Linking Phenotypic Models and Genotypic Data

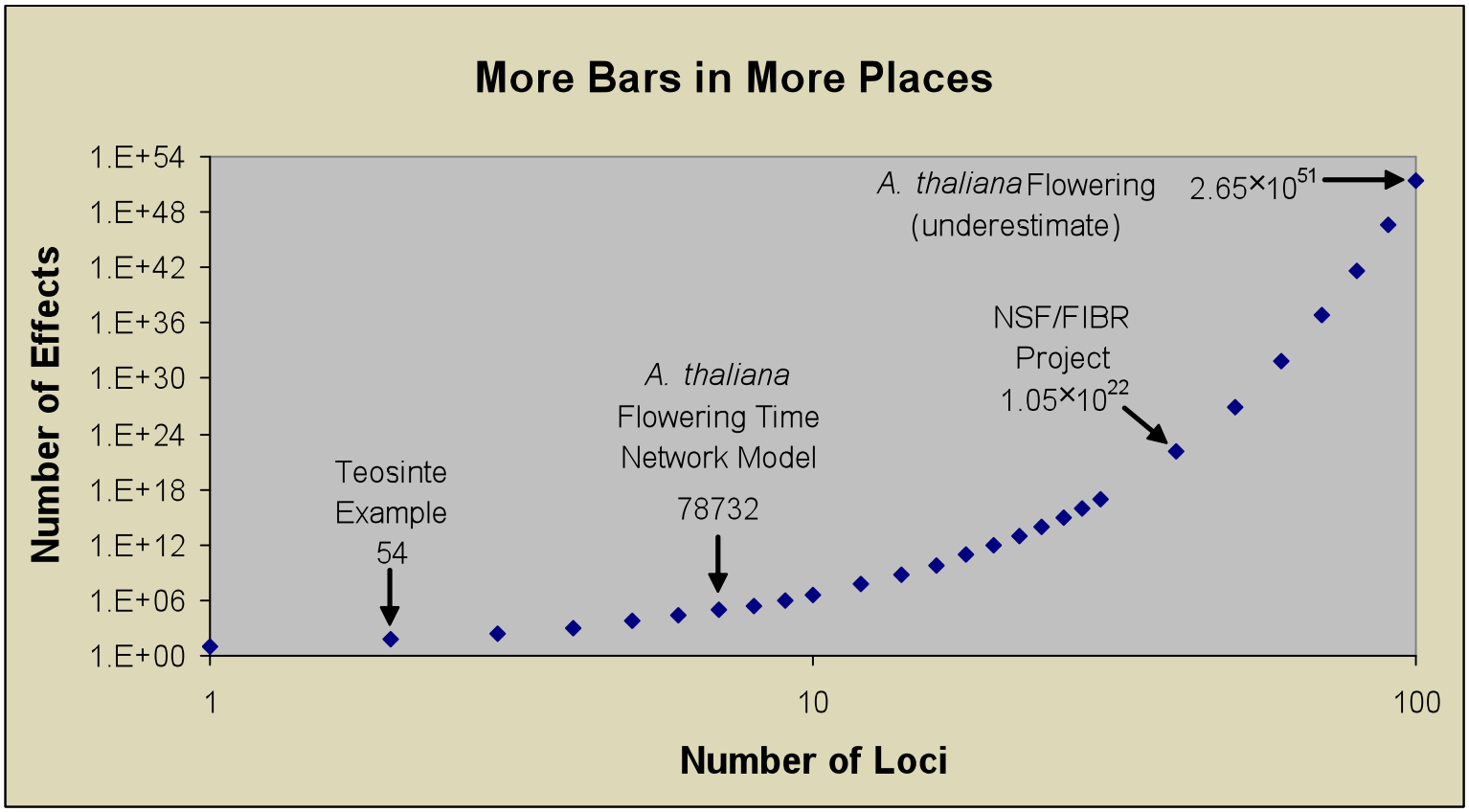
iPlant Collaborative™

Four modeling approaches (I)

- General linear models that relate phenotypes to genetic markers
  - Used to find markers that may be close to genes that control traits – i.e. does not require prior knowledge of relevant loci
  - Once mapped, markers can be used to predict phenotypes for novel genotypes not used in analysis
  - GLMs are not mathematically efficient at representing *G*×*E* or *G*×*G* interactions
- Ecophysiological crop models which relate phenotypes to genotypes by fitting crop coefficients to different cultivars
  - Can predict phenotypes in novel environments but cannot predict behaviors of genotypes not fit
  - Efficient representation of *G*×*E* interactions

iPlant Collaborative™

Representing networks as GLMs



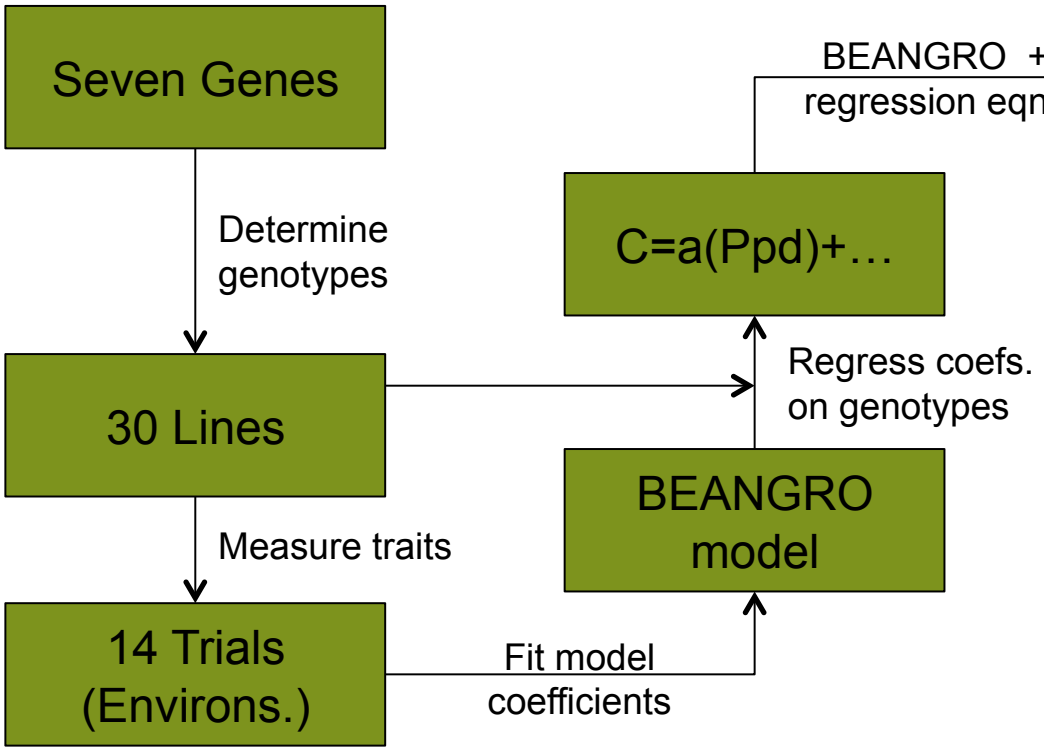
iPlant Collaborative

Four modeling approaches (II)

- Ecophysiological crop models whose coefficients are regressed on "candidate alleles"
  - Efficient representation of *G*×*E* and *G*×*G* interactions
  - Can make predictions for novel environments and novel genotypes
  - Require relevant loci and alleles to be known ahead of time
- Ecophysiological models whose coefficients are mapped to genetic markers
  - Can be used to find regions ("QTL") that bracket potentially relevant genes
  - Can predict outcomes for new genotypes in new environments
  - Do not require prior knowledge of relevant genes
  - Efficient representation of interactions

iPlant Collaborative

Genes + GLMs + crop models



Trait	% Var
Seed yld.	31
Seed wt.	58
Flowering	84
Maturity	85
Max LAI	52
Canopy dwt.	36
Harvest ind.	Zero!

White & Hoogenboom, 1996

iPlant Collaborative

Functional-structural models

Towards a 3-D genotype–phenotype model of rice

Lifeng Xu, Michael Henke, Jun Zhu, Winfried Kurth, Gerhard Buck-Sorlin

doi:10.1093/aob/mcq264

Annals of Botany, Special Issue, Apr 2011

Cieslak et al. 2011

http://ladies-with-bottle.blogspot.com/2010/06/chinese-fruits-kiwi-fruit.html

G2P model extender tool

Genetic candidates

QTL

Association mapping

$$\theta = \mu + p_1(\mathrm{Mk}_1) + p_2(\mathrm{Mk}_2)(\mathrm{Mk}_3)\ldots$$

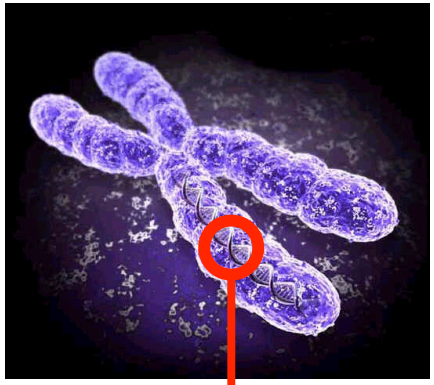
iPlant Collaborative

Example applications

- Given a gene network (G2P) model based on a few lab strains, expand it to a large number of genotypes
- Given a ecophysiological model, convert it to a G2P model by linking to candidate genes &/o QTL (incl. *de novo* detection of the latter)
- For any type of model, check for genotypic or environmental dependencies of putative constants that would signal model incompleteness or inadequacy

iPlant Collaborative™

A parametric correspondence principle



Polymorphic locus

Alleles: A1…An

Phenotypes: Y1….Yn

Model

Some numerical parameter

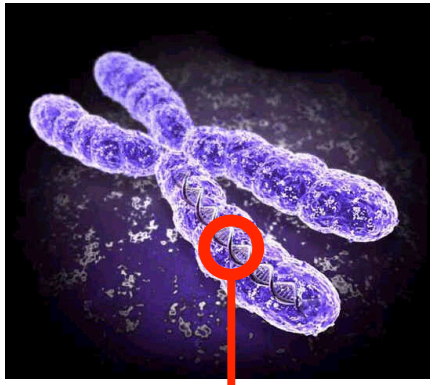
Range of values in a set of mapping lines

2.71…3.14

y1….yn : predictions

iPlant Collaborative

A parametric correspondence principle



Polymorphic locus

Alleles: A1…An

Model

Some numerical parameter

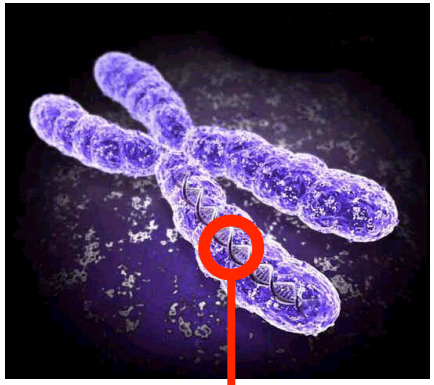
Range of values in a set of mapping lines

2.71…3.14

Phenotypes: Y1….Yn ⟷ A correspondence here… ⟷ y1….yn : predictions

iPlant Collaborative

A parametric correspondence principle



Polymorphic locus

Alleles: A1…An

…induces a correspondence here

Model

Some numerical parameter

Range of values in a set of mapping lines

2.71…3.14

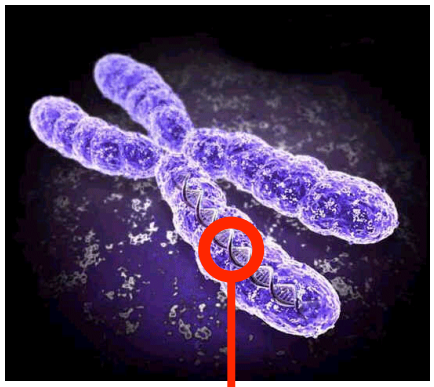
Phenotypes: Y1….Yn

A correspondence here…

y1….yn : predictions

iPlant Collaborative

A parametric correspondence principle



Polymorphic locus

Can be interrelated by regression

Model

Some numerical parameter

Range of values in a set of mapping lines

Alleles: A1…An

…induces a correspondence here

2.71…3.14

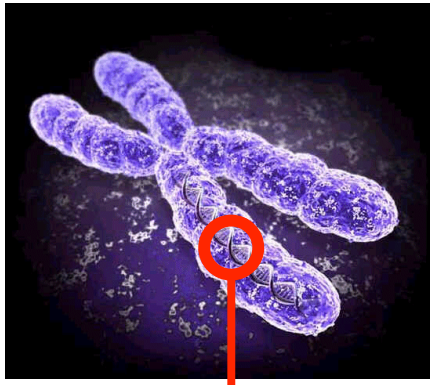
Phenotypes: Y1….Yn

A correspondence here…

y1….yn : predictions

iPlant Collaborative

A parametric correspondence principle



Polymorphic locus

No matter what the model's mathematical form !!!

Can be interrelated by regression

Model

Some numerical parameter

Range of values in a set of mapping lines

Alleles: A1…An

…induces a correspondence here

2.71…3.14

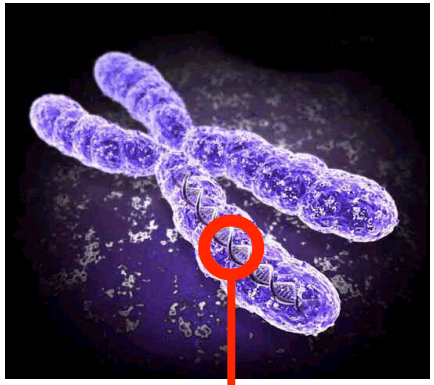
Phenotypes: Y1….Yn

A correspondence here…

y1….yn : predictions

iPlant Collaborative

A parametric correspondence principle



Polymorphic locus

No matter what the model's mathematical form !!!

Model

Can be interrelated by regression

Category values in a set of mapping lines

Alleles: A1…An

…induces a correspondence here

B, B, S, …

Can be categories rather than numbers

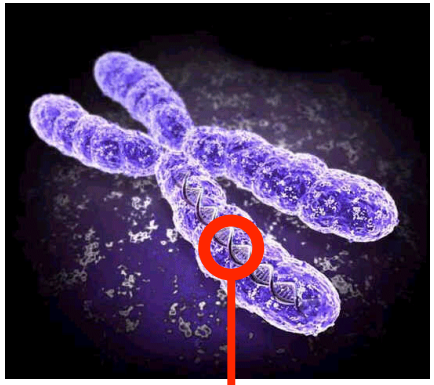
Phenotypes: Y1….Yn

A correspondence here…

y1….yn : predictions

iPlant Collaborative

A parametric correspondence principle



Polymorphic locus

No matter what the model's mathematical form !!!

Model

$y_i = \alpha X_{\text{B or S}} + \mu$

Category values in a set of mapping lines

B, B, S, …

Association – the flip side of regression!!!!

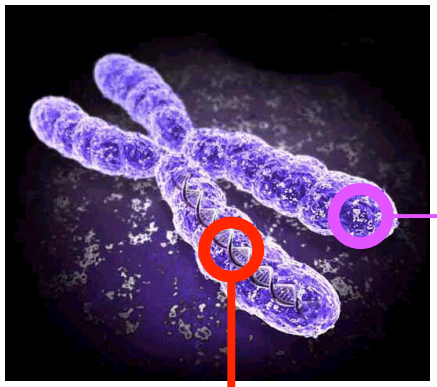
Phenotypes: Y1….Yn

A correspondence here…

y1….yn : predictions

iPlant Collaborative

A parametric correspondence principle



Unrelated marker

Marker near influential gene

Model

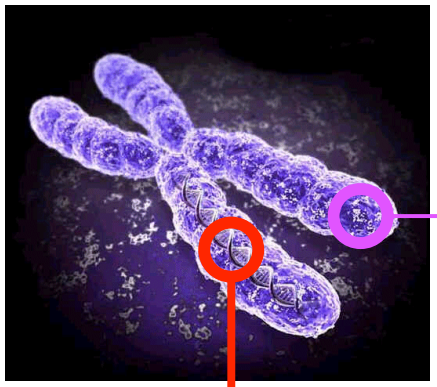
T1G11	S	B	S	B	B	S	S	B	B
F21M12	S	B	S	B	B	S	S	B	B
MSAT1.10	S	B	B	S	B	S	S	B	B
NGA248	B	S	S	S	B	S	B	S	S
T27K12	B	B	S	S	B	S	B	B	S
NGA128	B	B	S	S	B	S	B	B	S
F5I14	B	B	S	B	B	S	B	B	B
MSAT1.13	B	B	S	B	S	S	B	B	B

Each column is an offspring produced by crossing parent lines B & S

Phenotypes: Y1….Yn ⟷ A correspondence here… ⟷ y1….yn : predictions

iPlant Collaborative

A parametric correspondence principle



Model

$y_i = \alpha X_{\text{B or S}} + \mu$ → No skill

Unrelated marker

T1G11	S	B	S	B	B	S	S	B	B
F21M12	S	B	S	B	B	S	S	B	B
MSAT1.10	S	B	B	S	B	S	S	B	B
NGA248	B	S	S	S	B	S	B	S	S
T27K12	B	B	S	S	B	S	B	B	S
NGA128	B	B	S	S	B	S	B	B	S
F5I14	B	B	S	B	B	S	B	B	B
MSAT1.13	B	B	S	B	S	S	B	B	B

Marker near influential gene

$y_i = \alpha X_{\text{B or S}} + \mu$

Skilled predictor

Phenotypes: Y1….Yn

iPlant Collaborative

$$
\begin{cases}
\dfrac{dmPHYB}{dt} = -mPHYB + rmPHYB \cdot LHY^2_{t-\tau_1} \cdot (1 - rxPHYB \cdot CRY2) \\
\dfrac{dPHYB}{dt} = -rpPHYB \cdot (1 - rlPHYB \cdot LIGHT) \cdot (PHYB - mPHYB) \cdot TEMP \\
\dfrac{dmCRY2}{dt} = -mCRY2 + rmCRY2 \cdot LHY^2_{t-\tau_2} \\
\dfrac{dCRY2}{dt} = -rpCRY2 \cdot (1 - rlCRY2 \cdot LIGHT) \cdot (CRY2 - mCRY2) \cdot TEMP \\
\dfrac{dmLHY}{dt} = -mLHY + rmLHY \cdot (1 + rbLHY \cdot PHYB) \cdot TOC1_{t-\tau_3} \\
\dfrac{dLHY}{dt} = -rpLHY \cdot (1 + rlLHY \cdot LIGHT) \cdot (LHY - mLHY) \\
\dfrac{dmTOC1}{dt} = -mTOC1 + rmTOC1 / (1 + LHY^{nmTOC1}) \\
\dfrac{dTOC1}{dt} = -rpTOC1 \cdot (TOC1 - mTOC1) \\
\dfrac{dmGI}{dt} = (-\beta \cdot mGI + rpGI \cdot LHY^2_{t-\tau_4}) \cdot TEMP \\
\dfrac{dmCO}{dt} = (-\beta \cdot mCO + rpCO \cdot LHY^2_{t-\tau_5} \cdot (1 + rgCO \cdot mGI)) \cdot TEMP \\
\dfrac{dmFCA}{dt} = (betaFCA + betaFPA + betaFVE) \cdot TEMP \\
\dfrac{dmSOC1}{dt} = (-\beta \cdot mSOC1 + rpSOC1 \cdot (1 + rtSOC1 \cdot mFCA) \\
\qquad \cdot (1 + rcSOC1 \cdot mCO \cdot LIGHT)(1 + rySOC1 \cdot CRY2 \cdot LIGHT)) \cdot TEMP \\
\dfrac{dmLFY}{dt} = -mLFY + 1.0 / (1 + \exp(-rpLFY \cdot mSOC1 - biasLFY))
\end{cases}
$$

Time series: mRNA's Proteins Environment

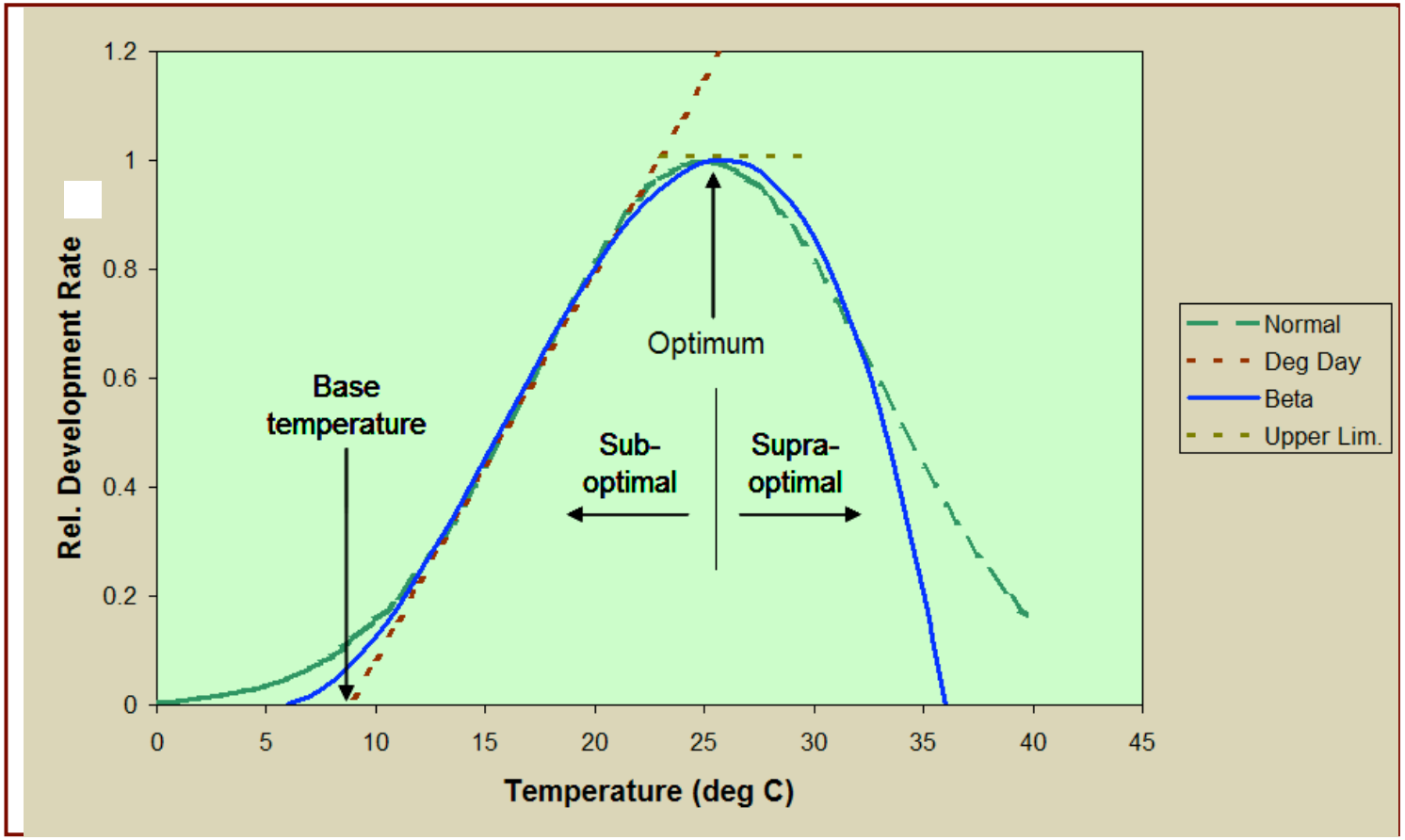
Dong, 2003

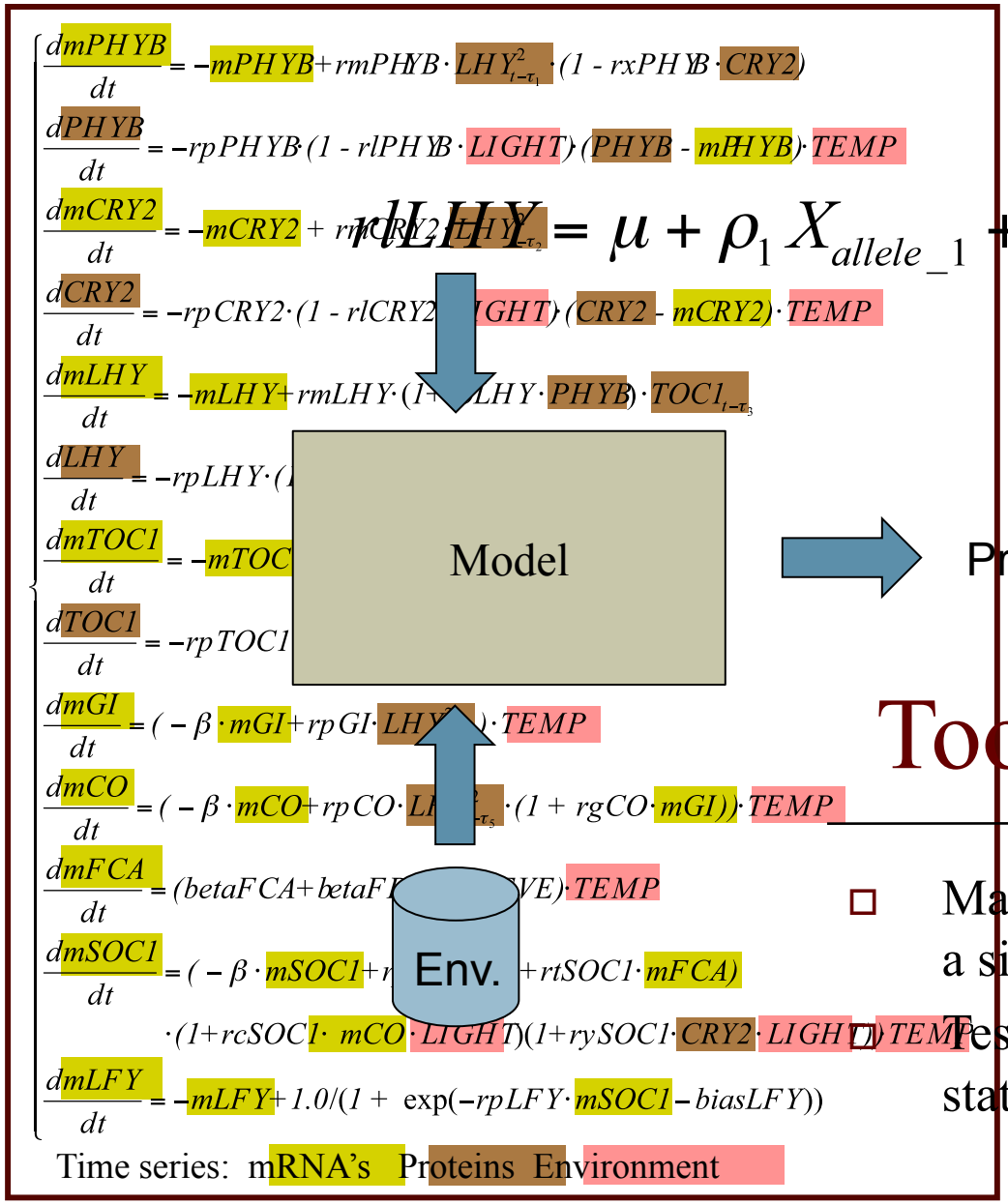
Where's the linearity?

- Within networks
  - Epistasis is totally trivial
  - Dominance is easy to explain
  - But additivity is a major mystery that requires
    - Linearity &
    - Is readily disrupted in transduction cascades

iPlant Collaborative

Quasi-linear regimes



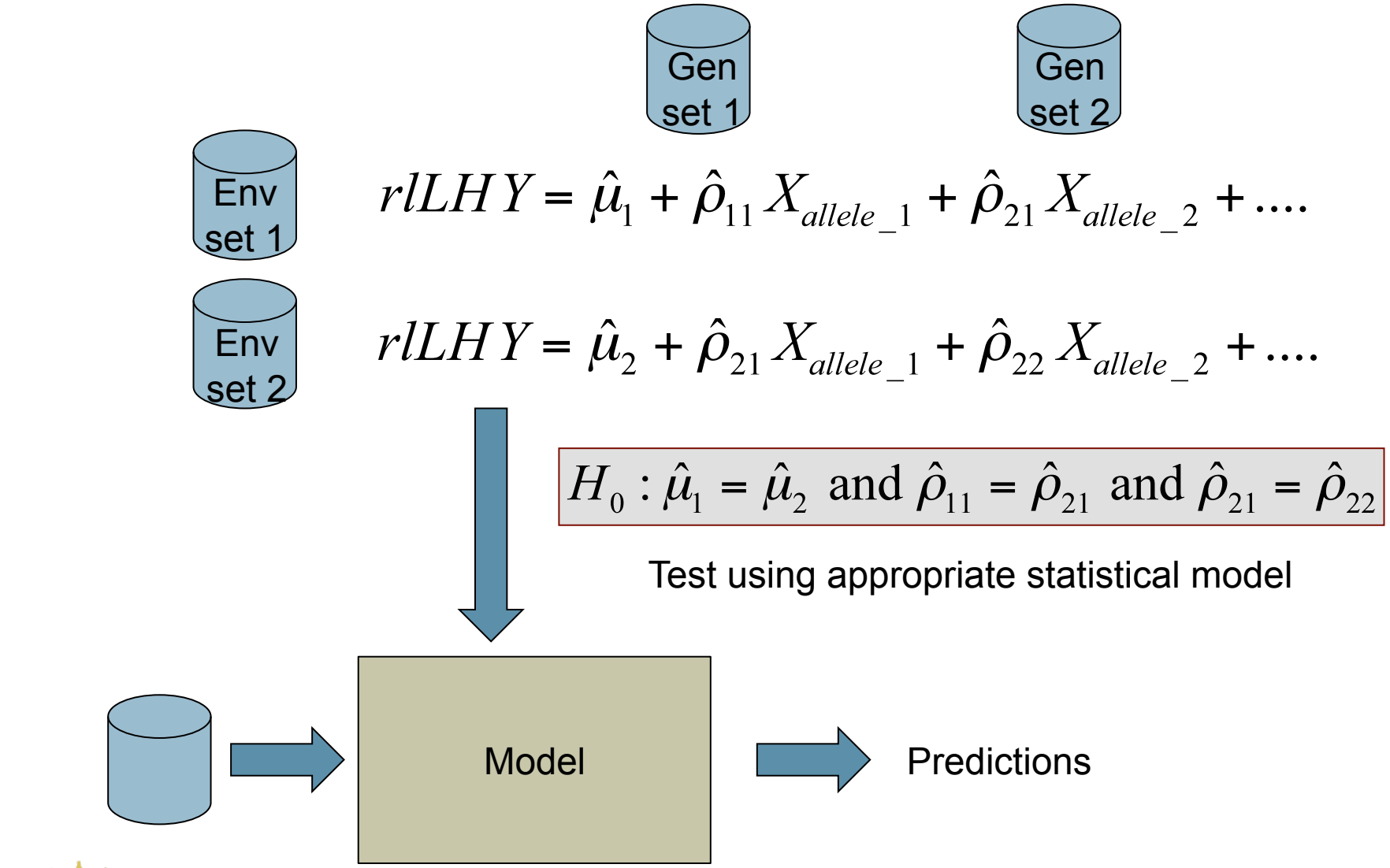


$$rmLHY = \mu + \rho_1 X_{allele\_1} + \rho_2 X_{allele\_2} + \ldots.$$

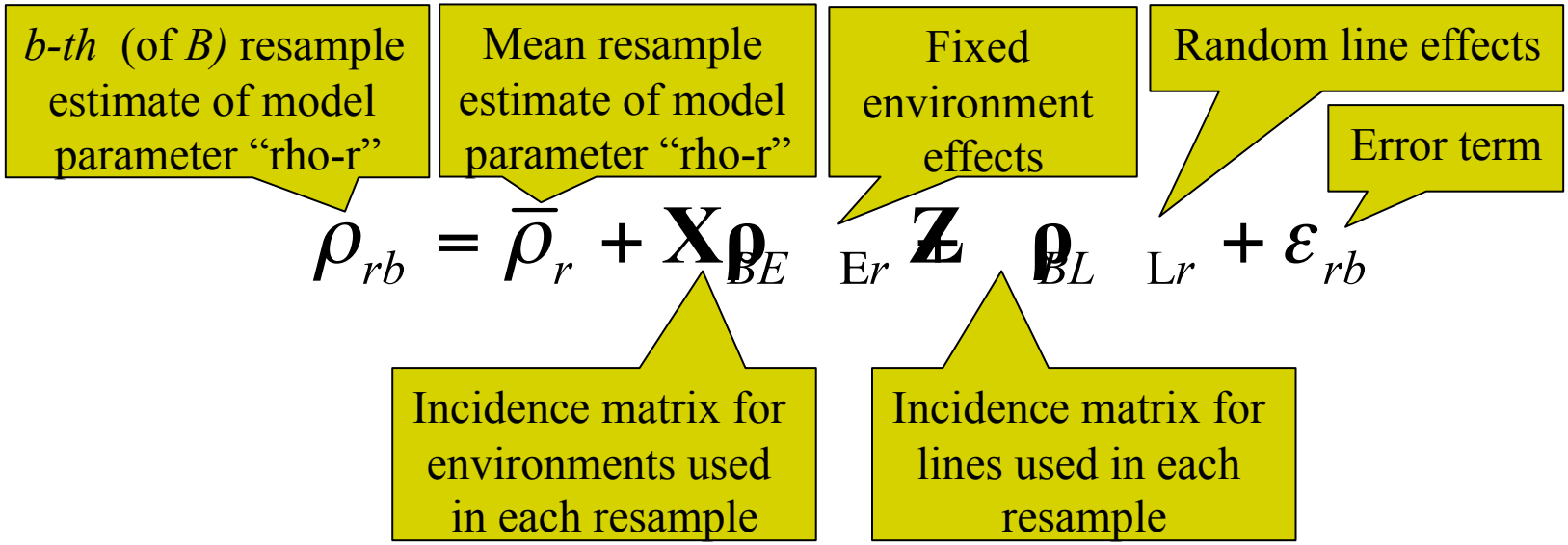
Tool aims

- Make models in a single step
- Test them statistically

iPlant Collaborative

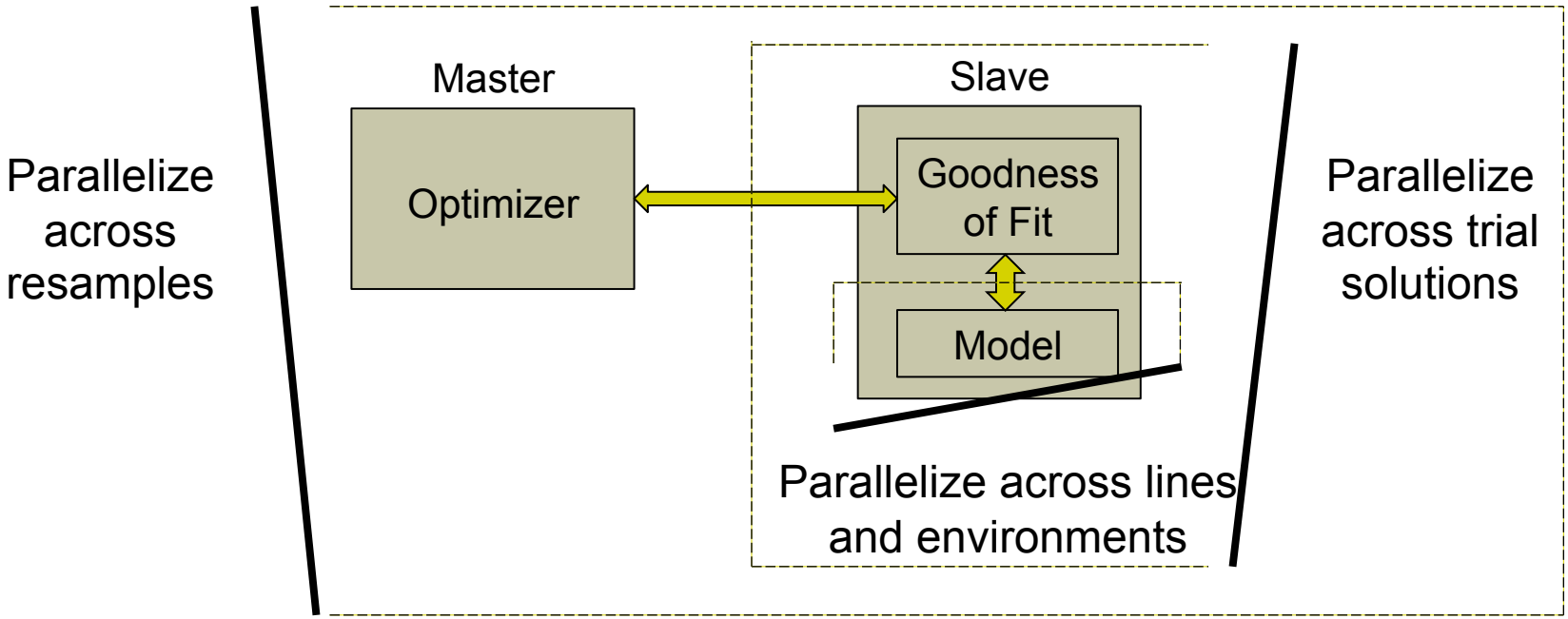

Gen set 1
Gen set 2
Env set 1
Env set 2
$rlLHY = \hat{\mu}_1 + \hat{\rho}_{11} X_{allele\_1} + \hat{\rho}_{21} X_{allele\_2} + \ldots.$
$rlLHY = \hat{\mu}_2 + \hat{\rho}_{21} X_{allele\_1} + \hat{\rho}_{22} X_{allele\_2} + \ldots.$
$H_0 : \hat{\mu}_1 = \hat{\mu}_2$ and $\hat{\rho}_{11} = \hat{\rho}_{21}$ and $\hat{\rho}_{21} = \hat{\rho}_{22}$
Test using appropriate statistical model
Model
Predictions

Statistical model of tool outputs



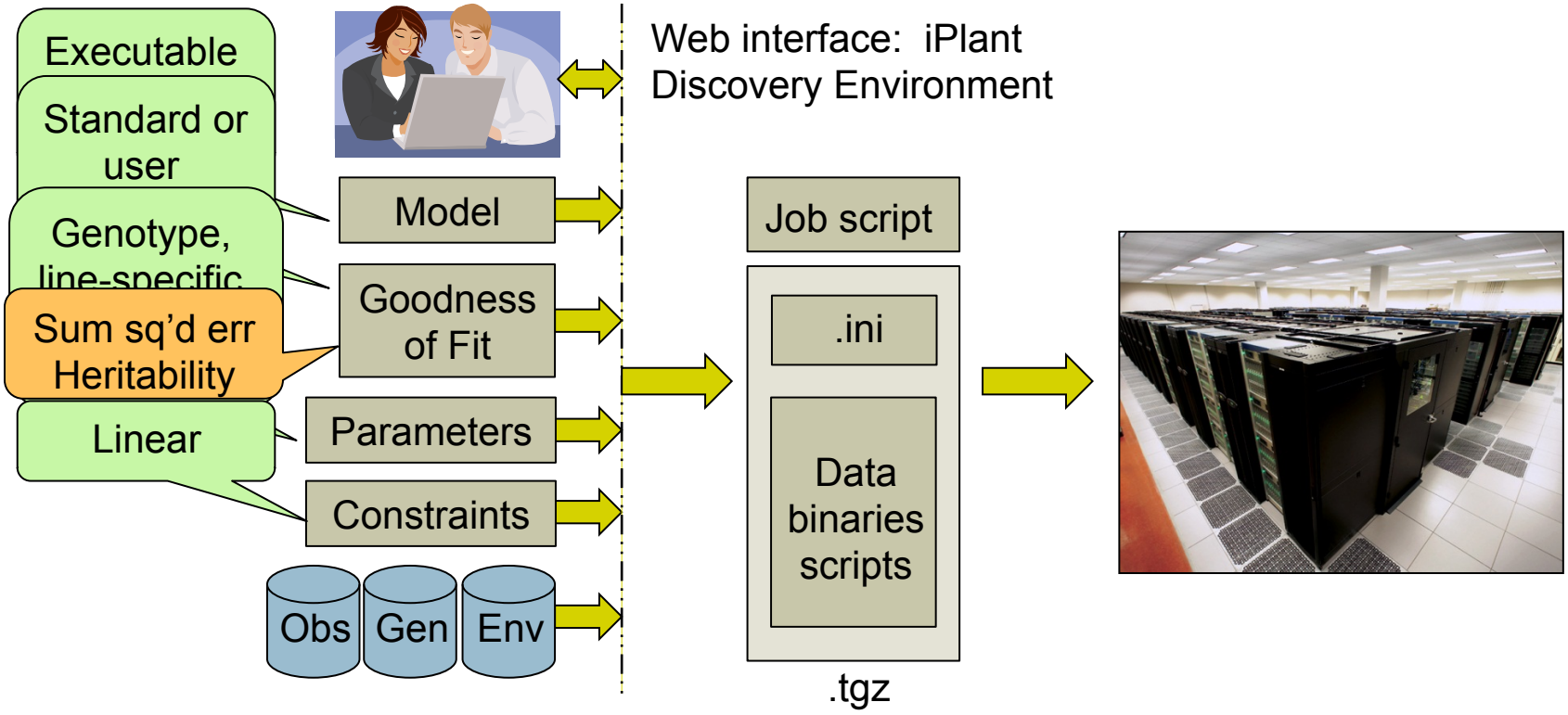
iPlant Collaborative

Architecture



iPlant Collaborative

Work flow I


Executable
Standard or user
Genotype, line-specific
Sum sq'd err Heritability
Linear
Model
Goodness of Fit
Parameters
Constraints
Obs
Gen
Env
Web interface: iPlant Discovery Environment
Job script
.ini
Data binaries scripts
.tgz

Work flow II

Web interface: iPlant
Discovery Environment

MLM test

Distribution of parameter
point estimates and
validation test of model

Test cases

- Schmitt Lab
  - Wilczek *et al.* 2009 model applied to 266 *A. thaliana* ecotypes in 10 European plantings
  - Extended model applied to 2 plantings of 106 Col/Kas RILs + 10 plantings of 34 RILs
- 218 plantings of 2 wheat lines using
  - A gene network model of anthesis date
  - An extant ecophysiological model of the same trait
- PGRI shade avoidance project – functional structural model of *B. rapa* (Weinig, Maloof, Das, Welch)
- *Cameacrista* model (Singer lab)

iPlant Collaborative™