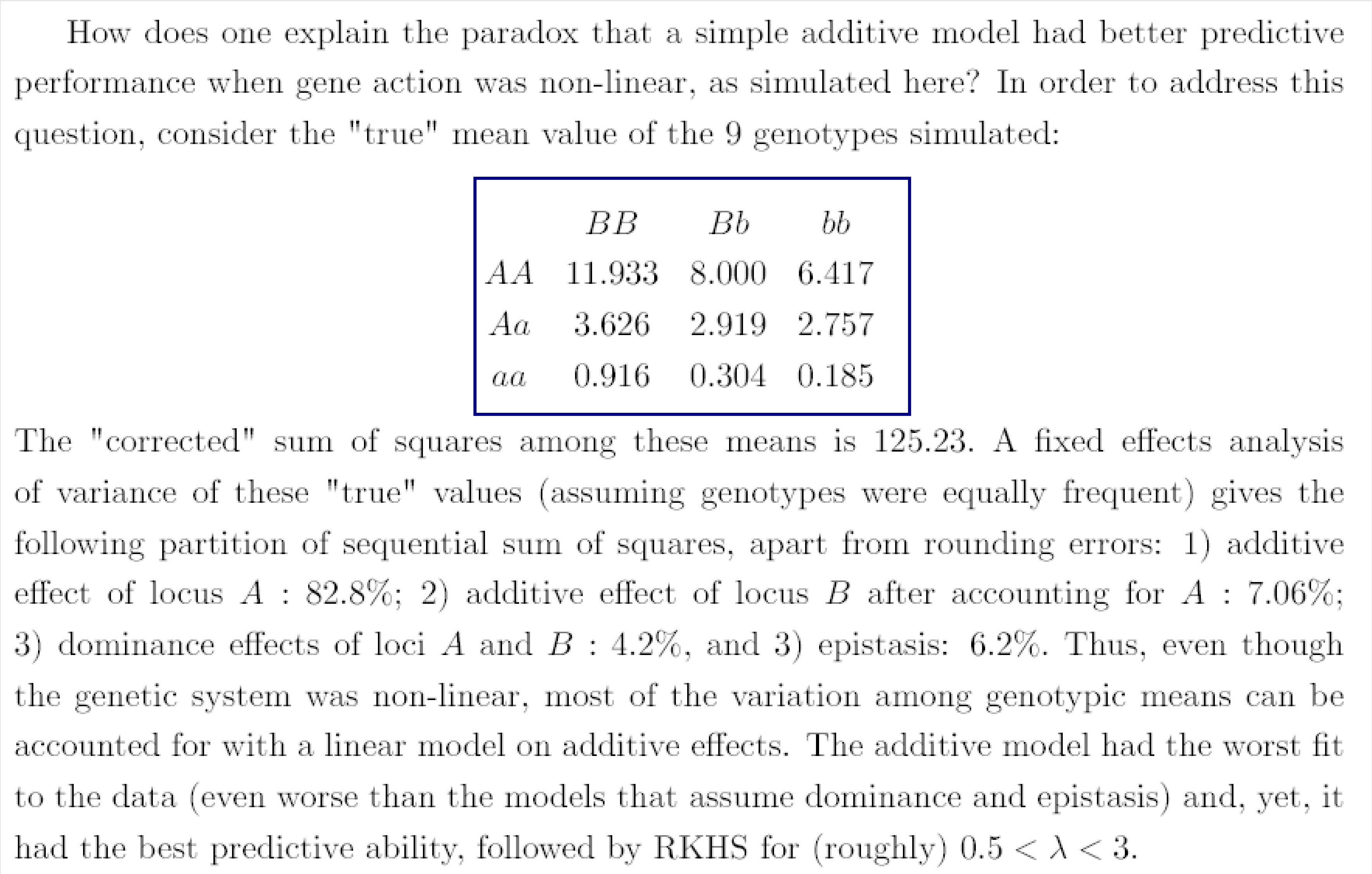


Figure 6. Average (over 100 samples with 45 realized observations in each) squared prediction error for four models fitted to the predictive sample (RKHS= reproducing kernel Hilbert spaces regression with Gaussian kernel and bandwidth= 1.75) for each value of the smoothing parameter λ.

TESTING
SET

Explanation of results



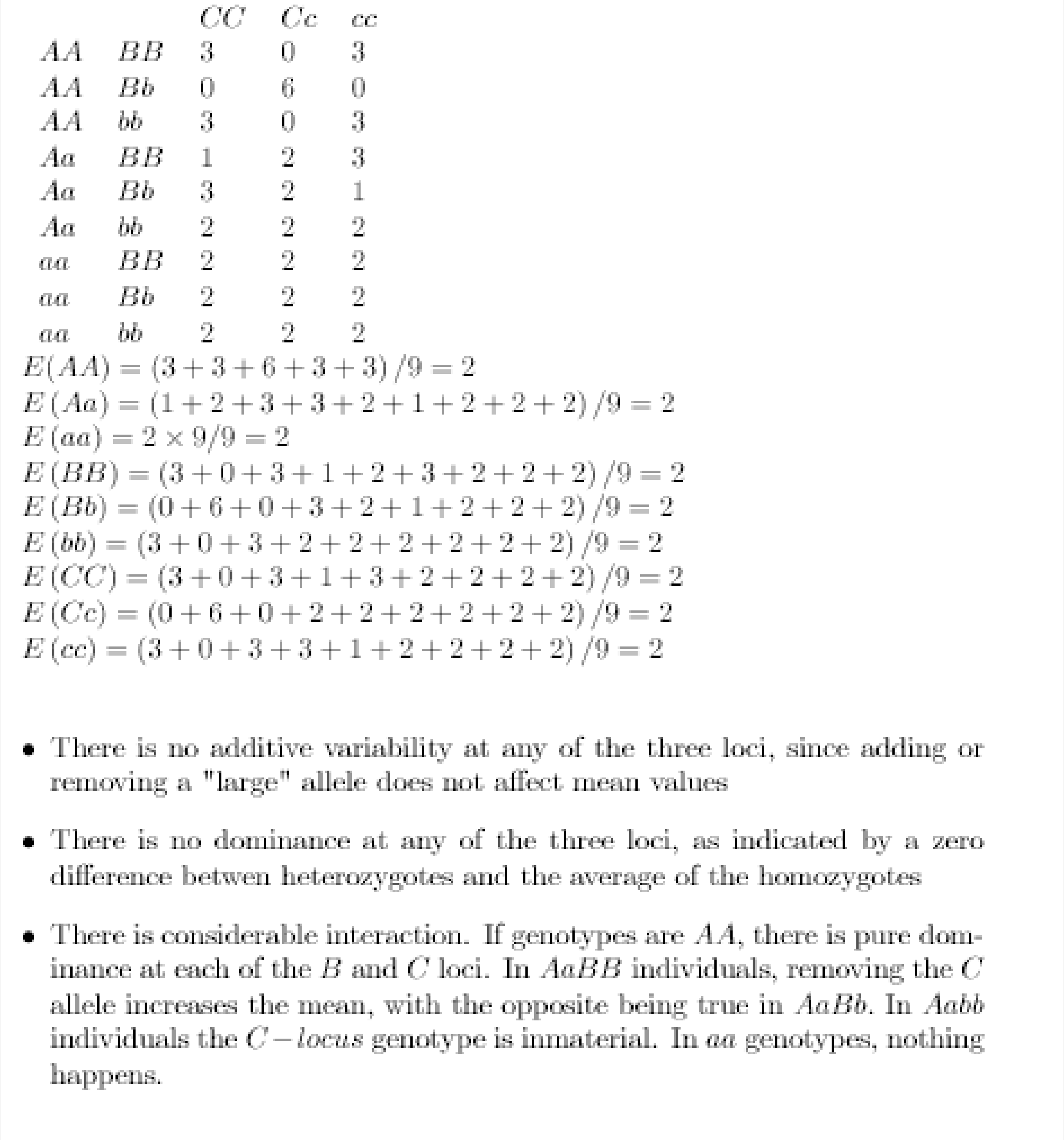
How does one explain the paradox that a simple additive model had better predictive performance when gene action was non-linear, as simulated here? In order to address this question, consider the "true" mean value of the 9 genotypes simulated:

	BB	Bb	bb
AA	11.933	8.000	6.417
Aa	3.626	2.919	2.757
aa	0.916	0.304	0.185

The "corrected" sum of squares among these means is 125.23. A fixed effects analysis of variance of these "true" values (assuming genotypes were equally frequent) gives the following partition of sequential sum of squares, apart from rounding errors: 1) additive effect of locus A : 82.8%; 2) additive effect of locus B after accounting for A : 7.06%; 3) dominance effects of loci A and B : 4.2%, and 3) epistasis: 6.2%. Thus, even though the genetic system was non-linear, most of the variation among genotypic means can be accounted for with a linear model on additive effects. The additive model had the worst fit to the data (even worse than the models that assume dominance and epistasis) and, yet, it had the best predictive ability, followed by RKHS for (roughly) $0.5 < \lambda < 3$.

!!

Example 3 of RKHS



		CC	Cc	cc
AA	BB	3	0	3
AA	Bb	0	6	0
AA	bb	3	0	3
Aa	BB	1	2	3
Aa	Bb	3	2	1
Aa	bb	2	2	2
aa	BB	2	2	2
aa	Bb	2	2	2
aa	bb	2	2	2

$E(AA) = (3+3+6+3+3)/9 = 2$

$E(Aa) = (1+2+3+3+2+1+2+2+2)/9 = 2$

$E(aa) = 2 \times 9/9 = 2$

$E(BB) = (3+0+3+1+2+3+2+2+2)/9 = 2$

$E(Bb) = (0+6+0+3+2+1+2+2+2)/9 = 2$

$E(bb) = (3+0+3+2+2+2+2+2+2)/9 = 2$

$E(CC) = (3+0+3+1+3+2+2+2+2)/9 = 2$

$E(Cc) = (0+6+0+2+2+2+2+2+2)/9 = 2$

$E(cc) = (3+0+3+3+1+2+2+2+2)/9 = 2$

- There is no additive variability at any of the three loci, since adding or removing a "large" allele does not affect mean values
- There is no dominance at any of the three loci, as indicated by a zero difference betwen heterozygotes and the average of the homozygotes
- There is considerable interaction. If genotypes are AA, there is pure dominance at each of the B and C loci. In $AaBB$ individuals, removing the C allele increases the mean, with the opposite being true in $AaBb$. In $Aabb$ individuals the $C-locus$ genotype is inmaterial. In aa genotypes, nothing happens.

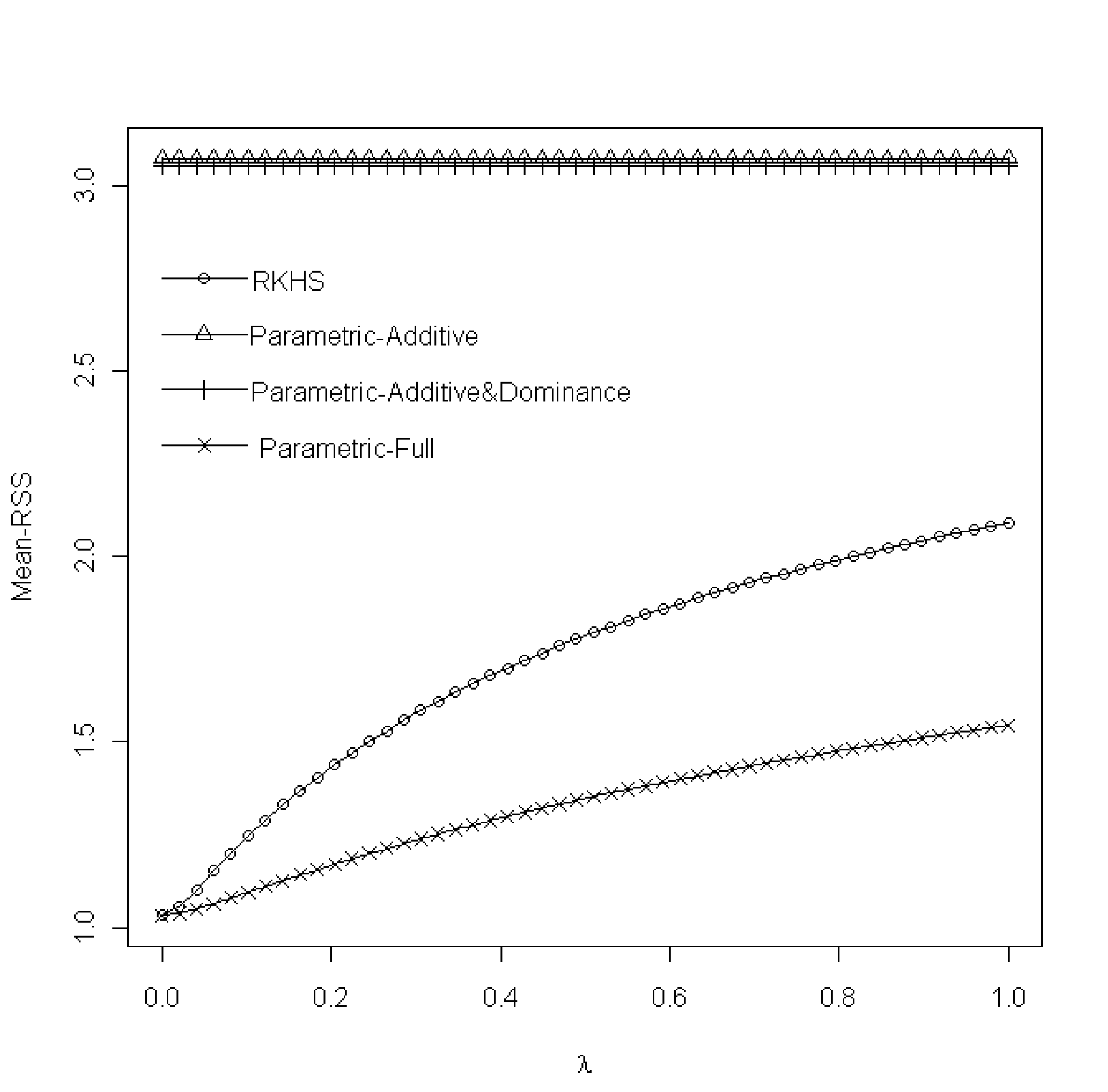
Source	**DF**	**Anova SS**	**Mean Square**	F Value	Pr > F
a	2	0.00000000	0.00000000	0.00	1.0000
b	2	0.00000000	0.00000000	0.00	1.0000
c	2	0.00000000	0.00000000	0.00	1.0000
a*b	4	0.00000000	0.00000000	0.00	1.0000
a*c	4	0.00000000	0.00000000	0.00	1.0000
b*c	4	13.33333333	3.33333333	1.00	0.4609
Error (a*b*c)	8	26.66666667	3.33333333		

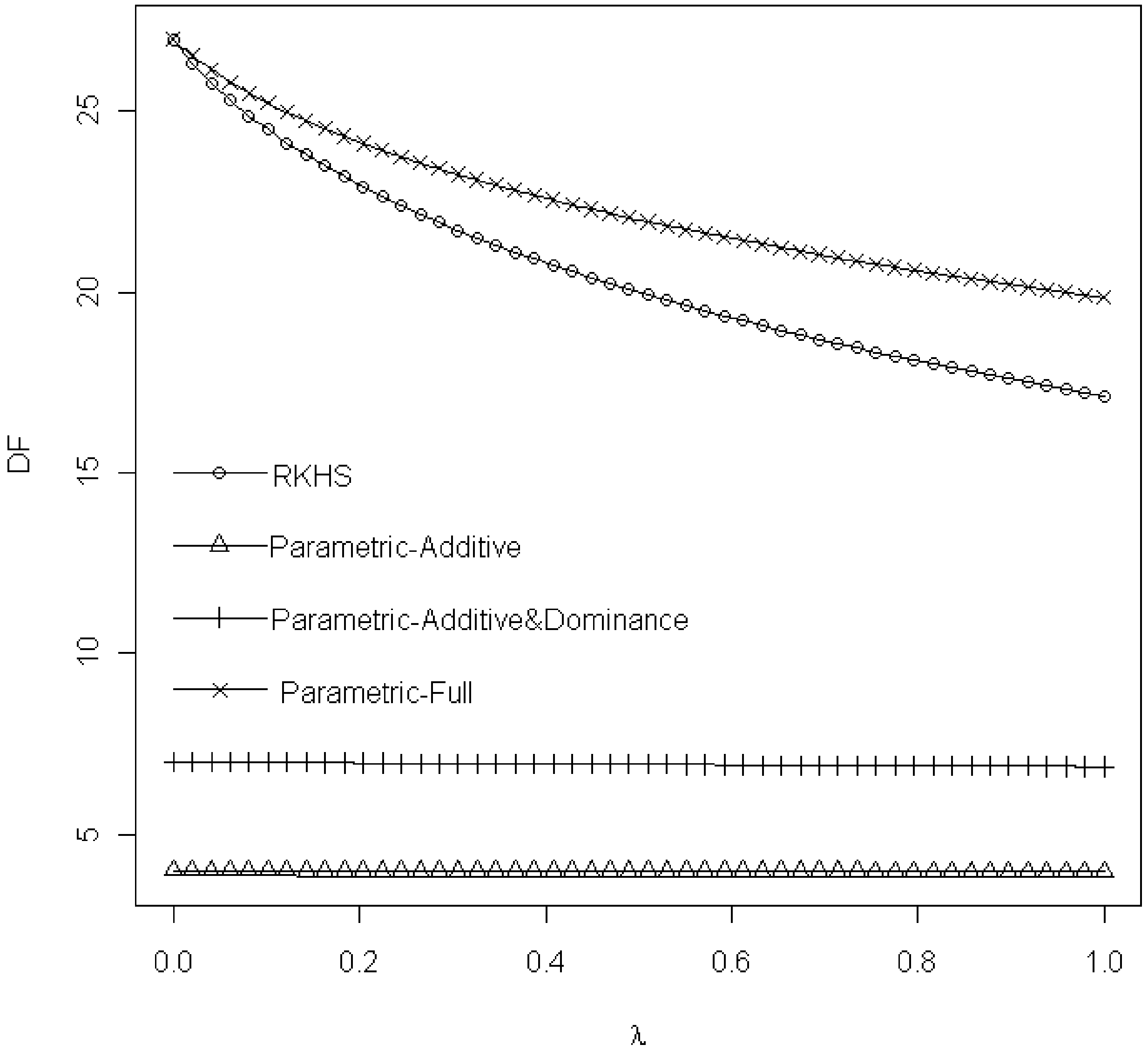
Variation between genotypic values is pure interaction!!!

Training set:
- 27 genotypes,
- 5 replicates per genotype,
- residual variance 1.5

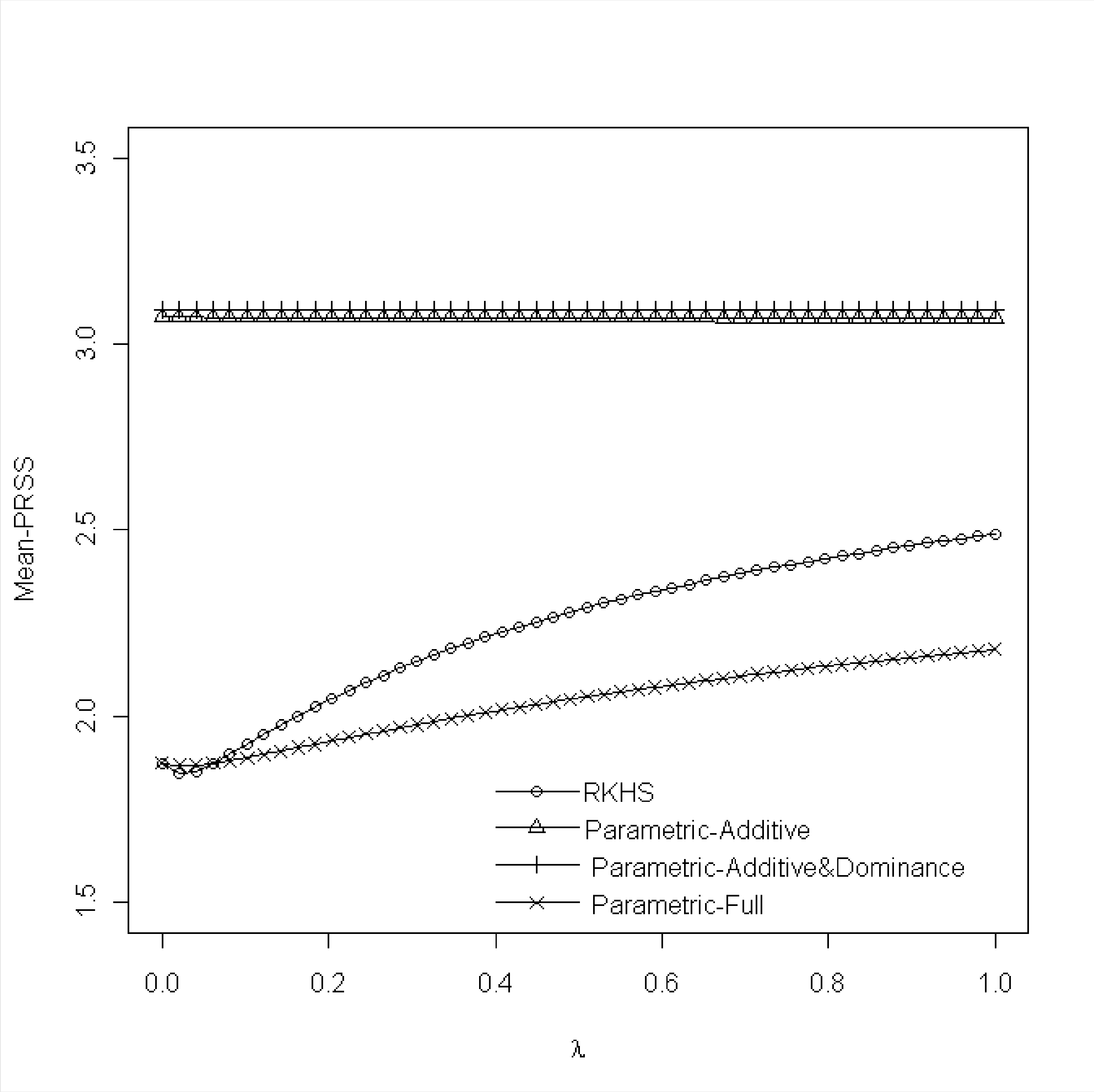
Testing set: 50 MC replicates, each as the training set.

Results in training set

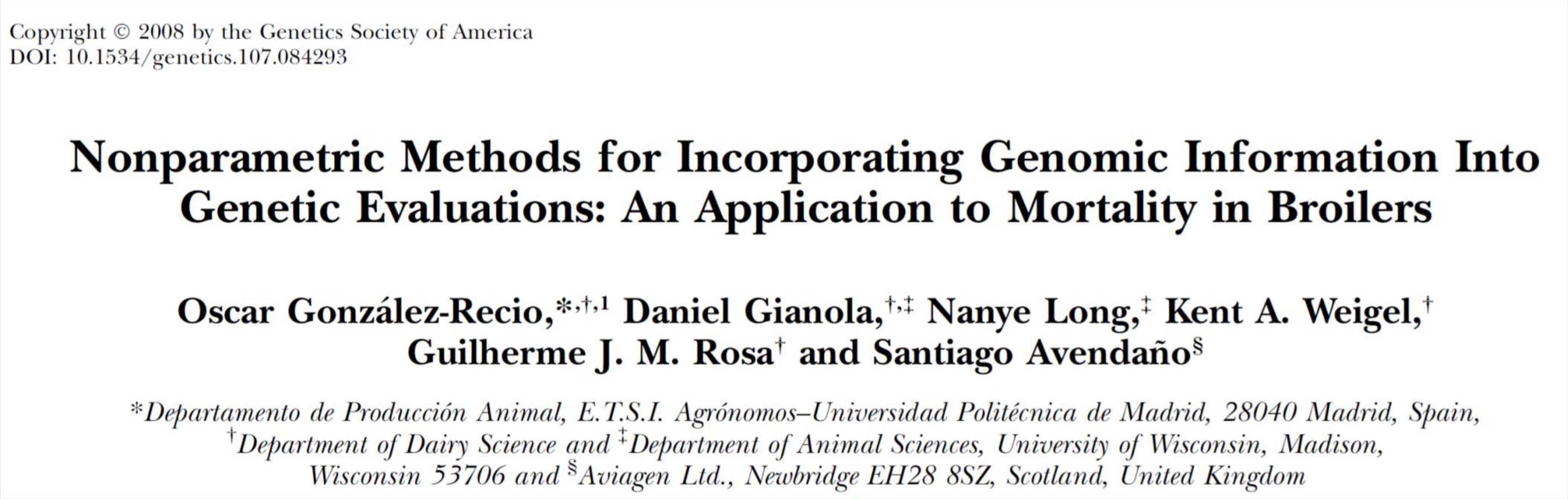



DF
25
20
15
10
5
0.0
0.2
0.4
0.6
0.8
1.0
λ
RKHS
Parametric-Additive
Parametric-Additive&Dominance
Parametric-Full

Results testing set.



FIRST APPLICATION OF RKHS IN ANIMAL BREEDING



DOI: 10.1534/genetics.107.084293

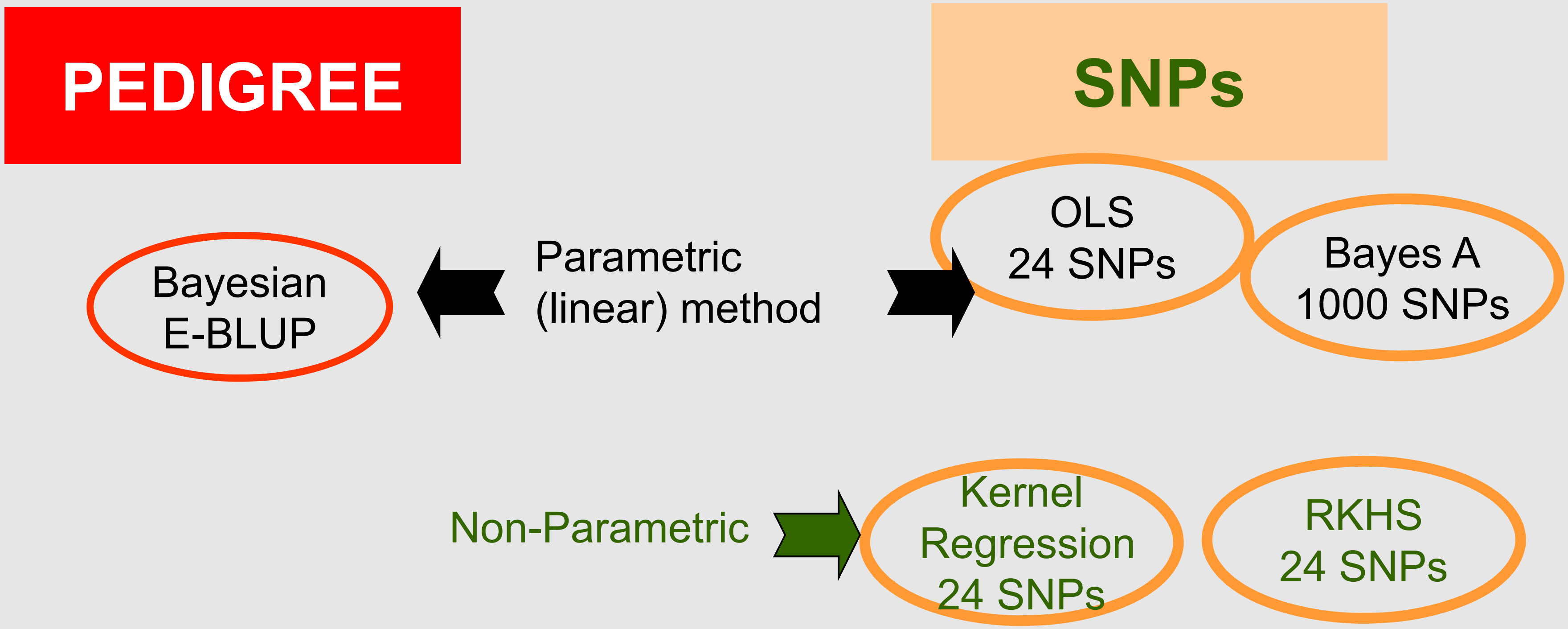
Nonparametric Methods for Incorporating Genomic Information Into Genetic Evaluations: An Application to Mortality in Broilers

**Oscar González-Recio,[*,†,1] Daniel Gianola,[†,‡] Nanye Long,[‡] Kent A. Weigel,[†]
Guilherme J. M. Rosa[†] and Santiago Avendaño[§]**

*[*]Departamento de Producción Animal, E.T.S.I. Agrónomos–Universidad Politécnica de Madrid, 28040 Madrid, Spain,
[†]Department of Dairy Science and [‡]Department of Animal Sciences, University of Wisconsin, Madison,
Wisconsin 53706 and [§]Aviagen Ltd., Newbridge EH28 8SZ, Scotland, United Kingdom*

- Average progeny “late mortality” (lm) in low hygiene environment for 200 sires of line29 (12,167 progenies).
  - Pre-corrected for hatch, age of dam and dam,
  - Standardized log-transformed means
- SNPs: filter and wrapper strategy (Long et al., 2007)
  - 24 SNPs selected out of over 5000 genotyped on sires

MODELS



PEDIGREE

SNPs

Bayesian
E-BLUP

Parametric
(linear) method

OLS
24 SNPs

Bayes A
1000 SNPs

Non-Parametric

Kernel
Regression
24 SNPs

RKHS
24 SNPs

Sequence alignment KERNEL

Dynamic programming algorithms

Similarity between two DNA sequences

Adapted to SNP sequences

$$K_h(\mathbf{x}-\mathbf{x}_i) = \exp\left[-\,Score\,(\mathbf{x}-\mathbf{x}_i)\right]$$

No need to tune *h*

(Delcher et al., 1999, 2002)

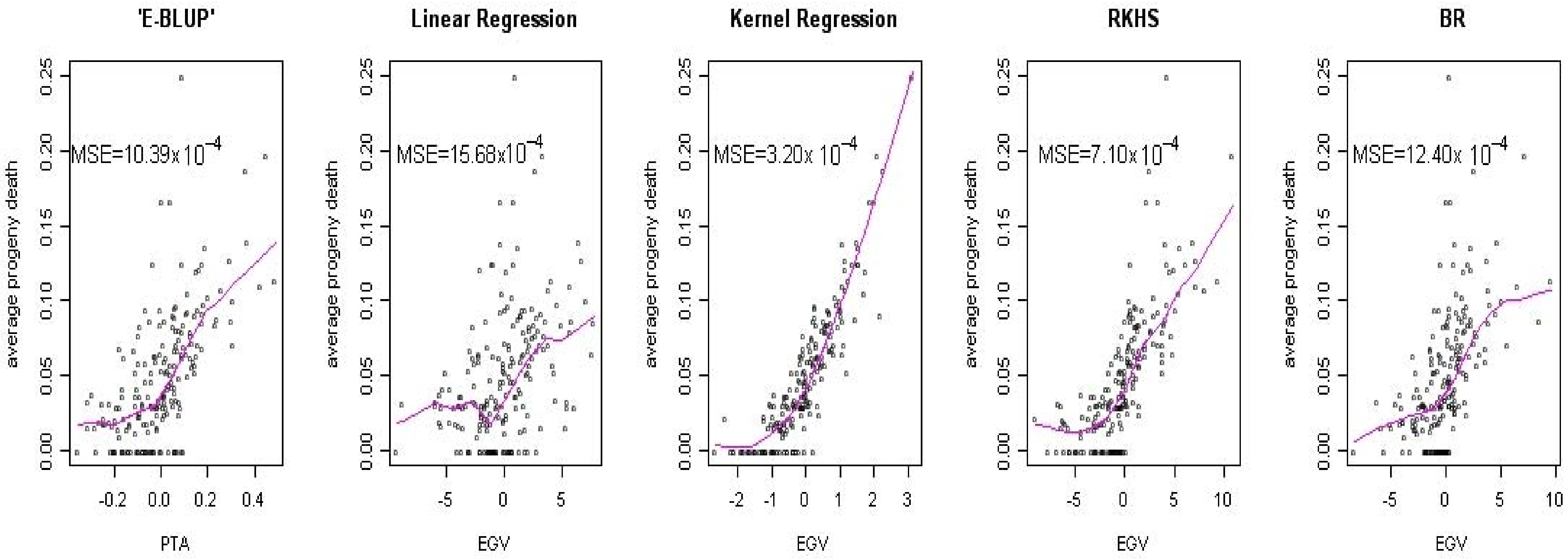
- Spearman (above diagonal) and Pearson correlations (below diagonal) between posterior means of sire effects

		E-BLUP		F-metric		Kernel		RKHS		BR
E-BLUP		…		0.52		0.77		0.84		0.91
F-metric		0.56				0.48		0.51		0.53
Kernel		0.66		0.38		…		0.93		0.76
RKHS		0.84		0.50		0.79		…		0.84
BR		0.92		0.57		0.58		0.80		…

- E-BLUP & Bayes A very similar.
- LR most different ranking.

MODEL FIT

- Regression of adjusted raw progeny LM on sire's PTA or EGV



MODEL FIT

- Less dispersion in non-parametric models
- Lower MSE for kernel regression
- Worst for Linear regression (F-metric model)

Still….which model predicts the data best ?

Predictive ability

- Cross validation

1. 5 subsets, letting 20% sire means missing each time at random

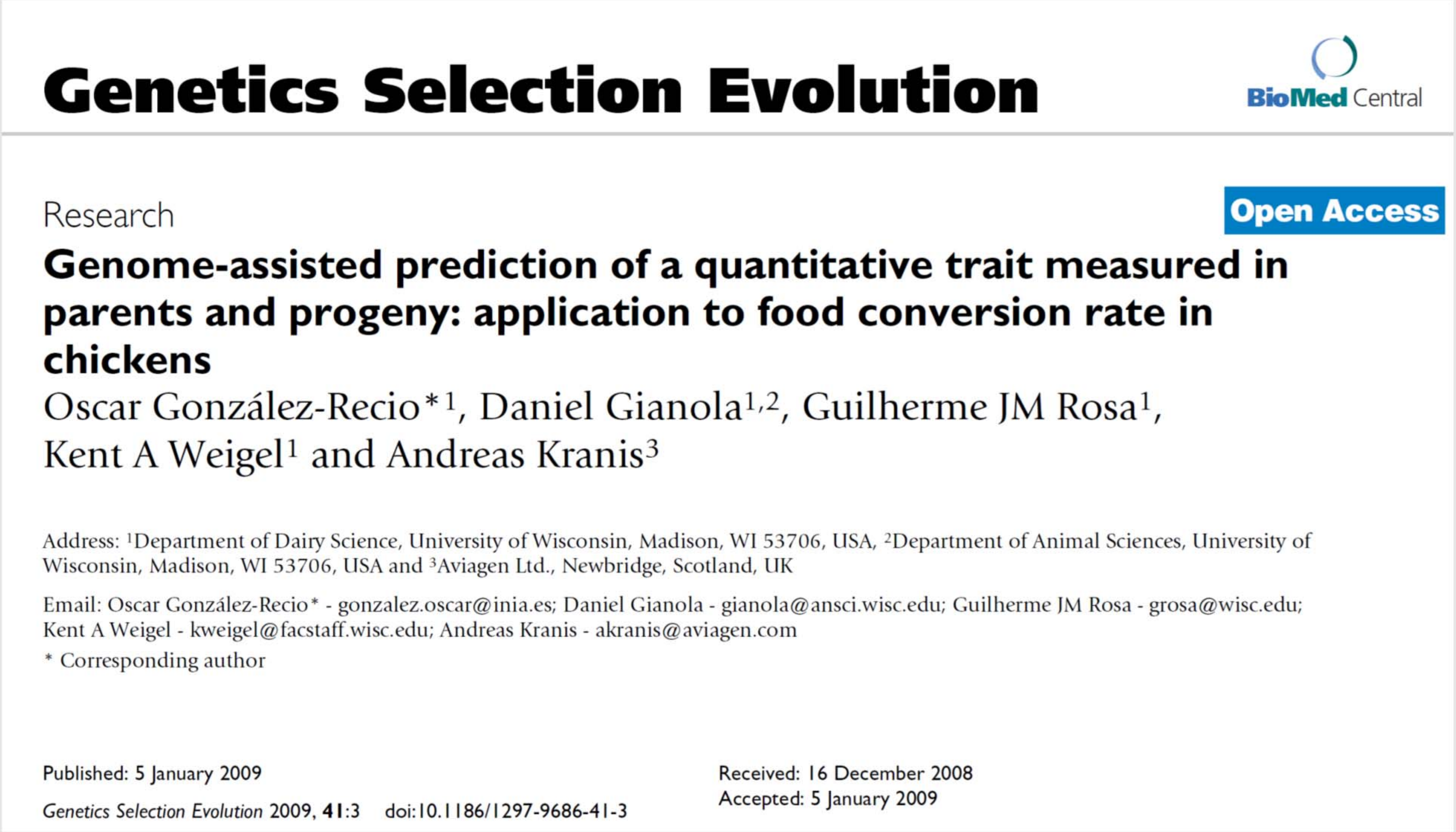
2. Calculate correlations between actual and inferred average progeny, for each method within subset.

Predictive ability

Subset		E-BLUP		F-metric		Kernel		RKHS		BR
1st		0.03		**0.27**		0.05		**0.27**		0.13
2nd		0.18		0.19		0.28		**0.37**		0.12
3rd		**0.18**		0.08		0.06		-0.01		0.17
4th		-0.04		0.07		0.13		**0.28**		0.15
5th		0.17		-0.12		0.23		0.15		**0.25**
GLOBAL		0.10		0.06		0.14		**0.20**		0.16

- •RKHS showed better predictive ability
  - –25% higher reliability than Xu’s method
  - –100% higher reliability than E-BLUP
  - –233% higher reliability than F-metric (linear regression on markers)
- •RKHS better than fixed or random regression on markers and E-BLUP.

APPLICATION TO FEED CONVERSION IN CHICKENS



Genetics Selection Evolution

BioMed Central

Research

Open Access

Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens

Oscar González-Recio*[1], Daniel Gianola[1,2], Guilherme JM Rosa[1], Kent A Weigel[1] and Andreas Kranis[3]

Address: [1]Department of Dairy Science, University of Wisconsin, Madison, WI 53706, USA, [2]Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA and [3]Aviagen Ltd., Newbridge, Scotland, UK

Email: Oscar González-Recio* - gonzalez.oscar@inia.es; Daniel Gianola - gianola@ansci.wisc.edu; Guilherme JM Rosa - grosa@wisc.edu; Kent A Weigel - kweigel@facstaff.wisc.edu; Andreas Kranis - akranis@aviagen.com

* Corresponding author

Published: 5 January 2009

Genetics Selection Evolution 2009, **41**:3 doi:10.1186/1297-9686-41-3

Received: 16 December 2008

Accepted: 5 January 2009

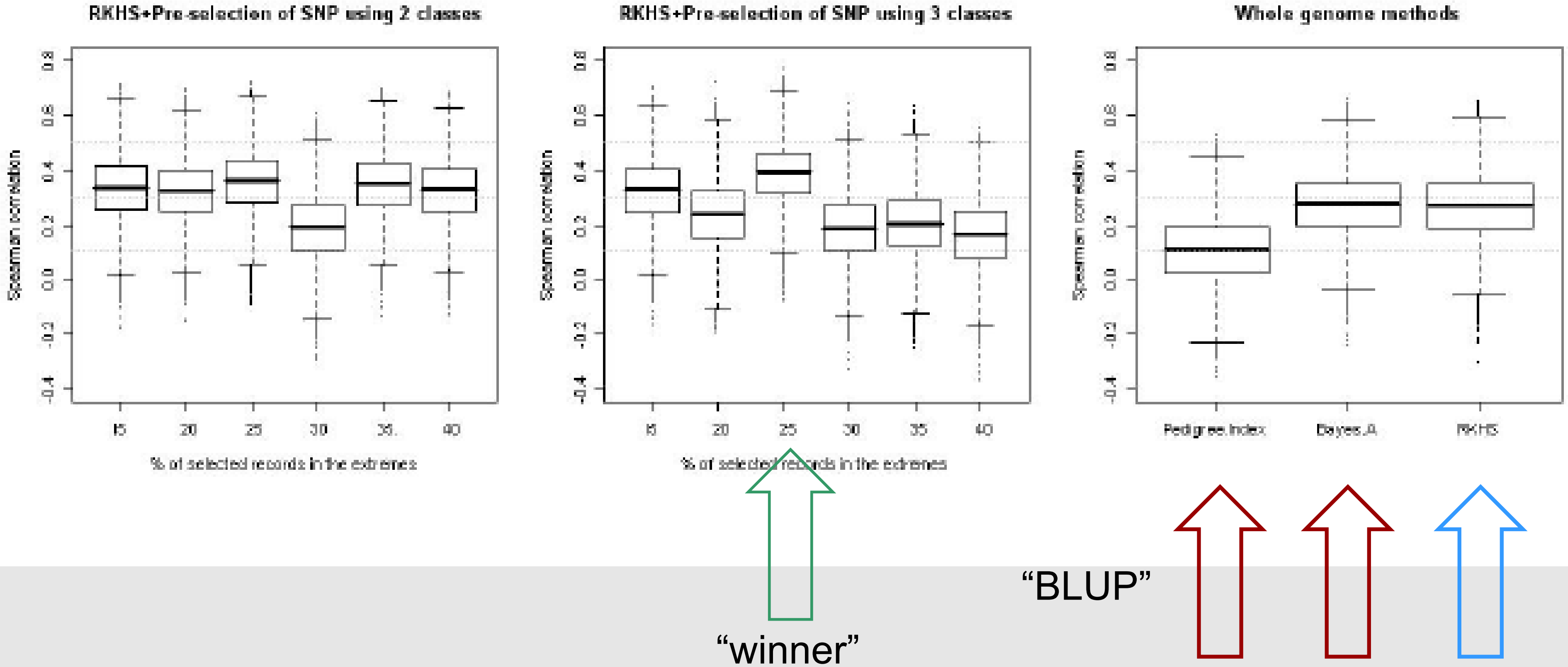
FCR measured on progeny of 333 sires with 3481 SNPs
FCR measured on progeny of 61 birds (sons of the above sires)

➔2- generation data set

BAYES A	--all markers
RKHS	--all markers
RKHS	--400 markers filtered using different INFOGAINS
BLUP (Bayes)	–pedigree information

Training set:	333 sires of sons
Predictive set:	61 sons of sires

Figure 2. Box plots for the bootstrap distribution of Spearman correlations between predicted and observed phenotype in the testing set (progeny) obtained with: RKHS on 400 pre-selected SNPs using 2 or 3 classes to classify sires with different percentiles (left and middle panels, respectively) and methods using pedigree or all available SNPs (right pannel).



"winner"

"BLUP"

"Bayes A"

RKHS

FIRST APPLICATION IN PLANTS

Predictive ability of models for genomic selection in Wheat [1]

	Predictive Correlation		Difference
Environment	BL	RKHS	(%)
E1	0.518	0.601	+16%
E2	0.493	0.494	0%
E3	0.403	0.445	+10%
E4	0.457	0.524	+15%

N= 599;

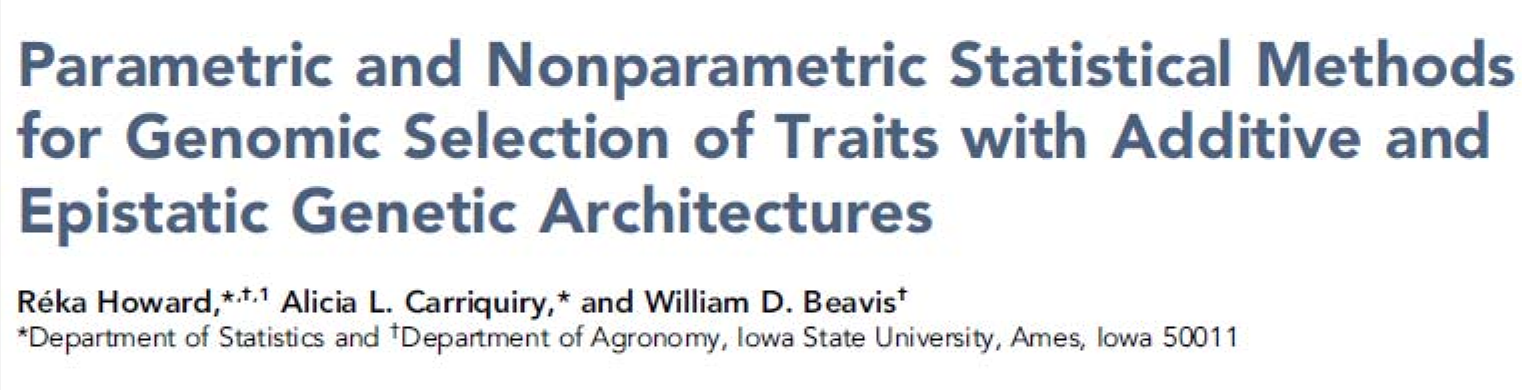
Trait: Grain Yield (4 environments);

Models: RKHS and Bayesian LASSO (BL)

[1] Crossa *et al.* (2010) Genetics.

SOME RECENT CASE STUDIES WITH RKHS

(some excitement from plant breeding)



Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures

Réka Howard,*,†,1 Alicia L. Carriquiry,* and William D. Beavis†
*Department of Statistics and †Department of Agronomy, Iowa State University, Ames, Iowa 50011

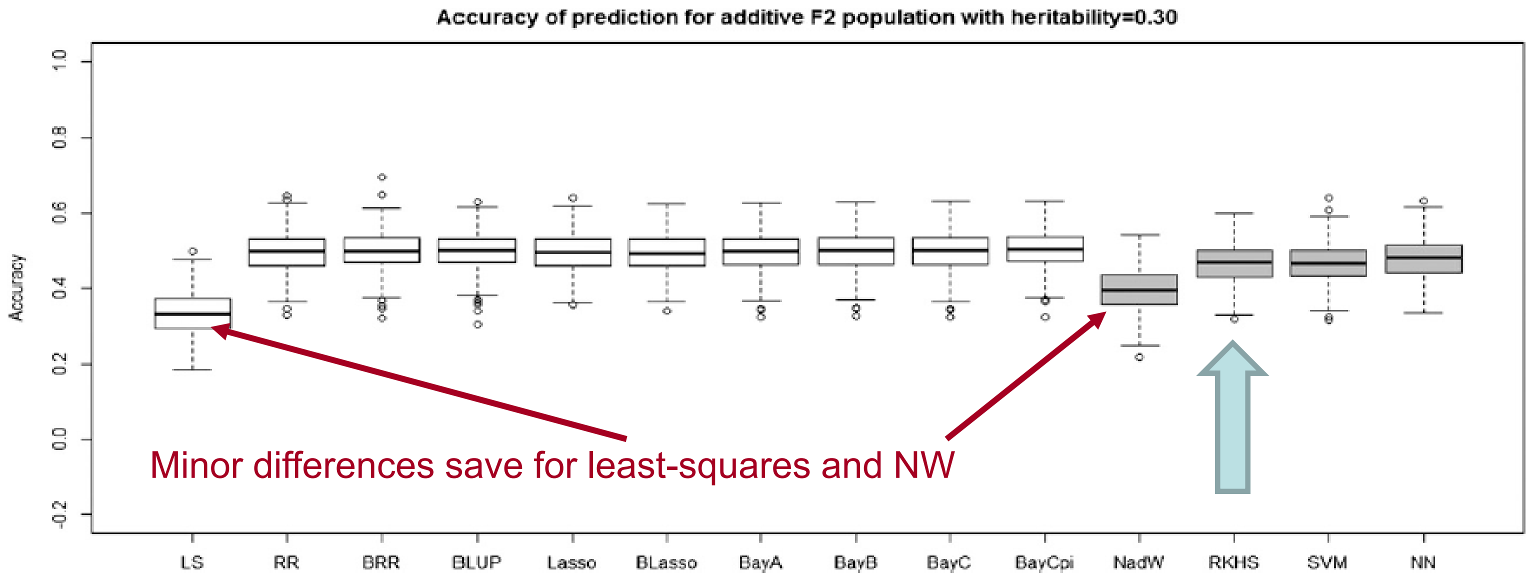
G3 2014

SIMULATED F2 POPULATIONS→ PLANT BREEDING

EXTREME ARCHITECTURES:

1. Completely additive : 10 ch-2 QTL/ch- 2000 markers

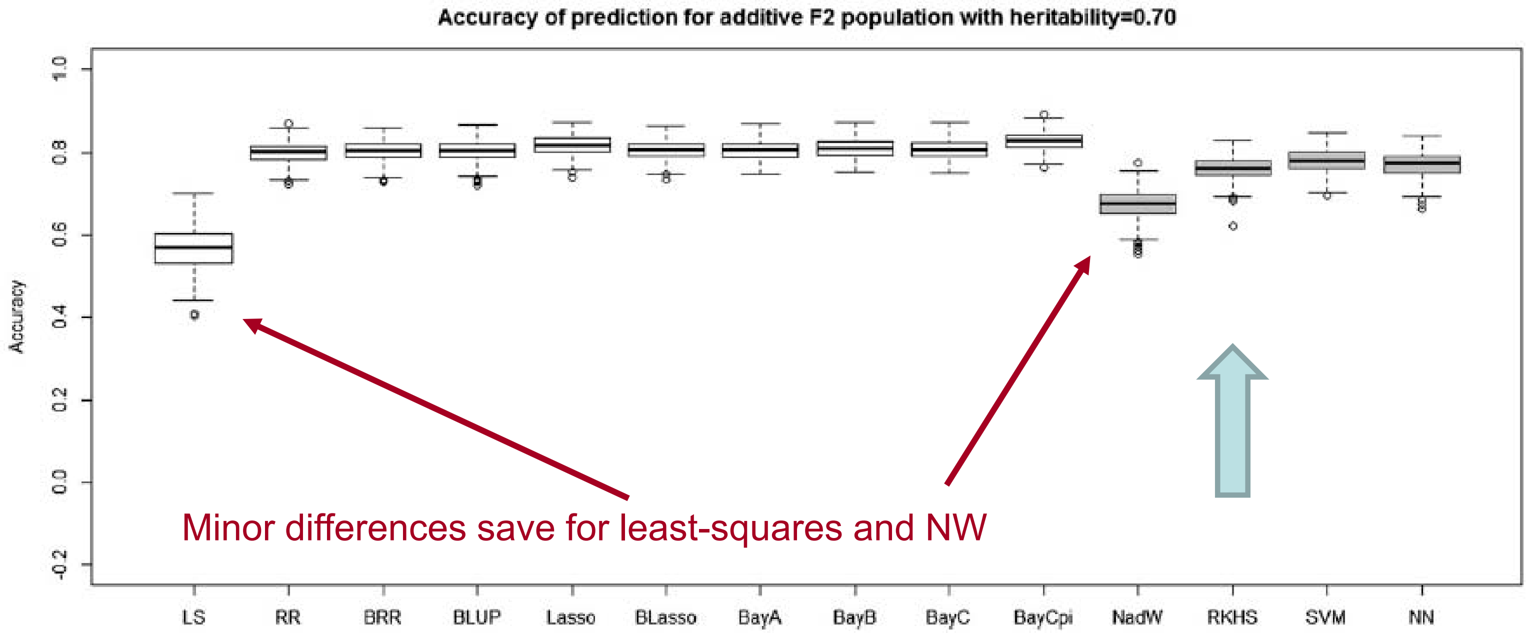
2. Completely epistatic : 10 ch-2 QTL/ch- 2000 markers
10 A X A epistatic interactions



Accuracy of prediction for additive F2 population with heritability=0.30

Minor differences save for least-squares and NW

Figure 7 The boxplots of accuracy of prediction for the F_2 population with additive genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last four (gray) boxplots correspond to the nonparametric methods.



Accuracy of prediction for additive F2 population with heritability=0.70

Minor differences save for least-squares and NW

Figure 5 The boxplots of accuracy of prediction for the F_2 population with additive genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last four (gray) boxplots correspond to the nonparametric methods.

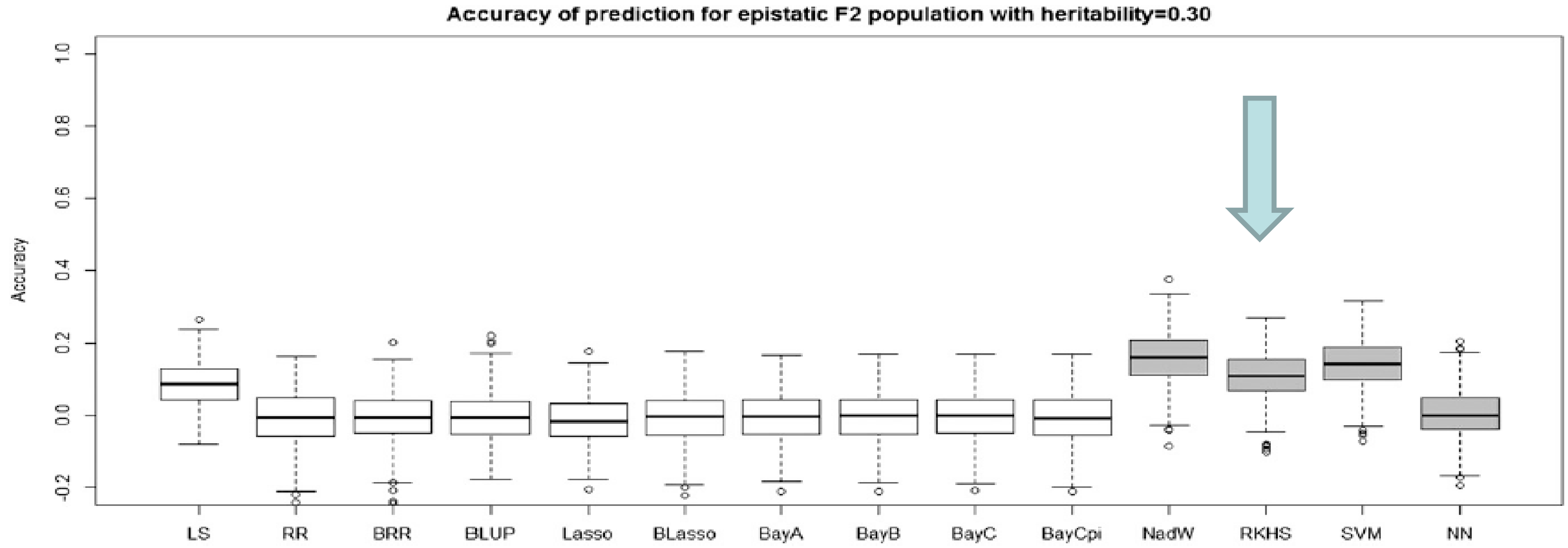


Figure 8 The boxplots of accuracy of prediction for the F_2 population with epistatic genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last four (gray) boxplots correspond to the nonparametric methods.

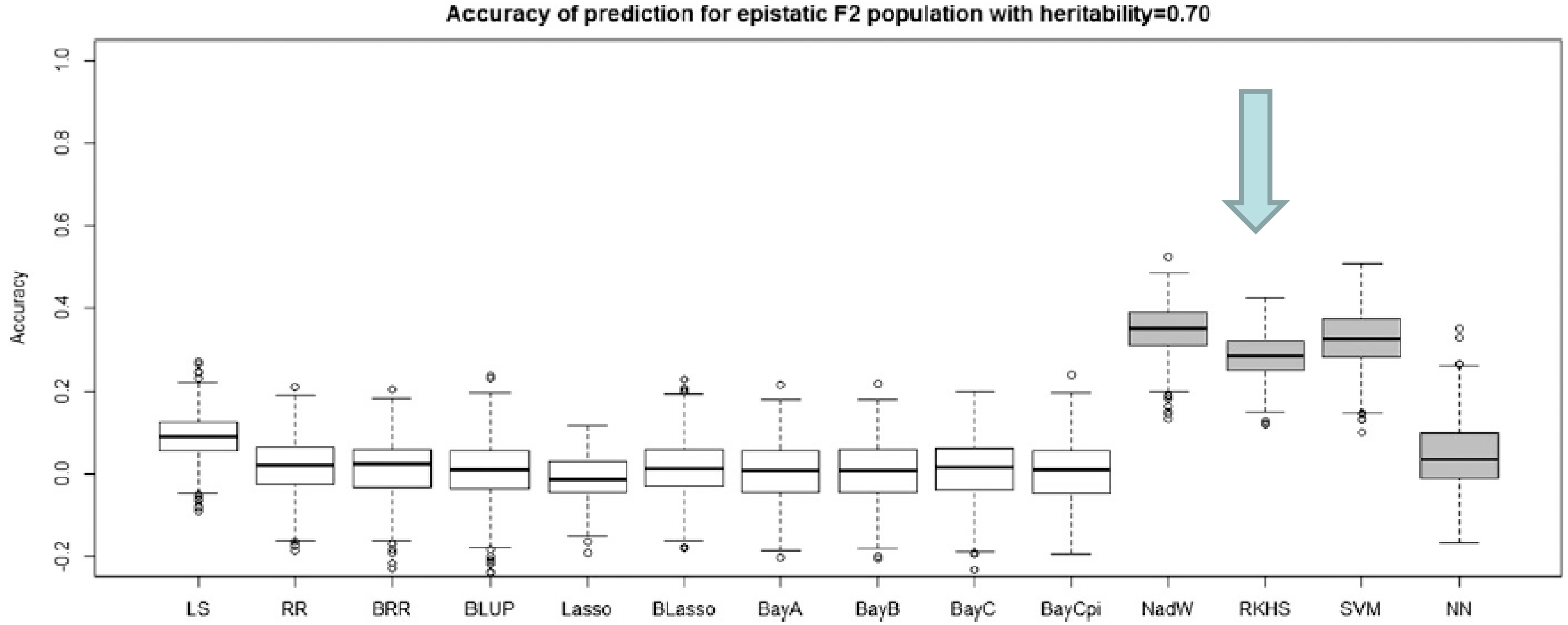


Figure 6 The boxplots of accuracy of prediction for the F_2 population with epistatic genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last four (gray) boxplots correspond to the nonparametric methods.

PEREZ et al. (2012, G3): wheat

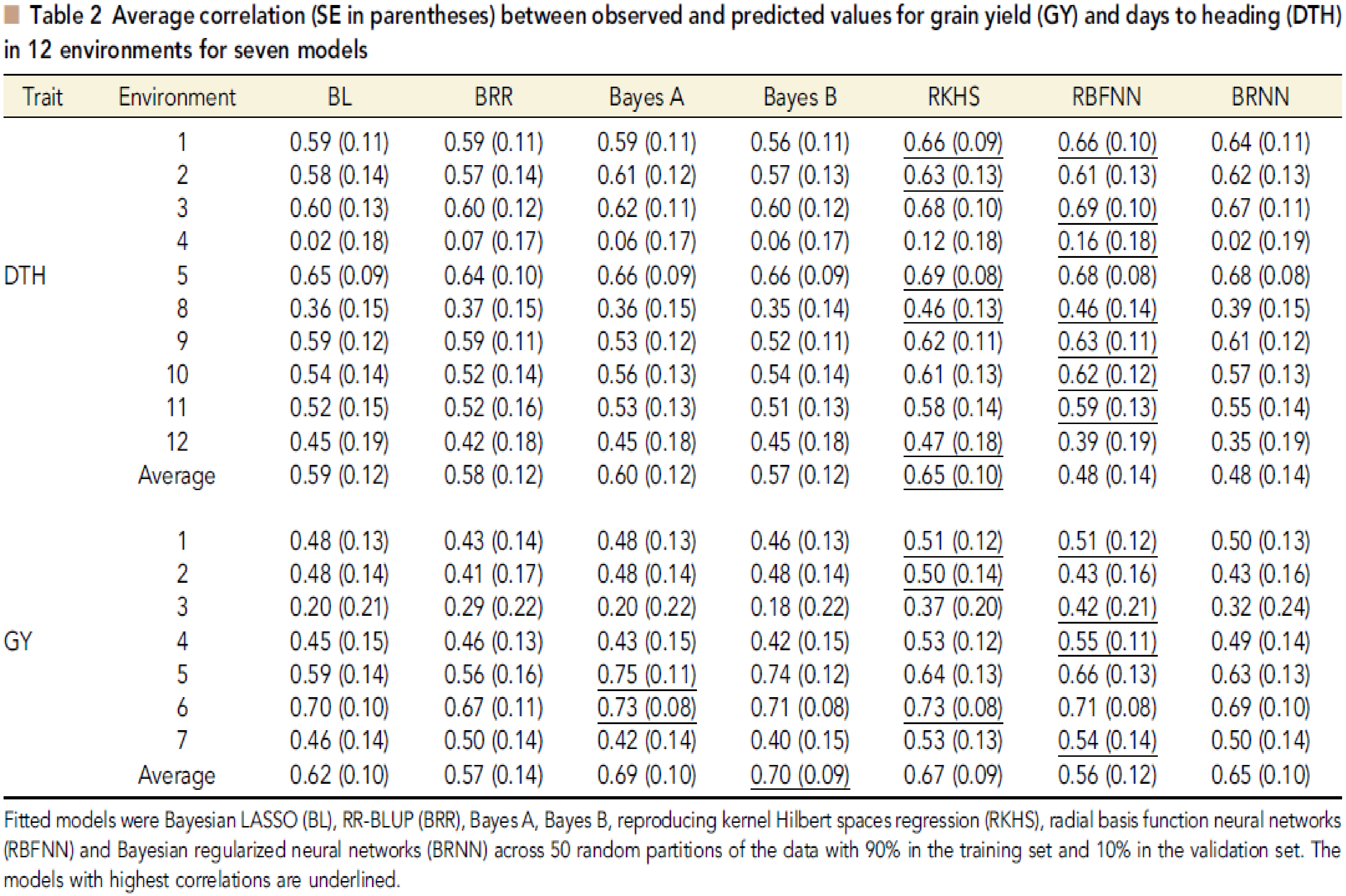
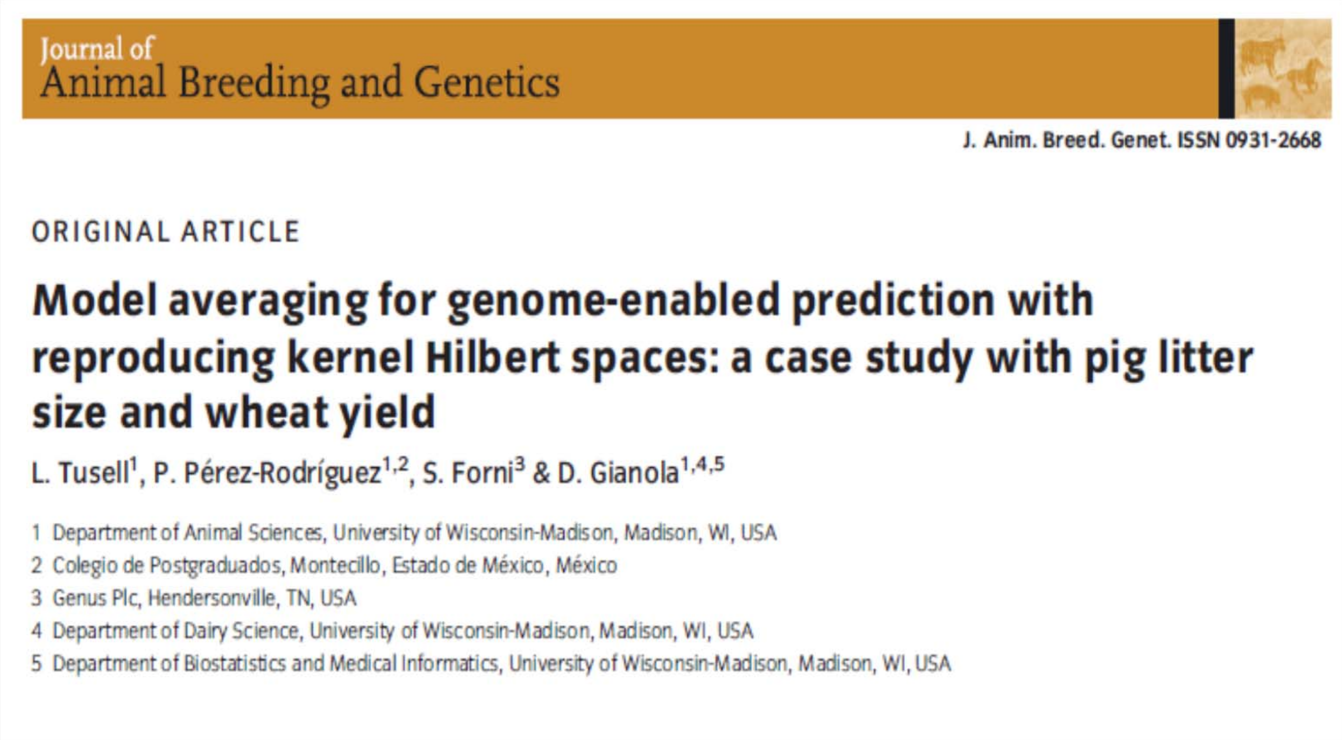


Table 2 Average correlation (SE in parentheses) between observed and predicted values for grain yield (GY) and days to heading (DTH) in 12 environments for seven models

Trait	Environment	BL	BRR	Bayes A	Bayes B	RKHS	RBFNN	BRNN
DTH	1	0.59 (0.11)	0.59 (0.11)	0.59 (0.11)	0.56 (0.11)	0.66 (0.09)	0.66 (0.10)	0.64 (0.11)
	2	0.58 (0.14)	0.57 (0.14)	0.61 (0.12)	0.57 (0.13)	0.63 (0.13)	0.61 (0.13)	0.62 (0.13)
	3	0.60 (0.13)	0.60 (0.12)	0.62 (0.11)	0.60 (0.12)	0.68 (0.10)	0.69 (0.10)	0.67 (0.11)
	4	0.02 (0.18)	0.07 (0.17)	0.06 (0.17)	0.06 (0.17)	0.12 (0.18)	0.16 (0.18)	0.02 (0.19)
	5	0.65 (0.09)	0.64 (0.10)	0.66 (0.09)	0.66 (0.09)	0.69 (0.08)	0.68 (0.08)	0.68 (0.08)
	8	0.36 (0.15)	0.37 (0.15)	0.36 (0.15)	0.35 (0.14)	0.46 (0.13)	0.46 (0.14)	0.39 (0.15)
	9	0.59 (0.12)	0.59 (0.11)	0.53 (0.12)	0.52 (0.11)	0.62 (0.11)	0.63 (0.11)	0.61 (0.12)
	10	0.54 (0.14)	0.52 (0.14)	0.56 (0.13)	0.54 (0.14)	0.61 (0.13)	0.62 (0.12)	0.57 (0.13)
	11	0.52 (0.15)	0.52 (0.16)	0.53 (0.13)	0.51 (0.13)	0.58 (0.14)	0.59 (0.13)	0.55 (0.14)
	12	0.45 (0.19)	0.42 (0.18)	0.45 (0.18)	0.45 (0.18)	0.47 (0.18)	0.39 (0.19)	0.35 (0.19)
	Average	0.59 (0.12)	0.58 (0.12)	0.60 (0.12)	0.57 (0.12)	0.65 (0.10)	0.48 (0.14)	0.48 (0.14)
GY	1	0.48 (0.13)	0.43 (0.14)	0.48 (0.13)	0.46 (0.13)	0.51 (0.12)	0.51 (0.12)	0.50 (0.13)
	2	0.48 (0.14)	0.41 (0.17)	0.48 (0.14)	0.48 (0.14)	0.50 (0.14)	0.43 (0.16)	0.43 (0.16)
	3	0.20 (0.21)	0.29 (0.22)	0.20 (0.22)	0.18 (0.22)	0.37 (0.20)	0.42 (0.21)	0.32 (0.24)
	4	0.45 (0.15)	0.46 (0.13)	0.43 (0.15)	0.42 (0.15)	0.53 (0.12)	0.55 (0.11)	0.49 (0.14)
	5	0.59 (0.14)	0.56 (0.16)	0.75 (0.11)	0.74 (0.12)	0.64 (0.13)	0.66 (0.13)	0.63 (0.13)
	6	0.70 (0.10)	0.67 (0.11)	0.73 (0.08)	0.71 (0.08)	0.73 (0.08)	0.71 (0.08)	0.69 (0.10)
	7	0.46 (0.14)	0.50 (0.14)	0.42 (0.14)	0.40 (0.15)	0.53 (0.13)	0.54 (0.14)	0.50 (0.14)
	Average	0.62 (0.10)	0.57 (0.14)	0.69 (0.10)	0.70 (0.09)	0.67 (0.09)	0.56 (0.12)	0.65 (0.10)

Fitted models were Bayesian LASSO (BL), RR-BLUP (BRR), Bayes A, Bayes B, reproducing kernel Hilbert spaces regression (RKHS), radial basis function neural networks (RBFNN) and Bayesian regularized neural networks (BRNN) across 50 random partitions of the data with 90% in the training set and 10% in the validation set. The models with highest correlations are underlined.

DOES MODEL AVERAGING HELP?
(in theory, MA expected to improve predictions)



Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield

L. Tusell[1], P. Pérez-Rodríguez[1,2], S. Forni[3] & D. Gianola[1,4,5]

1 Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI, USA
2 Colegio de Postgraduados, Montecillo, Estado de México, México
3 Genus Plc, Hendersonville, TN, USA
4 Department of Dairy Science, University of Wisconsin-Madison, Madison, WI, USA
5 Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA

Genus

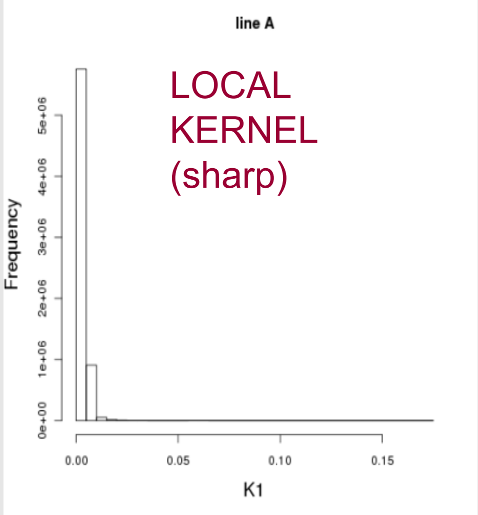
Line A

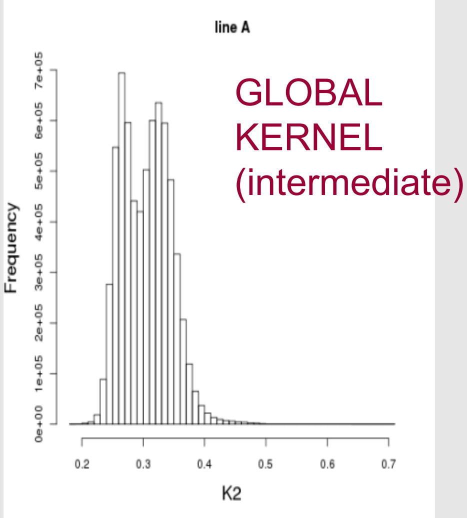
2,598 PB
46,855 SNPs

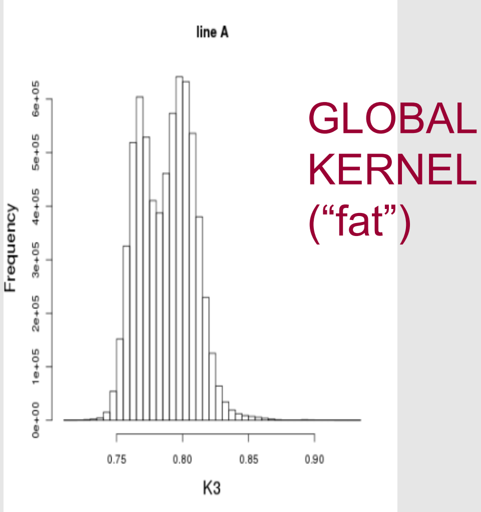
RKHS-KERNEL AVERAGING-MODEL AVERAGING

- For 3 bandwidths in Gaussian kernels, fitted:

1 : RKHS with K1
2 : RKHS with K2
3 : RKHS with K3
4 : RKHS-KA with K1, K2
5 : RKHS-KA with K1, K3
6 : RKHS-KA with K2, K3
7 : RKHS- KA with K1, K2, K3
8 : Average of predictions from models 1 to 7
8*: Weighted average from model 1 to 7 according to harmonic mean of $log\left[\hat{p}(\mathbf{y}|M_i)\right]$



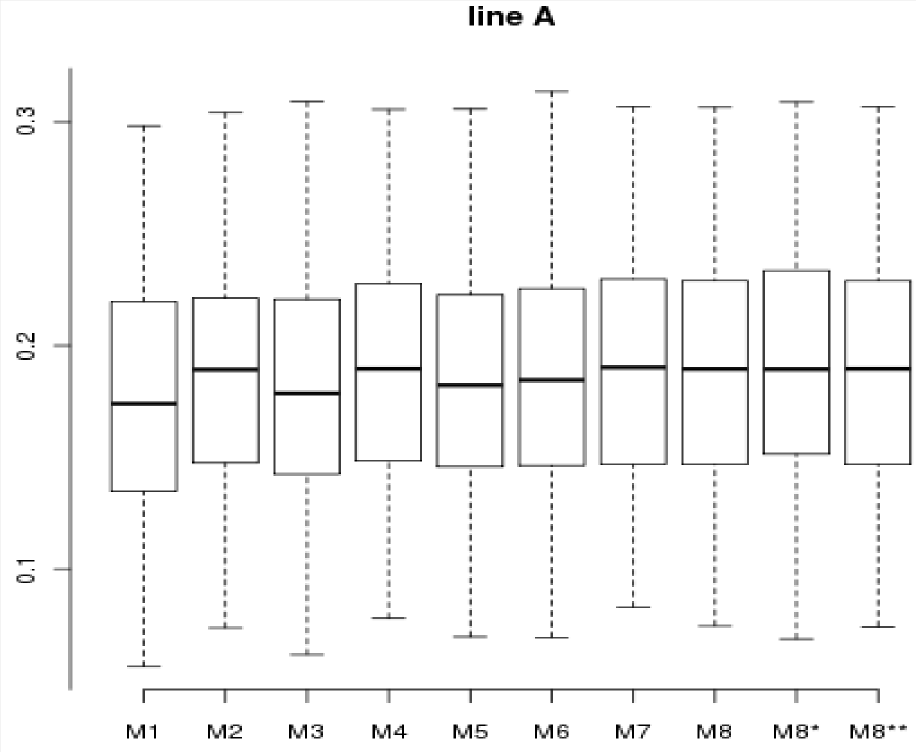




Multi-kernel: Histograms of entries of K={*K(x$_i$, x$_{i'}$)*}

Continued...

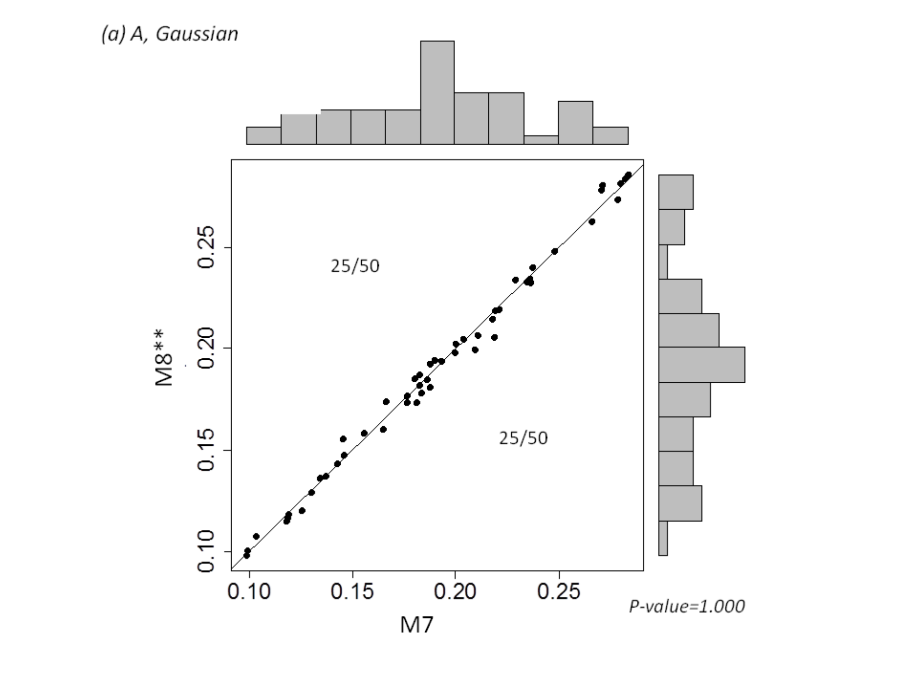
-50 random partitions: 90% of observations in training and 10% in testing
-Correlations between observed and predicted litter sizes



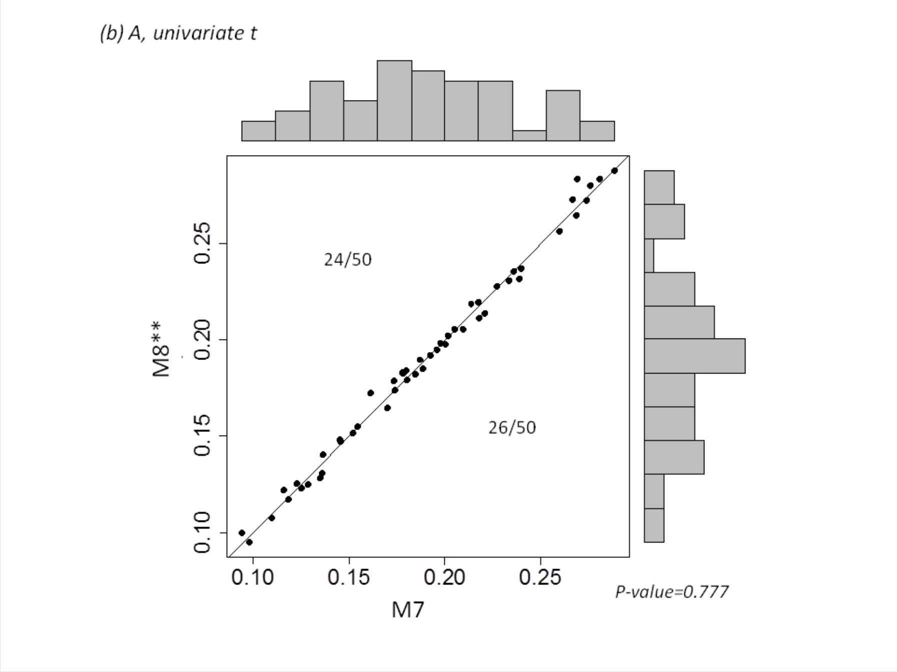
1-3 Single kernel
4-7 “multi-kernel”
8 Model averaging
8* Averaging using PMSE in a validation set, followed by testing
9 BMA

Distribution of correlations between observed and predicted phenotypes.

MULTI-KERNEL PREDICTIONS NOT WORSE THAN BMA; NOISY DATA

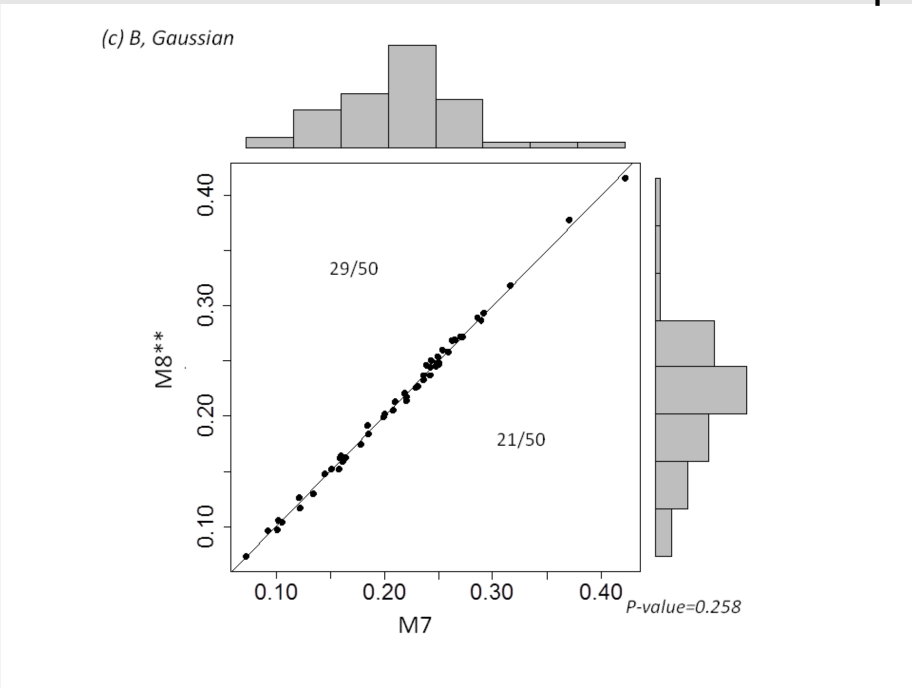


(a) A, Gaussian

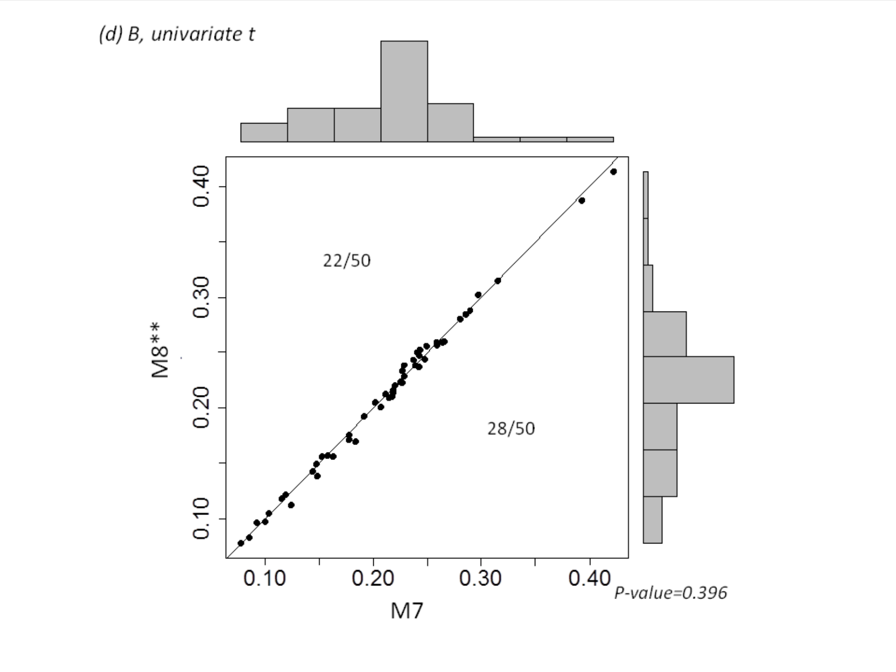


(b) A, univariate t

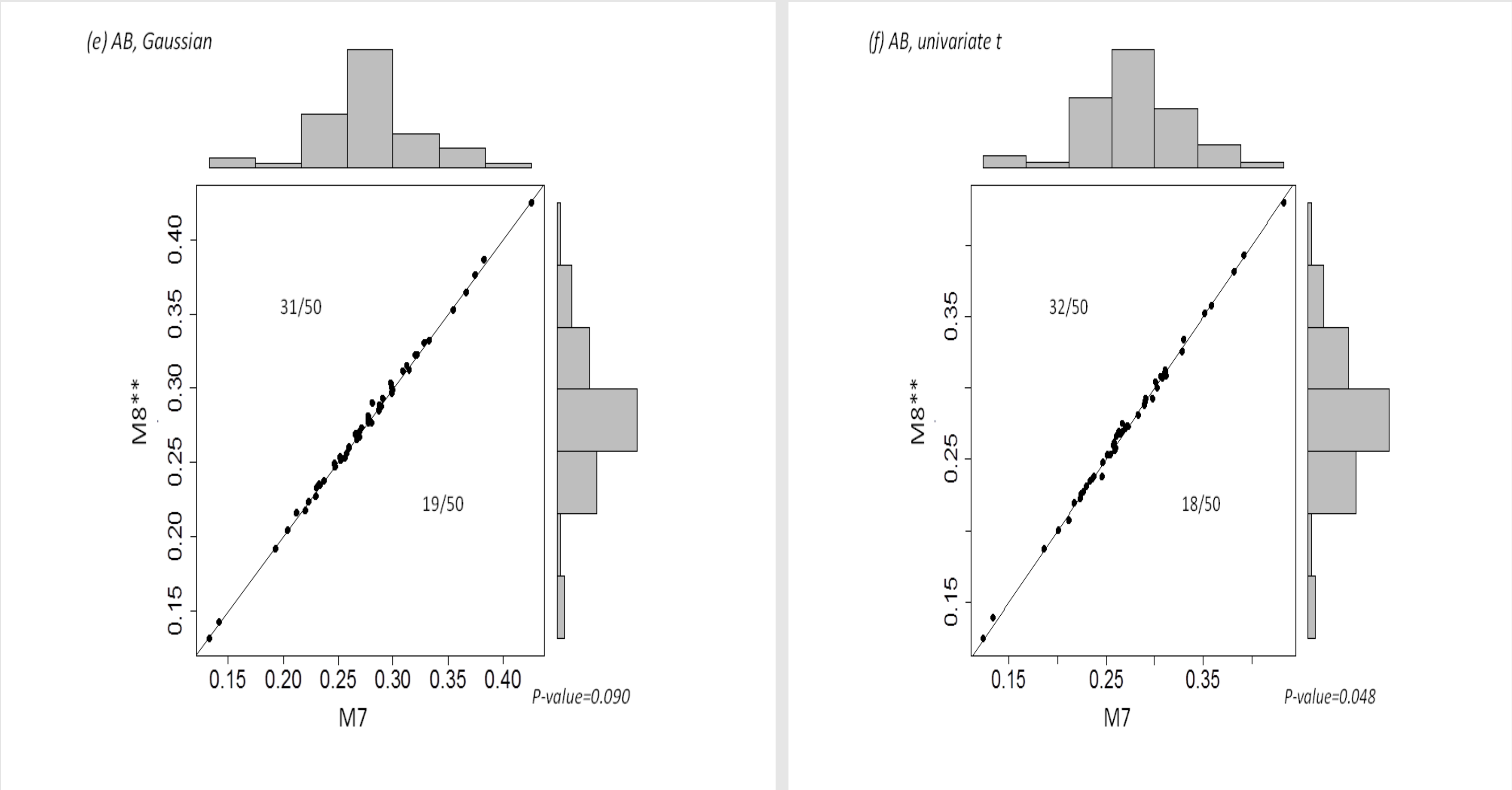
Tests: Friedman's non-parametric test for paired comparisons



(c) B, Gaussian



(d) B, univariate t



PICTURE: Results suggest that model averaging is slightly better than multi-kernel fitting in crossbreds (dominance?) but the extra work is not justified in purebreds. Anyhow, results are not clear cut.

Comparison among methods: plants (Heslot et al., 2012. Crop Science)

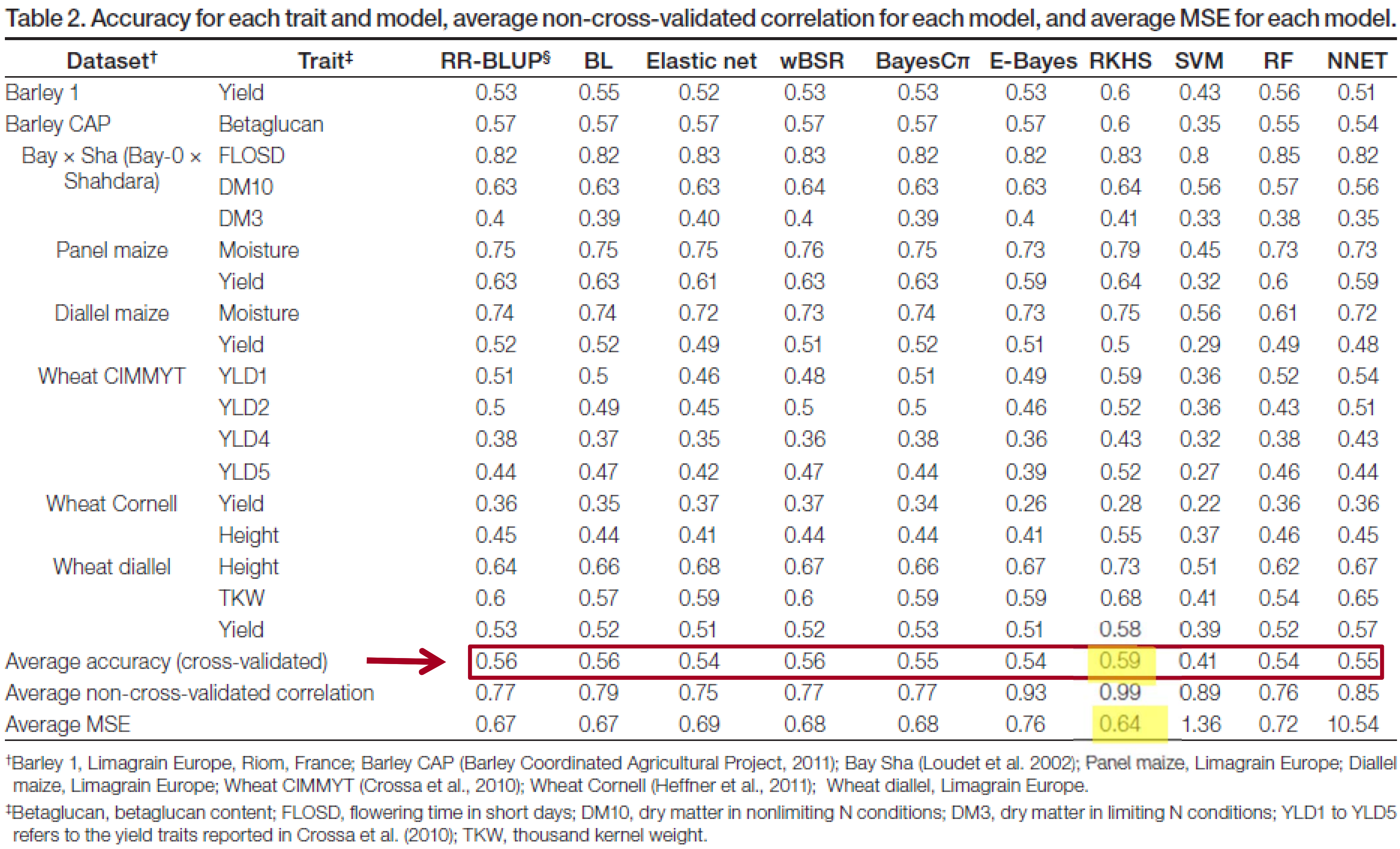


Table 2. Accuracy for each trait and model, average non-cross-validated correlation for each model, and average MSE for each model.

Dataset[†]	Trait[‡]	RR-BLUP[§]	BL	Elastic net	wBSR	BayesCπ	E-Bayes	RKHS	SVM	RF	NNET
Barley 1	Yield	0.53	0.55	0.52	0.53	0.53	0.53	0.6	0.43	0.56	0.51
Barley CAP	Betaglucan	0.57	0.57	0.57	0.57	0.57	0.57	0.6	0.35	0.55	0.54
Bay × Sha (Bay-0 × Shahdara)	FLOSD	0.82	0.82	0.83	0.83	0.82	0.82	0.83	0.8	0.85	0.82
	DM10	0.63	0.63	0.63	0.64	0.63	0.63	0.64	0.56	0.57	0.56
	DM3	0.4	0.39	0.40	0.4	0.39	0.4	0.41	0.33	0.38	0.35
Panel maize	Moisture	0.75	0.75	0.75	0.76	0.75	0.73	0.79	0.45	0.73	0.73
	Yield	0.63	0.63	0.61	0.63	0.63	0.59	0.64	0.32	0.6	0.59
Diallel maize	Moisture	0.74	0.74	0.72	0.73	0.74	0.73	0.75	0.56	0.61	0.72
	Yield	0.52	0.52	0.49	0.51	0.52	0.51	0.5	0.29	0.49	0.48
Wheat CIMMYT	YLD1	0.51	0.5	0.46	0.48	0.51	0.49	0.59	0.36	0.52	0.54
	YLD2	0.5	0.49	0.45	0.5	0.5	0.46	0.52	0.36	0.43	0.51
	YLD4	0.38	0.37	0.35	0.36	0.38	0.36	0.43	0.32	0.38	0.43
	YLD5	0.44	0.47	0.42	0.47	0.44	0.39	0.52	0.27	0.46	0.44
Wheat Cornell	Yield	0.36	0.35	0.37	0.37	0.34	0.26	0.28	0.22	0.36	0.36
	Height	0.45	0.44	0.41	0.44	0.44	0.41	0.55	0.37	0.46	0.45
Wheat diallel	Height	0.64	0.66	0.68	0.67	0.66	0.67	0.73	0.51	0.62	0.67
	TKW	0.6	0.57	0.59	0.6	0.59	0.59	0.68	0.41	0.54	0.65
	Yield	0.53	0.52	0.51	0.52	0.53	0.51	0.58	0.39	0.52	0.57
Average accuracy (cross-validated)		0.56	0.56	0.54	0.56	0.55	0.54	0.59	0.41	0.54	0.55
Average non-cross-validated correlation		0.77	0.79	0.75	0.77	0.77	0.93	0.99	0.89	0.76	0.85
Average MSE		0.67	0.67	0.69	0.68	0.68	0.76	0.64	1.36	0.72	10.54

[†] Barley 1, Limagrain Europe, Riom, France; Barley CAP (Barley Coordinated Agricultural Project, 2011); Bay Sha (Loudet et al. 2002); Panel maize, Limagrain Europe; Diallel maize, Limagrain Europe; Wheat CIMMYT (Crossa et al., 2010); Wheat Cornell (Heffner et al., 2011); Wheat diallel, Limagrain Europe.

[‡] Betaglucan, betaglucan content; FLOSD, flowering time in short days; DM10, dry matter in nonlimiting N conditions; DM3, dry matter in limiting N conditions; YLD1 to YLD5 refers to the yield traits reported in Crossa et al. (2010); TKW, thousand kernel weight.

TENTATIVE CONCLUSION: choice of method does not make a difference, in practice

Really?

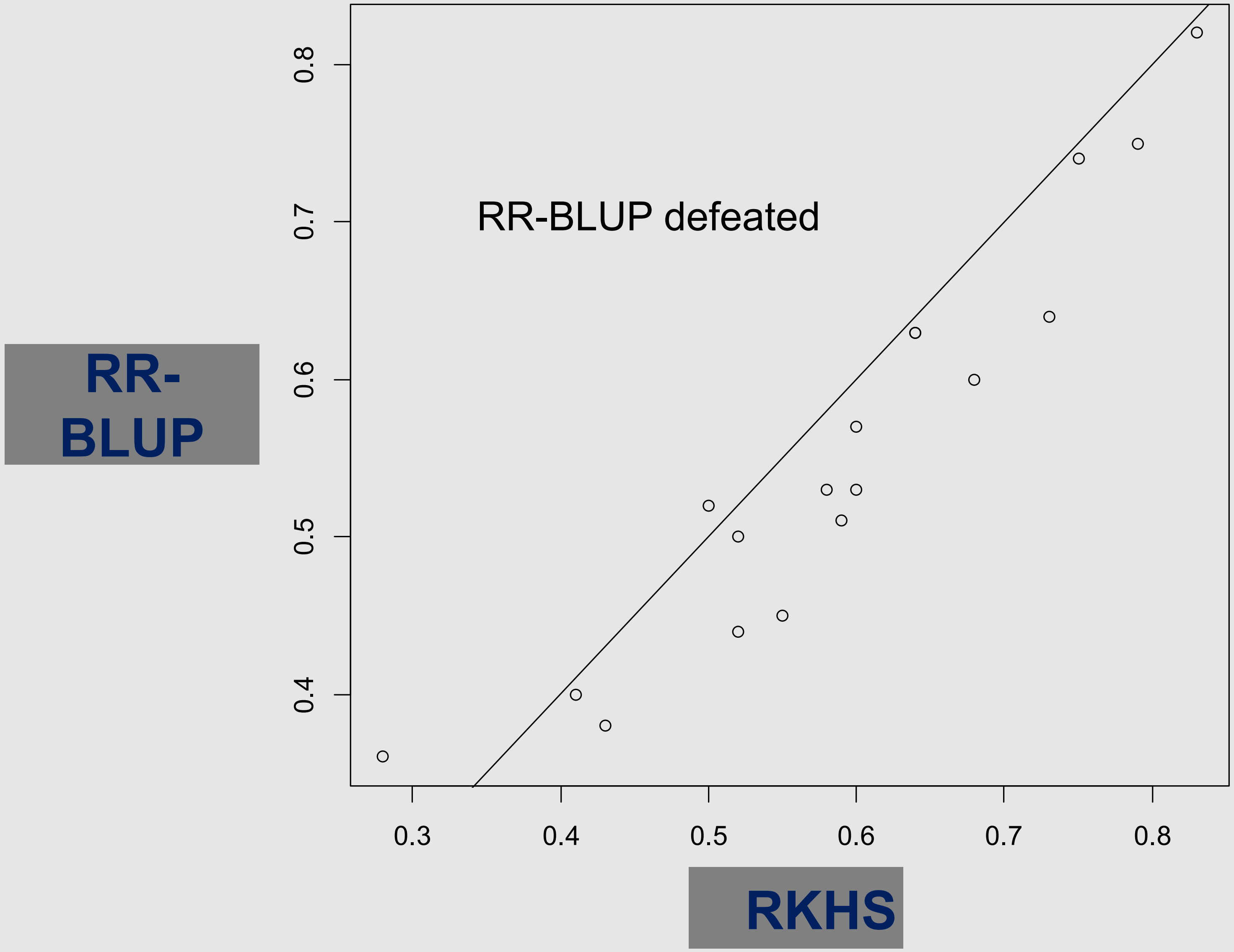
Single malt rankings:
DAILUAINE 14 YEARS CONNOISSEUR'S CHOIC
GLEN GARIOCH FOUNDER'S RESERVE (85)
BOWMORE 12 YEAR OLD (80)
OBAN 14 YEARS (73)...

WHISKY SECRETS:

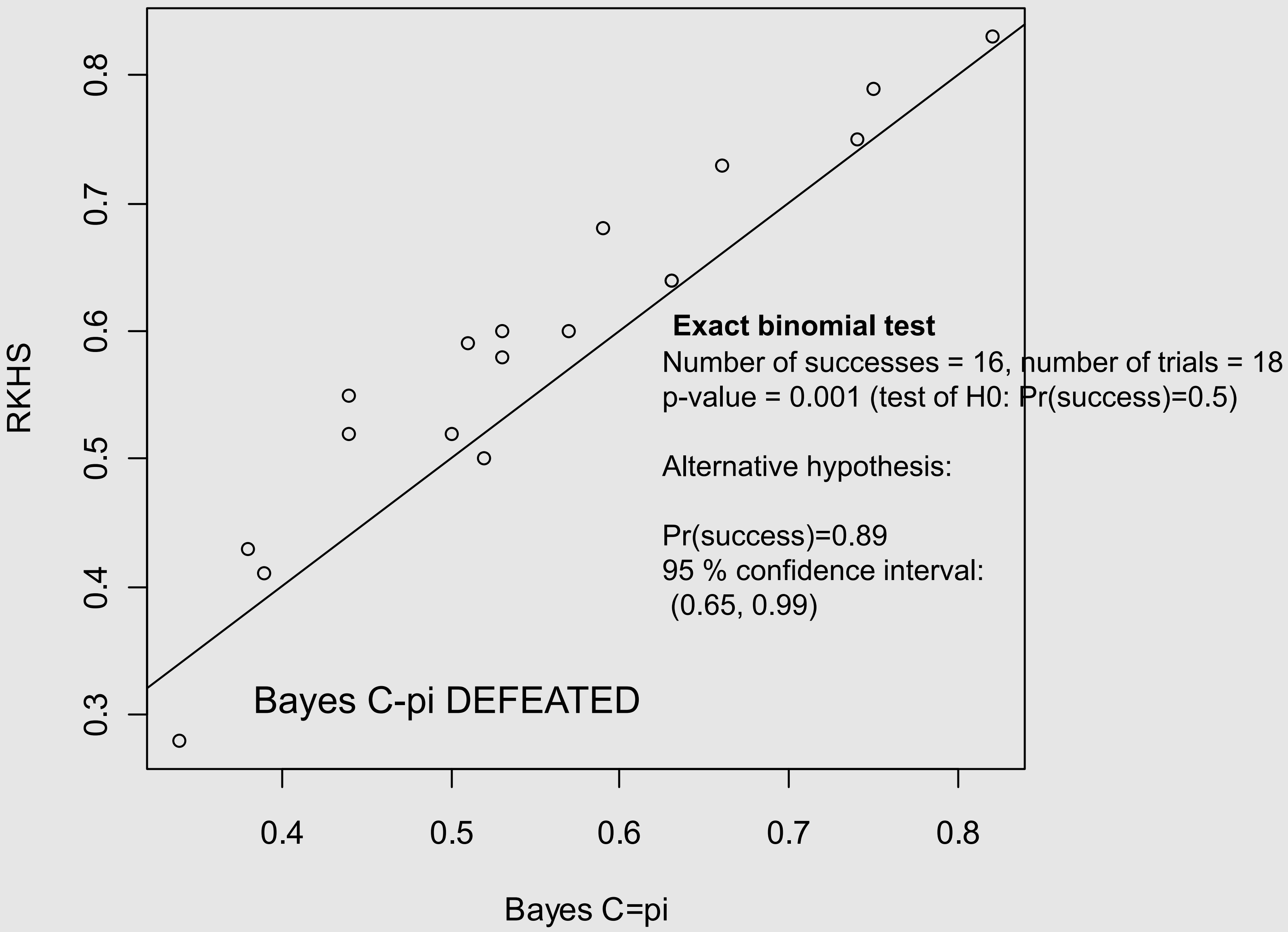
-Replicate comparisons

-Use paired comparisons (smaller variance of differences)

RKHS vs RR-BLUP:
18 comparisons of Heslot et al. (1982)



RKHS vs Bayes C-pi:
18 comparisons of Heslot et al. (2012)



Exact binomial test
Number of successes = 16, number of trials = 18
p-value = 0.001 (test of H0: Pr(success)=0.5)

Alternative hypothesis:

Pr(success)=0.89
95 % confidence interval:
(0.65, 0.99)

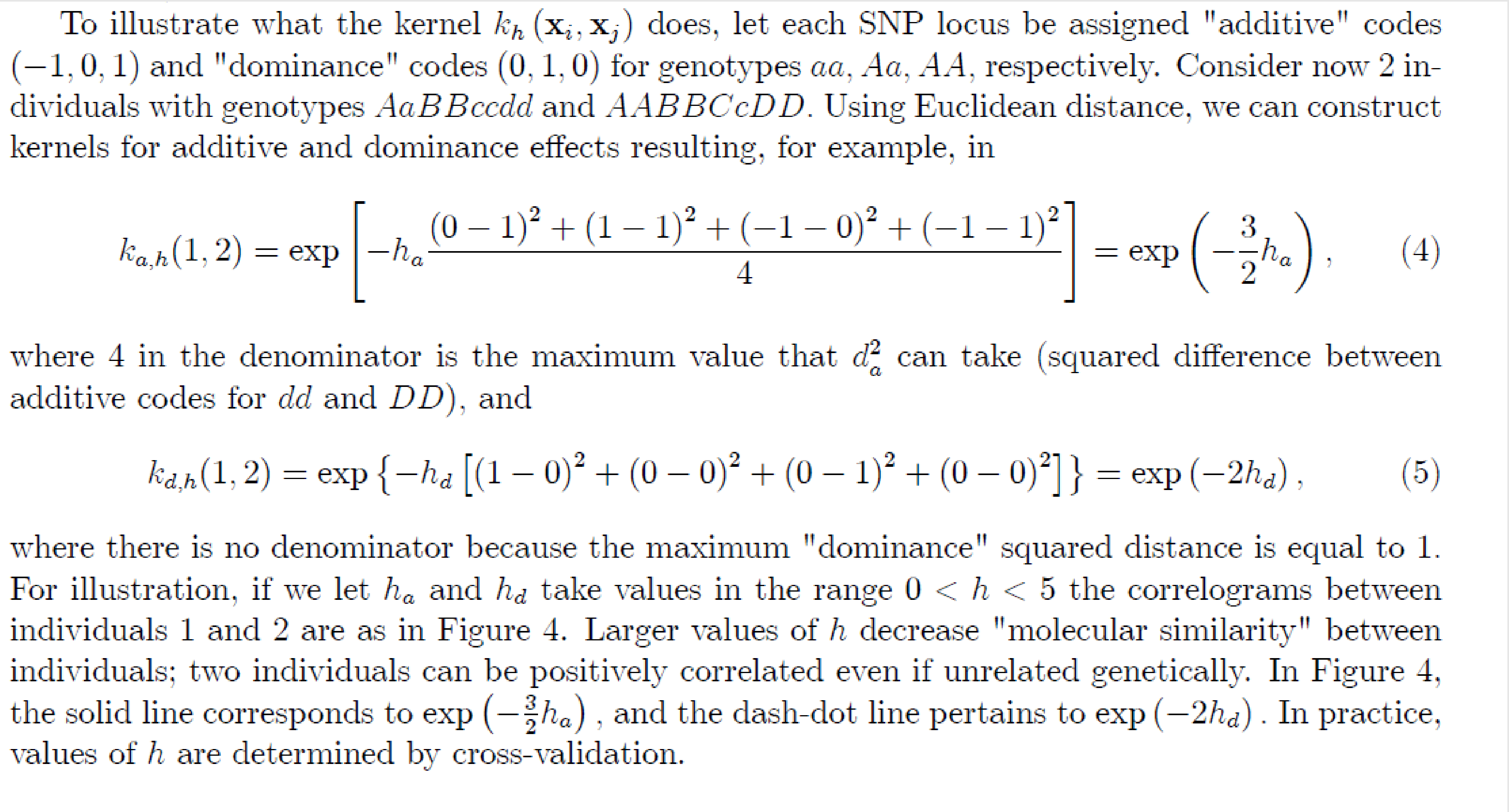
Bayes C-pi DEFEATED

FURTHER DOWN THE ROAD

ADDITIVE KERNELS ARE NOT THE ONLY POSSIBILITIES!!!

ADDITIVE, DOMINANCE AND ADDITIVE X DOMINANCE GAUSSIAN KERNELS

(FOR A X A: square kernel elements for A; for A X D: multiply kernel elements)



To illustrate what the kernel $k_h(\mathbf{x}_i, \mathbf{x}_j)$ does, let each SNP locus be assigned "additive" codes $(-1, 0, 1)$ and "dominance" codes $(0, 1, 0)$ for genotypes aa, Aa, AA, respectively. Consider now 2 individuals with genotypes $AaBBccdd$ and $AABBCcDD$. Using Euclidean distance, we can construct kernels for additive and dominance effects resulting, for example, in

$$k_{a,h}(1,2) = \exp\left[-h_a \frac{(0-1)^2 + (1-1)^2 + (-1-0)^2 + (-1-1)^2}{4}\right] = \exp\left(-\frac{3}{2}h_a\right), \qquad (4)$$

where 4 in the denominator is the maximum value that d_a^2 can take (squared difference between additive codes for dd and DD), and

$$k_{d,h}(1,2) = \exp\left\{-h_d\left[(1-0)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2\right]\right\} = \exp(-2h_d), \qquad (5)$$

where there is no denominator because the maximum "dominance" squared distance is equal to 1. For illustration, if we let h_a and h_d take values in the range $0 < h < 5$ the correlograms between individuals 1 and 2 are as in Figure 4. Larger values of h decrease "molecular similarity" between individuals; two individuals can be positively correlated even if unrelated genetically. In Figure 4, the solid line corresponds to $\exp\left(-\frac{3}{2}h_a\right)$, and the dash-dot line pertains to $\exp(-2h_d)$. In practice, values of h are determined by cross-validation.

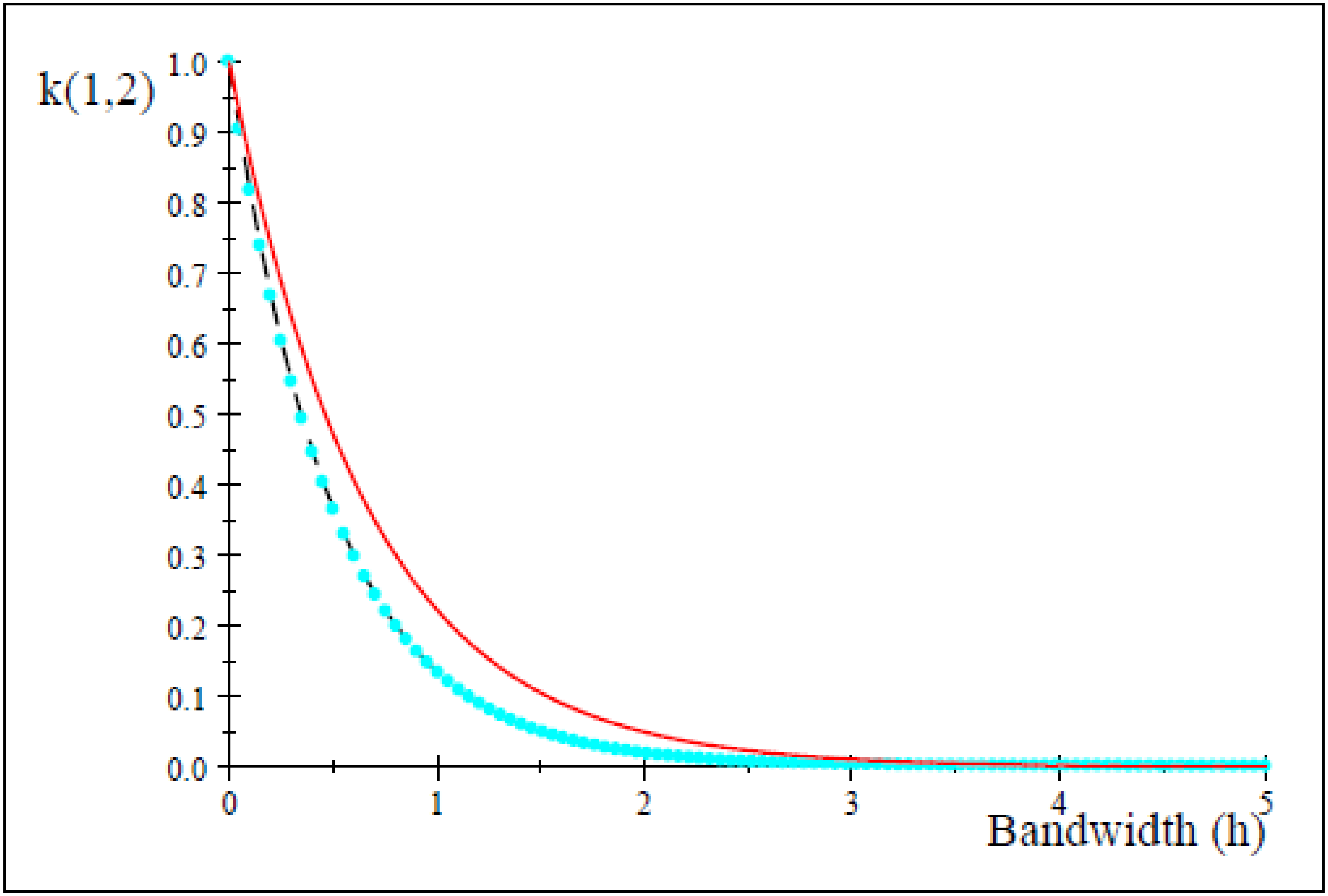


Figure 4. Correlogram between individuals 1 and 2 as a function of bandwidth (h). The solid line depicts correlation from an "additive" kernel; the dot-dash line corresponds to the "dominance" kernel.

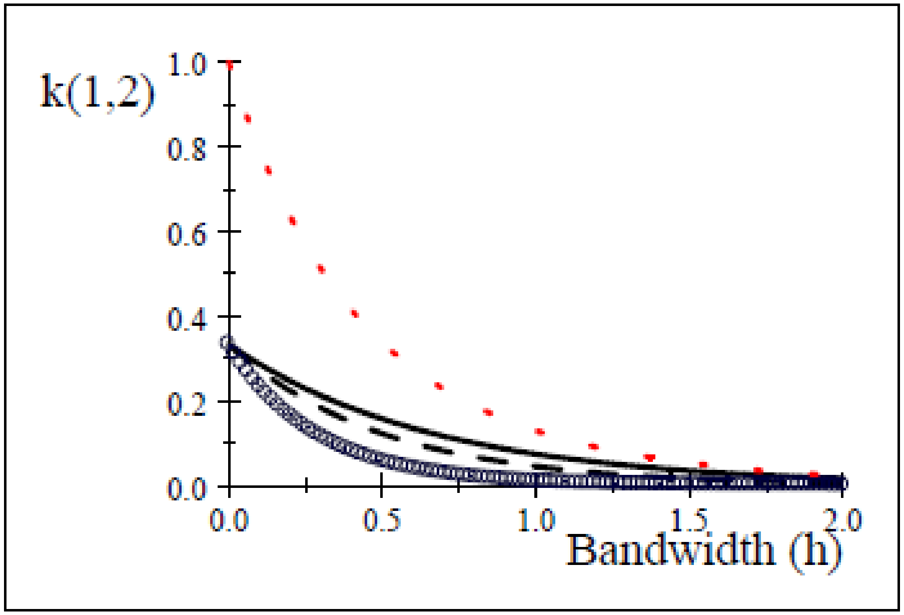


Figure 5A. Molecular similarity between two individuals as a function of bandwidth for variance partition $\sigma^2_{\alpha,a} = \sigma^2_{\alpha,d} = \sigma^2_{\alpha,ad} = 1$. Top line (dots) gives the total correlation, the solid, dashed and dot-dash lines give the contributions from the additive, dominance and additive × dominace kernels, respectively.

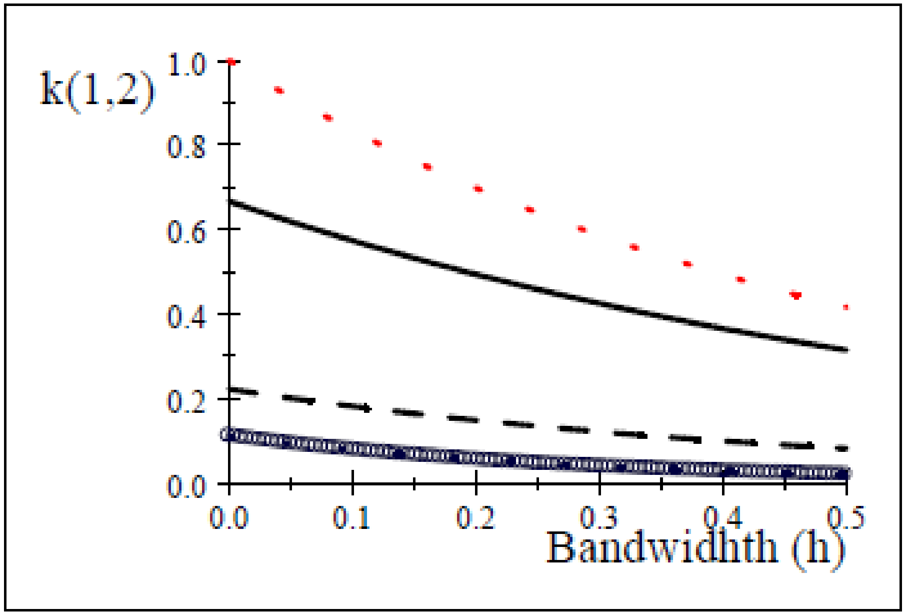
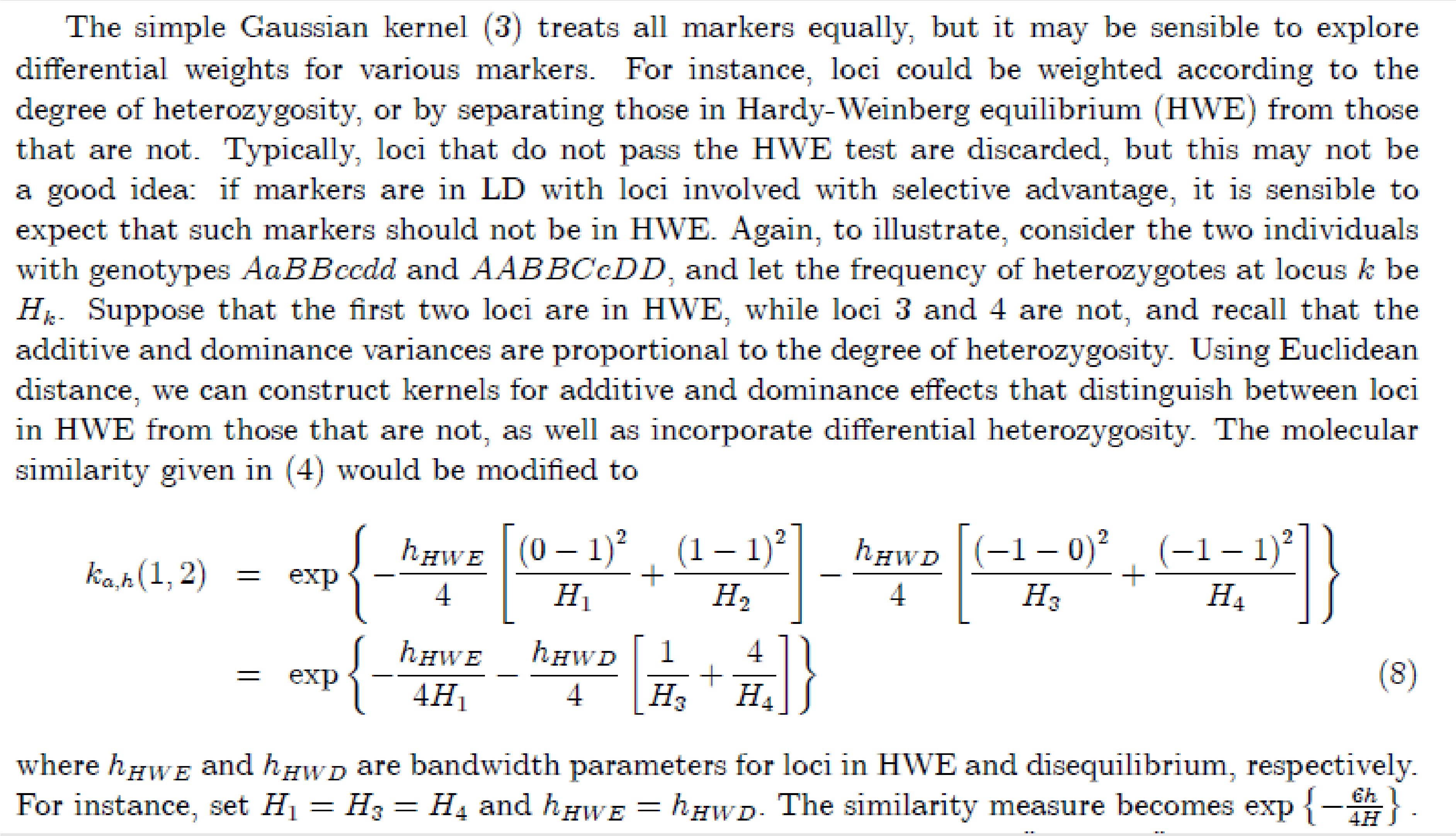


Figure 5B. Molecular similarity between two individuals as a function of bandwidth for variance partition $\sigma^2_{\alpha,a} = 2, \sigma^2_{\alpha,d} = \frac{2}{3}, \sigma^2_{\alpha,ad} = \frac{1}{3}$. Top line (dots) gives the total correlation, the solid, dashed and dot-dash lines give the contributions from the additive, dominance and additive × dominace kernels, respectively.

KERNELS WITH DIFFERENTIAL WEIGHTS TO SNPs



The simple Gaussian kernel (3) treats all markers equally, but it may be sensible to explore differential weights for various markers. For instance, loci could be weighted according to the degree of heterozygosity, or by separating those in Hardy-Weinberg equilibrium (HWE) from those that are not. Typically, loci that do not pass the HWE test are discarded, but this may not be a good idea: if markers are in LD with loci involved with selective advantage, it is sensible to expect that such markers should not be in HWE. Again, to illustrate, consider the two individuals with genotypes $AaBBccdd$ and $AABBCcDD$, and let the frequency of heterozygotes at locus k be H_k. Suppose that the first two loci are in HWE, while loci 3 and 4 are not, and recall that the additive and dominance variances are proportional to the degree of heterozygosity. Using Euclidean distance, we can construct kernels for additive and dominance effects that distinguish between loci in HWE from those that are not, as well as incorporate differential heterozygosity. The molecular similarity given in (4) would be modified to

$$\begin{aligned} k_{a,h}(1,2) &= \exp\left\{-\frac{h_{HWE}}{4}\left[\frac{(0-1)^2}{H_1}+\frac{(1-1)^2}{H_2}\right]-\frac{h_{HWD}}{4}\left[\frac{(-1-0)^2}{H_3}+\frac{(-1-1)^2}{H_4}\right]\right\} \\ &= \exp\left\{-\frac{h_{HWE}}{4H_1}-\frac{h_{HWD}}{4}\left[\frac{1}{H_3}+\frac{4}{H_4}\right]\right\} \qquad (8) \end{aligned}$$

where h_{HWE} and h_{HWD} are bandwidth parameters for loci in HWE and disequilibrium, respectively. For instance, set $H_1 = H_3 = H_4$ and $h_{HWE} = h_{HWD}$. The similarity measure becomes $\exp\left\{-\frac{6h}{4H}\right\}$.

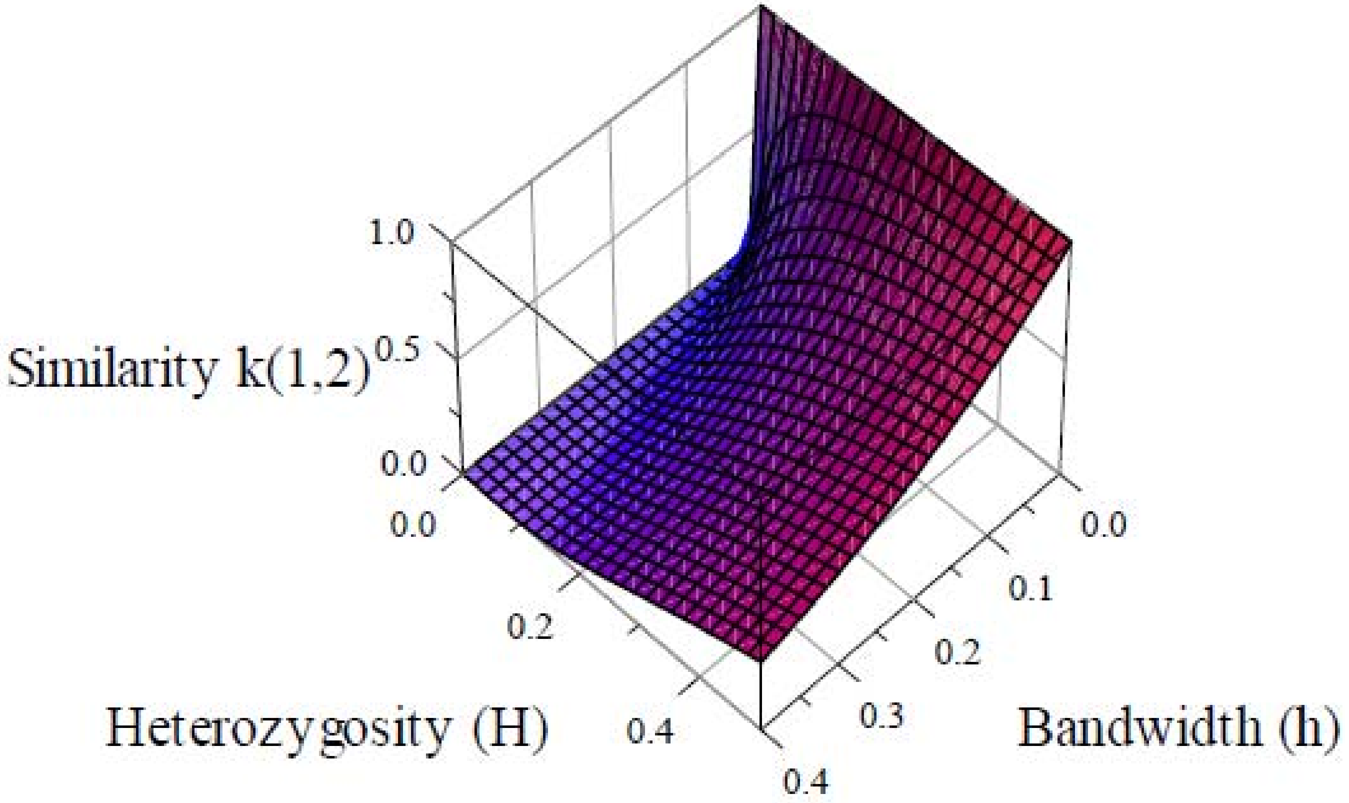
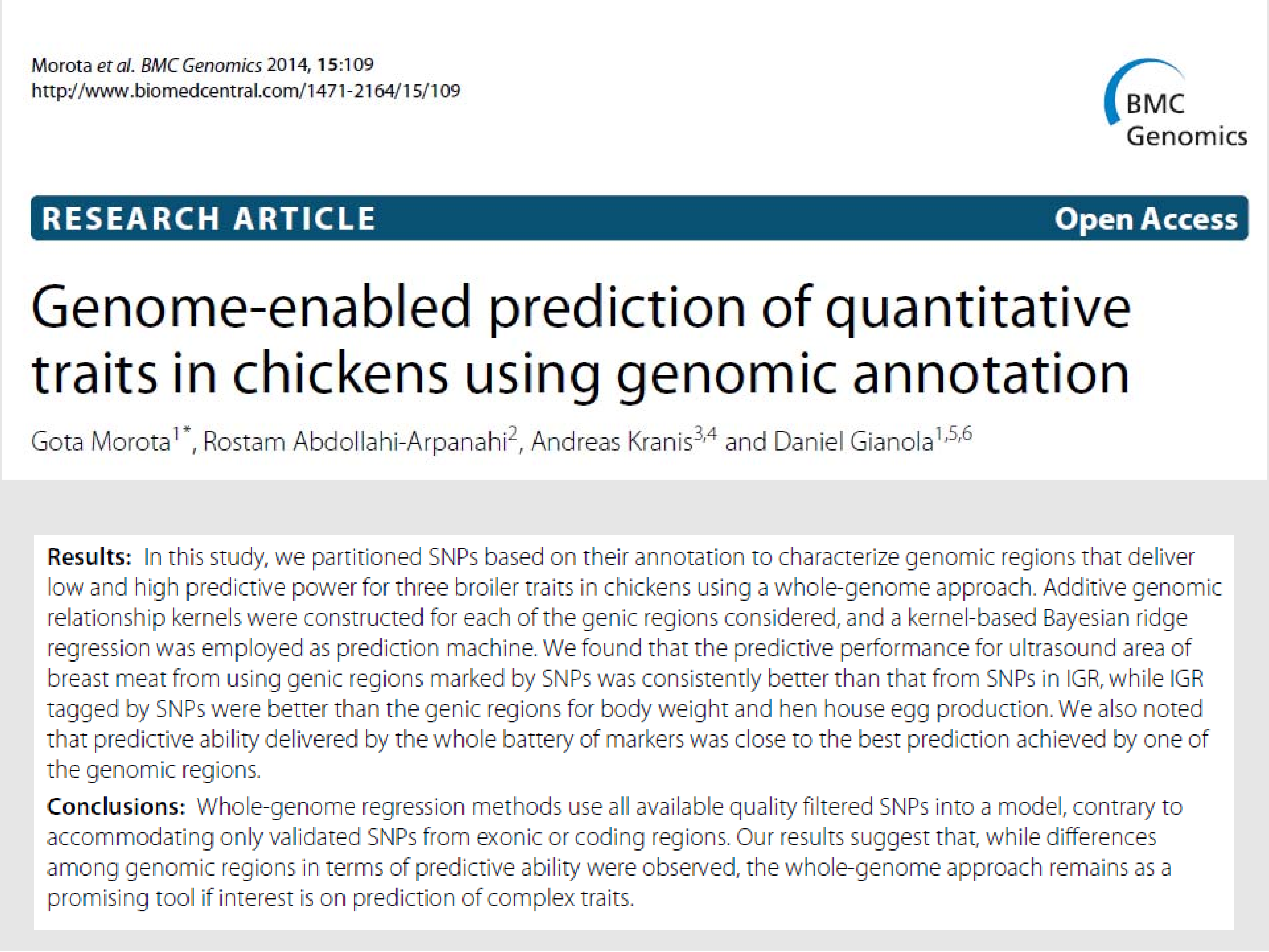


Figure 6. Similarity between 2 individuals created by an "additive" kernel as a funcion of frequency of heterozygosity and of the bandwidth parameter.

REFINING THE INFORMATION FROM MARKERS



Morota *et al. BMC Genomics* 2014, **15**:109
http://www.biomedcentral.com/1471-2164/15/109

BMC Genomics

RESEARCH ARTICLE **Open Access**

Genome-enabled prediction of quantitative traits in chickens using genomic annotation

Gota Morota[1*], Rostam Abdollahi-Arpanahi[2], Andreas Kranis[3,4] and Daniel Gianola[1,5,6]

Results: In this study, we partitioned SNPs based on their annotation to characterize genomic regions that deliver low and high predictive power for three broiler traits in chickens using a whole-genome approach. Additive genomic relationship kernels were constructed for each of the genic regions considered, and a kernel-based Bayesian ridge regression was employed as prediction machine. We found that the predictive performance for ultrasound area of breast meat from using genic regions marked by SNPs was consistently better than that from SNPs in IGR, while IGR tagged by SNPs were better than the genic regions for body weight and hen house egg production. We also noted that predictive ability delivered by the whole battery of markers was close to the best prediction achieved by one of the genomic regions.

Conclusions: Whole-genome regression methods use all available quality filtered SNPs into a model, contrary to accommodating only validated SNPs from exonic or coding regions. Our results suggest that, while differences among genomic regions in terms of predictive ability were observed, the whole-genome approach remains as a promising tool if interest is on prediction of complex traits.

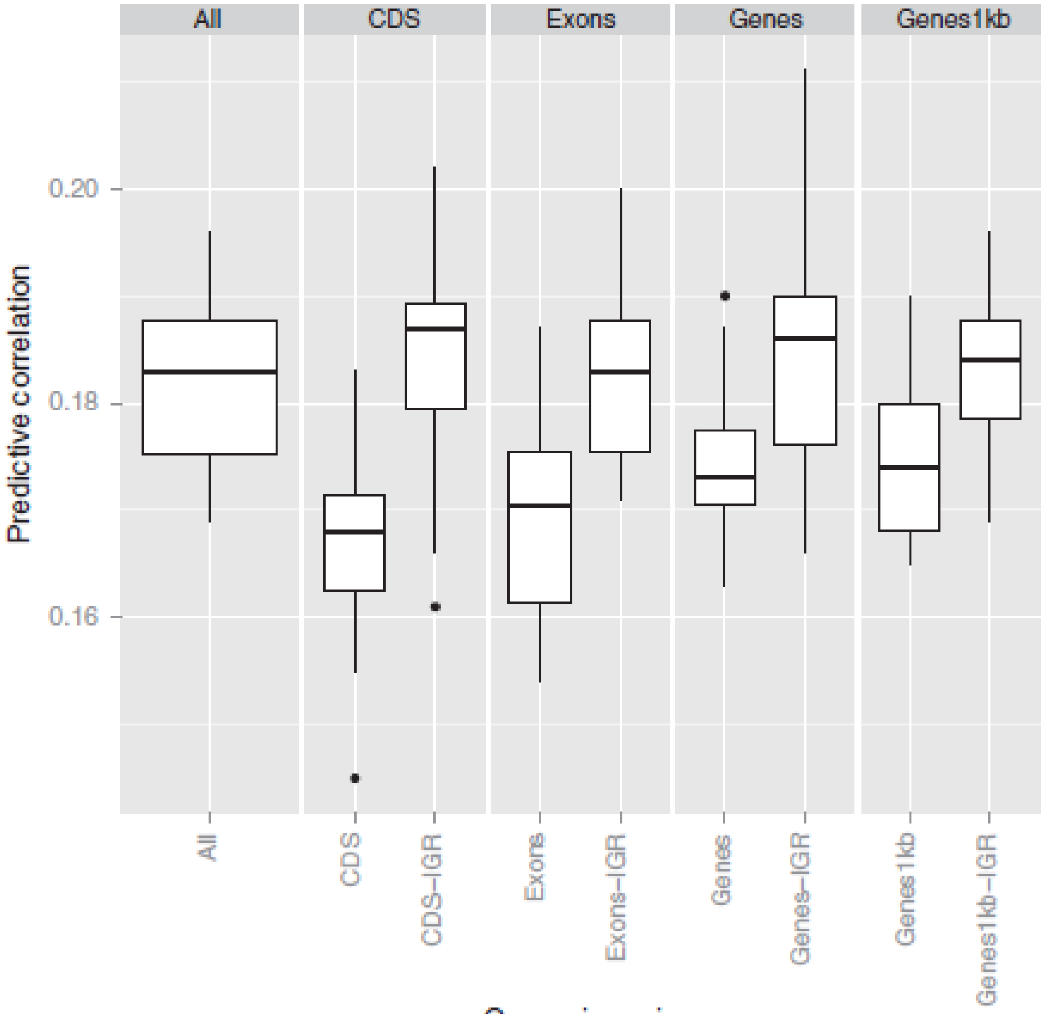


Figure 1 Predictive correlations comparing genic and non-genic regions for BW using kernel-based Bayesian ridge regression. The results were based on 10 fold cross-validation with 15 replications for each genomic region. Genic regions were coding DNA sequences (CDS), exons, genes, and genes with 1kb upstream and downstream. The genomic regions followed by the term "IGR" represent intergenic regions that contain equal SNP numbers to those of genic regions. "All" means all SNPs used for constructing **G**. Outliers denoted as black dots.

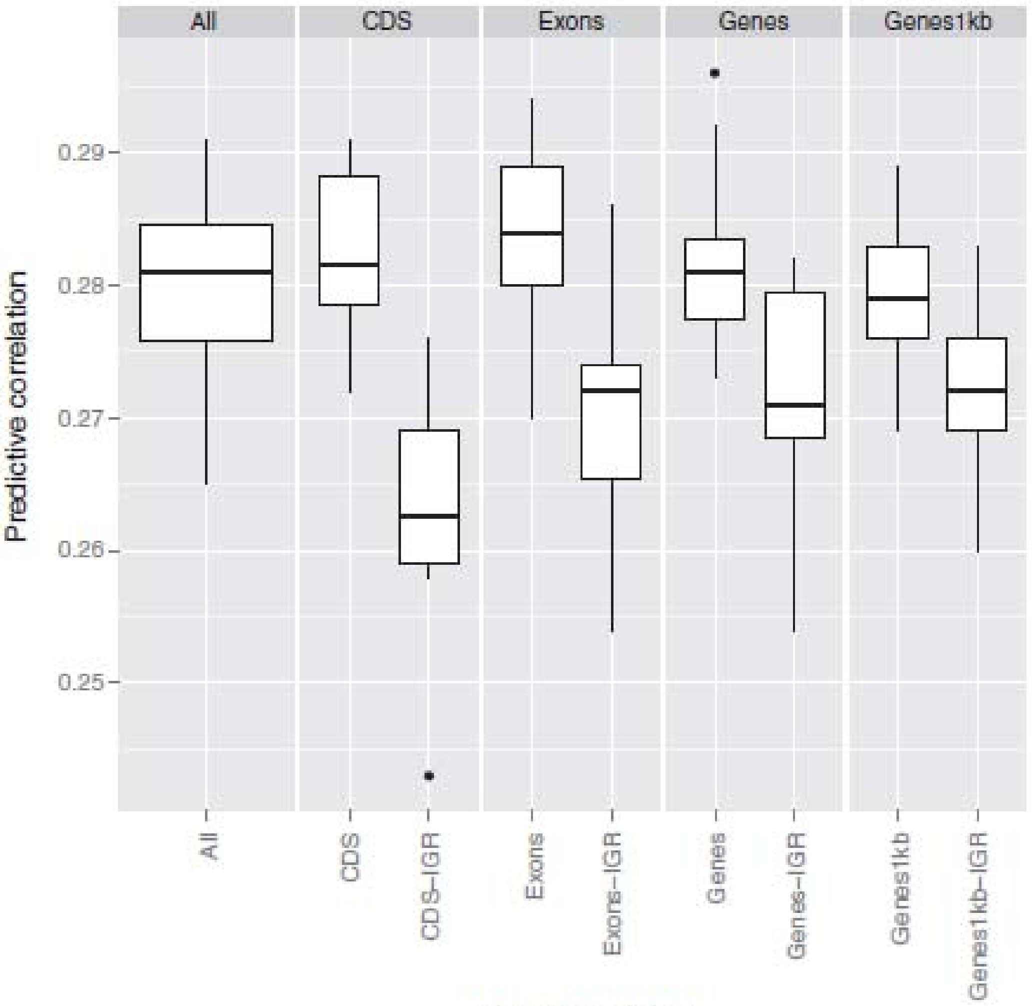


Figure 2 Predictive correlations comparing genic and non-genic regions for BM using kernel-based Bayesian ridge regression. The results were based on 10 fold cross-validation with 15 replications for each genomic region. Genic regions were coding DNA sequences (CDS), exons, genes, and genes with 1kb upstream and downstream. The genomic regions followed by the term "IGR" represent intergenic regions that contain equal SNP numbers to those of genic regions. "All" means all SNPs used for constructing **G**. Outliers denoted as black dots.

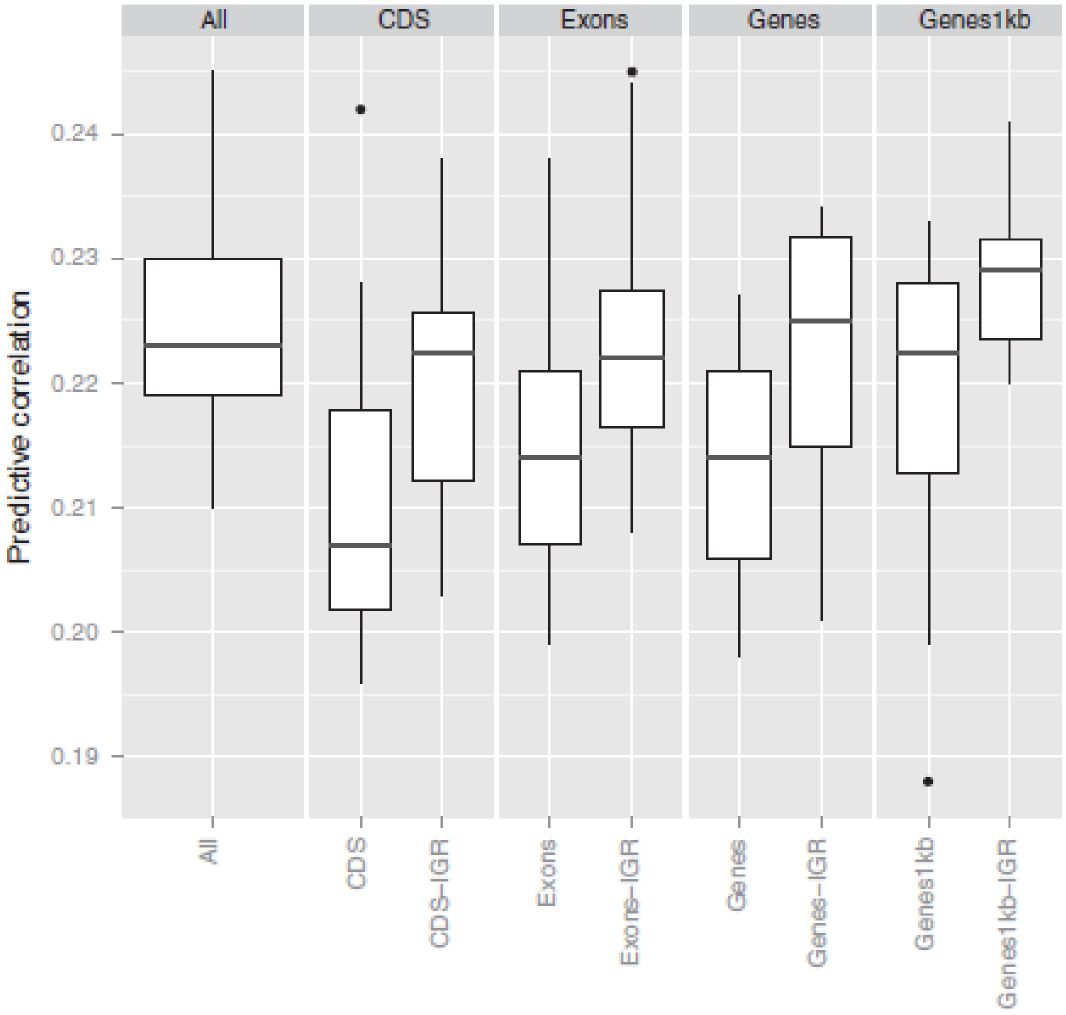
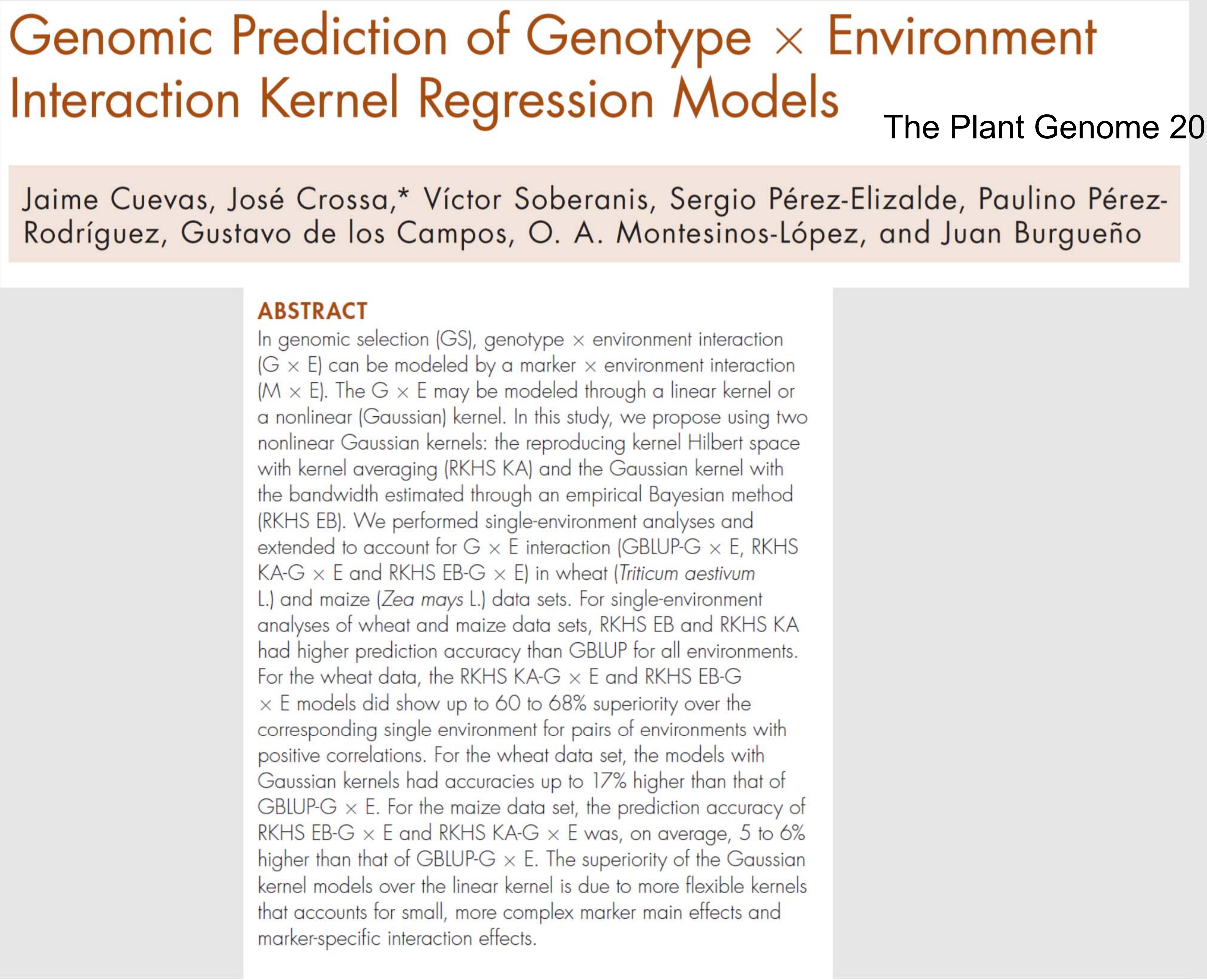


Figure 3 Predictive correlations comparing genic and non-genic regions for HHP using kernel-based Bayesian ridge regression. The results were based on 10 fold cross-validation with 15 replications for each genomic region. Genic regions were coding DNA sequences (CDS), exons, genes, and genes with 1kb upstream and downstream. The genomic regions followed by the term "IGR" represent intergenic regions that contain equal SNP numbers to those of genic regions. "All" means all SNPs used for constructing **G**. Outliers denoted as black dots.



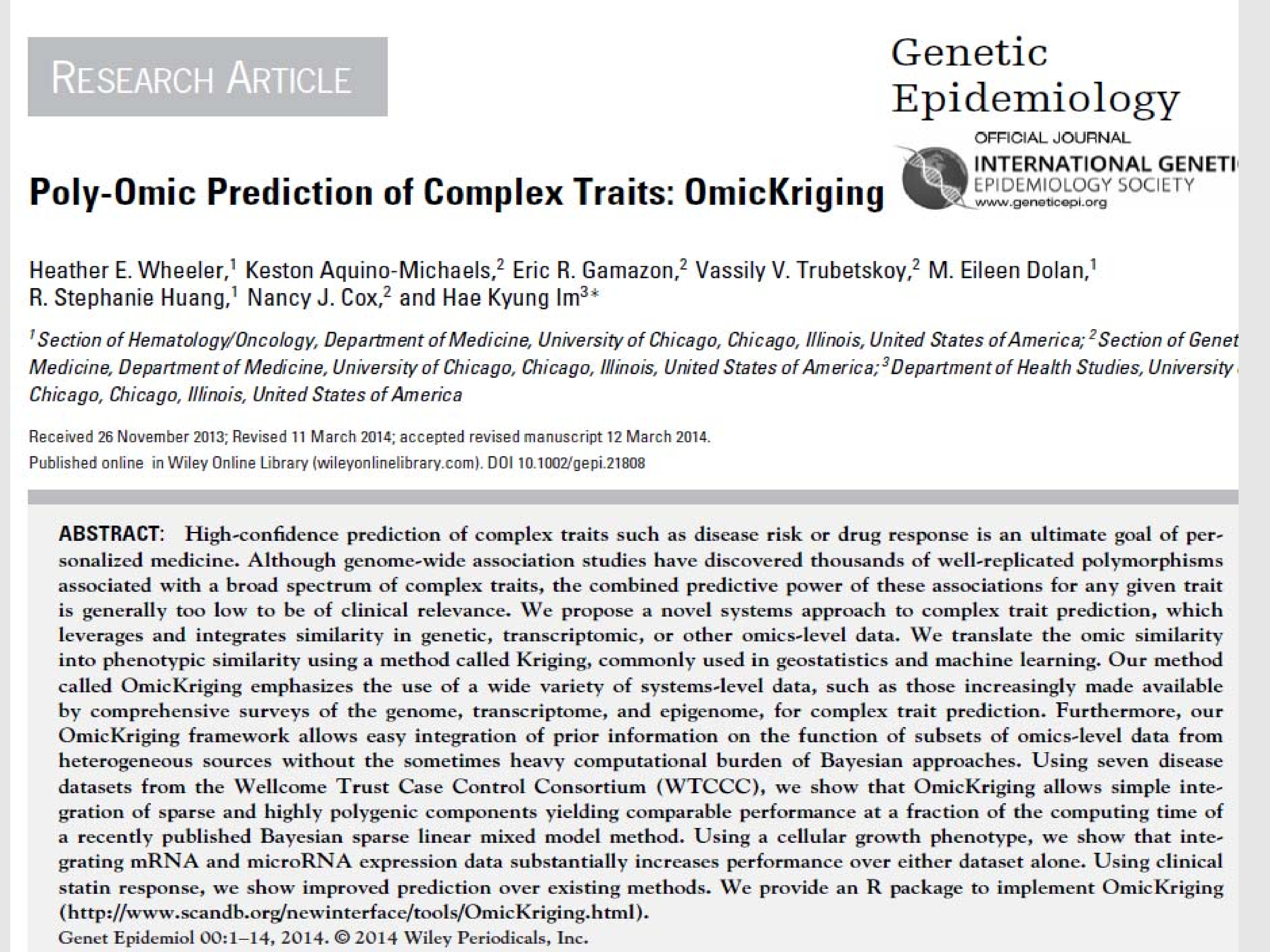
Genomic Prediction of Genotype × Environment Interaction Kernel Regression Models

The Plant Genome 2016

Jaime Cuevas, José Crossa,* Víctor Soberanis, Sergio Pérez-Elizalde, Paulino Pérez-Rodríguez, Gustavo de los Campos, O. A. Montesinos-López, and Juan Burgueño

ABSTRACT

In genomic selection (GS), genotype × environment interaction (G × E) can be modeled by a marker × environment interaction (M × E). The G × E may be modeled through a linear kernel or a nonlinear (Gaussian) kernel. In this study, we propose using two nonlinear Gaussian kernels: the reproducing kernel Hilbert space with kernel averaging (RKHS KA) and the Gaussian kernel with the bandwidth estimated through an empirical Bayesian method (RKHS EB). We performed single-environment analyses and extended to account for G × E interaction (GBLUP-G × E, RKHS KA-G × E and RKHS EB-G × E) in wheat (*Triticum aestivum* L.) and maize (*Zea mays* L.) data sets. For single-environment analyses of wheat and maize data sets, RKHS EB and RKHS KA had higher prediction accuracy than GBLUP for all environments. For the wheat data, the RKHS KA-G × E and RKHS EB-G × E models did show up to 60 to 68% superiority over the corresponding single environment for pairs of environments with positive correlations. For the wheat data set, the models with Gaussian kernels had accuracies up to 17% higher than that of GBLUP-G × E. For the maize data set, the prediction accuracy of RKHS EB-G × E and RKHS KA-G × E was, on average, 5 to 6% higher than that of GBLUP-G × E. The superiority of the Gaussian kernel models over the linear kernel is due to more flexible kernels that accounts for small, more complex marker main effects and marker-specific interaction effects.



RESEARCH ARTICLE

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

Poly-Omic Prediction of Complex Traits: OmicKriging

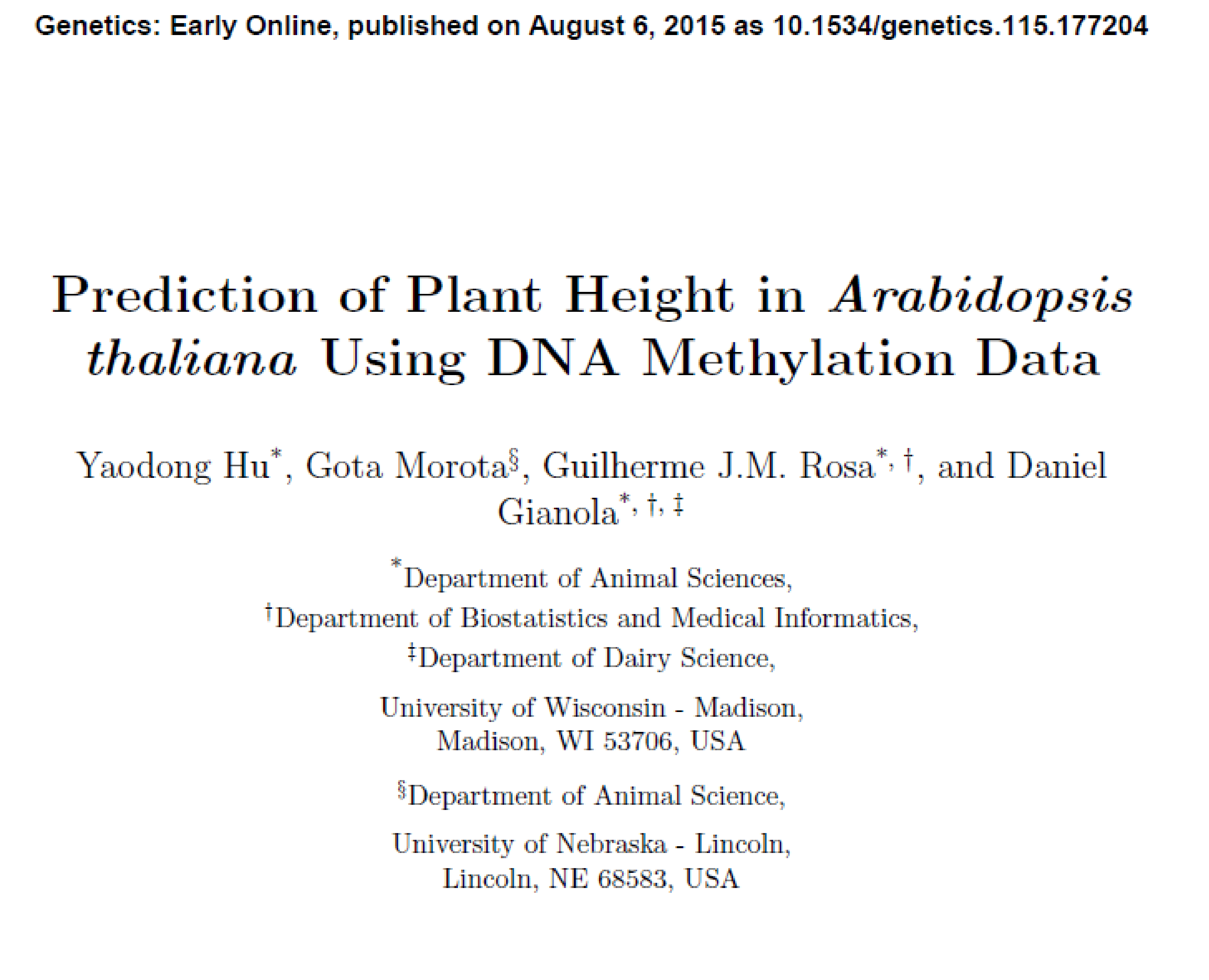
Heather E. Wheeler,[1] Keston Aquino-Michaels,[2] Eric R. Gamazon,[2] Vassily V. Trubetskoy,[2] M. Eileen Dolan,[1] R. Stephanie Huang,[1] Nancy J. Cox,[2] and Hae Kyung Im[3]*

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

ABSTRACT: High-confidence prediction of complex traits such as disease risk or drug response is an ultimate goal of personalized medicine. Although genome-wide association studies have discovered thousands of well-replicated polymorphisms associated with a broad spectrum of complex traits, the combined predictive power of these associations for any given trait is generally too low to be of clinical relevance. We propose a novel systems approach to complex trait prediction, which leverages and integrates similarity in genetic, transcriptomic, or other omics-level data. We translate the omic similarity into phenotypic similarity using a method called Kriging, commonly used in geostatistics and machine learning. Our method called OmicKriging emphasizes the use of a wide variety of systems-level data, such as those increasingly made available by comprehensive surveys of the genome, transcriptome, and epigenome, for complex trait prediction. Furthermore, our OmicKriging framework allows easy integration of prior information on the function of subsets of omics-level data from heterogeneous sources without the sometimes heavy computational burden of Bayesian approaches. Using seven disease datasets from the Wellcome Trust Case Control Consortium (WTCCC), we show that OmicKriging allows simple integration of sparse and highly polygenic components yielding comparable performance at a fraction of the computing time of a recently published Bayesian sparse linear mixed model method. Using a cellular growth phenotype, we show that integrating mRNA and microRNA expression data substantially increases performance over either dataset alone. Using clinical statin response, we show improved prediction over existing methods. We provide an R package to implement OmicKriging (http://www.scandb.org/newinterface/tools/OmicKriging.html).

Genet Epidemiol 00:1–14, 2014. © 2014 Wiley Periodicals, Inc.



Genetics: Early Online, published on August 6, 2015 as 10.1534/genetics.115.177204

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

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

[*]Department of Animal Sciences,
[†]Department of Biostatistics and Medical Informatics,
[‡]Department of Dairy Science,

University of Wisconsin - Madison,
Madison, WI 53706, USA

[§]Department of Animal Science,

University of Nebraska - Lincoln,
Lincoln, NE 68583, USA

Table 2: Comparison between prediction results using all probes and pre-selected probes.

Kernel		All Probes	Contents Rule Probes	CGI Rule Probes
Correlation	$Corr(\boldsymbol{y}, \hat{\boldsymbol{y}})$	0.384	0.398	0.395
	MSE	38.28	37.73	37.83
Gaussian	$Corr(\boldsymbol{y}, \hat{\boldsymbol{y}})$	0.531	0.532	0.531
	MSE	32.16	32.08	32.13

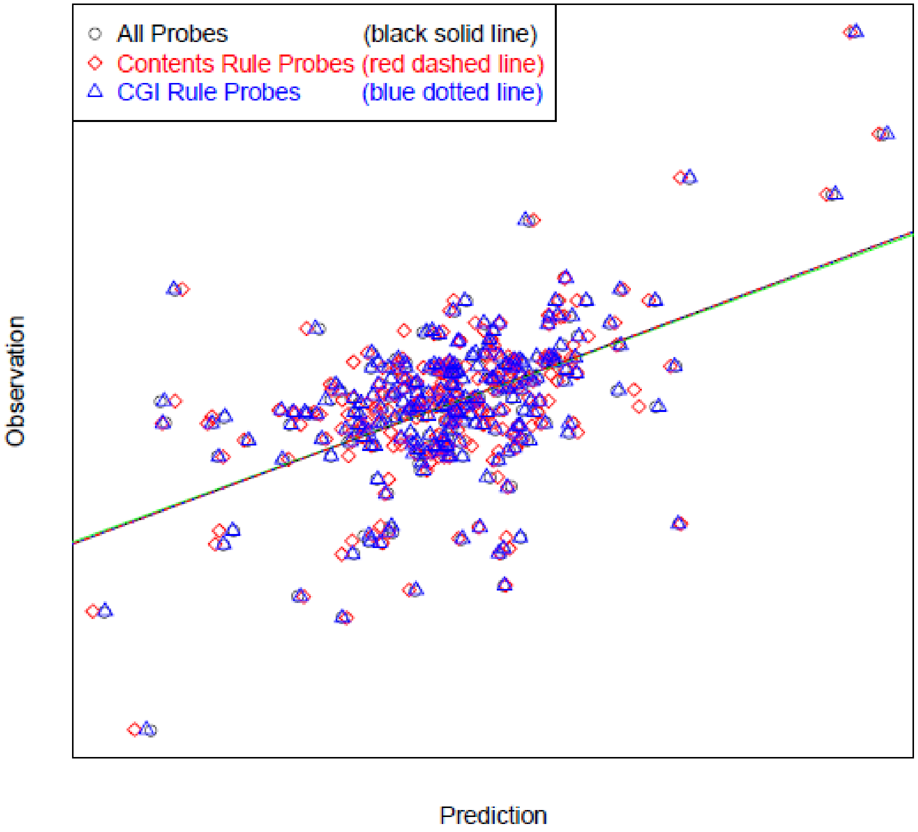


Figure 5: Graphical representation of prediction performance using different set of probes in a Gaussian kernel. The green solid line is a 45° line passing through the origin, the other three lines are fitted lines of regressing observation on predictions. No differences were observed for the three different sets of probes used in prediction.

ENVIRONMENTOMICS

Introducing Highly Dimensional Genomic and Environmental Covariate Data
into Models for Prediction of Complex Traits
[ACTUALLY A RKHS]

Jarquín et al. (Theoretical and Applied Genetics, 2014)

➔Effects of genes on traits modulated by environmental conditions (EC): G×E.

➔Model main and interaction effects of large numbers of markers and of many ECs using co-variance functions. Random effects: all markers, all the ECs and all interactions between markers and ECs.

➔139 wheat lines genotyped with 3,548 SNPs evaluated over 8 years and various locations in northern France. 130 ECs defined

➔Prediction accuracy of models with G X E higher (20%) than main effects only models

➔Capitalize on genomic and environmental information available

KAGEWASO!!

KERNEL- ASSISTED GENOME WIDE ASSOCIATION STUDY

LARSON & SCHAID (2013, Genet. Epidemiol.)

CHEN ET AL. (2012, Genet. Epidemiol.)

SCHIFANO ET AL. (2012, Genet. Epidemiol: pre-select SNP sets and test significance of set variance)

HE ET AL. (2012, Genetica)

HAN (2010, Genet. Epidemiology)

SCHAID ET AL. (2010, Human Heredity)

MUKHOPADHYAY ET AL. (2010, TESTS, Genet. Epidemiology)

PAN (2009 , Genet. Epidemiology, tests)

TENG ET AL. (2009, SIMILARITY METHODS, Biometrics)

KWEE ET AL. (2008, Am J. Human Genetics, tests)

LIU et al. (2007, 2008, Biometrics, BMC Bioinformatics)

Liu et al. (2007) and Schifano et al. (2012) use SNPs in some pathway and put these into a kernel. Then one has a random effects model (recall that α has zero mean)

$$
\begin{aligned}
y &= K\alpha + e \\
V &= KVar(\alpha)K + I\sigma_e^2 \\
&= K\sigma_\alpha^2 + I\sigma_e^2
\end{aligned}
$$

$$
l(\sigma_\alpha^2, \sigma_e^2) \propto |V|^{-\frac{1}{2}} \exp\left[-\frac{y'V^{-1}y}{2\sigma_\alpha^2} \right]
$$

Use score test for "significance"

If “significant”, then use SNPs in pathways for doing something, e.g., a genetic test

PROBLEM: a complex trait is probably affected by ALL pathways. An option might be a Multi-pathway KAGEWASO!

SEQUENCE INFORMATION?

- p/n ratio will go from 50 to 1000-2000
- “All causal mutations there” (Gurus et al., many papers)
- Bayesian alphabet may collapse computationally
- Regression coefficients will be tiny, effectively infinitesimal (can still predict signal, though)
- Advantage of n X n methods?
- “Neo-systems approach”: not very useful in absence of rate coefficients.
- Pigs do not fly (yet)

Remarks

- Inference and prediction: connected but different.
- Kernel based methods attractive not only for prediction but for more properly conducting GWAS.
- Many GWAS will be “GWASHED” away.
- Challenges to parametric methods posed by genomic and post-genomic data.
- Future: analytical shift? Semi-parametric and “machine learning” type techniques?