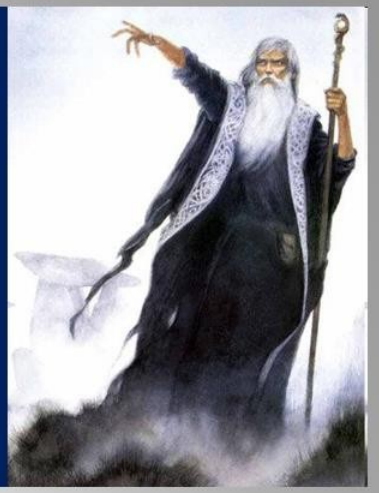
PREDICTION OF COMPLEX TRAITS WITH KERNEL METHODS

what have we learned?



Daniel Gianola

Hans Fischer Senior Fellow TUM-IAS München

Sewall Wright Professor of Animal Breeding and Genetics

University of Wisconsin

Dairy Science

UW–MADISON ANIMAL SCIENCES

Biostatistics & Medical Informatics

University of Wisconsin SCHOOL OF MEDICINE AND PUBLIC HEALTH

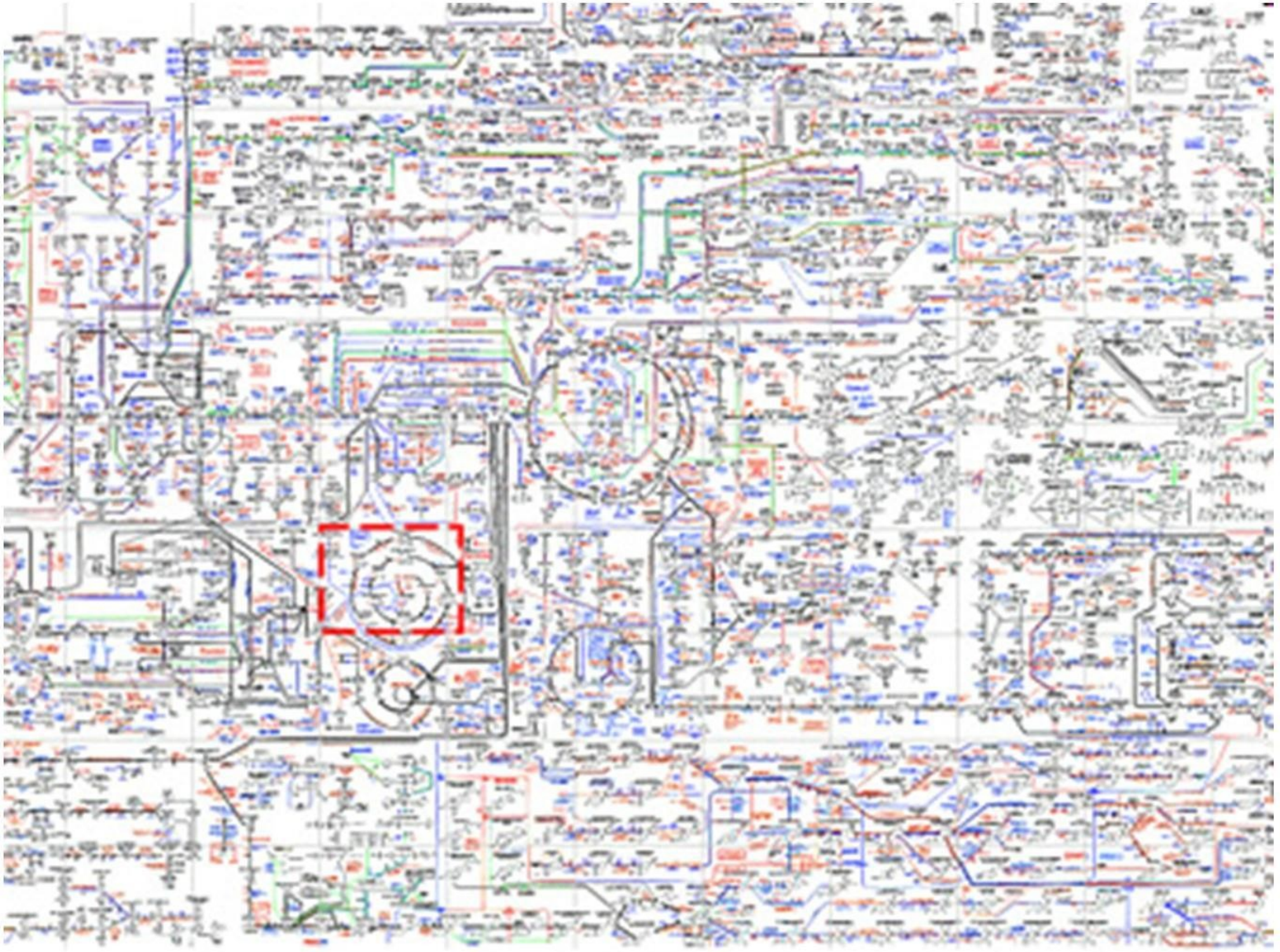
1. GENERAL CONSIDERATIONS IN THE DEVELOPMENT OF PREDICTION MACHINES

Proposition 1

It must be true that quantitative traits are “complex”, in any sense of the word.

Why?

A “complex” trait involves many metabolic pathways: Roche’s Chart

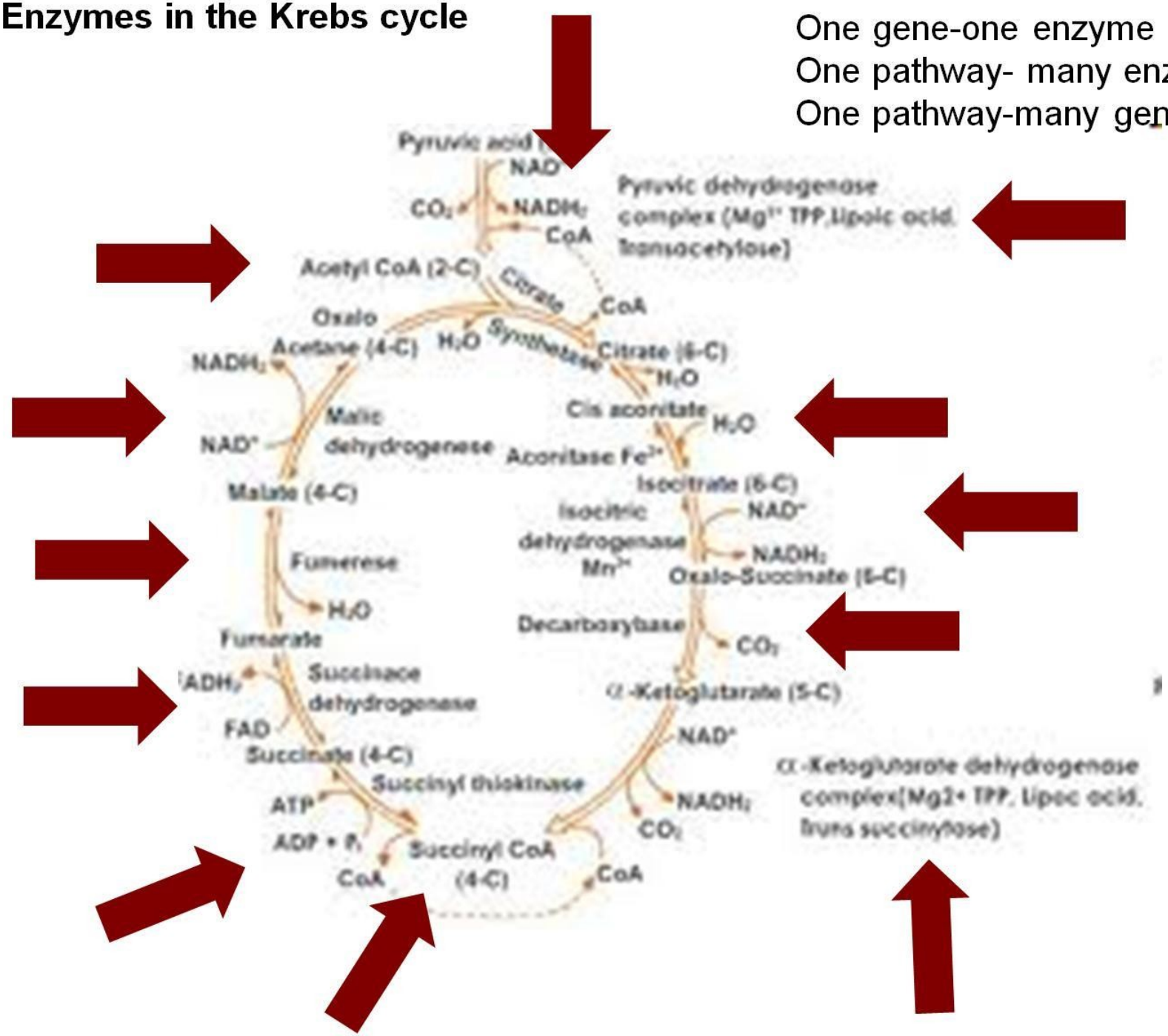


Proposition 2

It must be true that epistasis is pervasive

Enzymes in the Krebs cycle

One gene-one enzyme
One pathway- many enzymes
One pathway-many genes



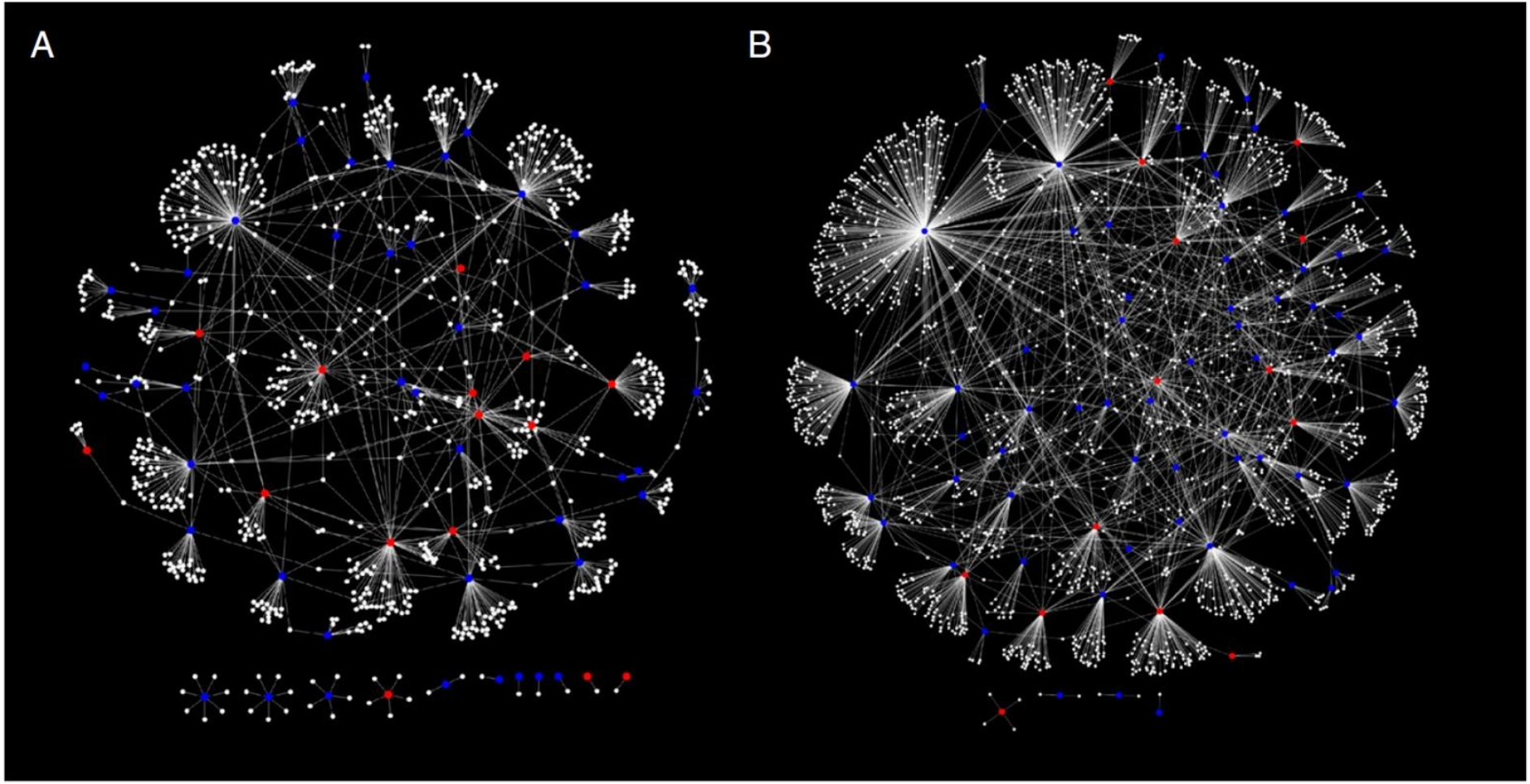


Fig. 5. Networks of epistatic interactions. Interaction networks are depicted for (*A*) starvation resistance and (*B*) chill coma recovery. Nodes depict genes, and edges significant interactions. Red nodes are genes containing significant SNPs from the Flyland analysis. Blue nodes are genes containing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

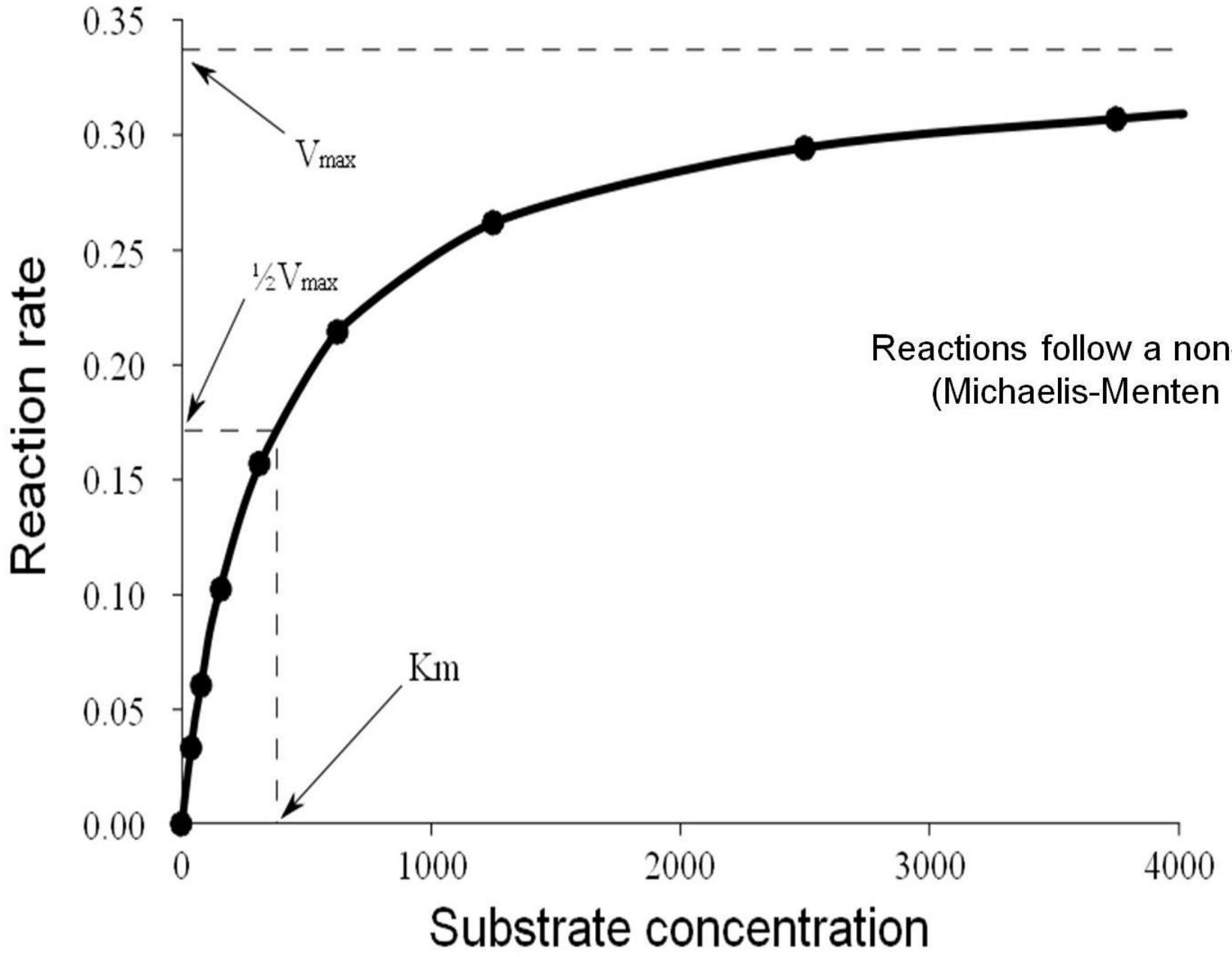
PNAS, 2012

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]

Proposition 3

A phenotype must be the result of a system involving epistasis and non-linearities of all sorts



Reactions follow a non-linear dynar
(Michaelis-Menten kinetics)

2. 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?

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

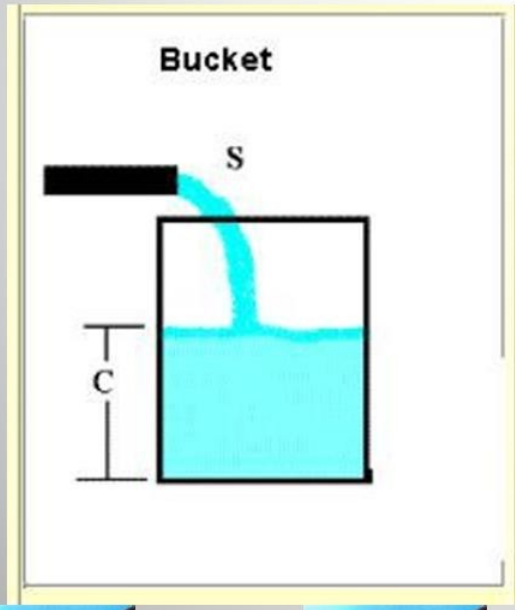
THE CENTRAL DOGMA OF QUANTITATIVE GENETICS:

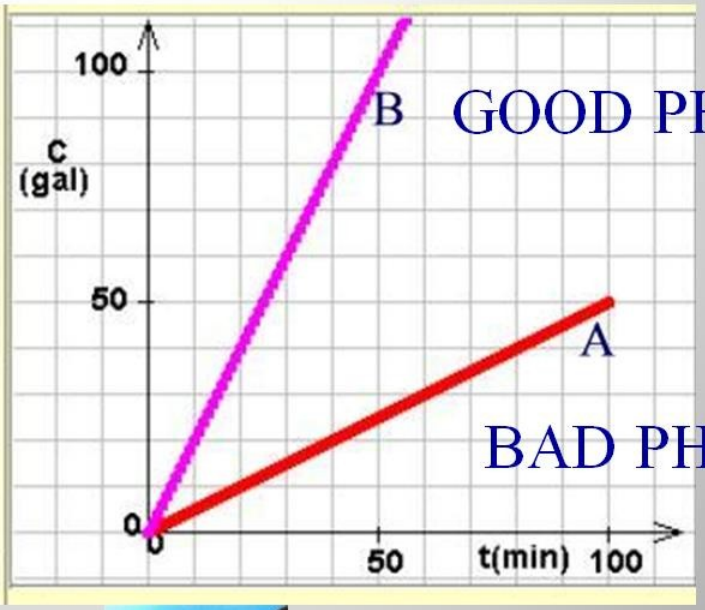
the additive genetic model

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

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

Genome





GOOD PHENOTYPE

BAD PHENOTYPE

= **'additive genetic value'**

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.

THEN, THIS PLACES AN UPPER LIMIT TO THE THEORY OF QUANTITATIVE GENETICS FOR DISCOVERY PURPOSES, AS EVERYTHING WILL TURN OUT TO BE ADDITIVE...

THE AGE OF INNOCENCE

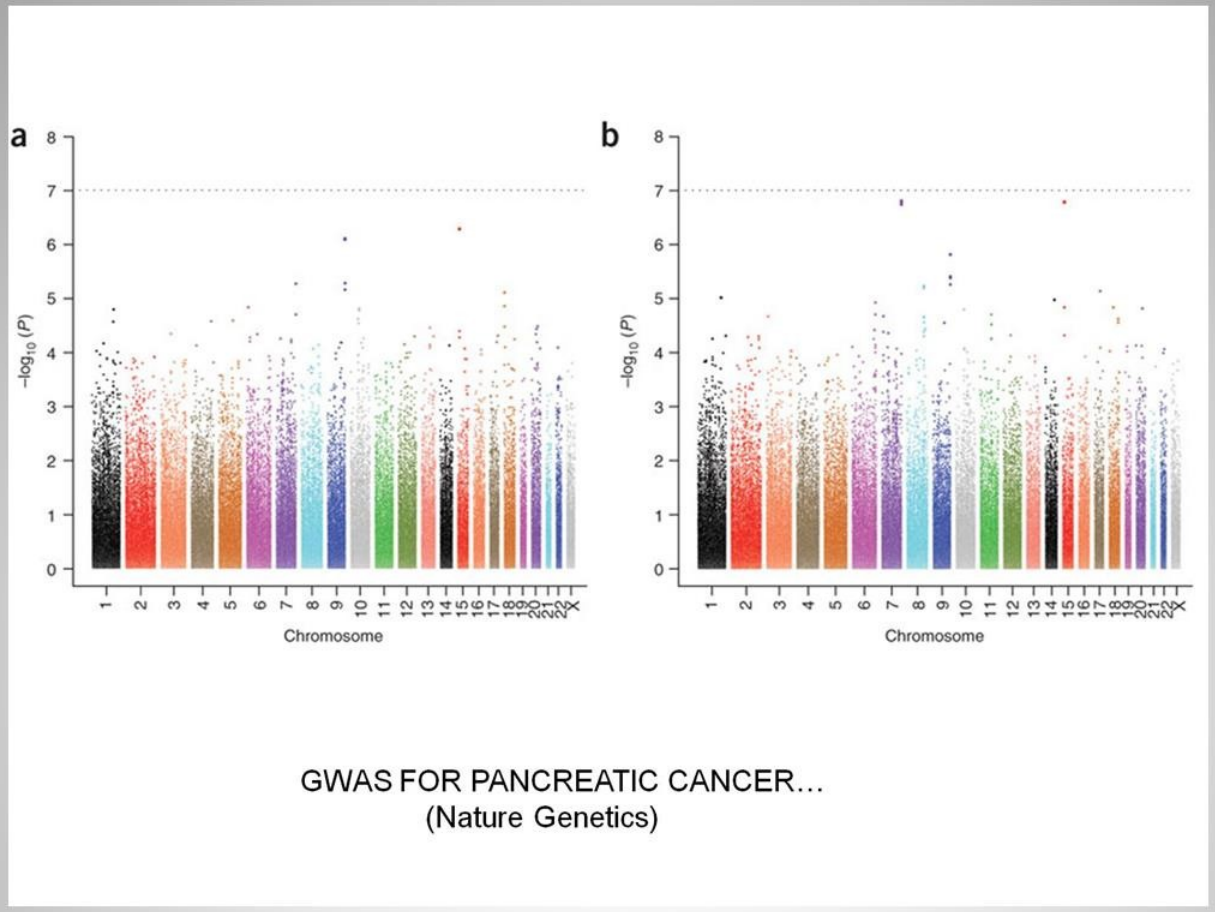
(issue)

Unraveling “genetic architecture”
with statistical models

ABSTRACTION PARADIGM 1

GWAS: ***search for association between some marker or genomic region and a phenotype***

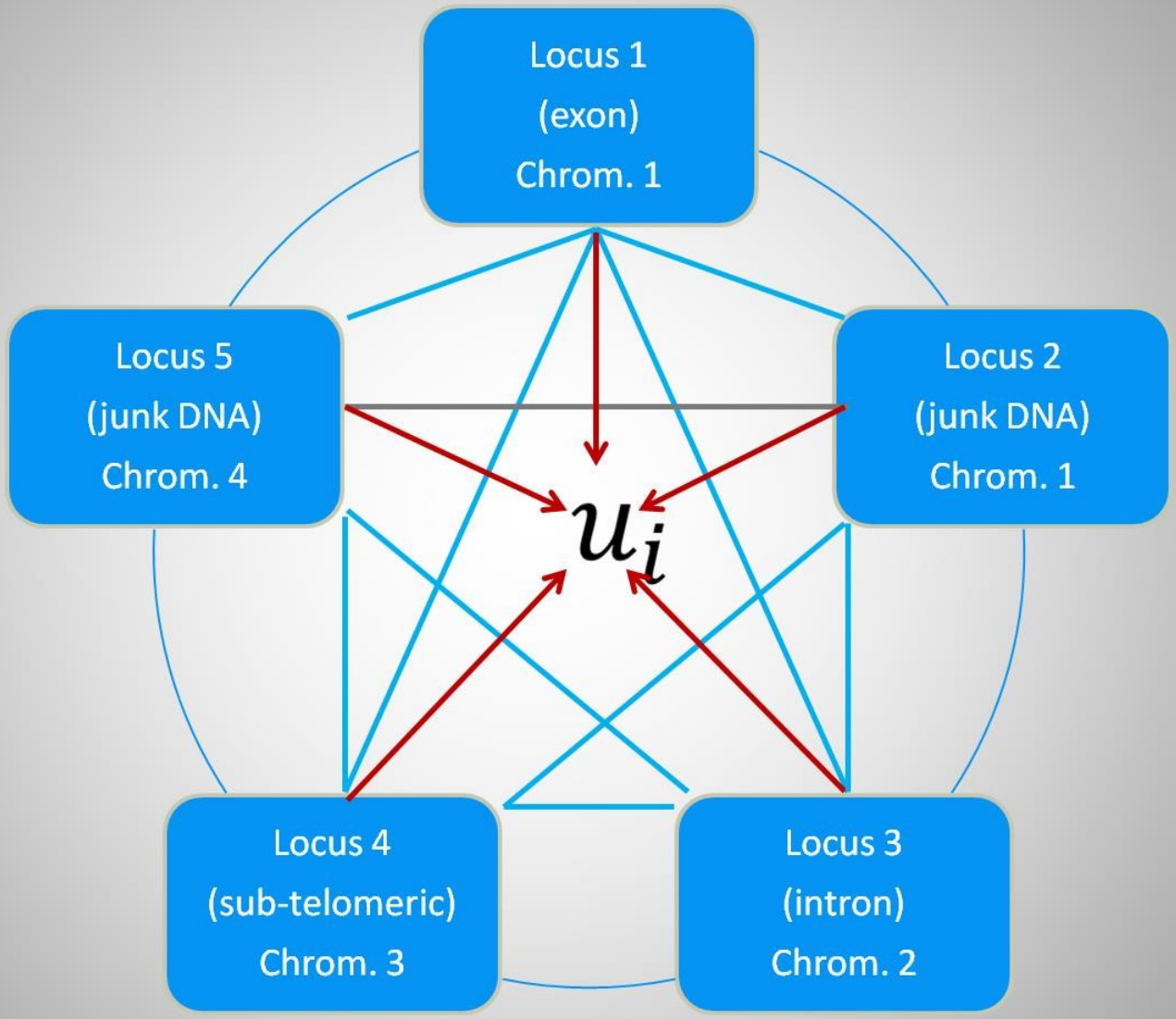
EXAMPLES



GWAS FOR PANCREATIC CANCER…
(Nature Genetics)

Genome-Wide Association Study to Identify Single Nucleotide Polymorphisms (SNPs) Associated With the Development of Erectile Dysfunction in African-American Men After Radiotherapy for Prostate Cancer. *International Journal of Radiation Oncology, Biology, Physics 2010*

Figure 1. Five locus system in linkage disequilibrium. Arrows represent direct effects on additive genetic value (*u*) ; undirected lines and arcs represent correlations between genotypes stemming from linkage disequilibrium.



$$u = QTL_1 + QTL_2 + \cdots + QTL_5$$

How many QTLs? “Honey I shrunk epistasis!”

SNPs

Statistical
QTL chaser

Gene

ABSTRACTION PARADIGM 2

Fisher's infinitesimal model of additive effects
(extended vectorially by C. R. Henderson, animal breeder)

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance.

Transactions of the Royal Society of Edinburgh, *52*, 399-433.

EMULATE FISHER'S MODEL USING MOLECULAR MARKERS

A (slightly) less naïve form of approximating G is the whole-genome linear model:

$$G = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_px_p$$

Where the x's are either pedigree relationships, or marker genotype codes or whatever the latest fad in genomic data is

Bayes A
Bayes B
Bayes C (with or without π)
Bayesian Lasso
NON-BAYESIAN REGULARIZED: Lasso, Elastic Net

LEADS TO (EXTRAORDINARILY) SHRUNKEN ESTIMATES OF EFFECTS, BUT GOOD PREDICTIONS OF "TOTAL SIGNAL"

PARADIGM 2 IS NAÏVE FOR DISCOVERY

-IT PRODUCES (CONDITIONALLY) BIASED AND INCONSISTENT ESTIMATES

-ORDER PIZZA FOR 500 AND 1 MILLION EAT…

-THERE IS AN IDENTIFICATION PROBLEM IN THE n<<p CASE. NOT TRUE THAT DIFFERENT BAYESIAN MODELS (A, B, C,…, ETC.) ARE INFORMATIVE ABOUT "GENETIC ARCHITECTURE"

- AT BEST PRODUCES A LOCAL APPROXIMATION TO EPISTASIS

Dealing with epistatic interactions and non-linearities

gene x gene

gene x gene x gene

gene x gene x gene x gene

.....

(Alice in Wonderland)

DO THESE ASSUMPTIONS HOLD?

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
- ❑Linkage equilibrium+ no linkage

ALL ASSUMPTIONS VIOLATED!

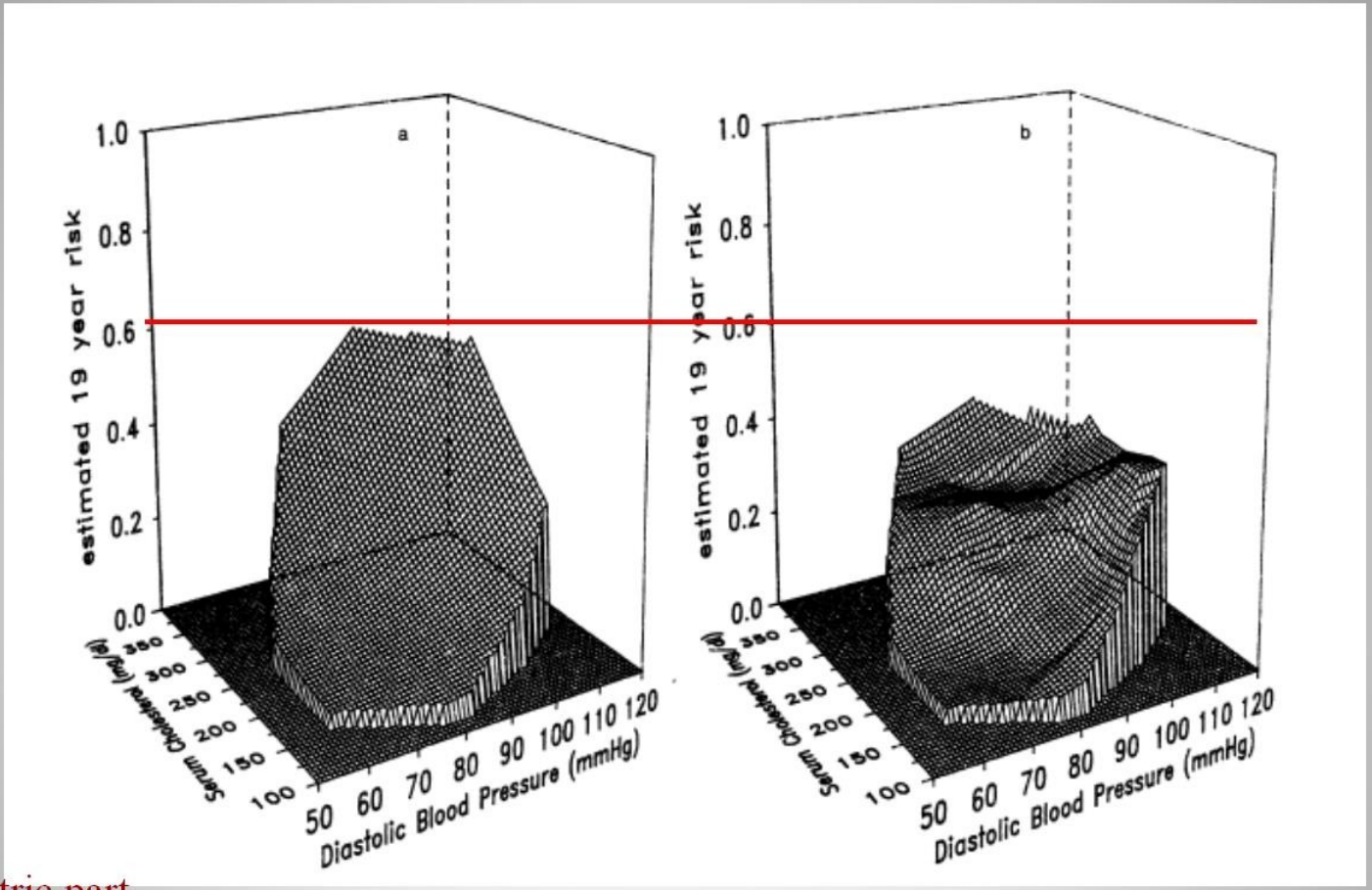
PARADIGM 3

(machine learning: 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

Logistic regression with thin-plate splines



parametric part

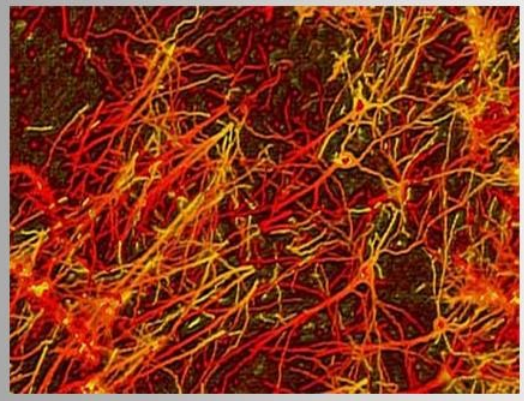
$$f(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \sum_{j=1}^{N} \alpha_j \left[\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 \right] \log\left[\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 \right]$$

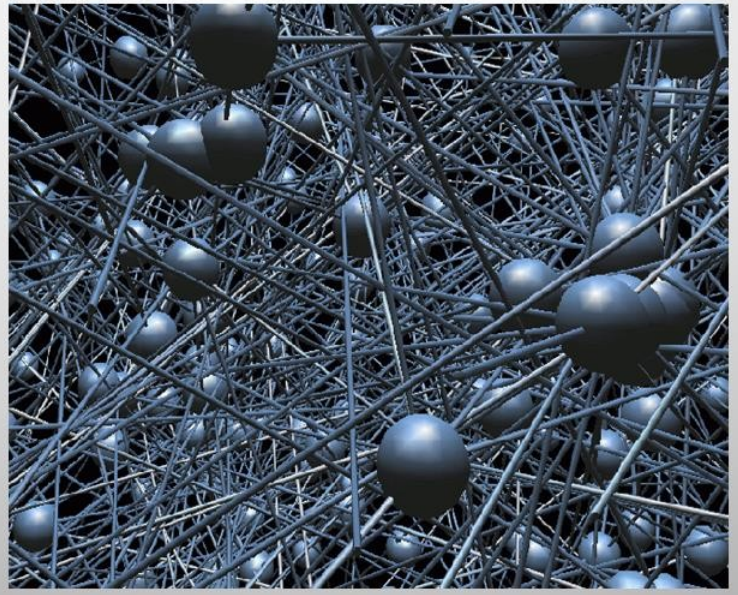
Risk of heart attack after 19 years as a function of cholesterol level and blood pressure. Left: logistic regression model. Right: thin plate spline fit. Wahba (2007)

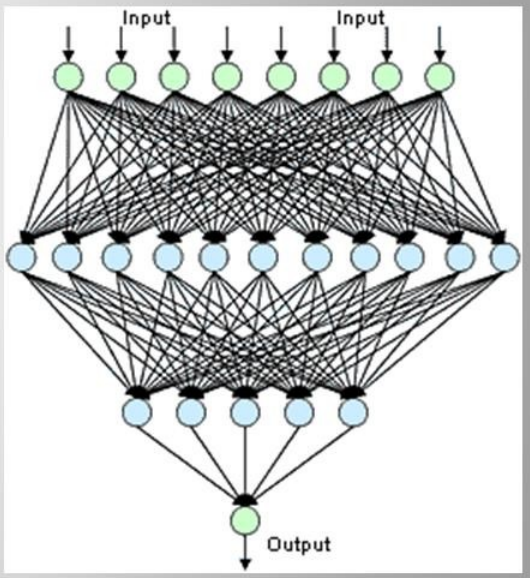
OUR EXPERIENCE WITH PARADIGM3

More universal prediction machines:

A. Regularized Neural Networks







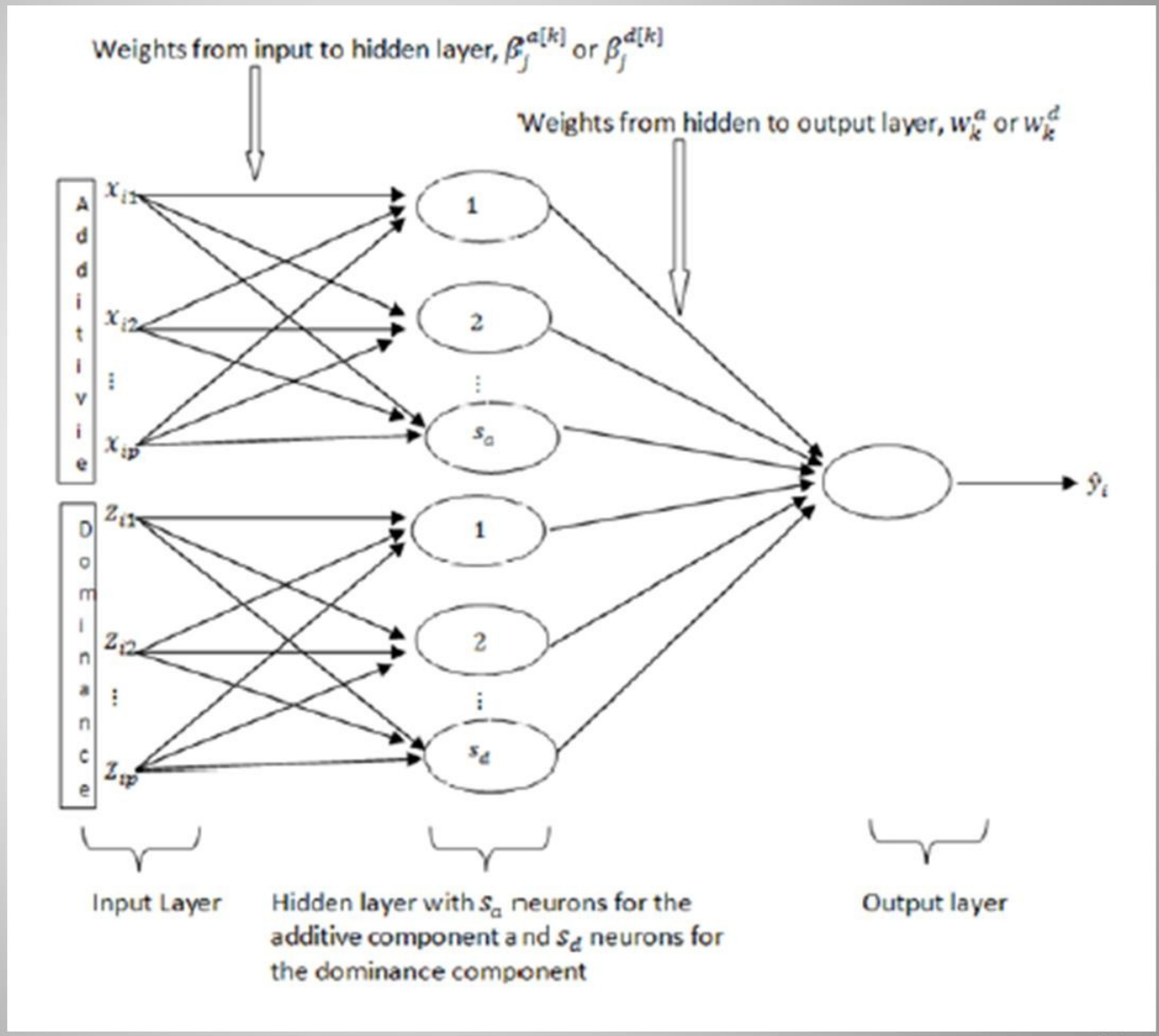
CROSS-VALIDATION (CV)

- Data available (genomic, phenotypic)
- Data generated according to unknown process
- Split into training (fitting)- testing (predictand) sets
- Fitting process essentially describes current data (model is typically wrong)
- Use training process to make statement about yet-to-be observed data (testing set)
- Prediction error (conditional and unconditional): point estimate
- Distribution of prediction errors (conditional or unconditional): interval estimate. For this, CV must be replicated

NETWORK DESIGN: THE POLI-PERCEPTRON (PPNN)
"Paulino Perez Neural Net"

Additive marker codes (or relationships)

Dominance marker codes (or relationships)



Including Additive and Dominance effects together in a NN.

- The data equation is given by:

$$y_i = \mu + c_a \sum_{k=1}^{s_a} w_k^a g_k \left(b_k^a + \sum_{j=1}^{p} a_{ij} \beta_j^{a[k]} \right) + c_d \sum_{k=1}^{s_d} w_k^d g_k \left(b_k^d + \sum_{j=1}^{p} d_{ij} \beta_j^{d[k]} \right) + e_i$$

Additive effects

ξ^2A=regularization of additive effects

Dominance effects

ξ^2D=regularization of dominance effects

Table 1: Correlations between observed and predicted values of milk yield in a testing set of Jersey cows.

fold	Additive		Additive + Dominance	
	brnn-1-2-1	fbm-1-2-1	brnn-1-2-2-1	fbm-1-2-2-1
1	0.3455	0.4795	0.2715	0.4438
2	0.5211	0.5860	0.5546	0.4681
3	0.2553	0.4086	0.4264	0.5407
4	0.5241	0.6878	0.5051	0.7222
5	0.1118	0.2707	0.0989	0.3444
6	0.3821	0.5105	0.4211	0.5551
7	-0.1365	-0.1662	-0.0756	-0.0758
8	0.4533	0.6542	0.5387	0.7401
9	0.1361	0.0400	0.1974	-0.0800
10	0.4334	0.5411	0.4920	0.6106
Avg. correlation	**0.3026**	**0.4012**	**0.3430**	**0.4269**
Avg. RMSE	**2498.17**	**1768.63**	**2392.10**	**1609.48**

- brnn-1-2-1 Bayesian Regularized Neural Network with 2 neurons for the additive component
- fmb-1-2-1 Flexible Bayesian Modeling (Neal's model) with two neurons for the additive component.
- brnn-1-2-2-1 Bayesian Regularized Neural Network with 2 neurons for each of the additive and dominance components.
- fbmnn-1-2-2-1 Flexible Bayesian Modeling (Neal's model) with two neurons for each of the additive and dominance components.

fmb better than brnn

Reasons?

-Failure of Laplace's approximation

-Stuck at local mode?

-MCMC explores the entire space, contrary to MAP, which uses a single point (dangerous if there is multimodality)

B. Reproducing Kernel Hilbert spaces mixed model regression

Function of molecular information **x (vector of SNP variables)**

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

Smoothing parameter (λ)

"Penalized sum of squares"

Some norm under Hilbert space (*H*) of functions

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

Solution to variational problem: linear function

No. individuals with molecular data

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

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

Regression coefficient

Reproducing kernel

Example of reproducing kernel:

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

Can have more than 1 bandwidth parameter

Penalized estimation

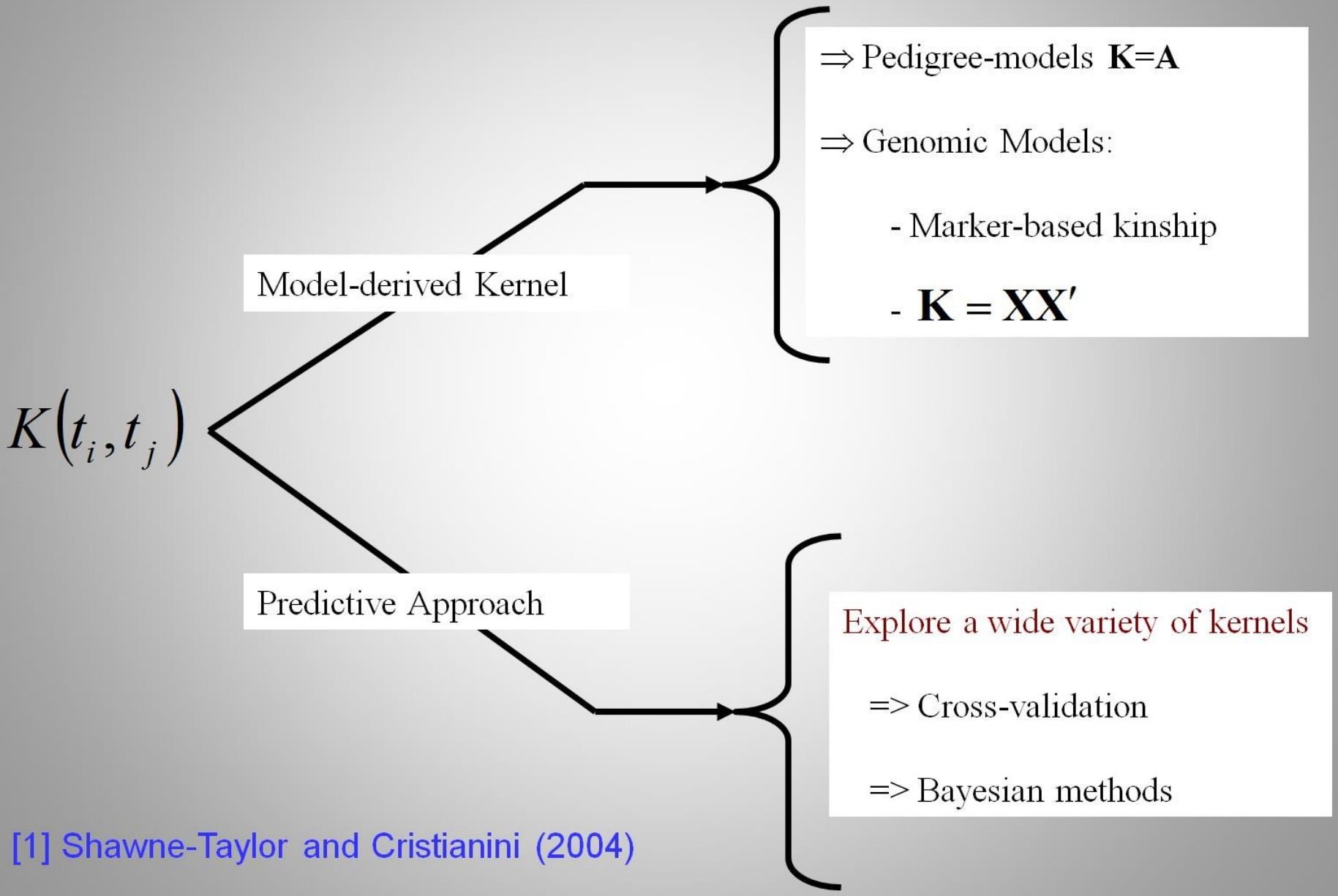
$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ (\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} \right\}$$

Bayesian View

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

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

How to Choose the Reproducing Kernel? [1]



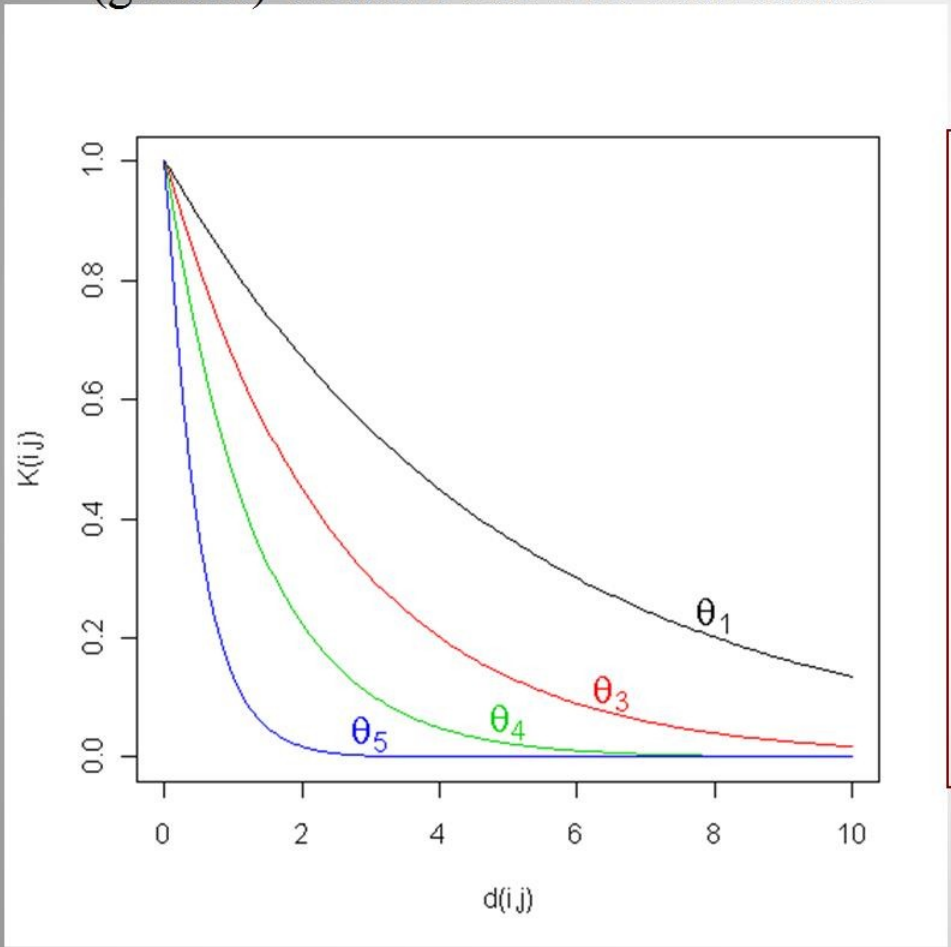
[1] Shawne-Taylor and Cristianini (2004)

Choosing the RK based on predictive ability

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

(genetic) distance between individuals

$$\Rightarrow \quad K(i,j|\theta) = Exp\{ -\theta \times d(\mathbf{x}_i, \mathbf{x}_j) \}$$



Strategies

- Grid of Values of θ + CV
- Fully Bayesian: assign a prior to θ (computationally demanding)
- Kernel Averaging [1] ("Multi-kernel")

$$K(i,j) = \alpha_1 K(i,j|\theta_1) + (1-\alpha_1) K(i,j|\theta_5)$$

Actually, this means:

$$y = K_1\alpha_1 + K_2\alpha_2 + e$$
$$\alpha_1 \sim N(0, inv(K_1)Var(\alpha_1))$$
$$\alpha_2 \sim N(0, inv(K_2)Var(\alpha_2))$$

[1] de los Campos et al. (2010) Genetics Research

SOME CASE STUDIES WITH RKHS

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.

REFINING THE INFORMATION FROM MARKERS

DIFFERENTIAL CONTRIBUTION OF MARKERS: GBLUP (also a kernel method) IN CHICKENS

Journal of
Animal Breeding and Genetics

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

ORIGINAL ARTICLE

Effect of allele frequencies, effect sizes and number of markers on prediction of quantitative traits in chickens

R. Abdollahi-Arpanahi[1], A. Nejati-Javaremi[1], A. Pakdel[1], M. Moradi-Shahrbabak[1], G. Morota[2], B.D. Valente[2,3], A. Kranis[4,5], G.J.M. Rosa[2,6] & D. Gianola[2,3,6]

1 Department of Animal Science, University College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran
2 Department of Animal Sciences, University of Wisconsin, Madison, WI, USA
3 Department of Dairy Science, University of Wisconsin, Madison, WI, USA
4 Aviagen Ltd, Midlothian, UK
5 The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, UK
6 Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA

Table 2 Predictive ability estimated by genomic best linear unbiased prediction (GBLUP) using single nucleotide polymorphisms (SNPs) binned based on minor allele frequency (MAF) for body weight (BW), ultra-sound of breast muscle (BM), and hen house egg production (HHP) traits in broiler chickens

	BW		BM		HHP	
MAF bin	COR1 ± SE2	h^2_{TRN}3	COR ± SE	h^2_{TRN}	COR ± SE	h^2_{TRN}
0.01–0.09	0.28[a]4 ± 0.002	0.29	0.30[b] ± 0.002	0.28	0.22[b] ± 0.002	0.18
0.09–0.20	0.26[c] ± 0.002	0.26	0.33[a] ± 0.001	0.31	0.23[a] ± 0.001	0.19
0.20–0.29	0.23[d] ± 0.002	0.23	0.29[c] ± 0.001	0.28	0.20[d] ± 0.002	0.17
0.29–0.40	0.21[e] ± 0.003	0.20	0.29[c] ± 0.001	0.27	0.18[e] ± 0.003	0.13
0.40–0.50	0.24[d] ± 0.004	0.19	0.29[c] ± 0.001	0.24	0.16[f] ± 0.003	0.14
All markers	0.27[b] ± 0.002	0.30	0.33[a] ± 0.001	0.33	0.21[c] ± 0.002	0.19

[1]Correlation between genomic predicted breeding values and corrected phenotypes in testing set.

[2]Standard error.

[3]'Genomic' heritability estimated in the data used to train the model.

[4]Different superscript letters indicate significant differences ($p < 0.05$).

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.

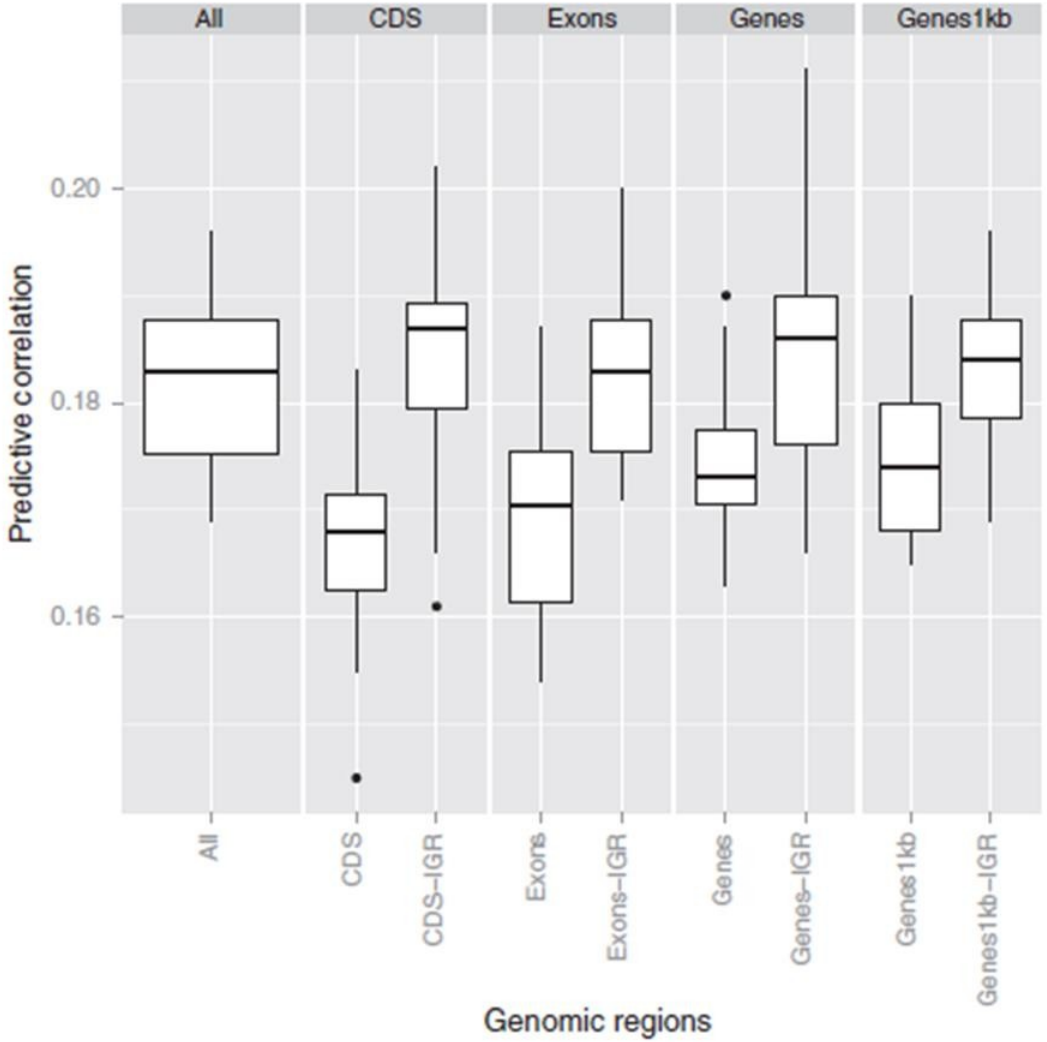


Figure 1 Predictive correlations comparing genic and non-genic regions for BW using kernel-based Bayesian ridge regression. The results were based on 10 fold cross-validation with 15 replications for each genomic region. Genic regions were coding DNA sequences (CDS), exons, genes, and genes with 1kb 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.

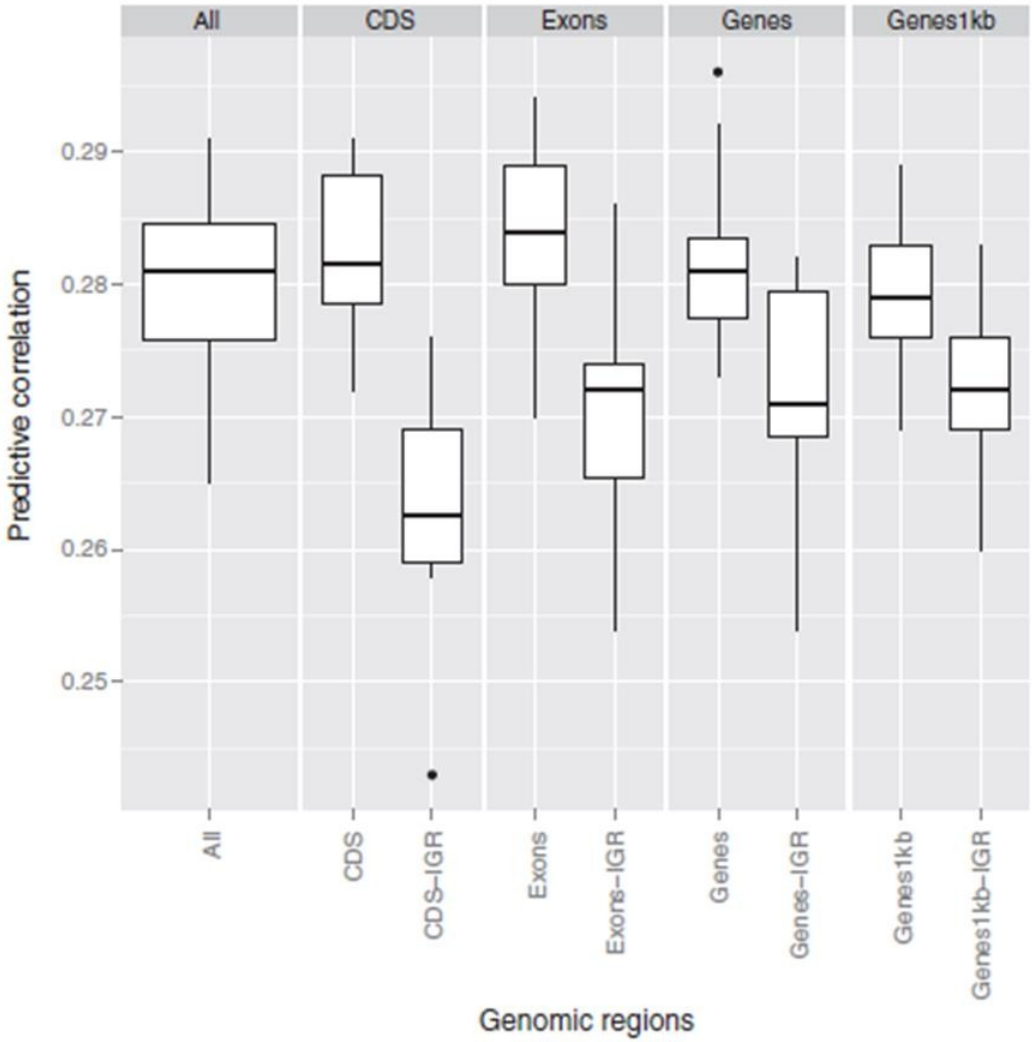


Figure 2 Predictive correlations comparing genic and non-genic regions for BM using kernel-based Bayesian ridge regression. The results were based on 10 fold cross-validation with 15 replications for each genomic region. Genic regions were coding DNA sequences (CDS), exons, genes, and genes with 1kb 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.

DOES MODEL AVERAGING HELP?

(in theory, this is 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

Comparison of methods for predicting litter size

Tusell et al. (2013, 2014)

Line A	Line B	Line AB
2,598 PB	1,604 PB	1,879 PB
46,855 SNPs	45,597 SNPs	50,151 SNPs

Phenotypic data

Average number of piglets born (PB) over parities

Pre-corrected by some environmental effects:

- farm*line*parity, farm*year*number of services,
- farm type, farm*month, age at first farrowing

Genomic data

Illumina PorcineSNP60 BeadChip.

SNPs excluded if:

- MAF < 0.05
- call rate > 0.95

Missing genotypes imputed from average allele frequencies at each locus.

11 methods compared including RKHS and NN (Neural nets)

RKHS-KERNEL AVERAGING-MODEL AVERAGING

• It is **theoretically** possible to enhance ANY predictive model by using Bayesian Model Averaging:

"Predictions obtained by averaging over models are better, on average, than predictions from single model, even the "best" ".

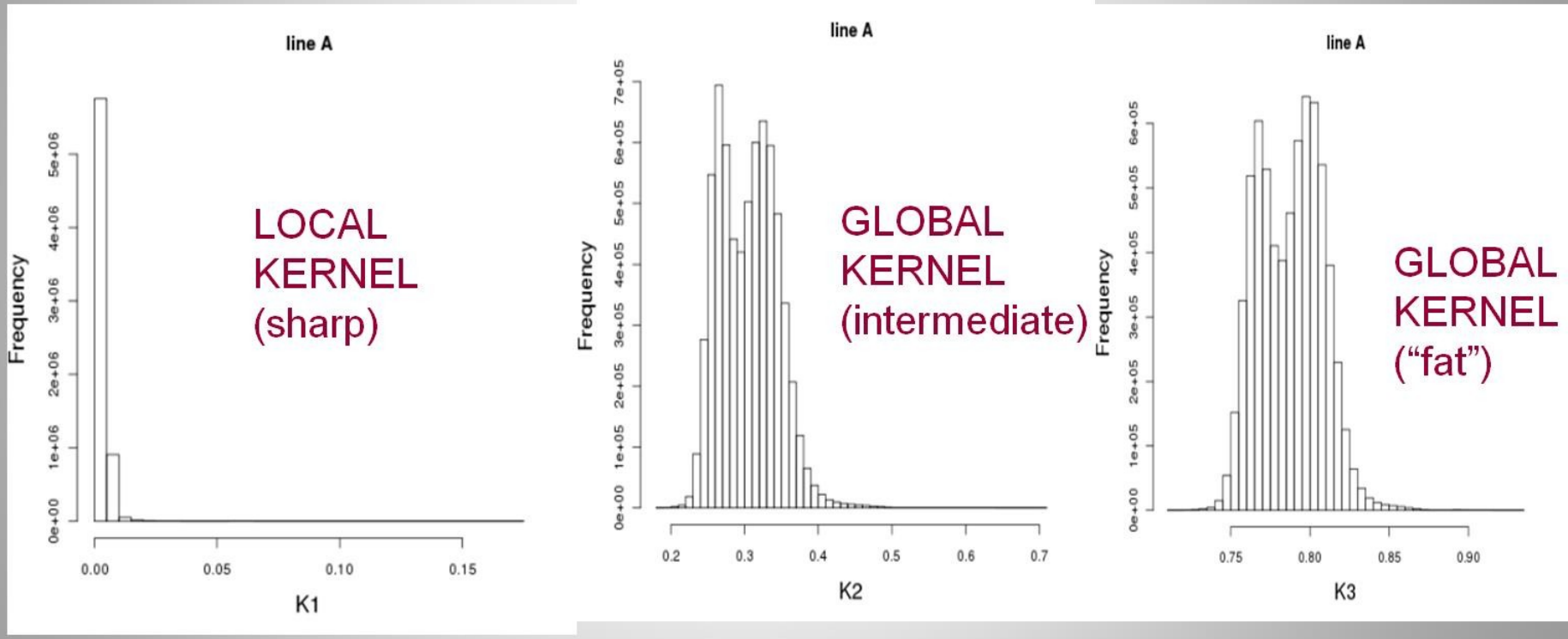
WELL KNOWN THEORETICAL RESULT IN BAYESIAN MODEL AVERAGING

•Example: with 3 bandwidths for Gaussian kernels, we can have predictions based on the following models:

1 : RKHS with K_1
2 : RKHS with K_2
3 : RKHS with K_3
4 : RKHS-KA with K_1, K_2
5 : RKHS-KA with K_1, K_3
6 : RKHS-KA with K_2, K_3
7 : RKHS- KA with K_1, K_2, K_3
8 : Average of predictions from models 1 to 7
8*: Weighted average from model 1 to 7 according to harmonic mean of $log\left[\hat{p}(\mathbf{y}|M_i)\right]$

(See Sorensen & Gianola, 2002.)

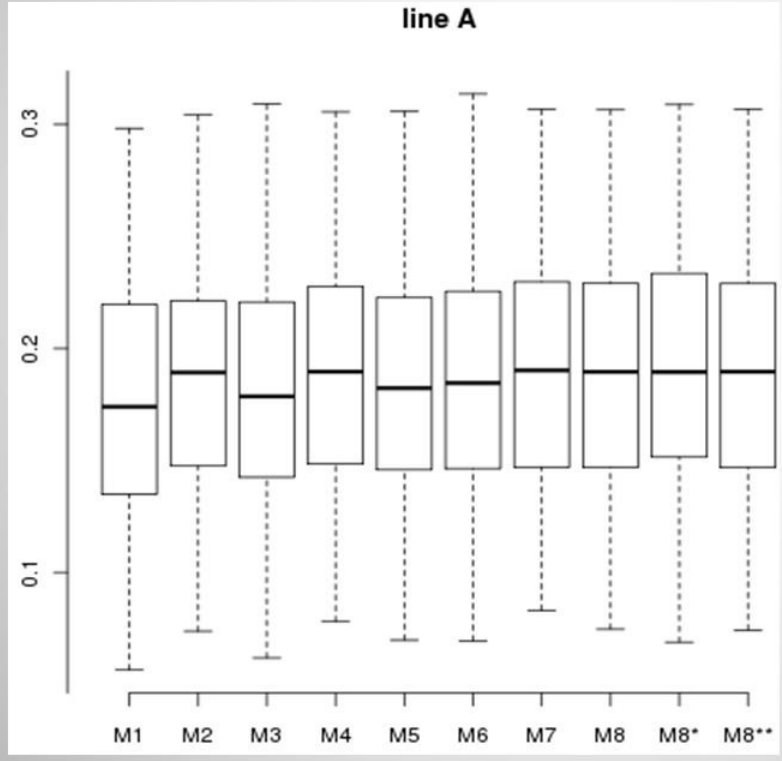
PIC dataset for line A (litter size): GAUSSIAN KERNEL AT 3 BANDWIDTHS



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

Continued...

Predictive ability:
-50 random partitions with 90% of observations in training and 10% in testing
-Correlations between observed and predicted phenotypes..



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

Distribution of correlations between observed and predicted phenotypes.

AVERAGING PREDICTIONS NOT WORSE THAN BMA; NOISY DATA

Comparison among methods in plants (Heslot et al., 2012)

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

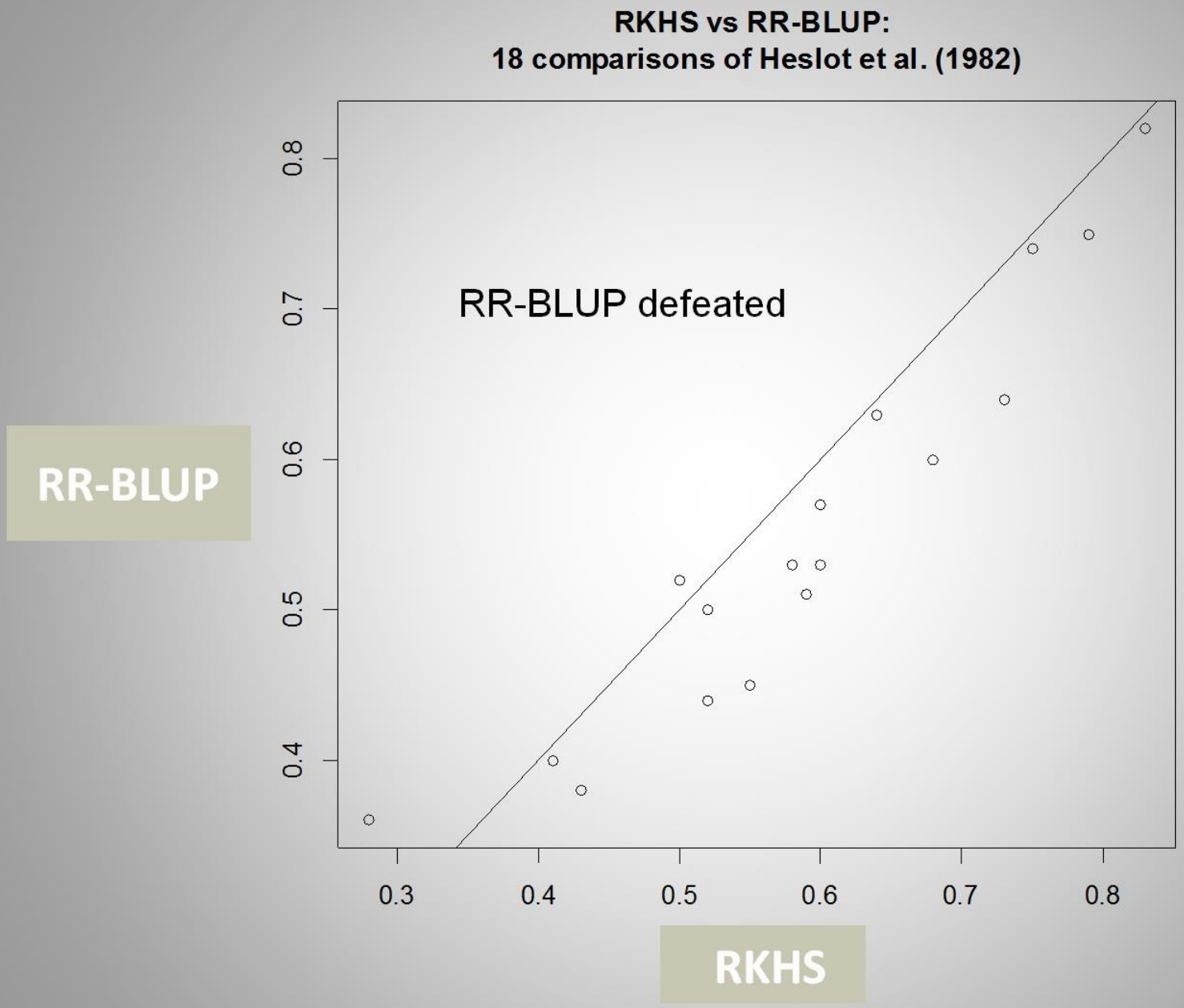
†Barley 1, Limagrain Europe, Riom, France; Barley CAP (Barley Coordinated Agricultural Project, 2011); Bay Sha (Loudet et al. 2002); Panel maize, Limagrain Europe; Diallel maize, Limagrain Europe; Wheat CIMMYT (Crossa et al., 2010); Wheat Cornell (Heffner et al., 2011); Wheat diallel, Limagrain Europe.

‡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.

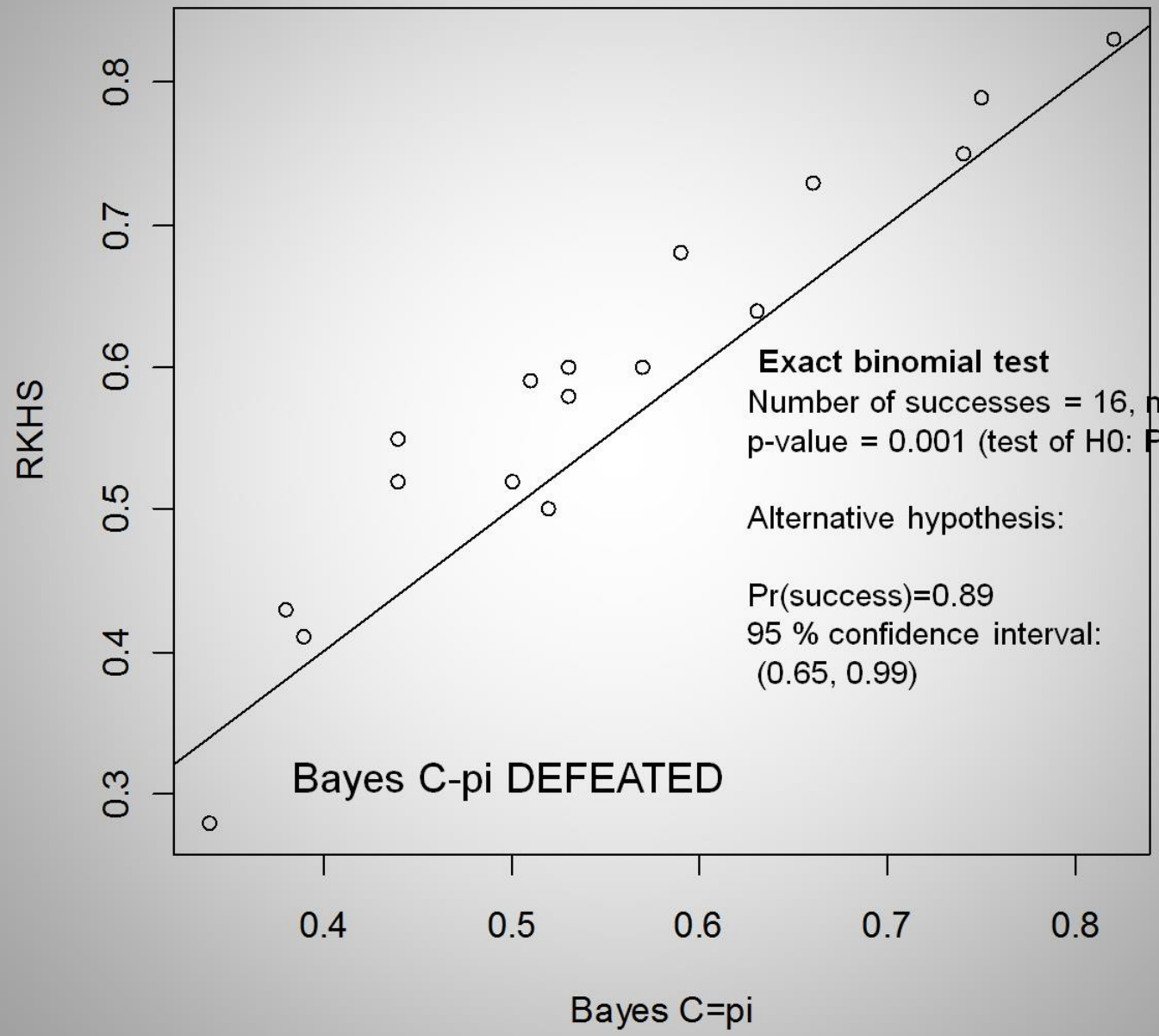
TENTATIVE CONCLUSION: choice of method does not make a difference, in practice

Really?

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

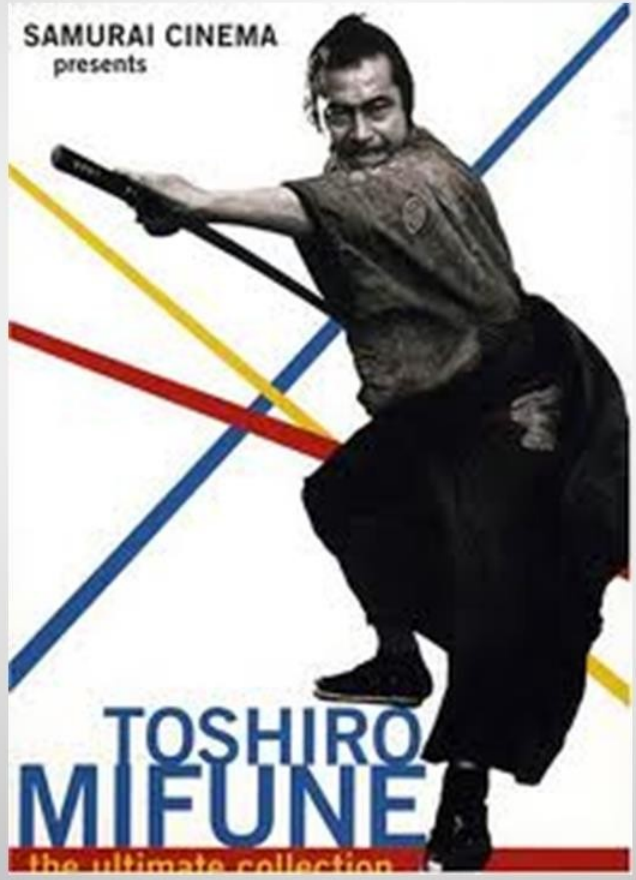

RKHS vs RR-BLUP:
18 comparisons of Heslot et al. (1982)
RR-BLUP defeated
RR-BLUP
RKHS
0.3
0.4
0.5
0.6
0.7
0.8
0.4
0.5
0.6
0.7
0.8

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



FURTHER DOWN THE ROAD

KAGEWASO!!



KAGEWASO: KERNEL- ASSISTED GENOME WIDE ASSOCIATION STUDY

- SCHIFANO ET AL. (2012, Genet. Epidemiology: pre-select SNP sets and test significance of set variance)
- HE ET AL. (2012, Genetica, suspiciously similar to SCHIFANO)
- 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)

- DE LOS CAMPOS ET AL. (2010, Gen. Res.)
- GIANOLA AND DE LOS CAMPOS (2008, Gen. Res.)
- GIANOLA AND VAN KAAM (2008, Genetics)
- GIANOLA ET AL. (2006, Genetics)

NOTE: WE DO NOT EXIST FOR STATISTICIANS, WITH EXCEPTIONS.

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.

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)

➔In most agricultural crops effects of genes on traits are modulated by environmental conditions, leading to large extents of genetic by environmental interaction (G×E).

➔Proposed modeling main and interaction effects of large numbers of genetic markers and of large numbers of ECs using co-variance functions. r random effects model on all the markers, all the ECs and all interactions between markers and ECs.

➔139 wheat lines genotyped with 3,548 SNPs and evaluated for grain yield over 8 years and various locations within northern France. A total of 130 ECs defined based on 5 phases of the phenology of the crop were recorded

➔Prediction accuracy of models including interaction terms was substantially higher (20%) than that of models based on main effects only

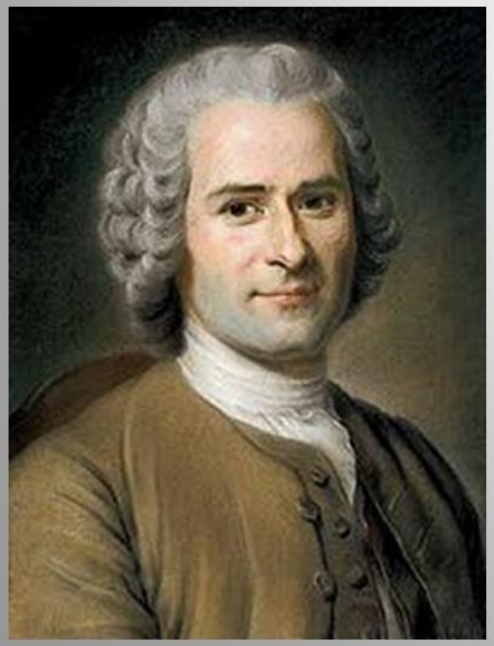
➔Capitalize upon wealth of genomic and environmental information available

SOME POSTERIOR THOUGHTS

- Cannot understand complexity ("genetic architecture") with parametric methods, especially if n<<p
- Prediction is a different ball game from inference
- For prediction, non-parametric methods almost as good as parametric ones even when assumptions hold and seemingly better otherwise
- Do not spend a lot of time inventing priors, or fancy models. A simple additive model may just do well…
- Spend more time in cross-validation and less in simulation (QTL saga…). Now there is data!!
- No universal prediction machine. Model performance varies with species, trait and environment.

ROUSSEAU ON THE ADDITIVE GENETIC MODEL

"…denier ce que est, et d'expliquer ce qui n'est pas…"
Rousseau "Nouvelle Heloise"



Geneve 1712- Ermenonville 1778

"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*

Conclusions

- Challenges to parametric methods posed by genomic and post-genomic data
- Many GWAS will be “GWASHED” away
- Kernel based methods are attractive not only for prediction but for properly conducting GWAS.
- Future: Shift in paradigm. Semi-parametric and “machine learning” type techniques?

FOULLEY ET MOI

Gianola D, Foulley JL. **Nonlinear prediction of latent genetic liability with binary expression: an empirical Bayes approach.**
In: **Proc. 2nd World Congr. Genet. Appl. Livest. Prod..** Madrid, Spain VII. 1982; p. 293

QUELQUES PUBLICATIONS ENSEMBLE

Prediction of breeding values when variances are not known D Gianola, JL Foulley, RL Fernando Génétique sélection évolution 18 (4), 485-498	84	1986

Computing aspects of a nonlinear method of sire evaluation for categorical data I Misztal, D Gianola, JL Foulley Journal of Dairy Science 72 (6), 1557-1568	110	1989

Génét. Sél. Evol., 1983, **15** (2), 201-224

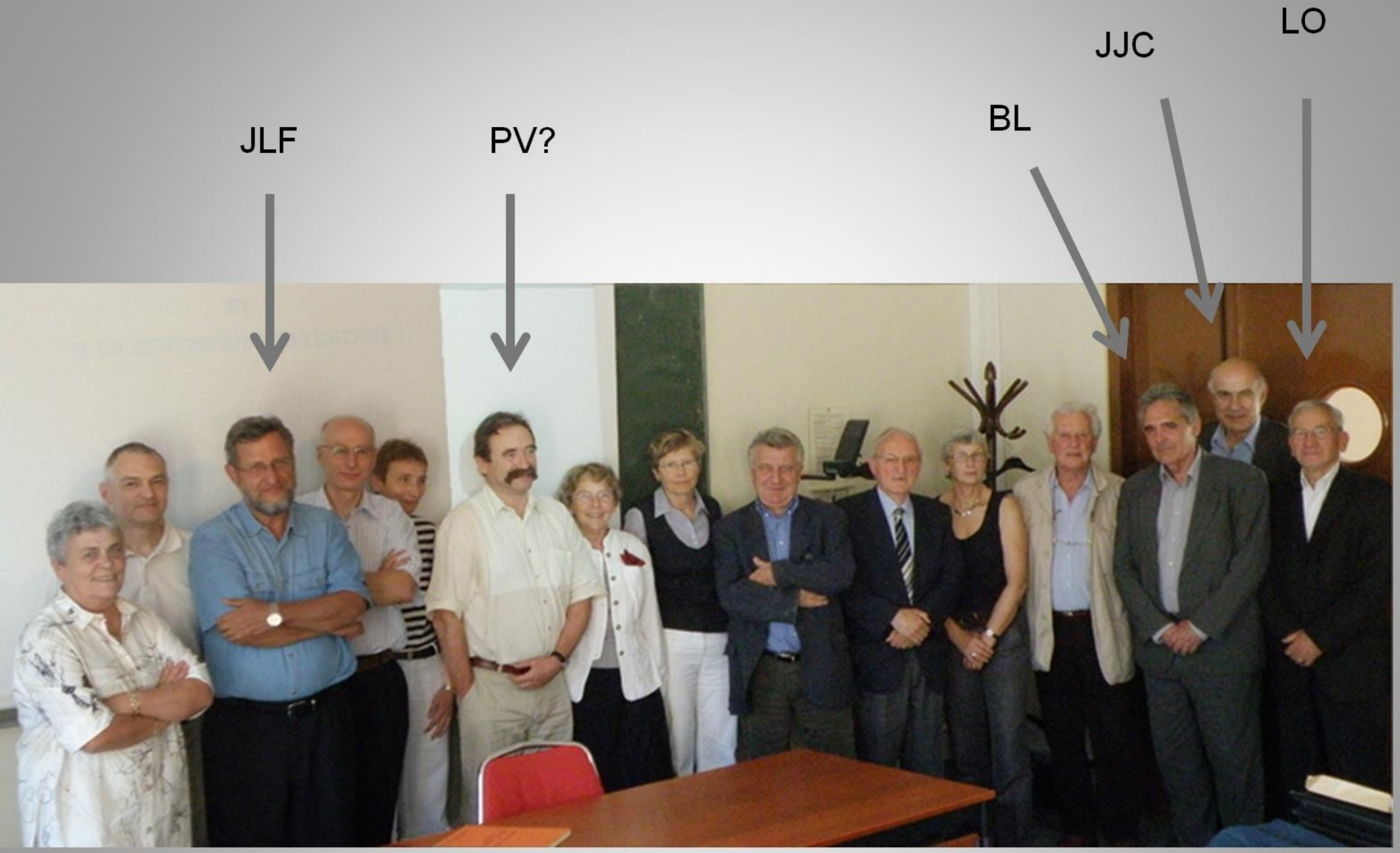
Sire evaluation for ordered categorical data with a threshold model

D. GIANOLA* and J.L. FOULLEY**

* *Department of Animal Science, University of Illinois, Urbana, Illinois 61801, U.S.A.*

** *I.N.R.A., Station de Génétique quantitative et appliquée, Centre de Recherches Zootechniques, F 78350 Jouy-en-Josas.*

485 citations


JLF
PV?
BL
JJC
LO