NC STATE UNIVERSITY

Study Genetic Basis and Pathways of Complex Traits

Zhao-Bang Zeng

Department of Statistics

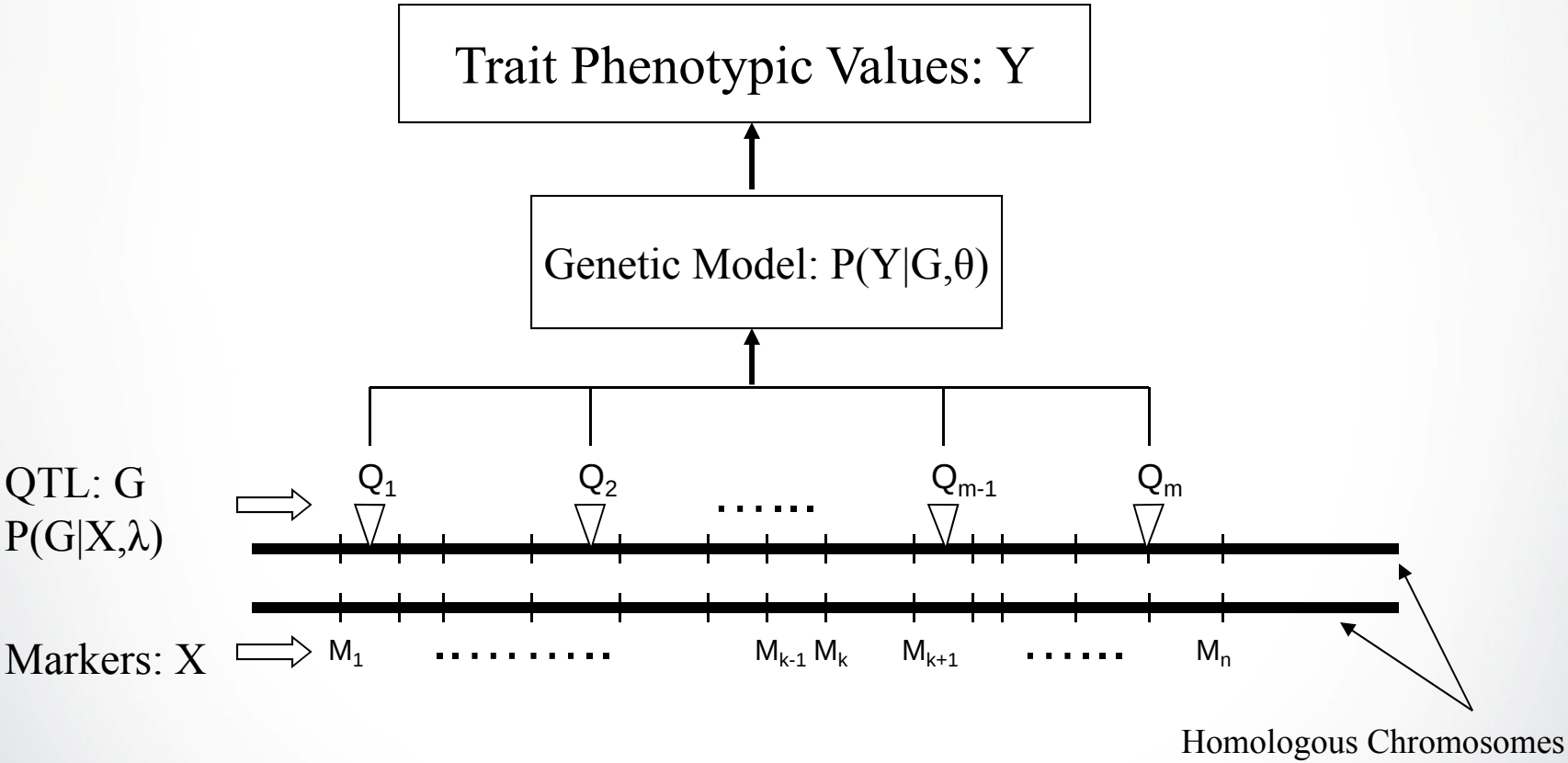
Department of Genetics

Bioinformatics Research Center

North Carolina State University

USA

QTL Mapping

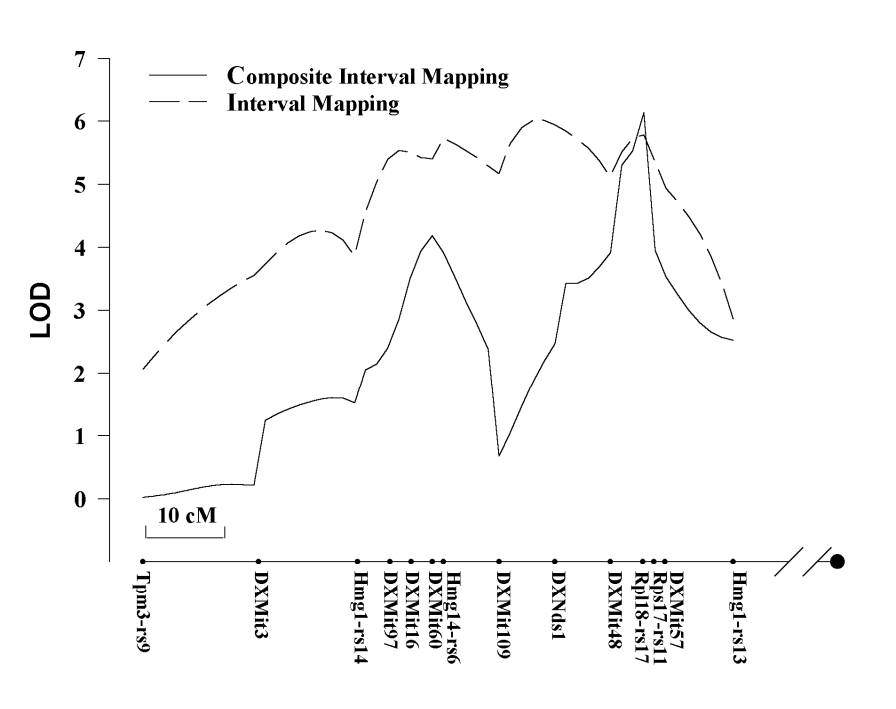


Likelihood of Data: $P(Y,X) = P(Y|X)\,P(X) = \sum_G P(Y|G,\theta)\,P(G|X,\lambda)\,P(X|\delta)$

Inferring the relationship between genotypes and phenotypes

Study issue: How to deal with linkage effects of multiple QTL?

Possibility of "ghost" QTL

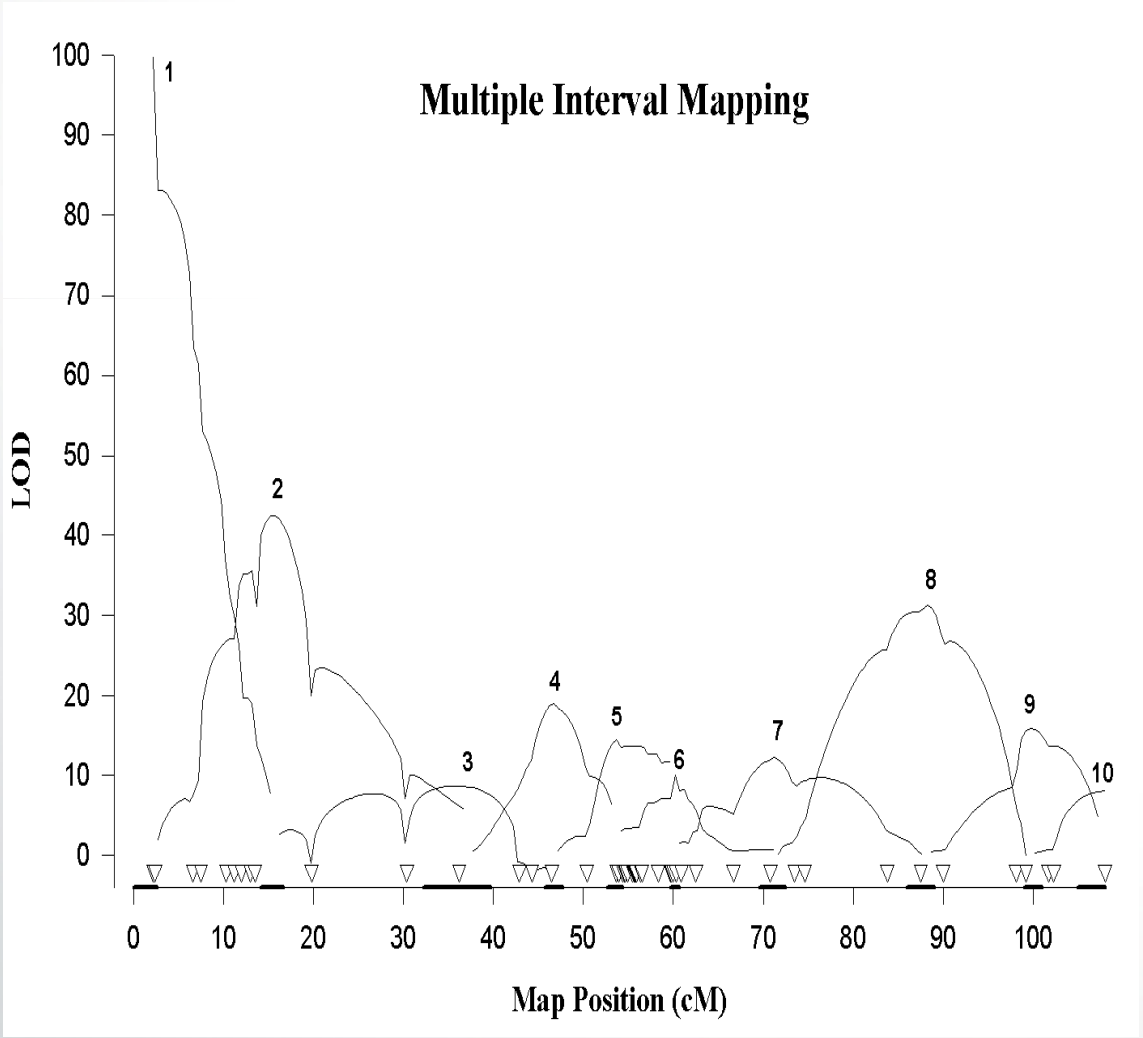


IM: Lander and Botstein (1989): CIM: Zeng (1993), (1994)

Solution 1: Use the information and a Markovian property of multiple markers to make the search in different regions *statistically independent*.

Solution 2: Directly search for multiple QTL

Study issue: How to search for multiple QTL?



Solution: Use multiple-step conditional sequential search (utilize the Markovian property and avoid multiple-dimensional search)

MIM: Kao et al (1999)

Study issue: How to map epistatic QTL?

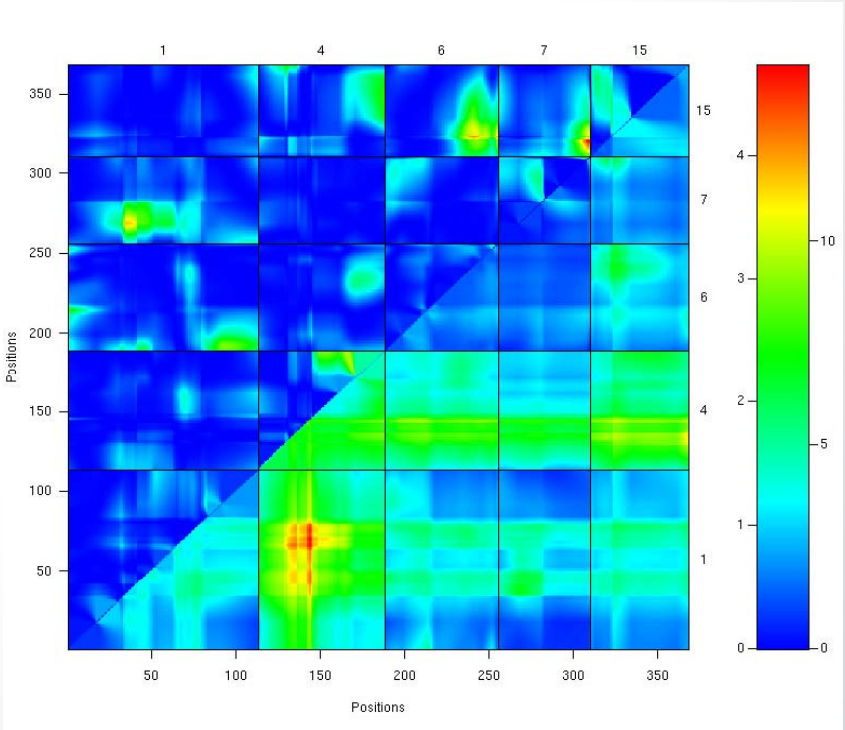
- QTL epistasis is ubiquitous
- It is difficult to find significant QTL epistasis:
  - Small sample size
  - Multiple-dimensional genome search: low statistical power
  - In reality, search for QTL is always biased for main-effect QTL, not much effort has yet been put for searching for epistatic QTL

But still how to map epistatic QTL, particularly those QTL that do not have significant main effects but significant epistatic effects?

Problems of current methods for searching epistatic QTL

- The 2D pattern can be misleading very easily
  - Due to complex linkage and epistatic structure of multiple QTL
- Low statistical power for this 2D genome scan
  - Genetic variance of other QTL is uncontrolled

Perform a 2D genome-scan and interpret the result on the face value



Our solution: A three-stage search strategy

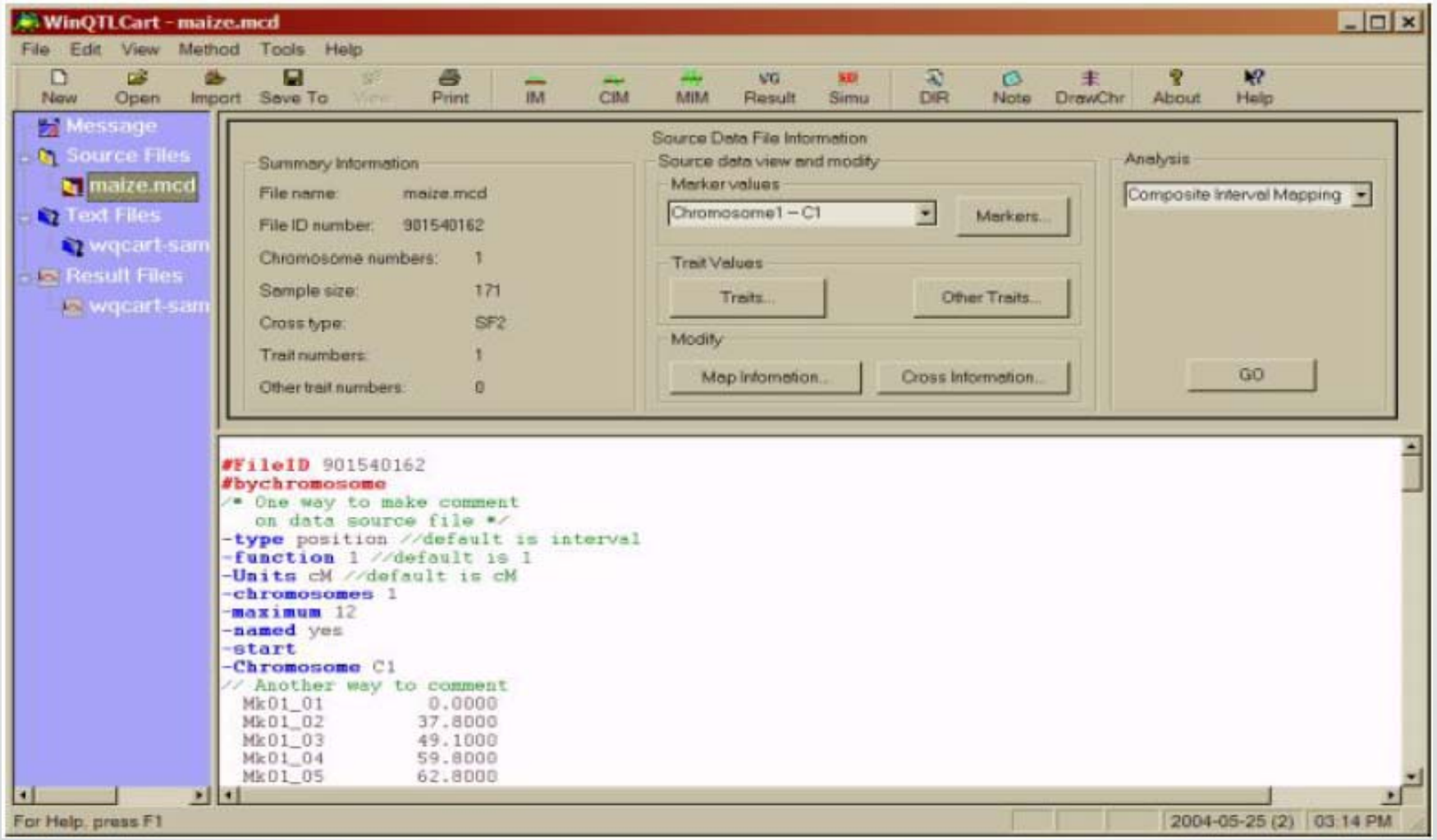
1. Search for main-effect QTL first (because most QTL effects are due to main effects)
2. Search for epistatic QTL that interact with main-effect QTL (1-D genome scan, those are likely the next most abundant)
3. Search for additional epistatic QTL pairs (2-D genome scan, looking for what is left)

An important point: Use of an orthogonal genetic model can help a robust inference of complex epistasis, a justification for the procedure

MIM with epistasis: Laurie et al (submitted)

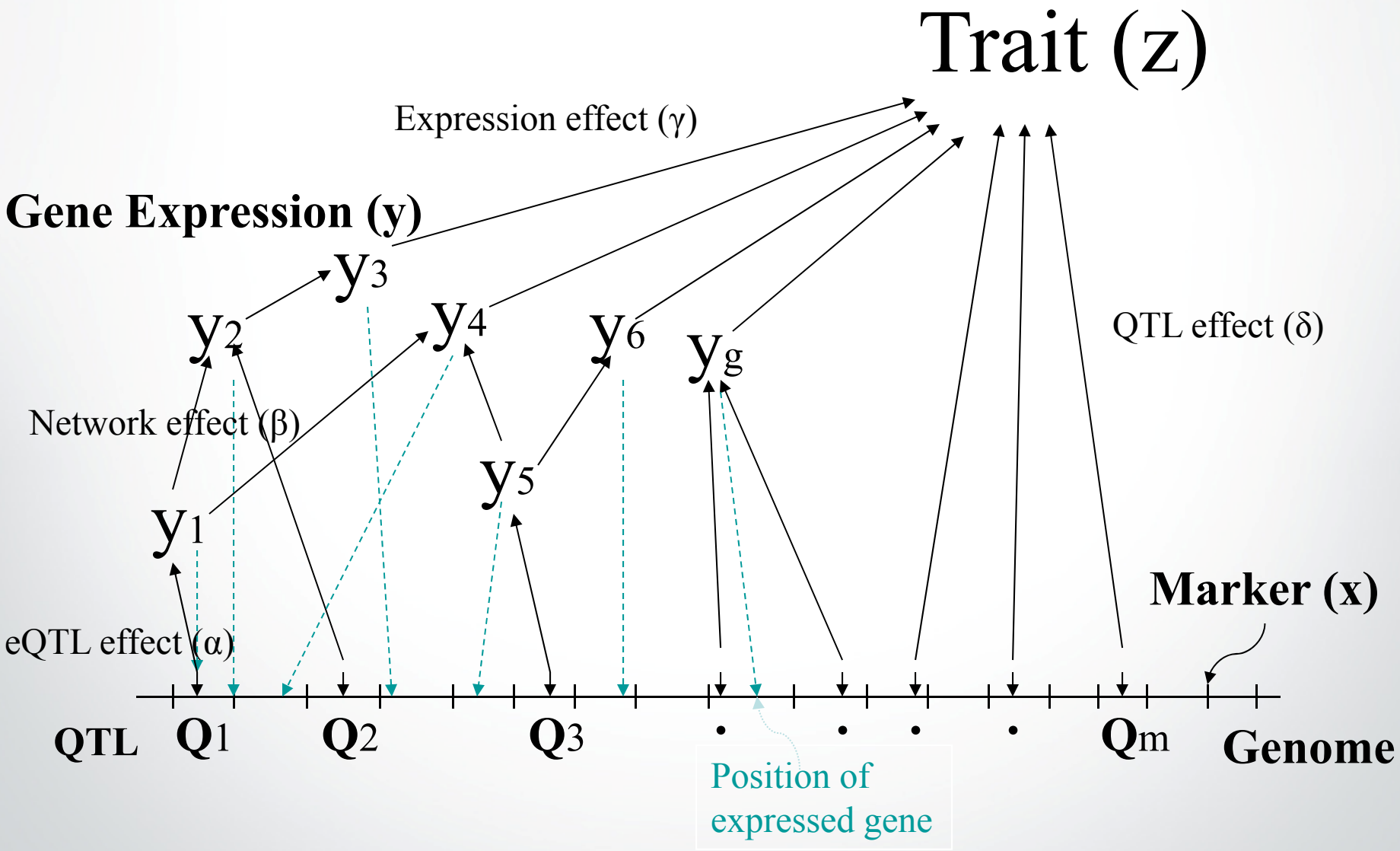
NC STATE UNIVERSITY

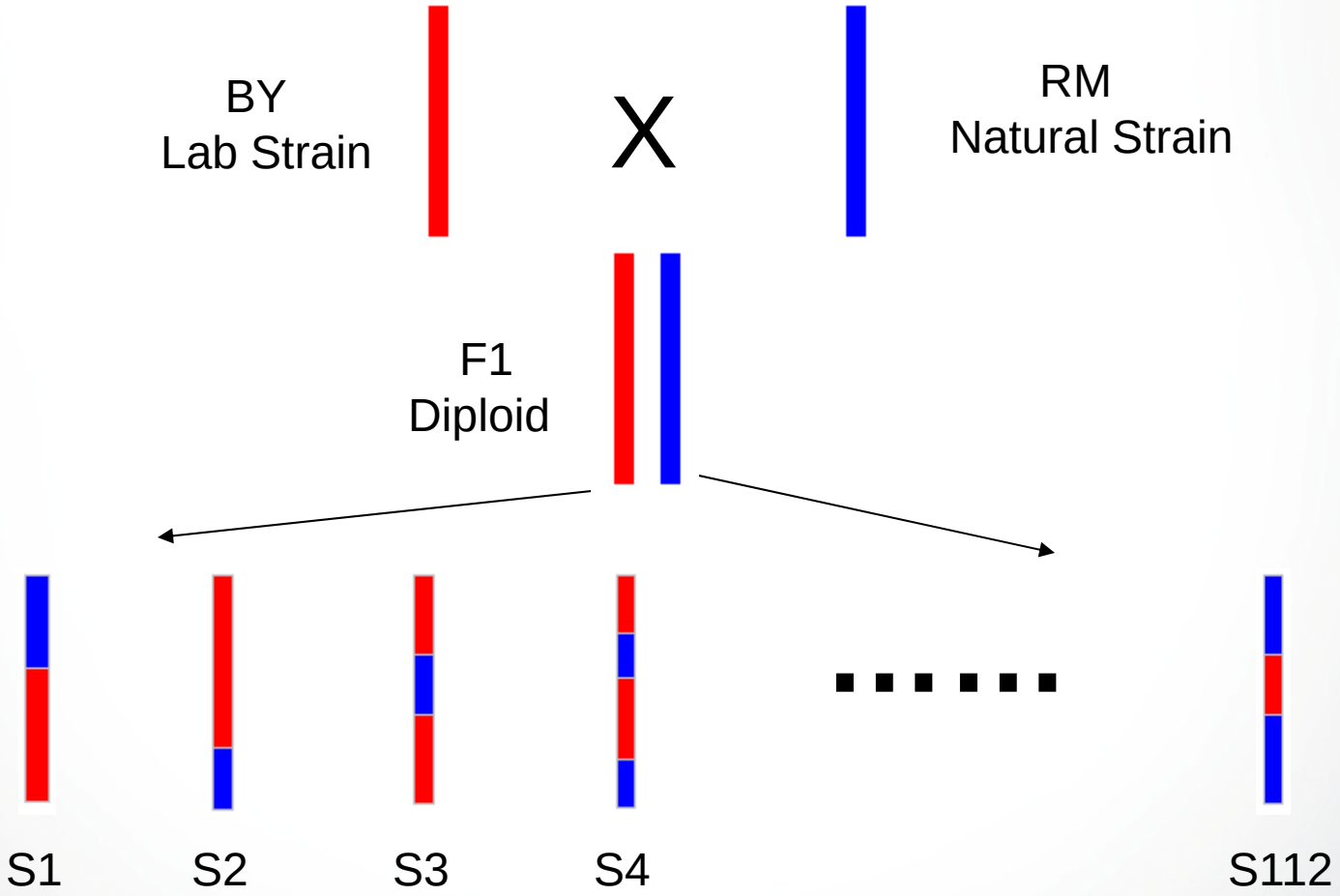
QTL Cartographer



QTL Cartographer: Basten, Zeng and Weir (1995-2005); Windows QTL Cartographer: Wang, Basten and Zeng (1999-2011)

Genetic Effect Network





An eQTL study on a yeast hybrid segregant population

Brem and Kruglyak (2005) PNAS 102:1572-1577

Experimental Design and Data

- **Sample:** BY (lab strain), RM (natural strain) and **112** F1 segregants.
- **Markers: 3312** using yeast oligoarrays
- **Gene expression:** Samples were labeled and hybridized to cDNA microarrays, containing **6215** open reading frames (ORF).
- **Reference design:** Each two-color experiment involved one sample and one reference, with the same BY RNA reference being used for all experiments.
- **Dye swap:** Two hybridizations were carried out for each sample, one with the sample labeled with Cy3 and the reference with Cy5, and one with the fluors reversed; for each gene, the two log ratios were averaged.

Multiple interval mapping for eQTL analysis

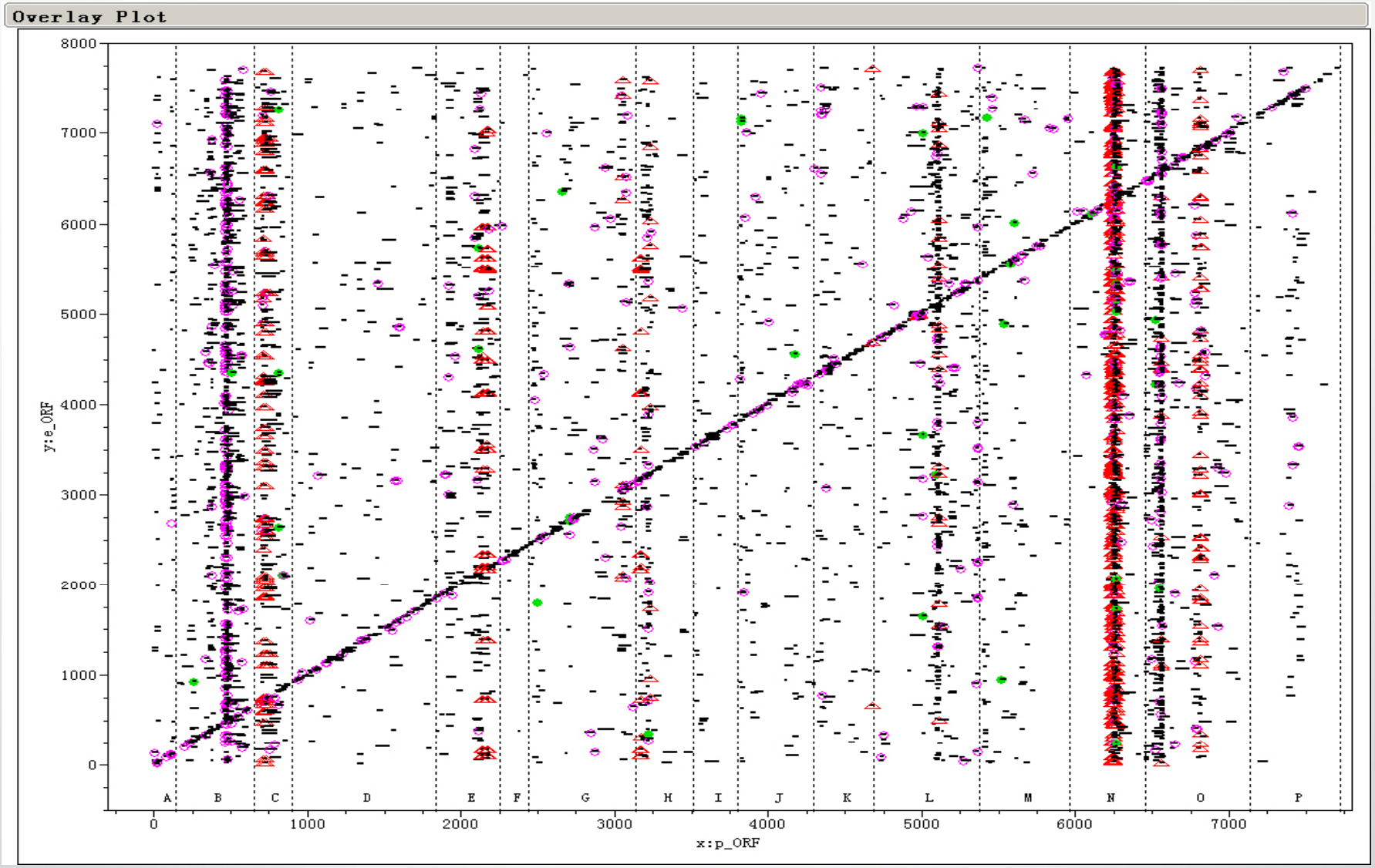
- Model:

$$y_{ik} = \alpha + \sum_{t} \beta_t x_{il_t} + \sum_{s<t} \gamma_{st} x_{il_s} x_{il_t} + e_i$$

- Sequential search for each eQTL conditional on the significance in the previous cycle for each eTrait.
- For each eTrait:
  - In cycle 1, if the max test statistic > threshold, the first eQTL is identified and continue the next step; otherwise stop the search.
  - In cycle t+1, if the conditional max test statistic > threshold, one more eQTL is added and continue the search; otherwise stop.
  - After the search for the main effects, epistatic effects of eQTL are tested based on the threshold and then added to the model.
  - Obtain 1.5-LOD support interval for each identified eQTL

MIM for eQTL: Zou and Zeng (2009)

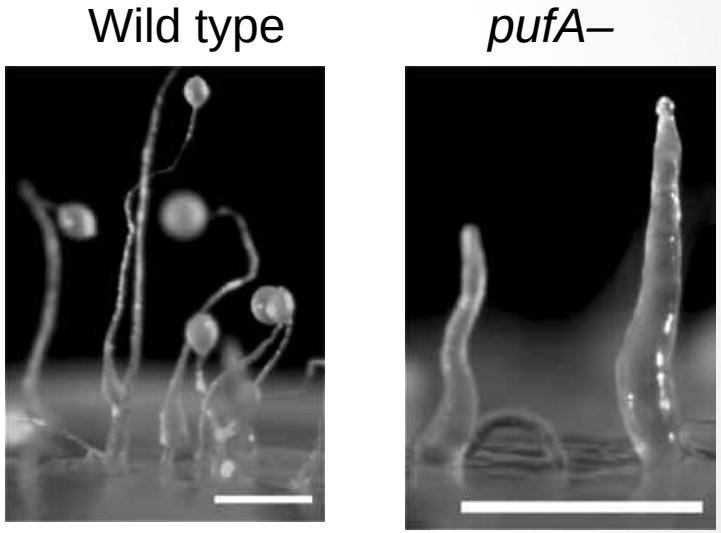
Re-analysis results (Zou and Zeng, 2009)

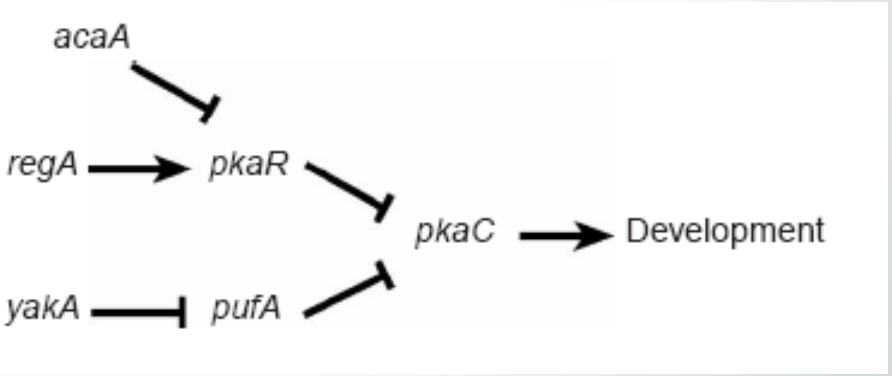


Genes, pathways and genetic model

A motivation example (*Dictyostelium*)

- Upon removal of nutrients, *D. discoideum* executes a developmental program in which single cells aggregate and form multicellular organisms.
- PKA pathway is known and important for the process. The pathway gene single and double knockout strains were created.
- Whole genome gene expression profiles were assayed and used to infer the pathway.





The protein kinase A (PKA) pathway

Van Driessche et al. (2005)

A model for inferring gene pathway from gene knockout experiment (Aylor and Zeng, 2008)

Classical genetic model and interpretation:

$$A^+B^+ : y = \mu + \varepsilon$$

$$A^-B^+ : y = \mu + \beta_A + \varepsilon$$

$$A^+B^- : y = \mu + \beta_B + \varepsilon$$

$$A^-B^- : y = \begin{cases} \mu + \beta_A + \varepsilon & \text{if } A \text{ is epistatic to } B \\ \mu + \beta_B + \varepsilon & \text{if } B \text{ is epistatic to } A \end{cases}$$

Mostly applied for lethal phenotype or viability

A quantitative genetic model

$$y = \mu + \beta_A x_A + \beta_B x_B + \beta_I x_A x_B + \varepsilon$$

$$x_A = \begin{cases} 0 \text{ for } A^+ \\ 1 \text{ for } A^- \end{cases} \qquad x_B = \begin{cases} 0 \text{ for } B^+ \\ 1 \text{ for } B^- \end{cases}$$

$\downarrow$

A^+B^+: $y = \mu + \varepsilon$

A^+B^-: $y = \mu + \beta_A + \varepsilon$

A^-B^+: $y = \mu + \beta_B + \varepsilon$

A^-B^-: $y = \mu + \beta_A + \beta_B + \beta_I + \varepsilon$

$\longrightarrow$

Model 1 : $y = \mu + \beta_A + \varepsilon$

Model 2 : $y = \mu + \beta_B + \varepsilon$

Model 3 : $y = \mu + \beta_I + \varepsilon$

Model 4 : $y = \mu + \beta_A + \beta_B + \varepsilon$

Model 5 : $y = \mu + \beta_A + \beta_I + \varepsilon$

Model 6 : $y = \mu + \beta_B + \beta_I + \varepsilon$

Model 7 : $y = \mu + \beta_A + \beta_B + \beta_I + \varepsilon$

Model 8 : $y = \mu + \varepsilon$

What is its relationship to gene pathway and interpretation?

Gene pathway interpretation

- We considered all combinations of gene order and action within simple ON/OFF models and then predicted the hypothetical effect of deleting genes on each of them.
- There are four points of variation to model for each gene pair relationship.
  - The first is the identity of the upstream gene, i.e. the gene order.
  - Secondly, the upstream gene will turn the downstream gene either on (enhance) or off (repress).
  - Thirdly, the downstream gene can enhance or repress the expression of a target gene for which expression is observed.
  - Lastly, we consider that the upstream gene itself will be enhanced or repressed by some initiating factor such as a developmental cue or environmental perturbation.

Linking quantitative model to pathway interpretation: An example

Genotype	Upstream Gene	Gene Action	Target Gene Expression	Regression Model
A^+B^+	ON	ON A ⊣ B → X OFF	μ	$\mu + \beta_A + \beta_I$
A^-B^+		~~A~~ ⊣ ON B → ON X	$\mu + \beta_A$	
A^+B^-		ON A ⊣ ~~B~~ → X OFF	μ	
A^-B^-		~~A~~ ⊣ ~~B~~ → X OFF	μ	
A^+B^+	OFF	A OFF ⊣ ON B → ON X	μ	$\mu + \beta_B$
A^-B^+		~~A~~ ⊣ ON B → ON X	μ	
A^+B^-		A OFF ⊣ ~~B~~ → X OFF	$\mu + \beta_B$	
A^-B^-		~~A~~ ⊣ ~~B~~ → X OFF	$\mu + \beta_B$	

Another example

Genotype	Upstream Gene	Gene Action	Target Gene Expression	Regression Model
A^+B^+	ON	ON A → ON B → ON X	μ	$\mu + \beta_A + \beta_B + \beta_I$
A^-B^+		~~A~~ → B OFF → X OFF	$\mu + \beta_A$	
A^+B^-		ON A → ~~B~~ → X OFF	$\mu + \beta_B$	
A^-B^-		~~A~~ → ~~B~~ → X OFF	$\mu + \beta_B$	
A^+B^+	OFF	A OFF → B OFF → X OFF	μ	μ
A^-B^+		~~A~~ → B OFF → X OFF	μ	
A^+B^-		A OFF → ~~B~~ → X OFF	μ	
A^-B^-		~~A~~ → ~~B~~ → X OFF	μ	

Match quantitative models with pathway interpretations

a. Hierarchical Relationships

Upstream Gene	A upstream of B		B upstream of A	
	ON	OFF	ON	OFF
Repressor	$\mu+\beta_A+\beta_I$ [5]	$\mu+\beta_B$ [2]	$\mu+\beta_B+\beta_I$ [6]	$\mu+\beta_A$ [1]
Enhancer	$\mu+\beta_A+\beta_B+\beta_I$ [7]	μ [8]	$\mu+\beta_A+\beta_B+\beta_I$ [7]	μ [8]

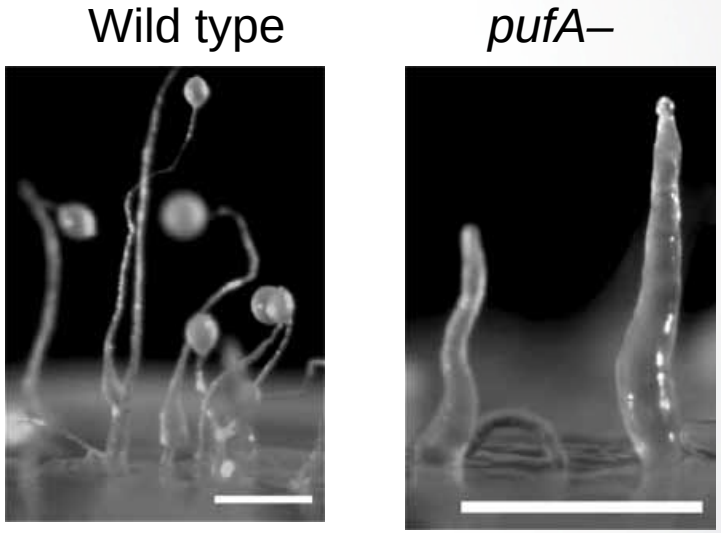
b. Non-hierarchical Relationships

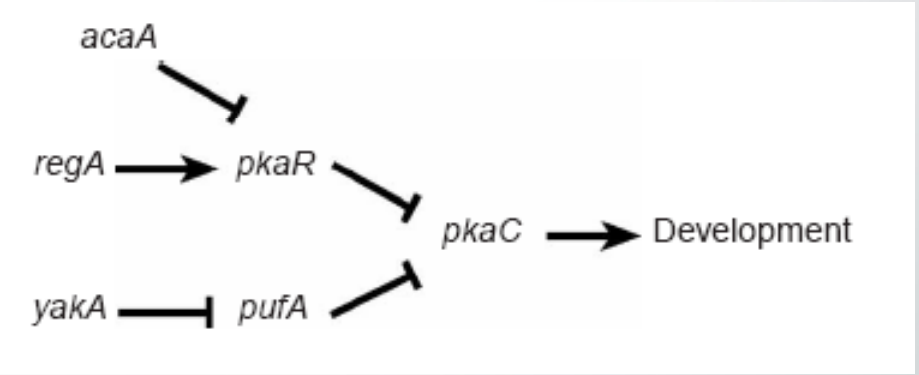
State of A/B	ON/ON	ON/OFF	OFF/ON	OFF/OFF
Enhancer/Enhancer	$\mu+\beta_I$ [3]	$\mu+\beta_A$ [1]	$\mu+\beta_B$ [2]	μ [8]
Enhancer/Repressor Or Repressor/Enhancer	$\mu+\beta_A+\beta_B$ [4]			
Repressor/Repressor	$\mu+\beta_I$ [3]			

Aylor and Zeng (2008)

Back to the *Dictyostelium* experiment (Van Driessche et al. 2005)

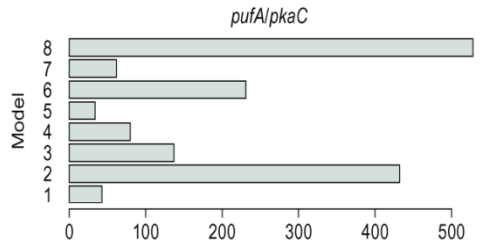
- Upon removal of nutrients, *D. discoideum* executes a developmental program in which single cells aggregate and form multicellular organisms.
- PKA pathway is known and important for the process. The pathway gene single and double knockout strains were created.
- Whole genome gene expression profiles were assayed and used to infer the pathway.

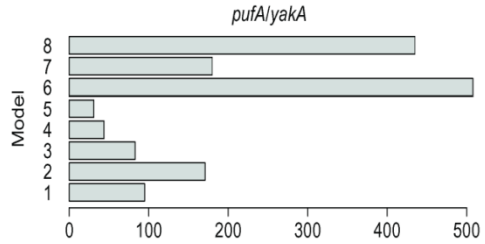


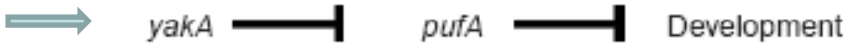


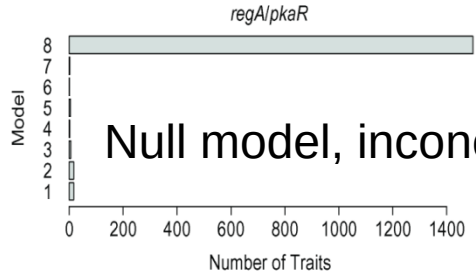
The protein kinase A (PKA) pathway

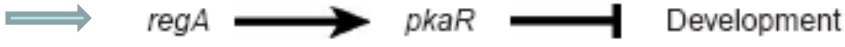
Analysis Results







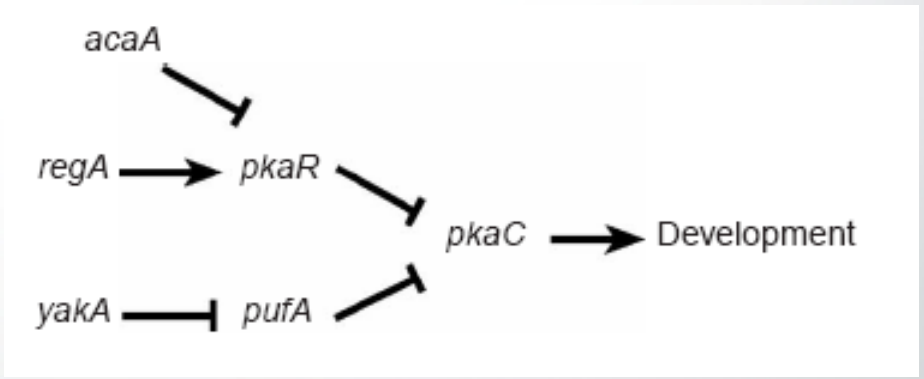




Or others

Null model, inconclusive

Recall: The protein kinase A (PKA) pathway



Aylor and Zeng (2008)

Acknowledgement

- Wei Zou
- David Aylor
- Christine Duarte
- Shengchu Wang
- Luciano Da Costa Silva
- Cecelia Laurie (U. Alabama-Tuscaloosa)

A remark

- Many biologists, particularly molecular biologists, avoid complex systems and statistics.
- But biological systems are complex and sadly  mathematics and statistics are unavoidable.
- Biologically relevant mathematical and statistical models can help us to disentangle complex information and infer complex phenomena and structures, but could also carry the risk of misleading us if not handled properly.
- Also mathematical and statistical models have to be guided by our understanding of biological process in order to be biologically relevant.