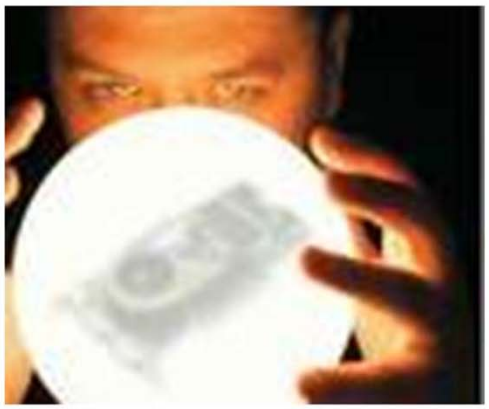


PREDICTOLOGY:

From pedigrees and DNA
To complex phenotypes

Daniel Gianola

Sewall Wright Professor of Animal Breeding and Genetics

University of Wisconsin

UW–MADISON ANIMAL SCIENCES

Dairy Science

Biostatistics & Medical Informatics

University of Wisconsin SCHOOL OF MEDICINE AND PUBLIC HEALTH

Universitetet for miljø- og biovitenskap

mat • natur • helse

NORWEGIAN UNIVERSITY OF LIFE SCIENCES MDCCCLIX

Georg-August-Universität Göttingen

Department of Animal Sciences: **Prof. Dr. Henner Simianer**

DFG Graduate Research School: **“Scaling problems in Statistics”**

Deutsche Forschungsgemeinschaft

DFG

Mercator-Gastprofessuren, 2006

Alexander von Humboldt

Stiftung / Foundation

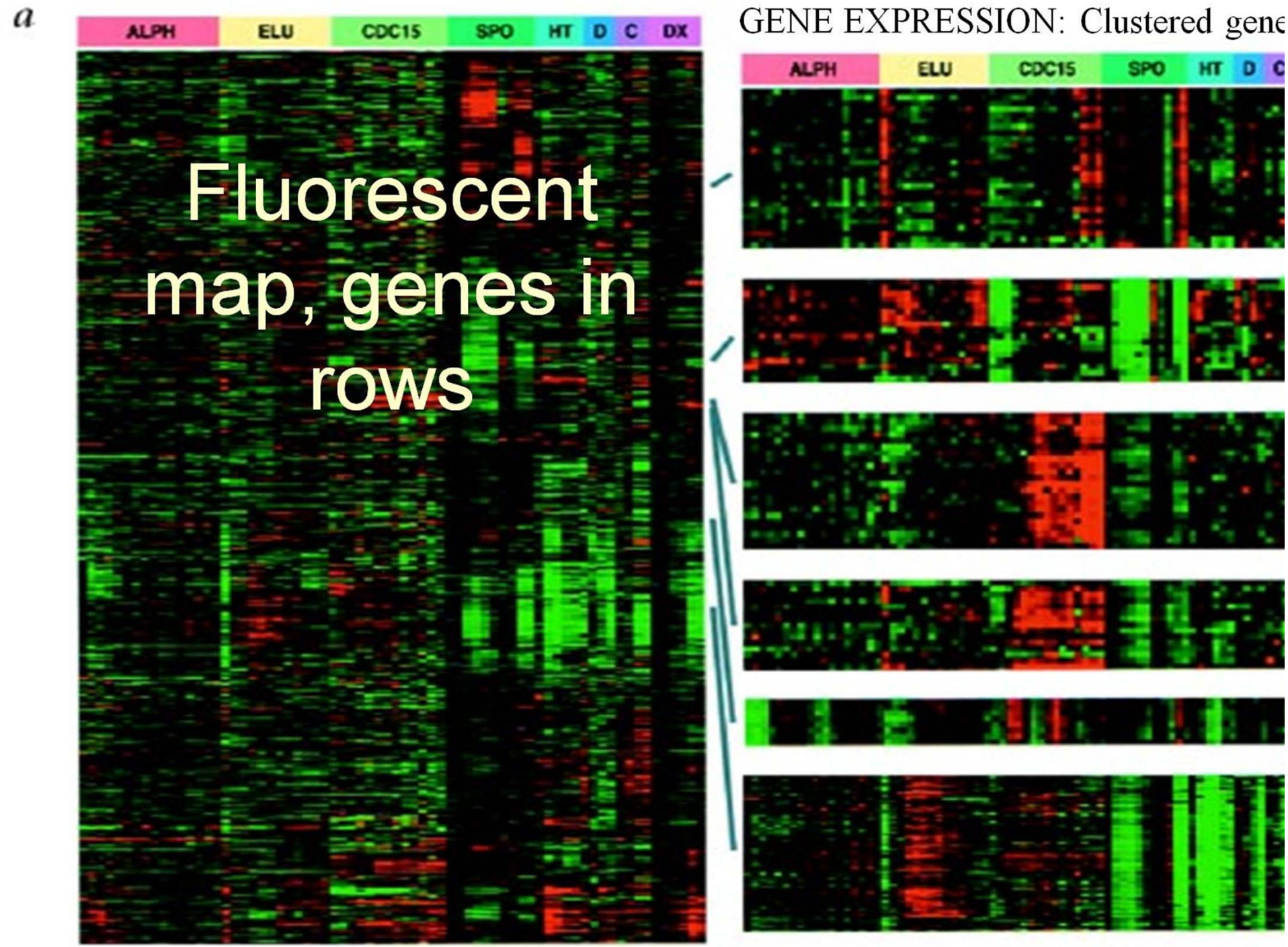
The Phenomic data (phenotypes+genomic)

1) Massive phenotypic data exist
2) Massive genomic data increasingly available

Example: SNPs (also gene expression)

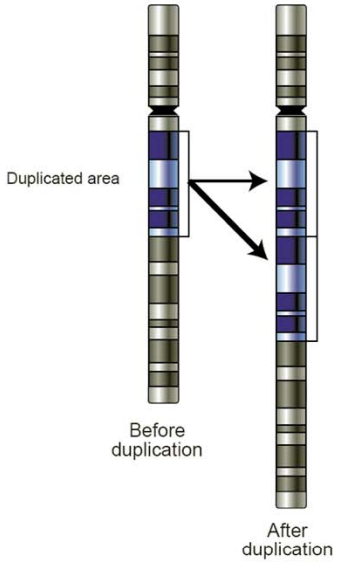
- 10^7 SNPs dbSNP 124 (Nat. Center Biotechnology)
- Perlegen: 1.58 million SNPs
- Animals:

-Wong et al. (2004) -- chicken genetic variation map with **2.8** million SNPs
-Hayes et al. (2004) -- **2500** SNPs in salmon genome
-Poultry breeding companies-- Thousands of SNPs on sires/dams
-USA (2008) -- >**50,000** SNPs in over **3000** Holstein sires
-All over developed world -- chips with 800,000 SNPs


a
ALPH
ELU
CDC15
SPO
HT
D
C
DX
Fluorescent map, genes in rows
GENE EXPRESSION: Clustered gene
ALPH
ELU
CDC15
SPO
HT
D
C

Copy number (CNV) of copy number polymorphisms (CNP): other source of information about genetic variation

- Individuals vary in number of copies of genomic regions
- Disease genes located in CNV regions

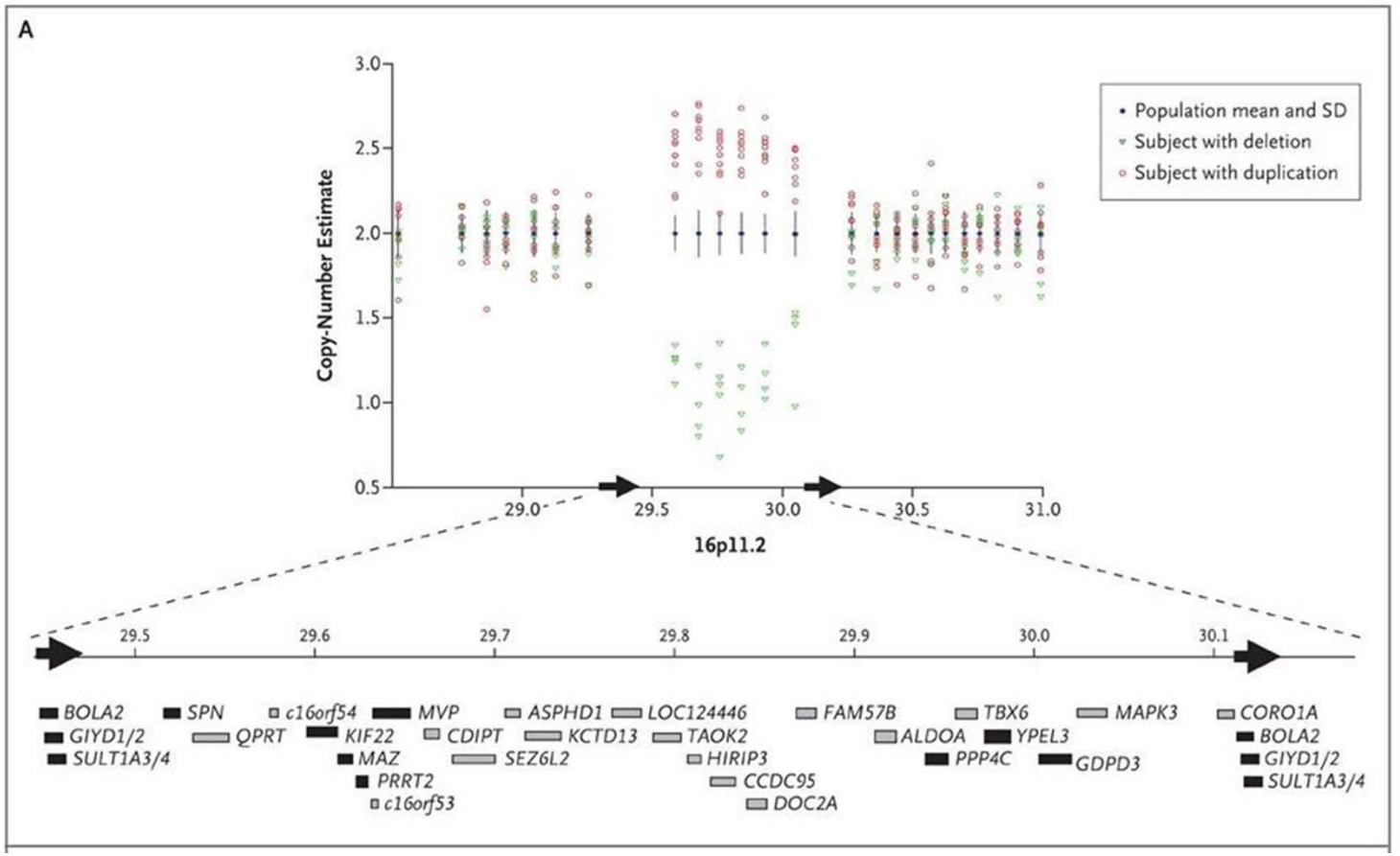


1 Kb~>1 Mb

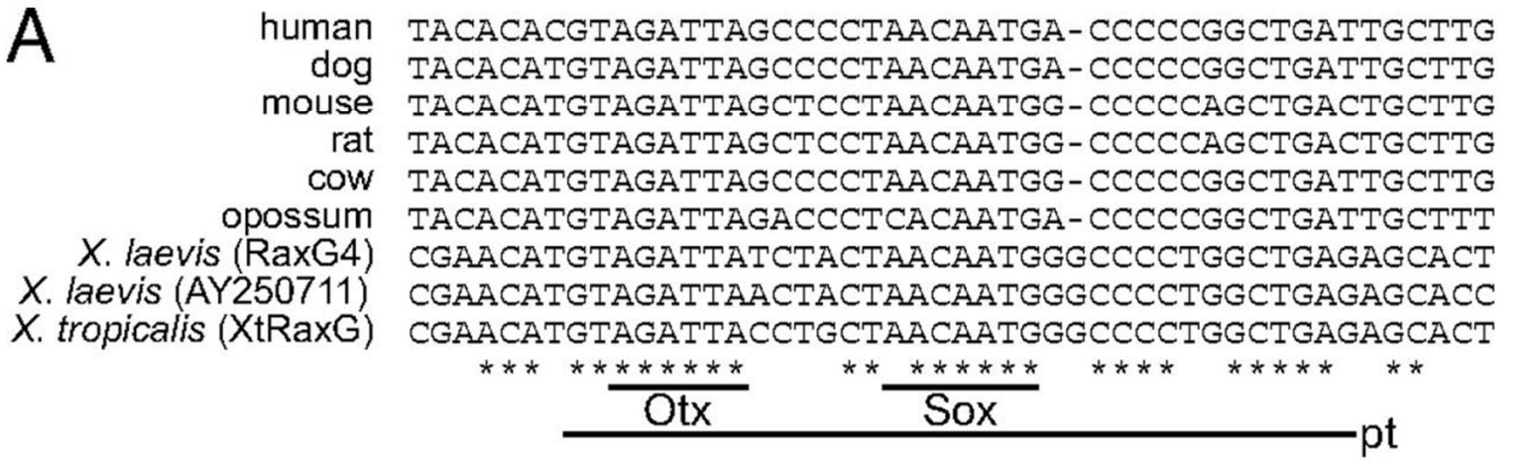
Higher number:

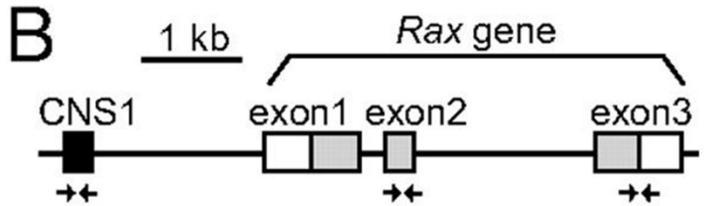
-Cancer cells

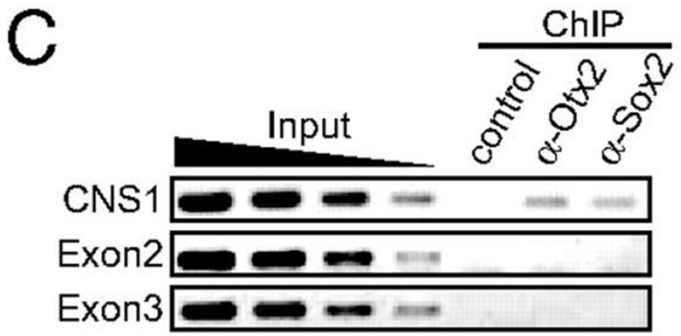
-liability to HIV

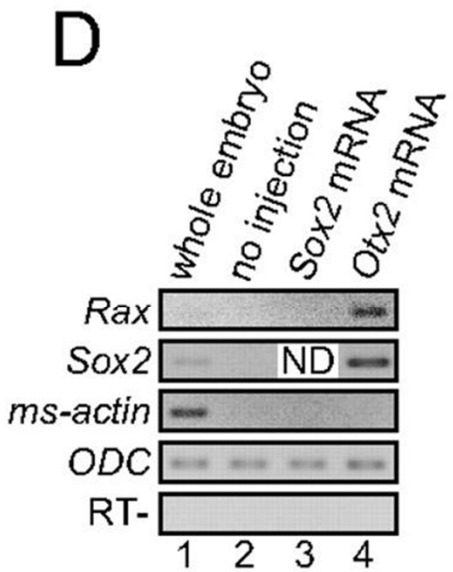


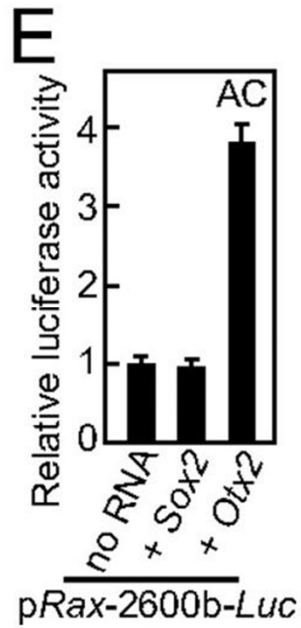
SEQUENCES FOR THOUSANDS OF ANIMALS (WITHIN SPECIES) COMING SOON

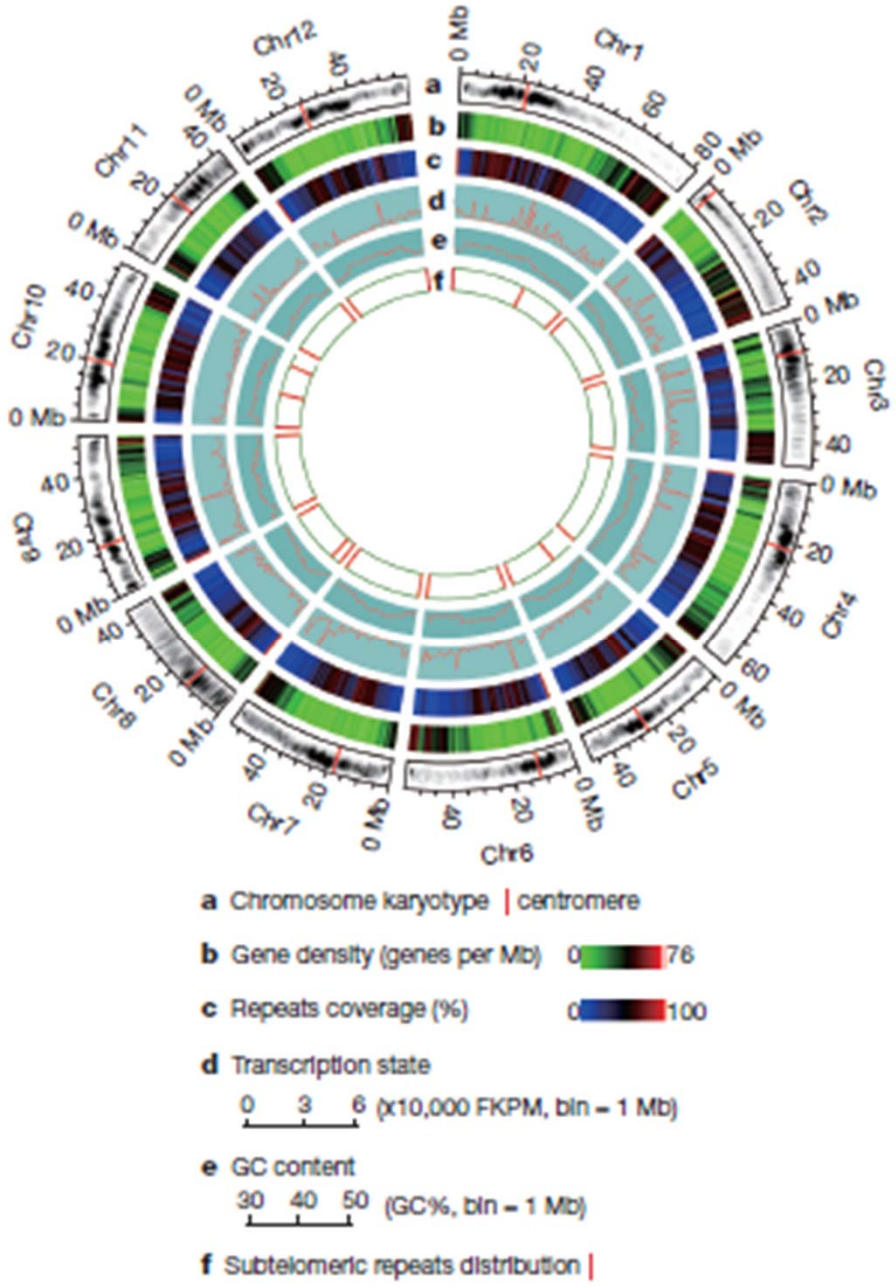












POTATO GENOME

(Nature 2011)

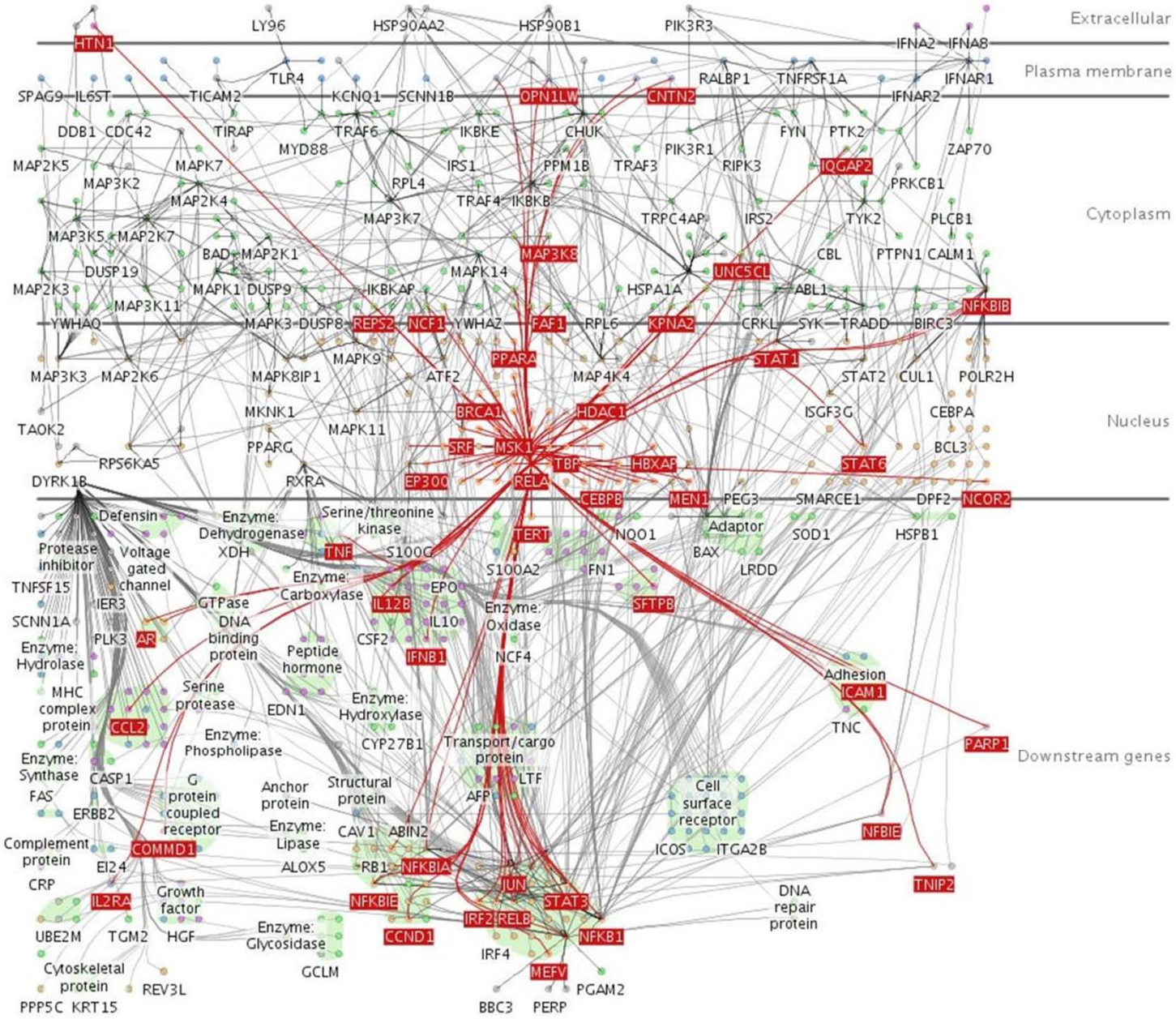
- Final assembly 727 Mb
- Genome size 844 Mb
- 1 SNP every 40 bp
- 1 indel every 394 bp (average 12.8 bp)
- 24,051 genes cluster with at least one of 11 genomes

Proposition 1

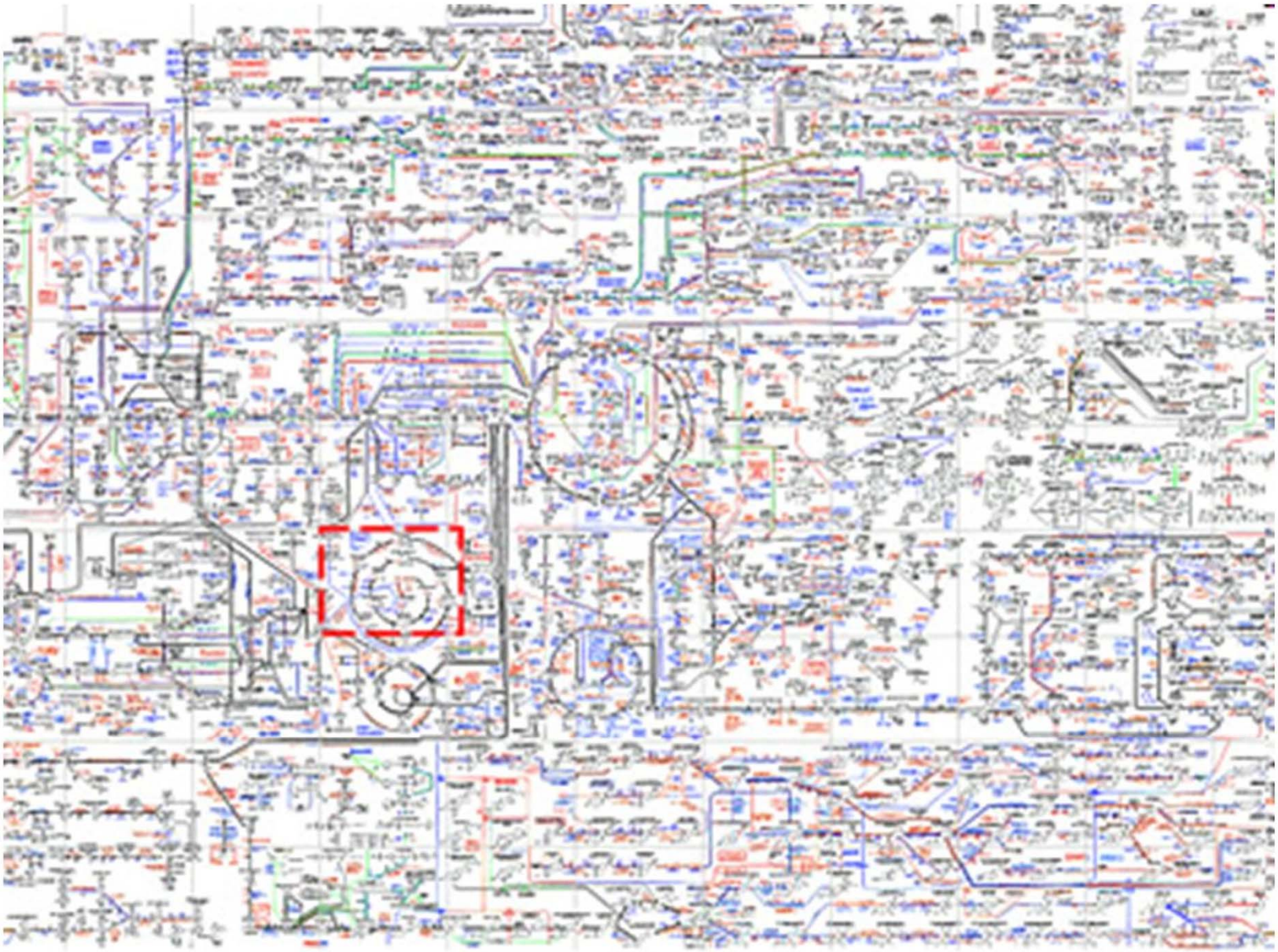
It must be true that quantitative traits are “complex”, in any sense of the word.

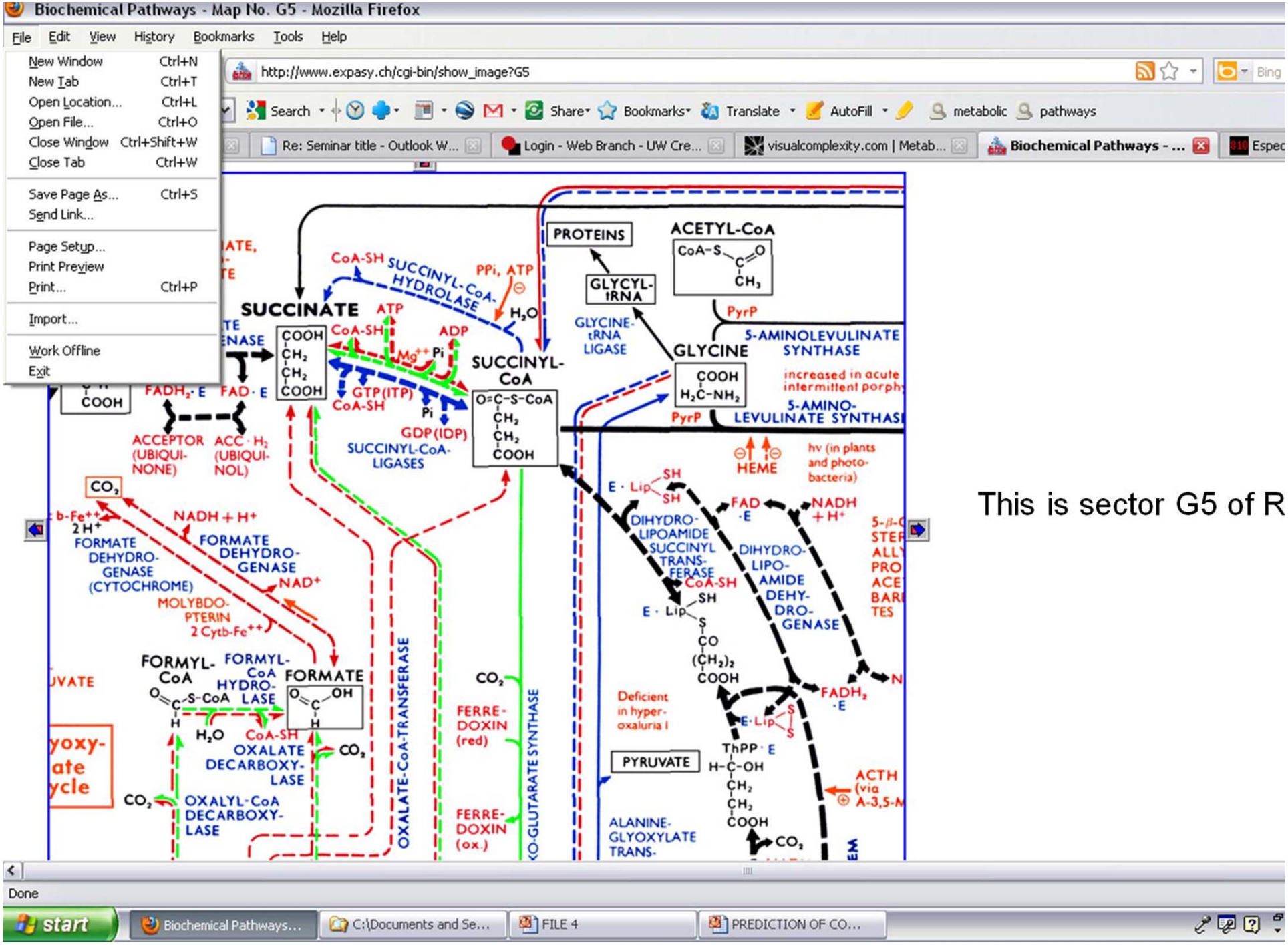
Why?

A SYSTEMS BIOLOGY MAP OF THE BRAIN



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





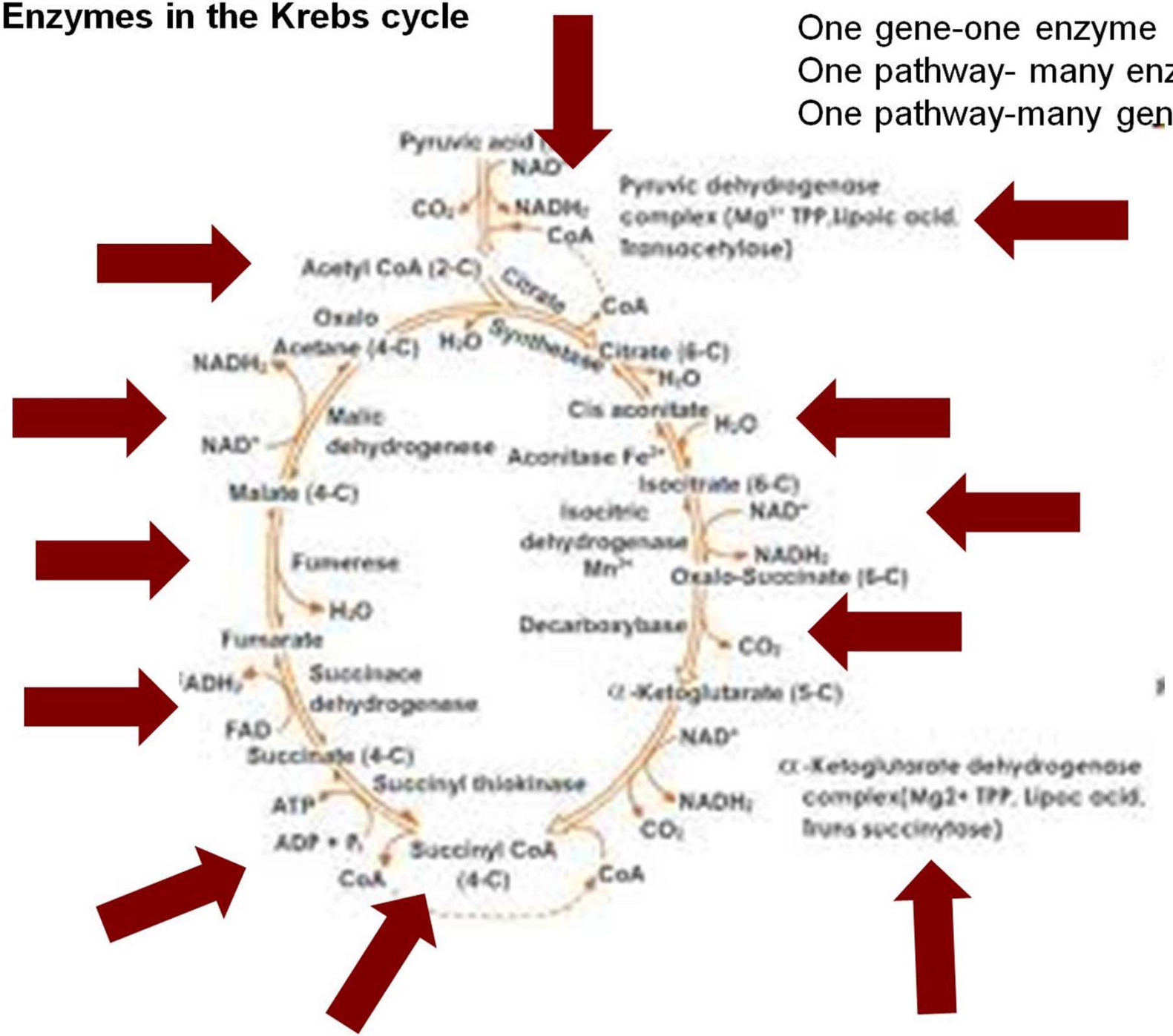
This is sector G5 of R

Proposition 2

It must be true that epistasis is pervasive

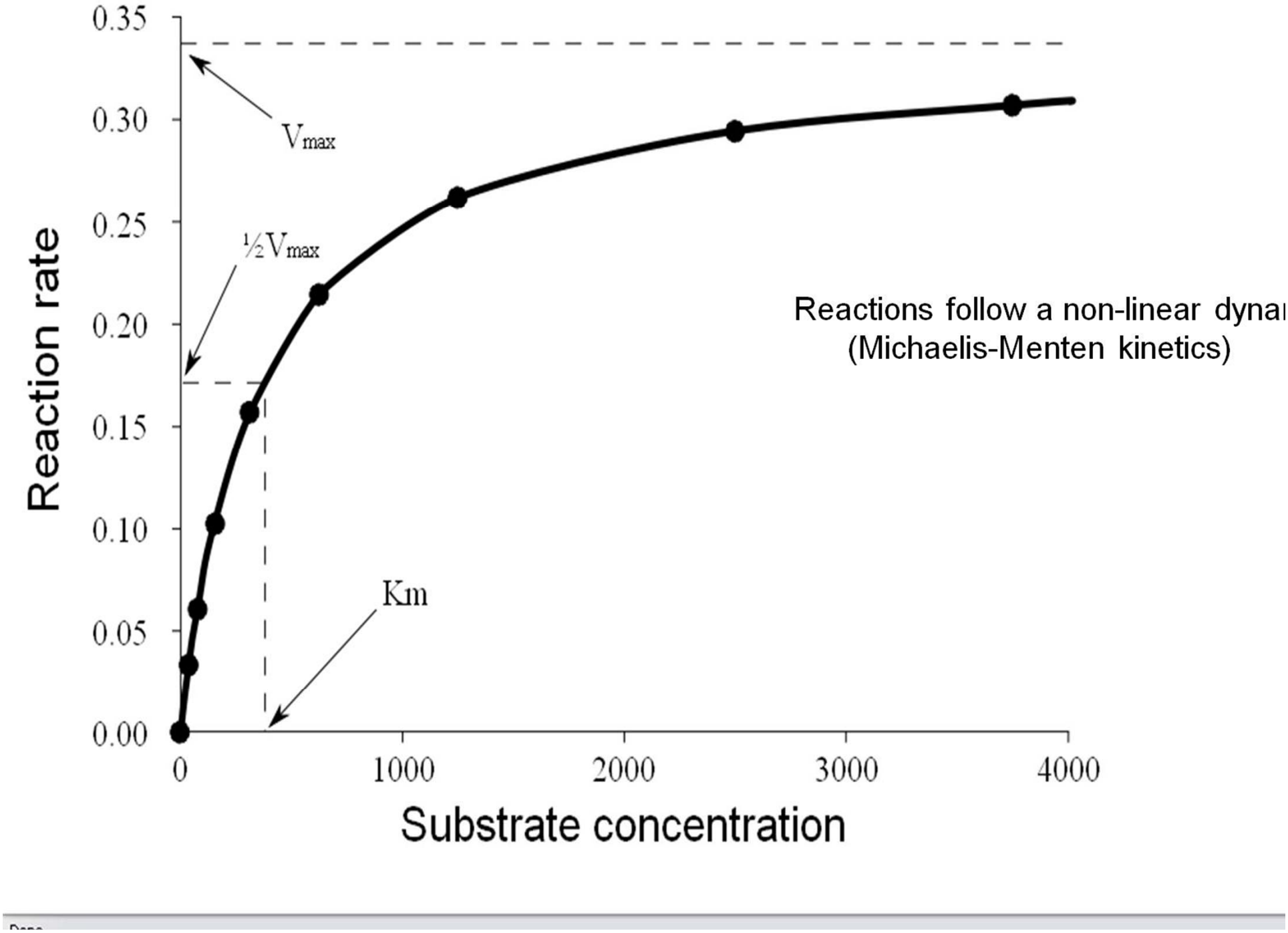
Enzymes in the Krebs cycle

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



Proposition 3

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


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Reaction rate
Vmax
½Vmax
Km
0
1000
2000
3000
4000
Substrate concentration
Reactions follow a non-linear dynar
(Michaelis-Menten kinetics)

Proposition 4

- It is unlikely that one could arrive to any reasonable mechanistic model satisfactory to understand, explain, learn and predict outcomes
- Hence, welcome to the world of abstractions

Coping with complexity

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 BIGGEST SHOW ON EARTH

can the lion be tamed?

The additive genetic model

A prevailing view (Hill et al., 2008; Crow, 2010; Hill, 2010)

Another show: “Les Idiots Savants”

(much less popular)

- If everything behaves as additive even with epistasis, can additive models allow learning about “genetic architecture”?
- If phenotypic prediction is crucial (medicine, precision mating) can exploitation of interaction have added value?
- Ideally, search for machine that
  - --captures additivity (breeding), interaction (medicine)
  - --has reasonably good predictive ability
  - --general and flexible with respect to input data
  - --does not fail if system is linear and non-interacting

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

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

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \ldots.$$

Linear approximation

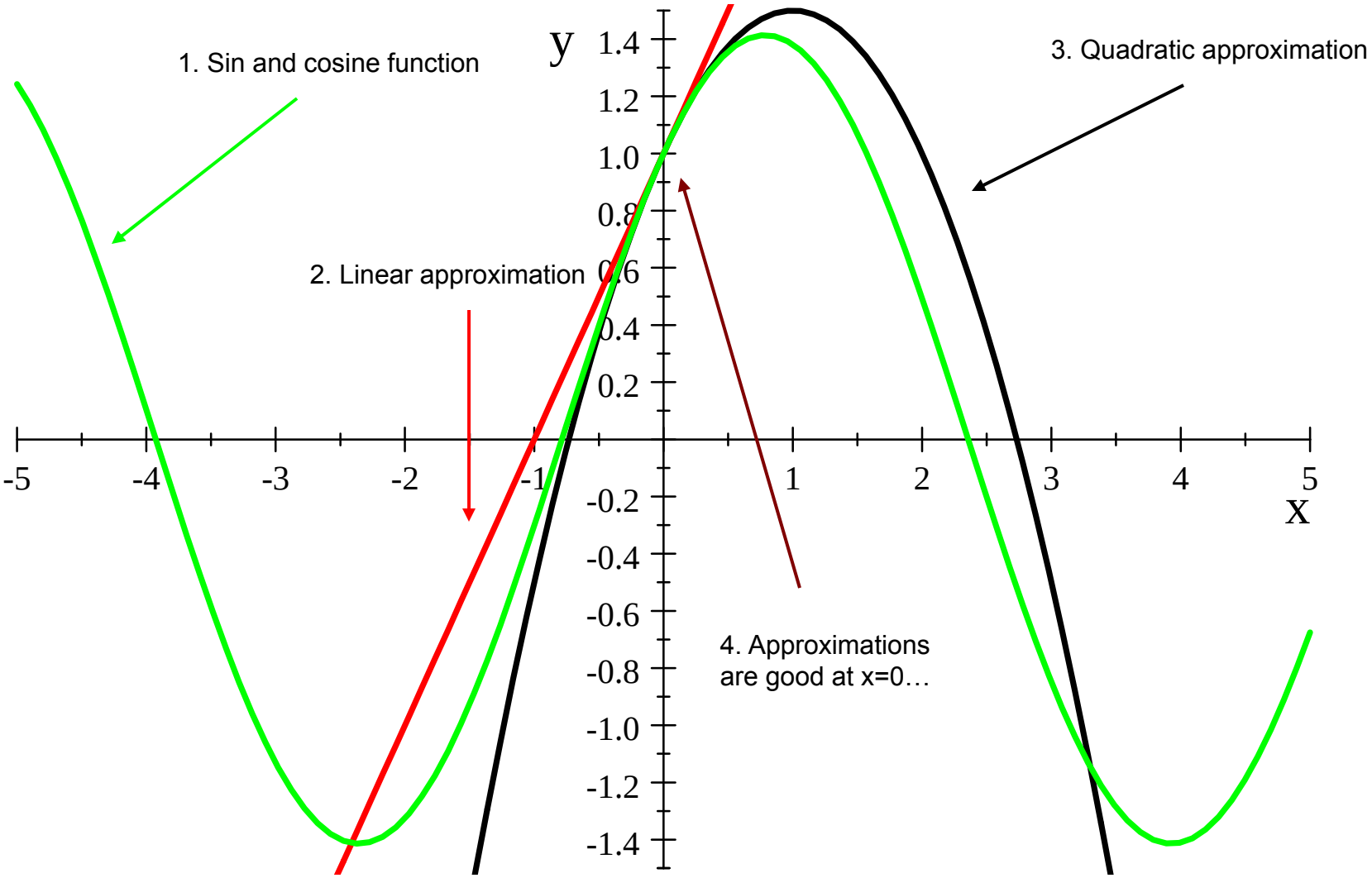
Quadratic approximation

n^{th} order approximation

FELDMAN and LEWONTIN (1975)
CHEVALET (1994)

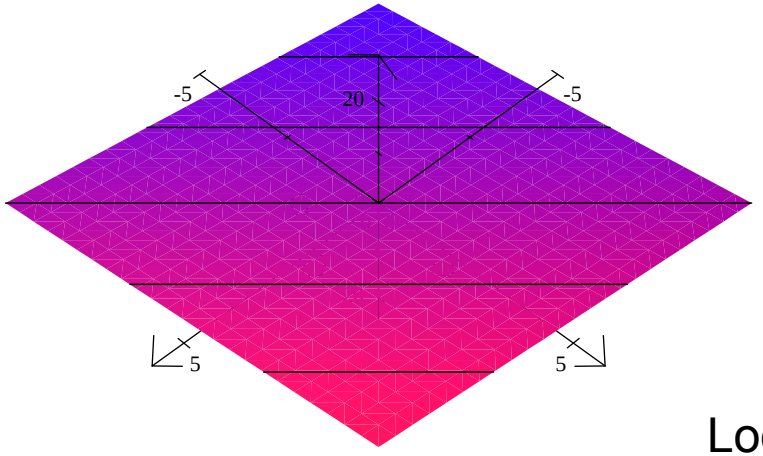
How good are linear and quadratic approximations? A Taylor series provides a local approximation only…

$$y = g(x) + e \qquad g(x) = \sin(x) + \cos(x)$$



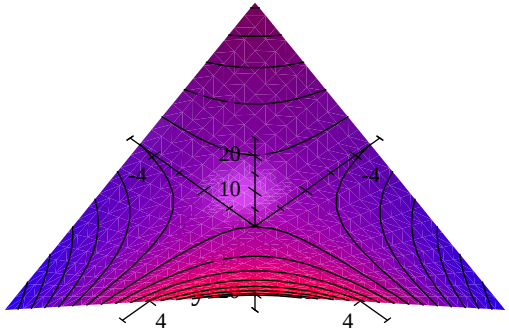
“TWO-LOCUS” ADDITIVE MODEL

$$x_1 + x_2$$



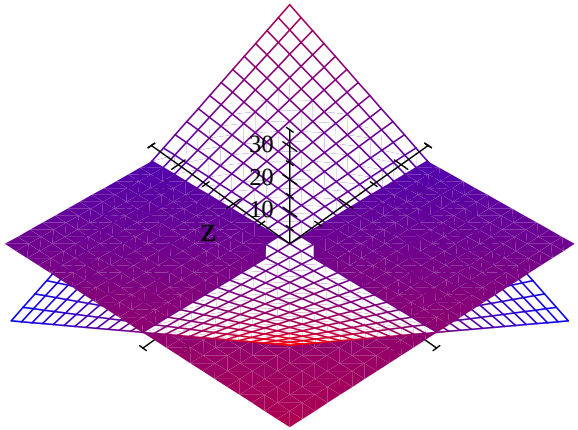
“TWO-LOCUS” EPISTASIS MODEL

$$x_1 + x_2 + x_1x_2$$

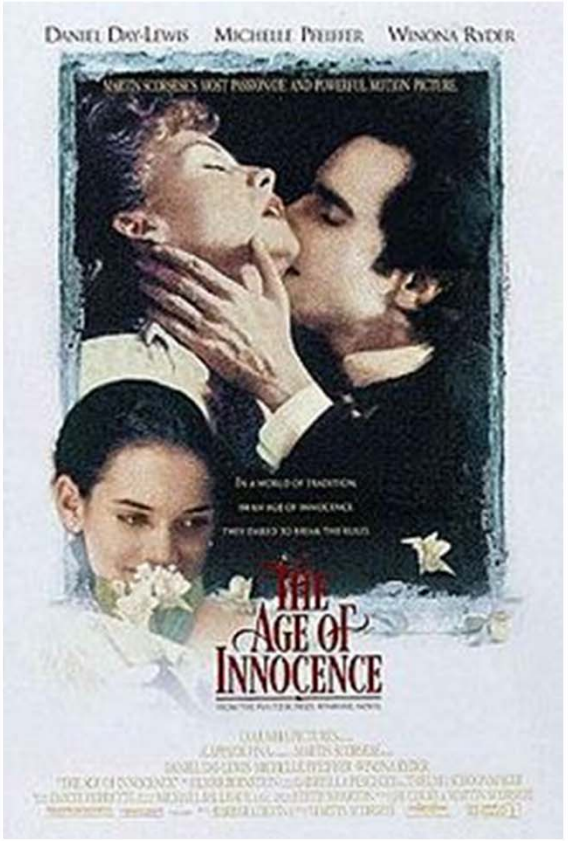


Look at the very different contours

Together



THE ADDITIVE MODEL IS NAÏVE AND INFLEXIBLE



THE AGE OF INNOCENCE

Unraveling “genetic architecture”
with statistical models

SINGLE MARKER REGRESSION WITH ORDINARY LEAST-SQUARES

n (#number of observations <<p (# markers)

"Full model"

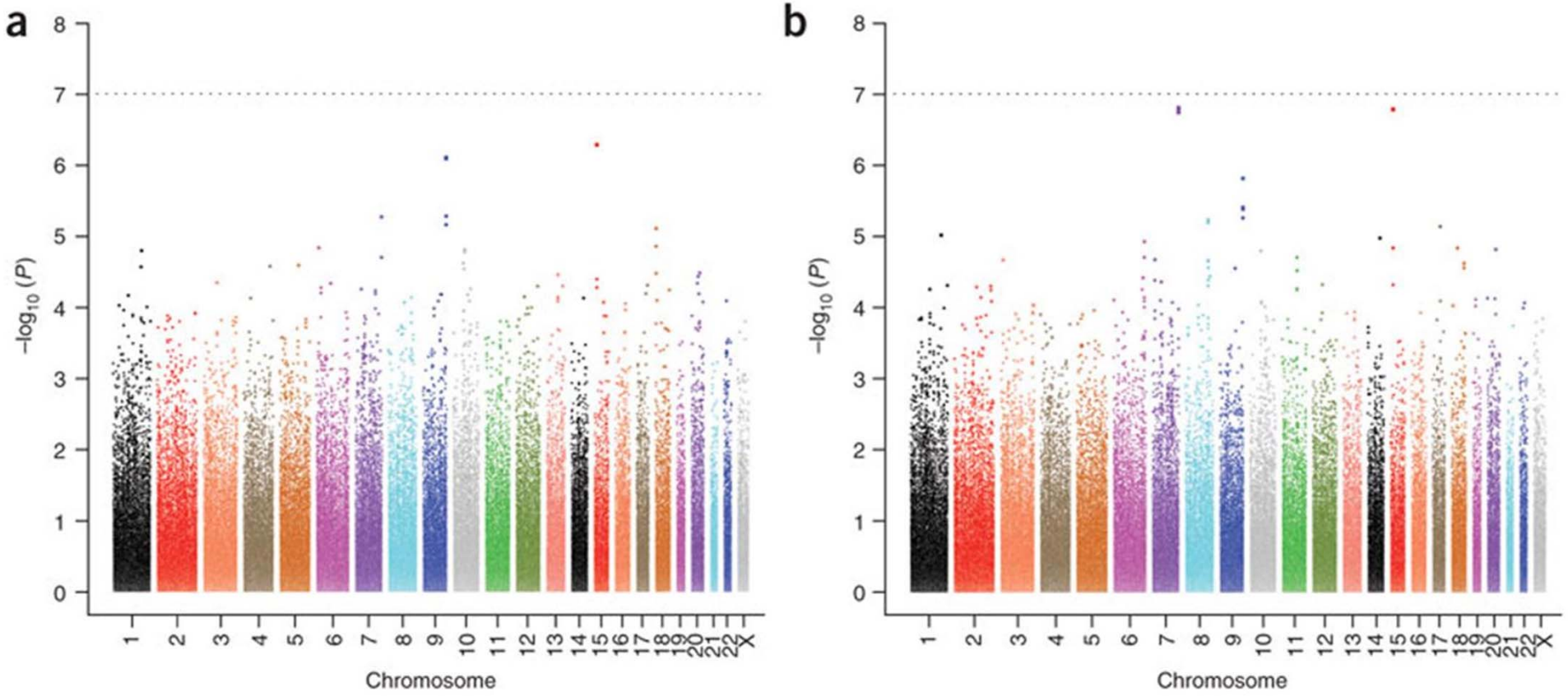
$$y = X\beta + e$$
$$= X_1\beta_1 + X_2\beta_2 + e$$

"marked phenotype"

"OLS" is biased If full model holds and one fits "smaller" model (e.g., single marker Regressions)

$$y = X_1\beta_1 + e$$
$$E(\tilde{\beta}_1|X_1) = (X_1'X_1)^{-1}E(y)$$
$$= (X_1'X_1)^{-1}[X_1\beta_1 + X_2\beta_2]$$
$$= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$$

EXTRAORDINARILY NAÏVE, YET….

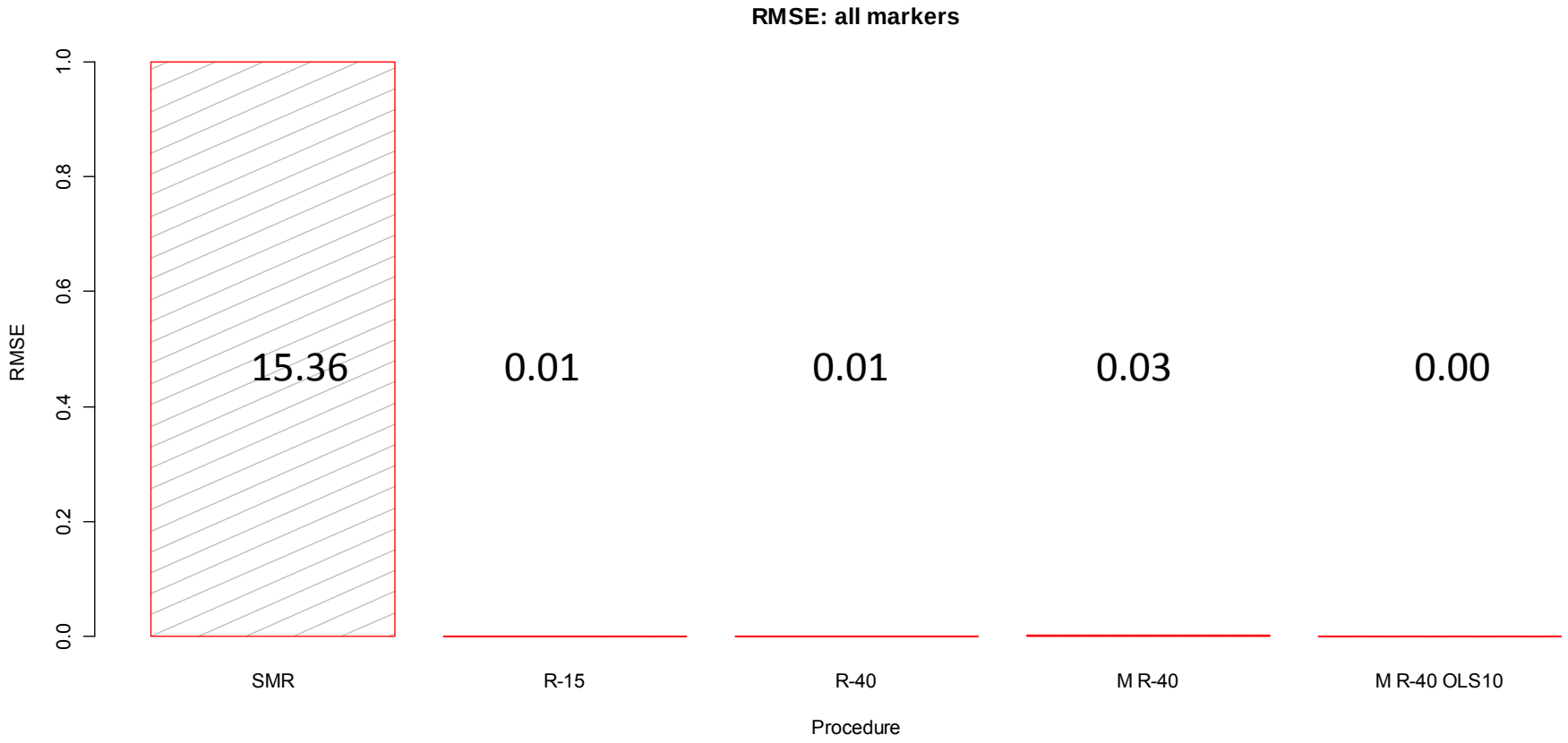


GWAS FOR PANCREATIC CANCER…
(Nature Genetics)

SINGLE MARKER REGRESSION: A DISASTER

N=100, 1000 binary markers, 5 first are signal, LD~1/3

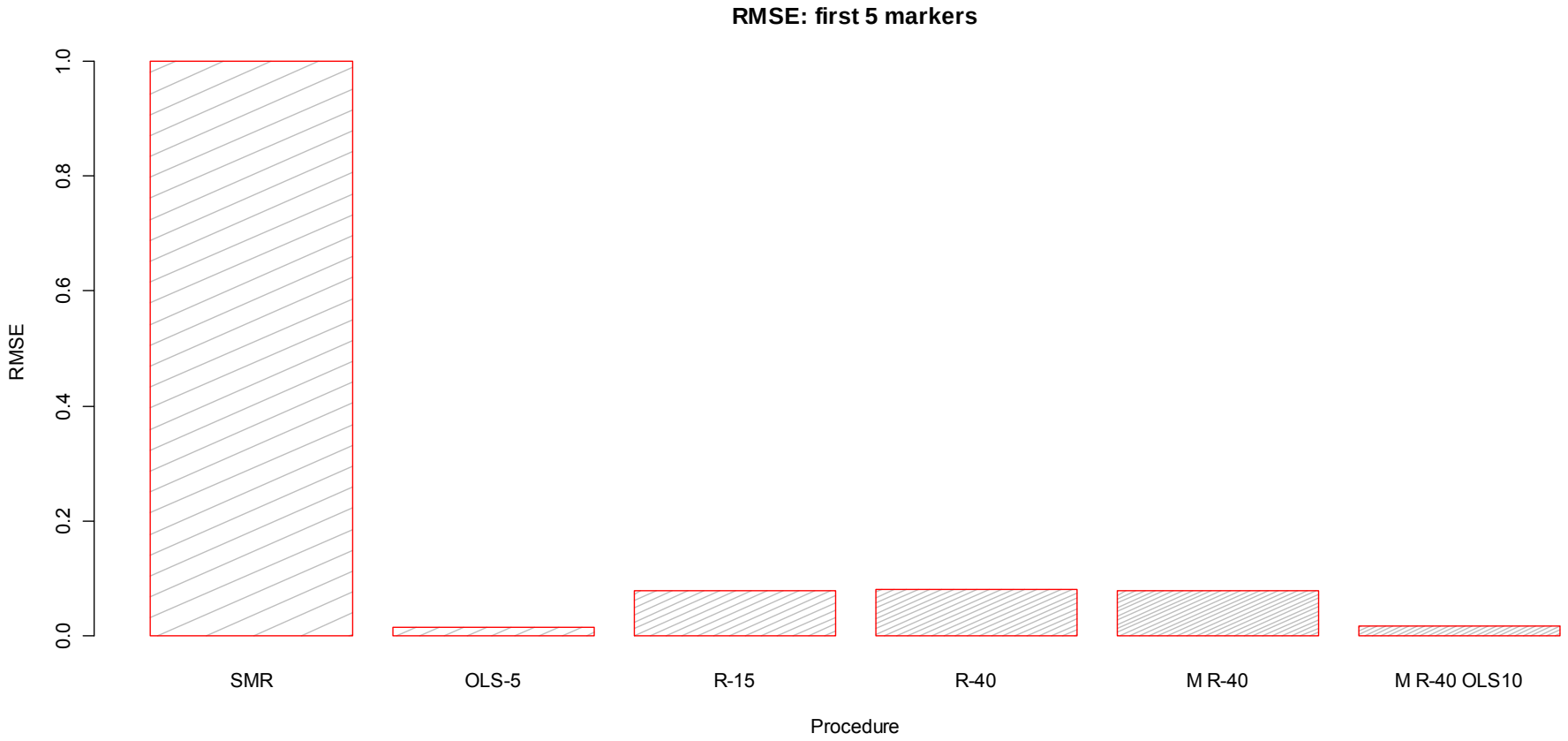
RELATIVE MEAN-SQUARED ERROR (ALL MARKERS)



SINGLE MARKER REGRESSION: A DISASTER

N=100, 1000 binary markers, 5 first are signal, LD~1/3

RELATIVE MEAN-SQUARED ERROR (FIRST FIVE 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"

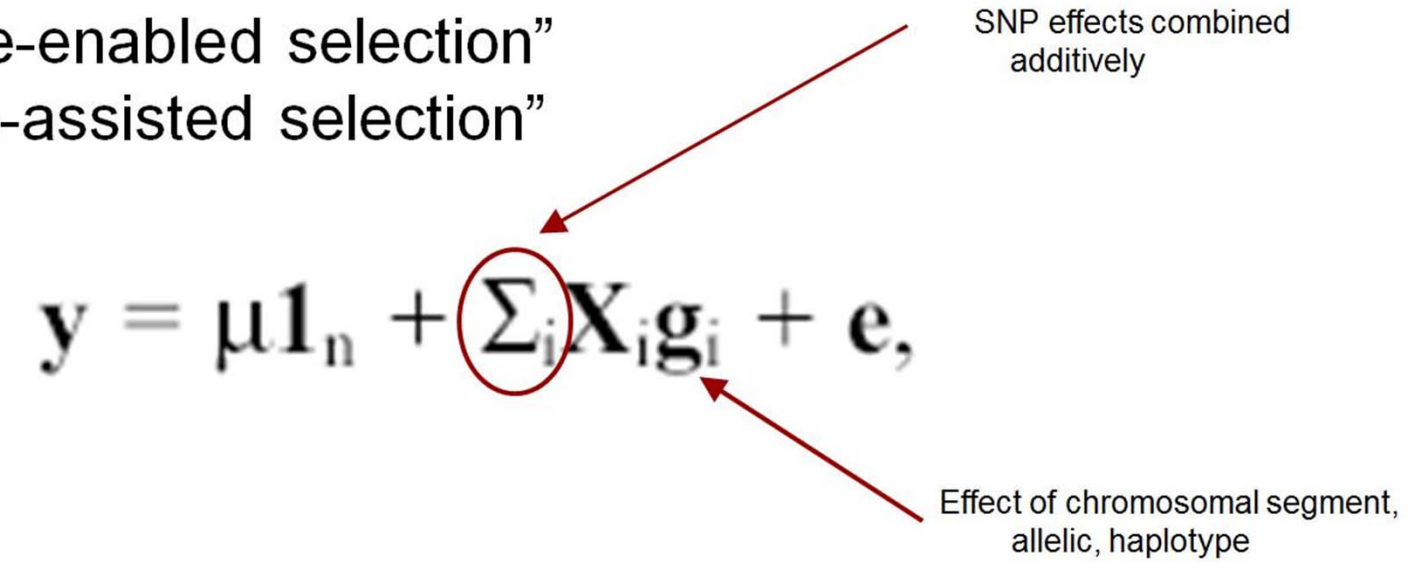
Meuwissen, Hayes and Goddard (2001)

"Genomic selection"

Better terms:

"Genome-enabled selection"

"Genome-assisted selection"



ANIMAL BREEDING:

USE ALL SNP MARKERS IN MODELS FOR GENOMIC-ASSISTED EVALUATION

QUESTION: BYE-BYE QTLS, PEDIGREES, GENES?.

Essentials of genome-enabled selection

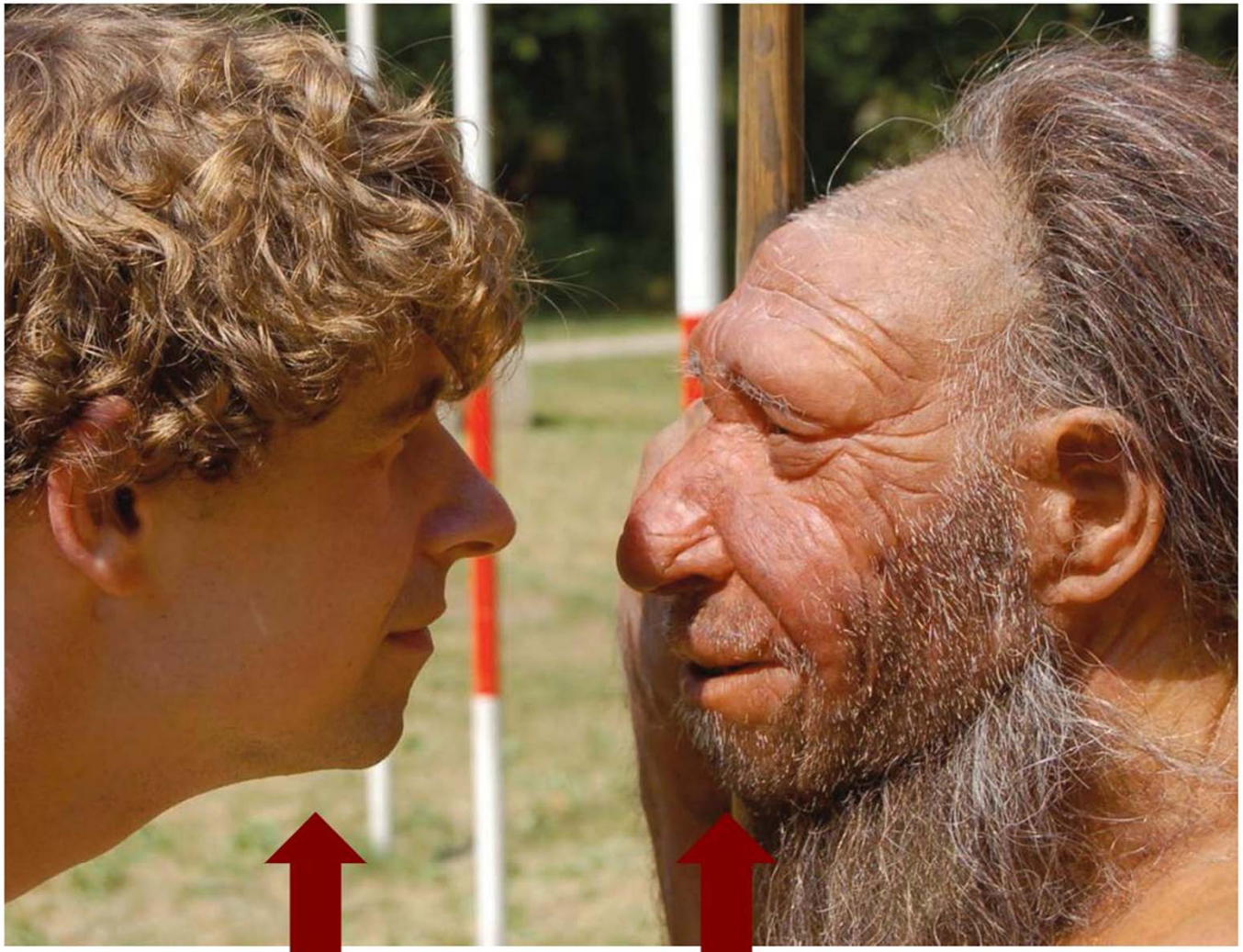
- Fit (train) some regression model (typically Bayesian) to a data set with markers and phenotypes
- Estimate marker effects
- Predict marked genetic value or phenotype in a new sample (testing or validation sample) for which only DNA information is available
- Once phenotype (or something related to phenotype) is observed, asses quality of prediction. For example, calculate predictive correlation or mean squared error of prediction
- Objective: gain reliability and if new sample is of juveniles, reduce generation interval. Dispense with progeny testing? Reduce frequency of phenotyping?

Schaeffer (2006):

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

CLOSE ENCOUNTERS OF THE PREHISTORIC KIND

Homo sapiens



Neanderthal

GENOMICS AND COMPLEX BIOLOGY

NO! THE ADDITIVE GENETIC MODEL

Arguably, one could do better than with linear Bayesian (regularized) linear models!

Reproducing Kernel Hilbert spaces mixed model

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

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

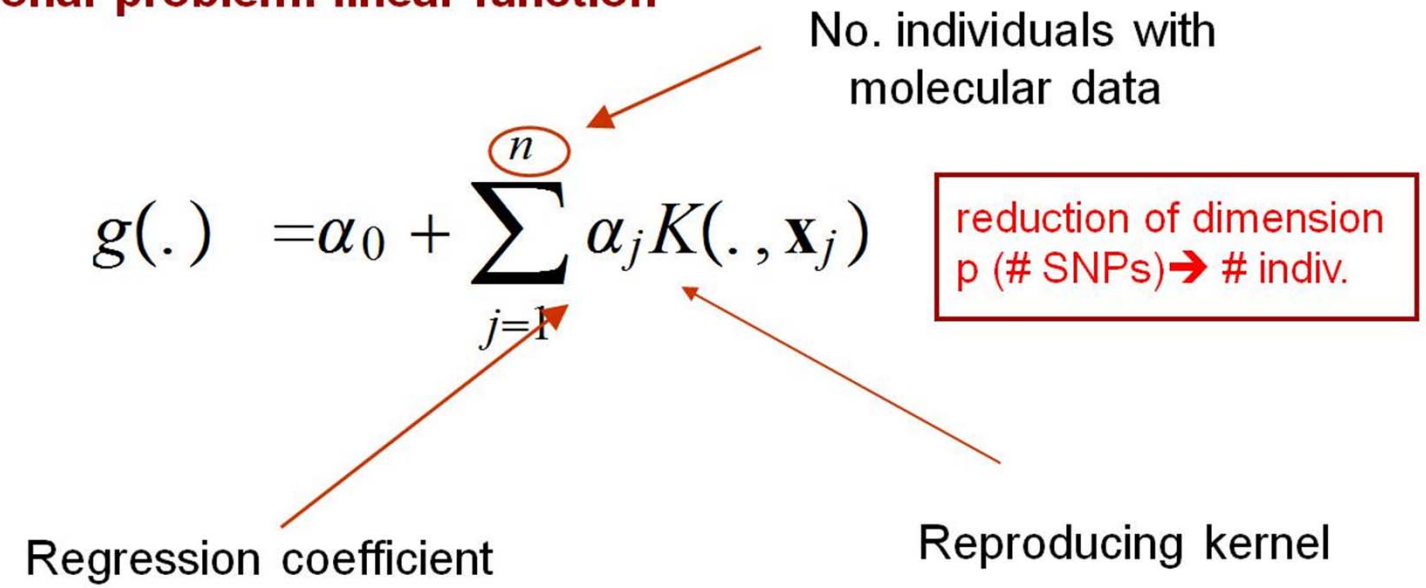
Smoothing parameter (λ)

"Penalized sum of squares"

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

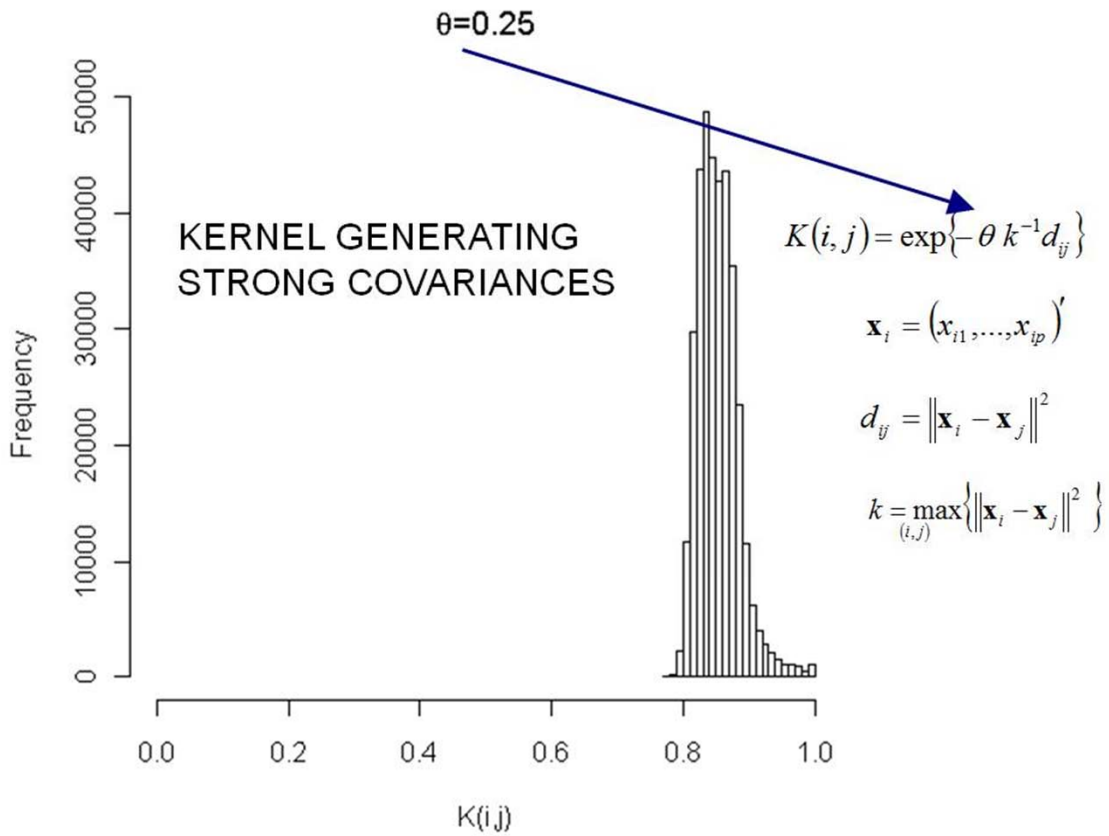
Variational problem: find *g(**x**)* over entire space of functions minimizing SS(.)

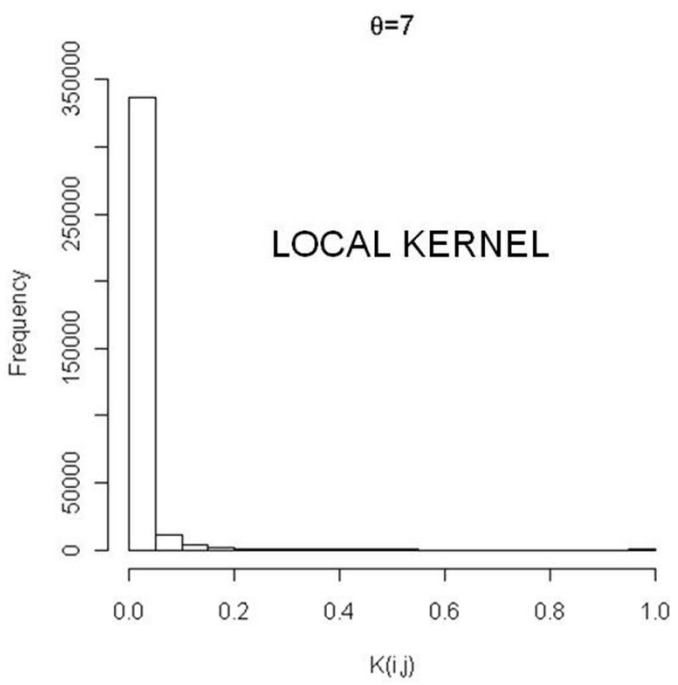
Solution to variational problem: linear function



Example of reproducing kernel:

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





Histogram of evaluations of Gaussian kernel by value of 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).

Mixed model representation (enhancing pedigrees…)

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

Define row vector

$$\mathbf{t}_i'(h) = \left\{ \exp\left[-\frac{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}{h} \right] \right\}$$

$$\mathbf{T}(h) = \begin{bmatrix} \mathbf{t}_1'(h) \\ \mathbf{t}_2'(h) \\ . \\ \mathbf{t}_n'(h) \end{bmatrix}$$

$\boldsymbol{t}'_i(h) = \boldsymbol{k}'_i(h)$

$\boldsymbol{T}(\boldsymbol{h}) = \boldsymbol{K}(\boldsymbol{h})$

Then:

Bandwidth parameter

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

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

Do:

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

Smoothing parameter

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

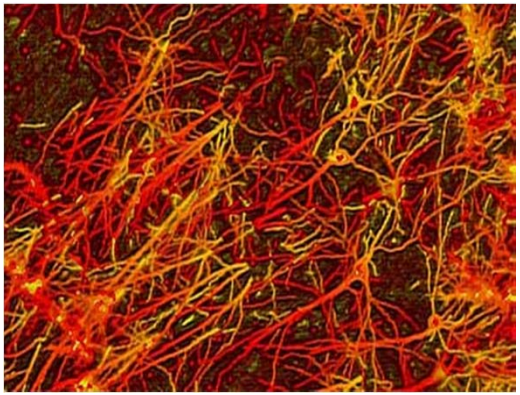
h assumed known here

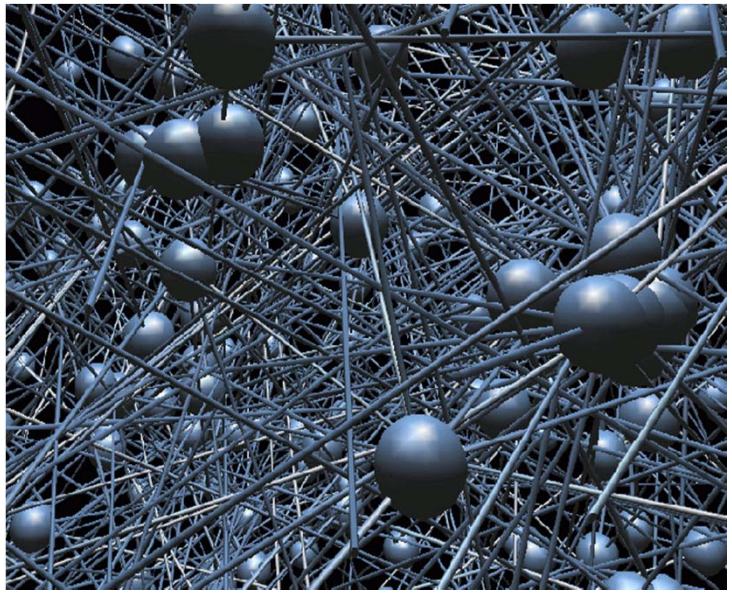
RKHS! GREAT MATH BUT:

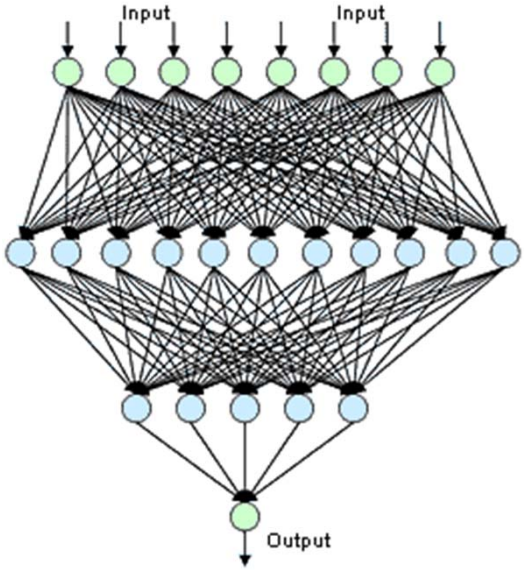
-Why should penalizing likelihoods lead to an exalted state of knowledge?

-The world outside of the hyper-cube

A perhaps more universal learning machine: Regularized Neural Networks







Kolmogorov's Theorem

For any continuous function $g(x_1, x_2, \ldots, x_p)$ of p variables there exists continuous functions h_j in $[0,1]$ a continuous function f in $[0,1]$ such that

$$g_i(x_{i1}, x_{i2}, \ldots, x_{ip}) = \sum_{q=1}^{2p+1} f\left[\sum_{j=1}^{p} w_j h_q(x_{i1}, ix_2, \ldots, ix_p) \right]$$

weights

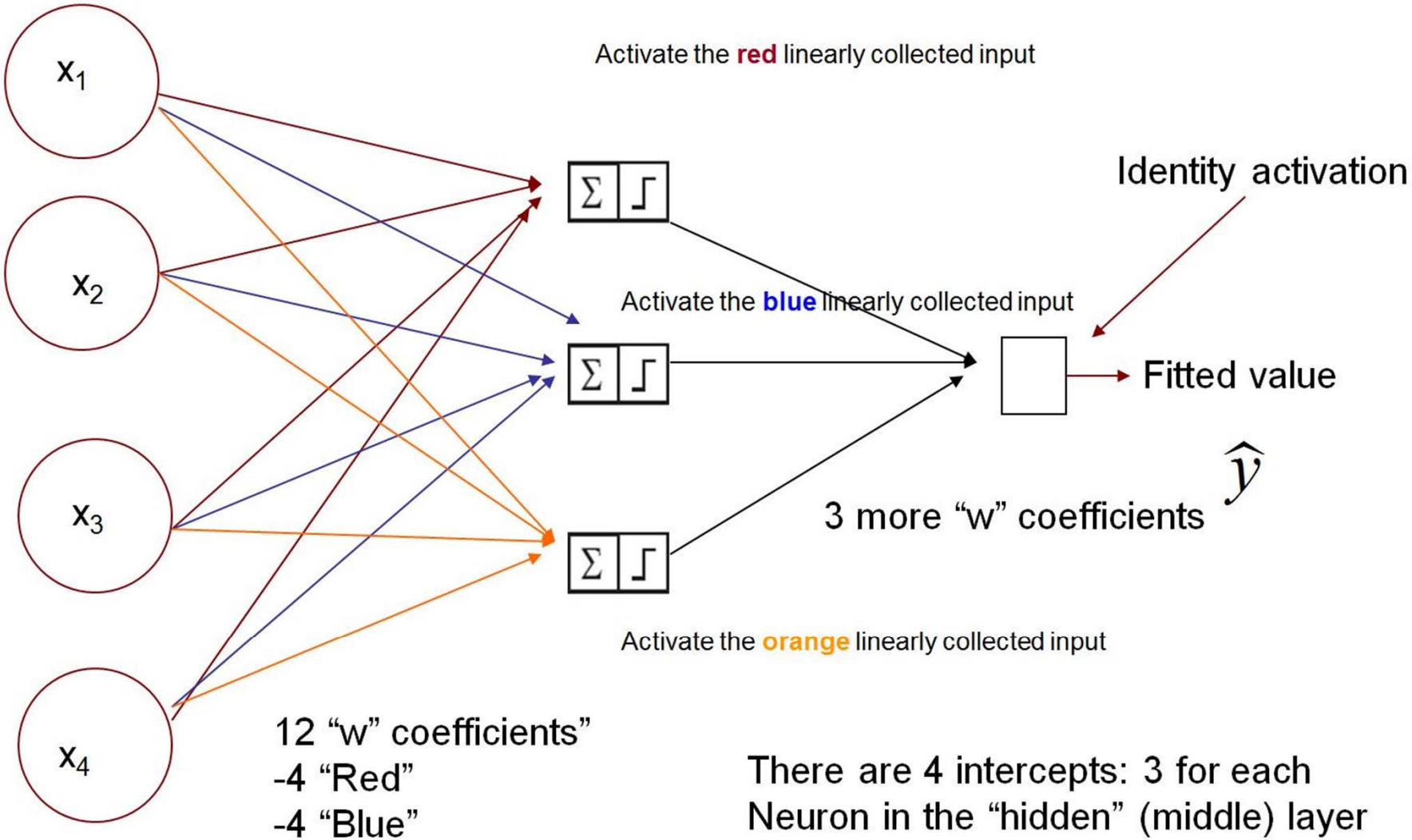
Linear or nonlinear transformation

Linear or on-linear transformation of inputs

The subscript indicates an evaluation on a given configuration of the input

KOLMOGOROV'S THEOREM CAN BE REPRESENTED AS AN ARTIFICIALNEURAL NETWORK

Illustration of a multi-layer model for **regression** with logistic activation function before emission to the output layer (X's may be SNPs or sequence data)



12 "w" coefficients"
-4 "Red"
-4 "Blue"
-4 "Orange

There are 4 intercepts: 3 for each Neuron in the "hidden" (middle) layer and 1 in the outer layer

Algebraically, the model looks like

$$y = \beta_0 + \beta_1 \frac{1}{1+\exp\left(w_0^{[1]} + w_1^{[1]}x_1 + w_2^{[1]}x_2 + w_3^{[1]}x_3 + w_4^{[1]}x_4\right)}$$ **RED**

$$+ \beta_2 \frac{1}{1+\exp\left(w_0^{[2]} + w_1^{[2]}x_1 + w_2^{[2]}x_2 + w_3^{[2]}x_3 + w_4^{[2]}x_4\right)}$$ **BLUE**

$$+ \beta_3 \frac{1}{1+\exp\left(w_0^{[3]} + w_1^{[3]}x_1 + w_2^{[3]}x_2 + w_3^{[3]}x_3 + w_4^{[3]}x_4\right)} + e$$

ORANGE

4 BETAS+ 15 w's= 19 regressions to estimate

- Bayesian regularization needed for $n<<p$
- Must avoid MCMC if job is to be done prior to death
- Can solve system of non-linear equations rapidly to find conditional posterior modes
- Can tune regularization parameters using Laplacian approximation to marginal likelihood

Bayesian regularization (need to cope with p>>n)

$$p(D \mid b, \mathbf{w}, \sigma^2, M) = \prod_{i=1}^{n} N\left(t_i \mid b, \mathbf{w}, \sigma^2, M\right)$$

Likelihood

A network Architecture (number of neurons and activation functions)

Prior $$p\left(\mathbf{w} \mid \sigma^2{}_w\right) = N\left(0, \mathbf{I}\sigma^2{}_w\right)$$

(This assumes that all w coefficients are shrunken to the same extent. This is probably not a good assumption, but convenient)

Conditional posterior

$$P(\mathbf{w} \mid D, \sigma^2, \sigma_w^2, M) = \frac{P(D \mid \mathbf{w}, \sigma^2, M) P(\mathbf{w} \mid \sigma_w^2, M)}{P(D \mid \sigma^2, \sigma_w^2, M)}$$

Marginal density of the data (used to assess variance components)

$$P(D \mid \sigma^2, \sigma_w^2, M) = \int P(D \mid \mathbf{w}, \sigma^2, M) P(\mathbf{w} \mid \sigma_w^2, M) d\mathbf{w}$$

$$p(D \mid \sigma^2, \sigma_w^2, M) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \left(\frac{1}{2\pi\sigma_w{}^2} \right)^{\frac{m}{2}} \times$$

Integral not in closed form in non-linear networks

$$\int \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(t_i - b - \sum_{k=1}^{S} w_k g_k (b_k + \sum_{j=1}^{n} a_{ij} u^{**[k]}{}_j) \right)^2 - \frac{1}{2\sigma_w{}^2} \mathbf{w}'\mathbf{w} \right] d\mathbf{w}$$

$$F(\alpha, \beta) = \beta \sum_{i=1}^{n} \left(t_i - b - \sum_{k=1}^{S} w_k g_k (b_k + \sum_{j=1}^{n} a_{ij} u^{**[k]}{}_j) \right)^2 + \alpha \mathbf{w}'\mathbf{w} = \beta E_D + \alpha E_w$$

"penalized" sum of squares

$1/2\sigma^2$ $1/2\sigma_w{}^2$

Laplacian approximation yields

Remember Smith and Graser (1986); Graser et al. (1987); Tempelman and Gianola (1993)

$$\log\left[p(D\mid\alpha,\beta,M)\right]\approx K+\frac{n}{2}\log(\beta)+\frac{m}{2}\log(\alpha)-\left|\beta E_D+\alpha E_w\right|_{w^{map}(\alpha,\beta)}-\frac{1}{2}\log\left\|\mathbf{H}\right\|_{w^{map}(\alpha,\beta)}$$

Hessian of F

$$\alpha_{new}=\frac{m}{2\left(w^{MAP}{}'w^{MAP}+trH^{-1}_{MAP}\right)}$$

$$\beta_{new}=\frac{n-m+2\alpha_{MAP}trH^{-1}_{MAP}}{2\sum_{i=1}^{n}\left(t_i-b-\sum_{k=1}^{S}w_k g_k(b_k+\sum_{j=1}^{n}a_{ij}u^{**[k]}{}_j)+e_i\right)^2_{MAP}}$$

Effective number of parameters $\gamma=m-2\alpha_{MAP}trH^{-1}_{MAP}$

THE INFINITESIMAL MODEL AS A REGRESSION ON RELATIONSHIPS

$$\mathbf{y} = \mathbf{u} + \mathbf{e}$$

$$\mathbf{u} \sim (\mathbf{0}, \mathbf{A}\sigma_a^2)$$

$$\mathbf{y} = \mathbf{A}\mathbf{A}^{-1}\mathbf{u} + \mathbf{e}$$

$$= \mathbf{A}\mathbf{u}^* + e$$

$$y_i = \sum_{j=1}^{N} a_{ij} u_j^* + e_i$$

Use elements of
A (or **G**) as inputs
(covariates) in a regression
Model with random effects

Recall
A=CC' (Cholesky)

The infinitesimal model as a regression on a pedigree

1) $\mathbf{t} = \mathbf{C}\mathbf{z}\sigma_u + \mathbf{e} = \mathbf{C}\mathbf{u}^* + \mathbf{e}$ $\qquad \mathbf{u}^* = \mathbf{z}\sigma_u \sim (\mathbf{0}, \mathbf{I}\sigma^2{}_u)$

$$t_i = g(\sum_{j=1}^{n} c_{ij} u^*{}_j) + e_i, \qquad \text{Identity activation}$$

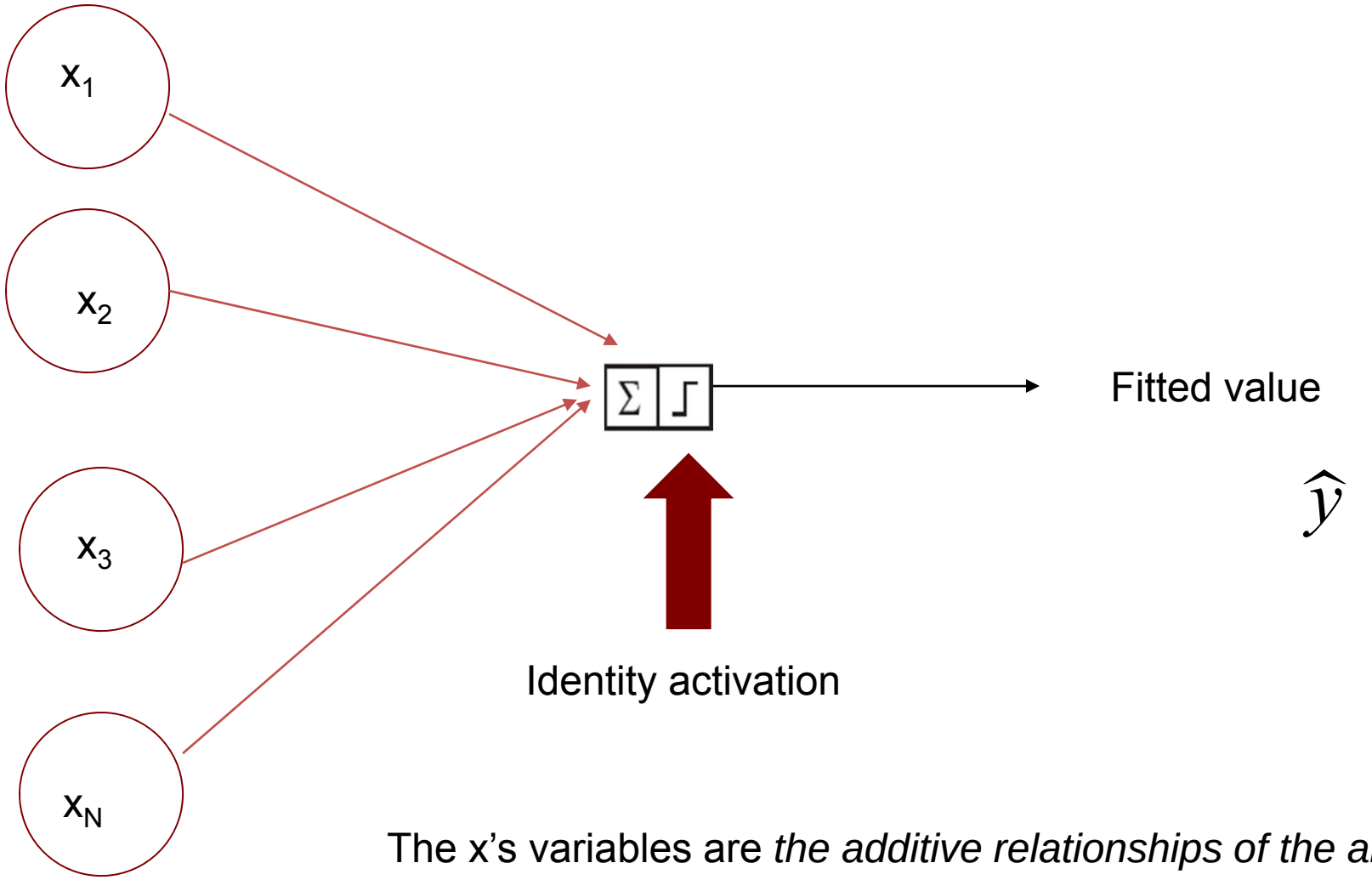
2) $\mathbf{t} = \mathbf{A}\mathbf{A}^{-1}\mathbf{u} + \mathbf{e} = \mathbf{A}\mathbf{u}^{**} + \mathbf{e}, \qquad \mathbf{u}^{**} = \mathbf{A}^{-1}\mathbf{u} \sim (\mathbf{0}, \mathbf{A}^{-1}\sigma^2{}_u)$

$$t_i = g(\sum_{j=1}^{n} a_{ij} u^{**}{}_j) + e_i, \qquad \text{Identity activation}$$

3) $\mathbf{t} = \mathbf{A}^{-1}\mathbf{A}\mathbf{u} + \mathbf{e} = \mathbf{A}^{-1}\mathbf{u}^{**} + \mathbf{e}, \qquad \mathbf{u}^{***} = \mathbf{A}\mathbf{u} \sim (\mathbf{0}, \mathbf{A}^3\sigma^2{}_u)$

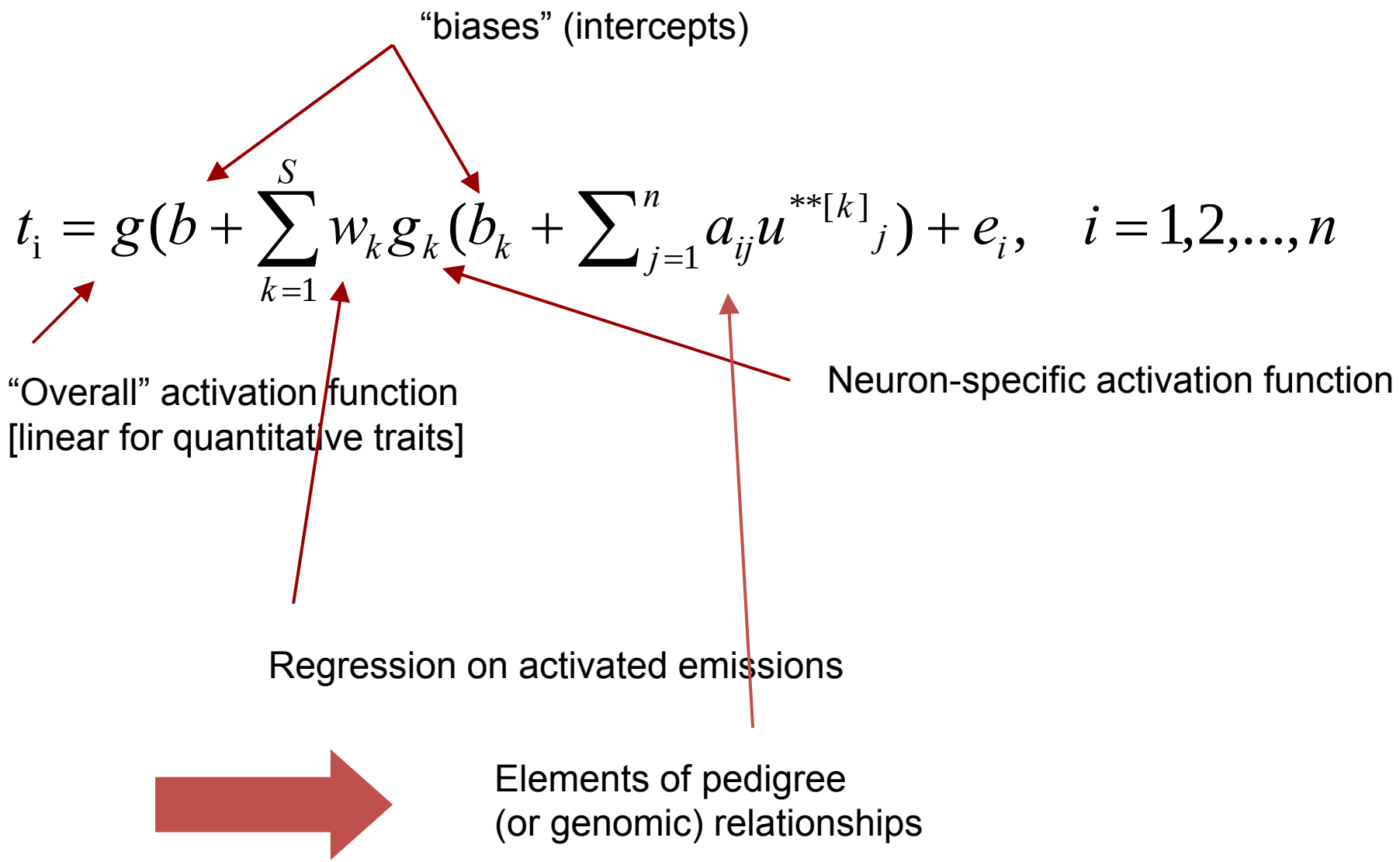
$$t_i = g(\sum_{j=1}^{n} a^{ij} u^{***}{}_j) + e_i, \qquad \text{Identity activation}$$

The infinitesimal model as a linear neural network



The x's variables are *the additive relationships of the animal phenotyed to ALL other individuals in the pedigree*

Other than a naïve theory (the infinitesimal additive model) nothing precludes using what might be a better approximation (Kolmogorov)



Data

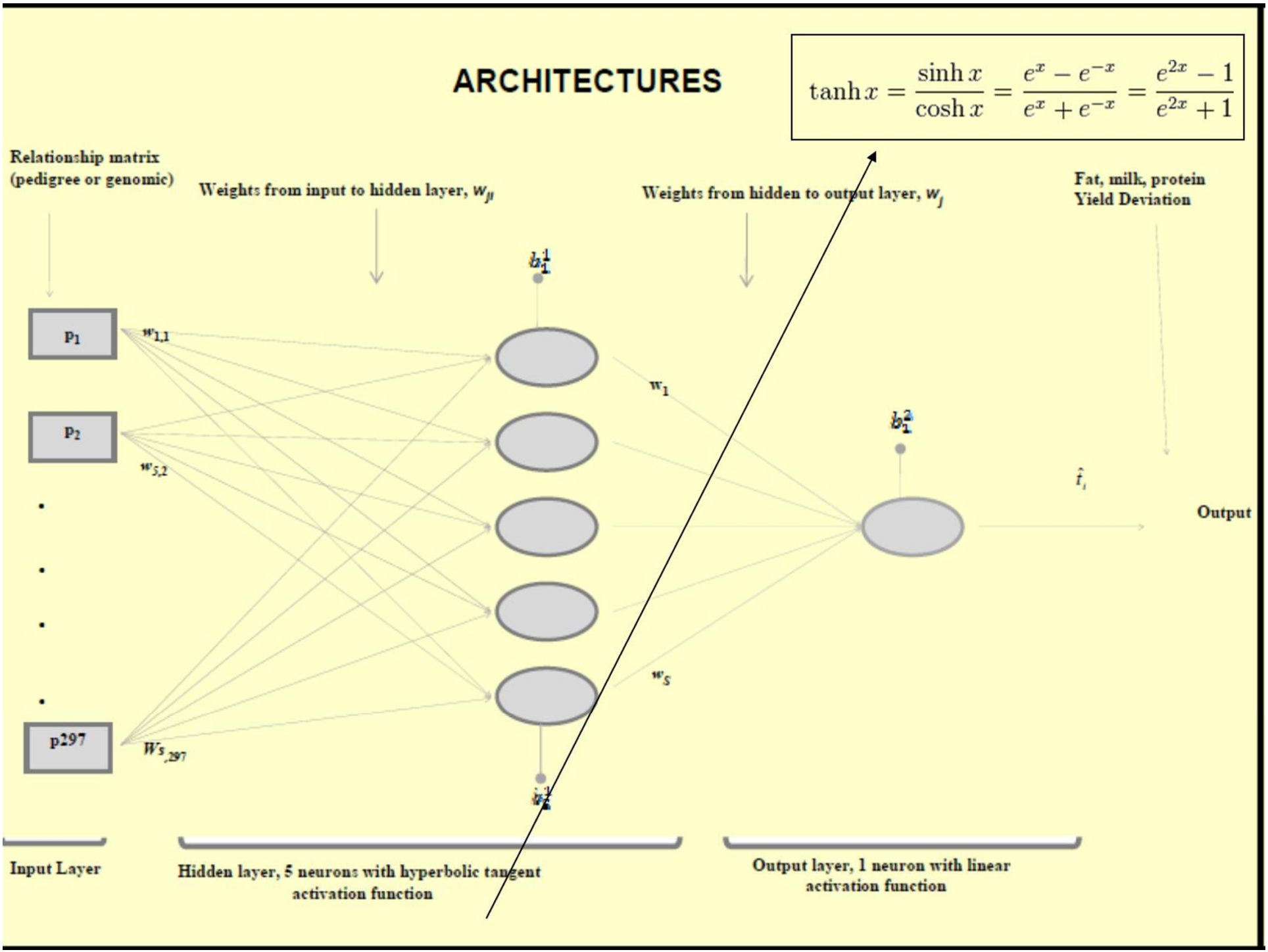
(297 Jersey cows)

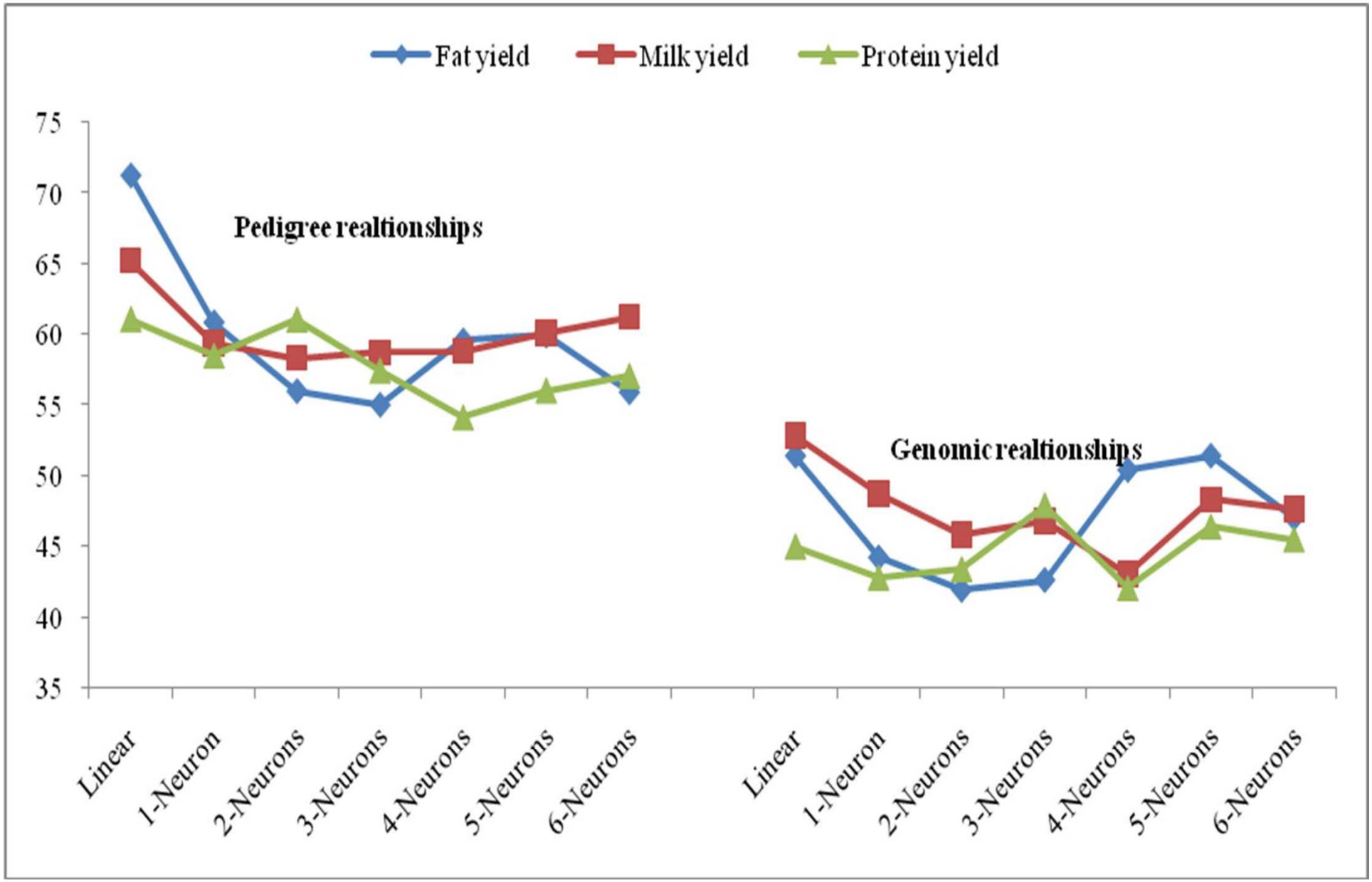
- **Target :** Fat Yield Deviation
  Milk Yield Deviation
  Protein Yield Deviation
- **Inputs :** Elements of Relationship Matrix
  (Pedigree or Genomic, or both)
- **Rationale (again)**

$$\mathbf{y} = \mathbf{u} + \mathbf{e}$$
$$\mathbf{u} \sim (\mathbf{0}, \mathbf{A}\sigma_a^2)$$
$$\mathbf{y} = \mathbf{A}\mathbf{A}^{-1}\mathbf{u} + \mathbf{e}$$
$$= \mathbf{A}\mathbf{u}^* + e$$
$$y_i = \sum_{j=1}^{N} a_{ij} u_j^* + e_i$$

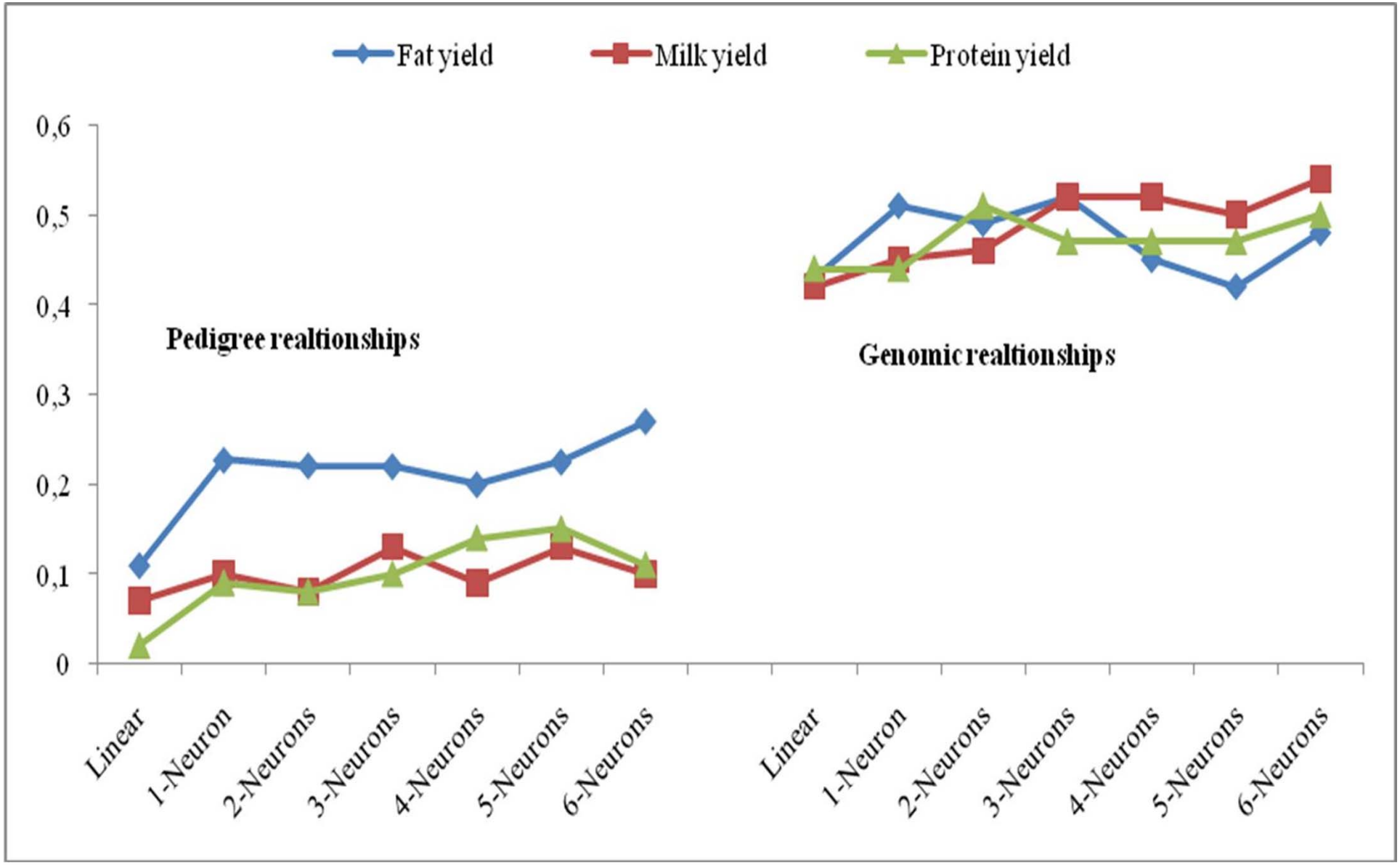
Use elements of
A (or **G**) as inputs in NN

35,798 SNPs used to build **G**
as in Van Raden (2008)


ARCHITECTURES
$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1}$
Relationship matrix (pedigree or genomic)
Weights from input to hidden layer, w_{ji}
Weights from hidden to output layer, w_j
Fat, milk, protein Yield Deviation
b_1^1
p_1
$w_{1,1}$
p_2
$w_{5,2}$
w_1
b_1^2
$\hat{t}_i$
Output
p297
$w_{5,297}$
w_5
b_5^1
Input Layer
Hidden layer, 5 neurons with hyperbolic tangent activation function
Output layer, 1 neuron with linear activation function

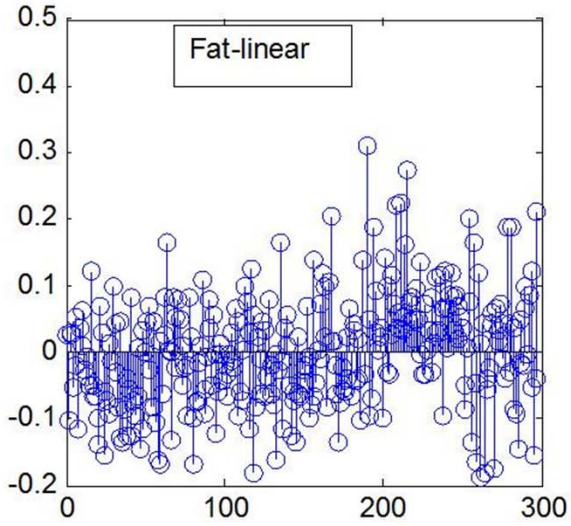


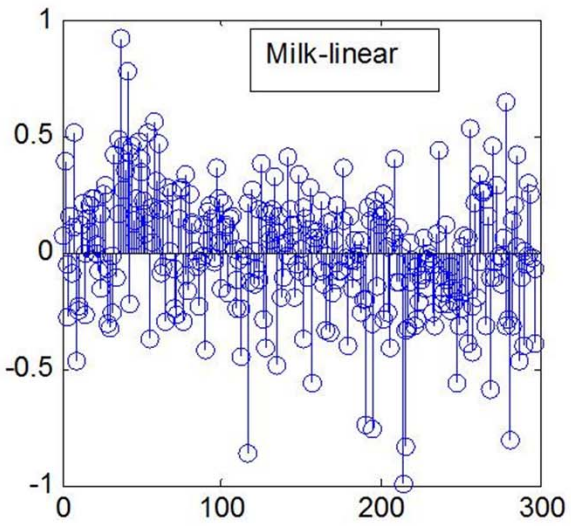
Sum of squared prediction errors in testing set

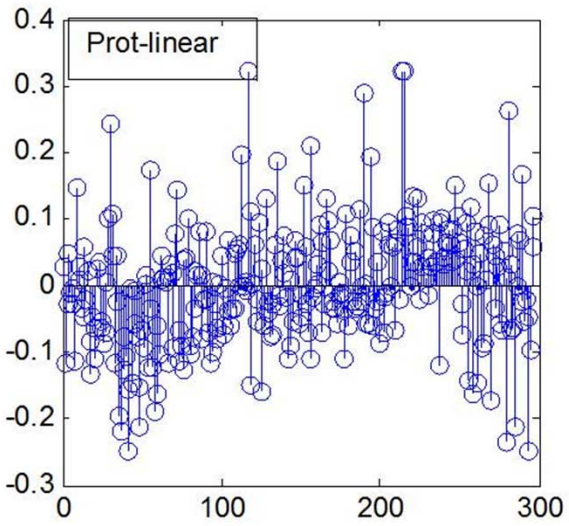


Correlations in testing set

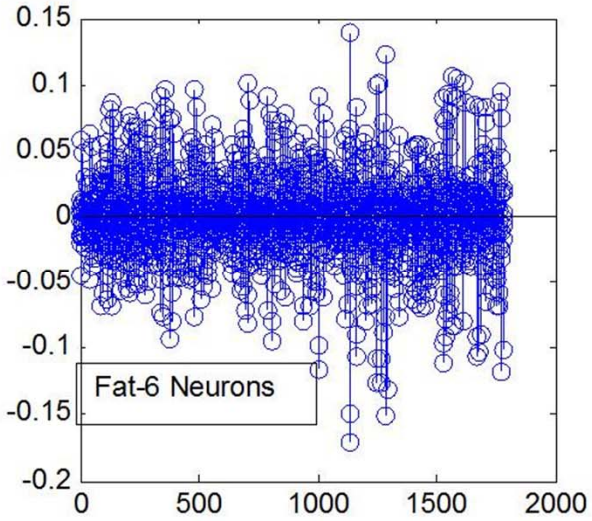
Values of weights (regressions) for the linear and "best" NN (pedigree relationships only

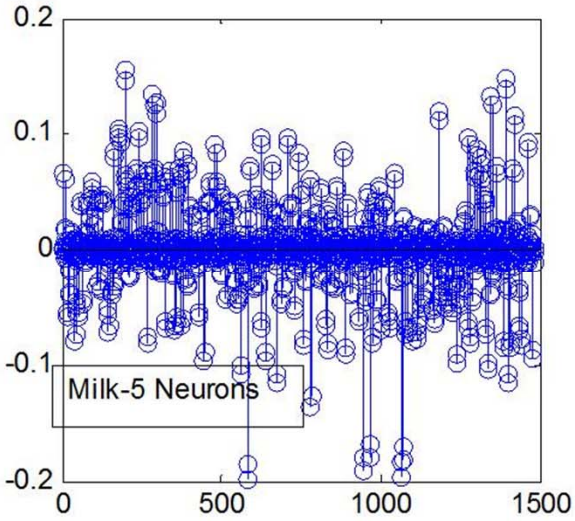


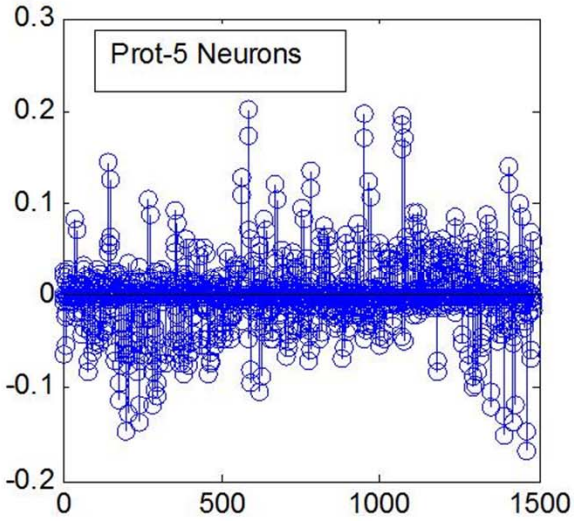




Note the differences in number of weights and in their sizes







REGULARIZATION

WHEAT DATA SET: 599 lines (480 training-119 testing, 50 random repeats)
1279 binary markers

ANN architectures	Linear	1 neuron	2 neurons	3 neurons	4 neurons
Criterion					
Effective number of parameters	299±5.5	260±6.1	253±5.9	238±5.5	220±2.8
Correlations in testing set	0.48±0.03	0.54±0.03	056±0.02	0.57±0.02	0.59±0.02
Mean squared error in testing set	0.99±0.04	0.77±0.03	0.74±0.03	0.71±0.02	0.72±0.02

BENCHMARKS: BAYESIAN LASSO 0.50 4 SVM MODELS 0.50-0.58

ANALYSIS IN PROGRESS BY CROSSA ET AL. (CIMMYT)

Maize corn-flowering Data used in Crossa et al. (2010)

Trait-environment	M-BL	M-RKHS	M-RBFNN
SS-ASI	0.5425	**0.5926**	0.5821
SS-FLF	0.7417	0.6132	**0.7460**
SS-FLM	0.7404	0.6453	**0.7678**
WW-ASI	0.5153	**0.5580**	0.5365
WW-FLF	0.7268	0.5372	**0.7869**
WW-FLM	0.7428	0.5743	**0.7981**
SS-GY	0.4743	**0.5318**	0.5174
WW-GY	**0.5634**	0.5459	0.5586

Maize disease - - GLS -- high density 55k

Sites	M-BL	M-RKHS	M-RBFNN
1	0.2188	0.2099	**0.2604**
2	0.4174	0.4131	**0.4308**
3	**0.5899**	0.5691	0.5823
4	**0.5215**	0.5044	0.5058
5	0.3419	0.3064	**0.3442**
6	**0.2842**	0.2535	0.2775

Maize under 2 level of drought -- high density 55k

Environment	M-BL	M-RKHS	M-RBFNN
GY-Moderate drought	0.6333	0.5591	**0.6531**
GY-Severe drought	**0.4104**	0.3652	0.3910

Wheat trait 1

Sites	M-BL	M-RKHS	M-RBFNN
1	0.5969	**0.6630**	0.6581
2	0.6861	**0.7278**	0.7069
3	0.6224	**0.6943**	0.6866
4	0.0673	0.1419	**0.1840**
5	0.6481	**0.6824**	0.6744
6	0.3798	**0.4659**	0.4586
7	0.5984	0.6235	**0.6284**
8	0.5493	0.6054	**0.6100**
9	0.5374	0.5821	**0.5827**
10	0.4775	**0.5024**	0.4274
11	0.7721	0.7422	**0.8039**

Wheat trait2

Site	M-BL	M-RKHS	M-RBFNN
1	0.4830	**0.5216**	0.5149
2	0.6928	0.6753	**0.7085**
3	0.2285	**0.3889**	0.3827
4	0.4610	0.5508	**0.5557**
5	0.7509	0.7147	**0.7880**
6	0.8101	0.8031	**0.8399**
7	0.4695	**0.5374**	0.5285
8	0.8345	0.8261	**0.8657**

PUNCH LINE:
over 35 trials, the winner is…

M-BL	**M-RKHS**	**M-RBFNN**
14%	34%	52%
5	12	18

Any concerns about the predictive ability of non-parametric methods, relative to those that "*help to understand genetic architecture*"?

CONCLUSION

- The mechanistic value of the additive model is dubious in the face of complexity of biological systems
- Refocus on prediction of behavior of complex systems
- RKHS methods?
- Neural networks as universal approximators?
- The infinitesimal model is a RKHS regression
- The infinitesimal model is a naïve network
- The new bag of tricks is the “mother of the bags” (it includes the old tricks in the bag)