
Microsoft® Research
Faculty Summit
10
YEAR ANNIVERSARY

Computational Challenges in Analyzing Complex Traits

Jun Zhu

Institute of Bioinformatics

Zhejiang University

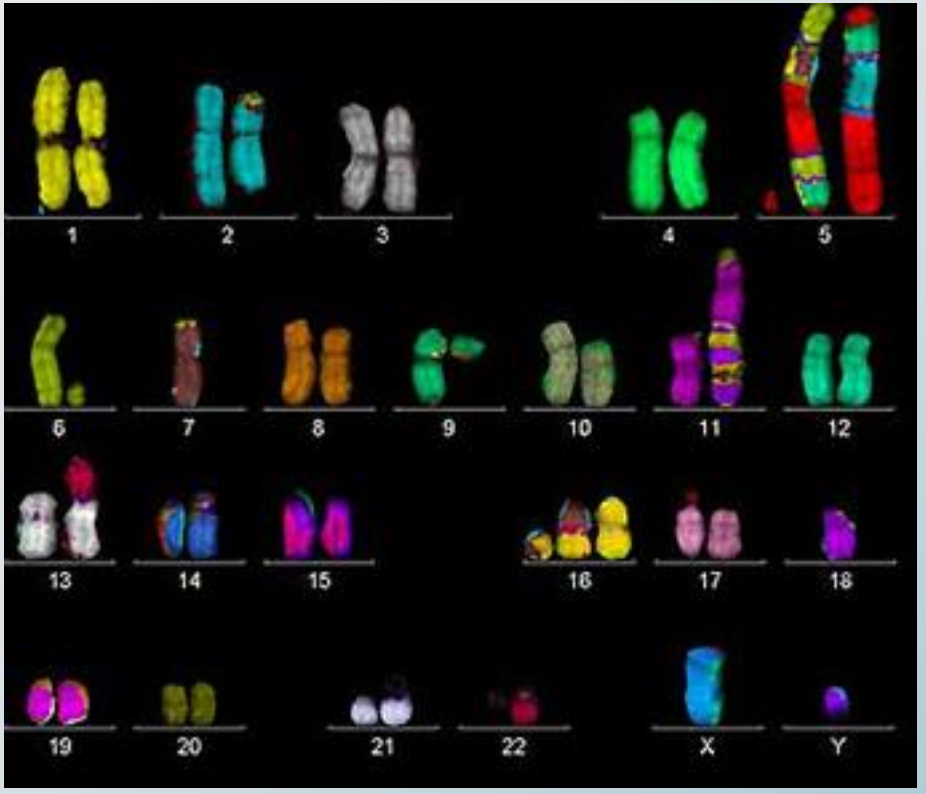
Mendelian Traits

- Phenotypes controlled by singular genes
- No epistasis
- No GE interaction

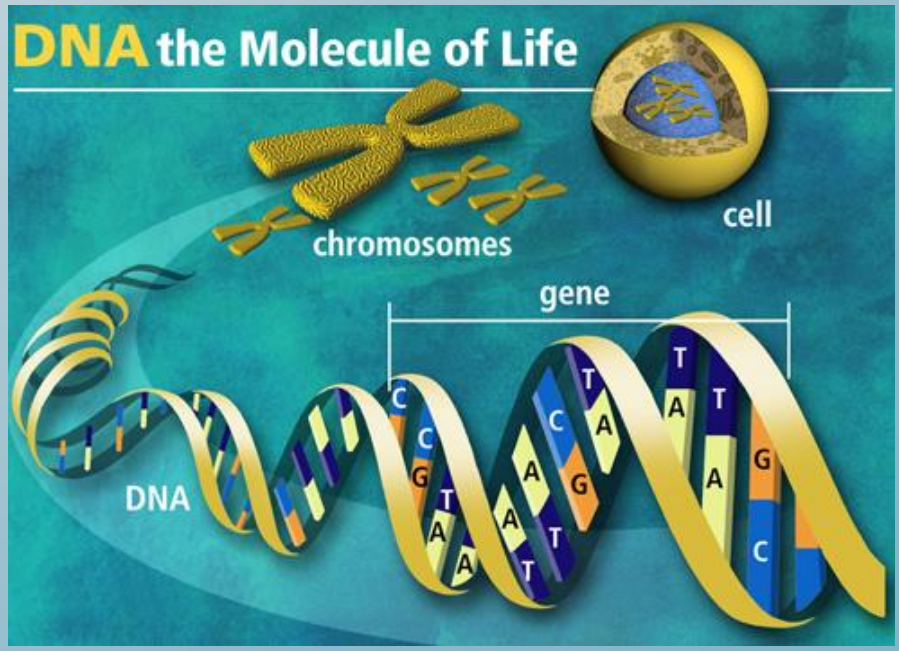
Complex Traits

- Phenotypes controlled by multiple genes
- Epistasis (gene-gene interaction)
- Gene-environment interaction
- Genetic pleiotropy and heterogeneity
- Low heritability
- Limited statistical power

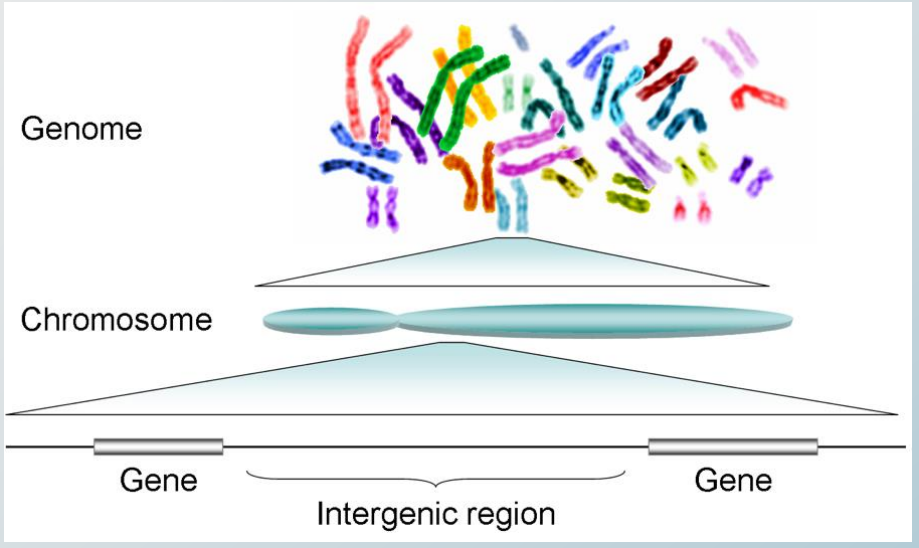
- **Human Chromosomes**



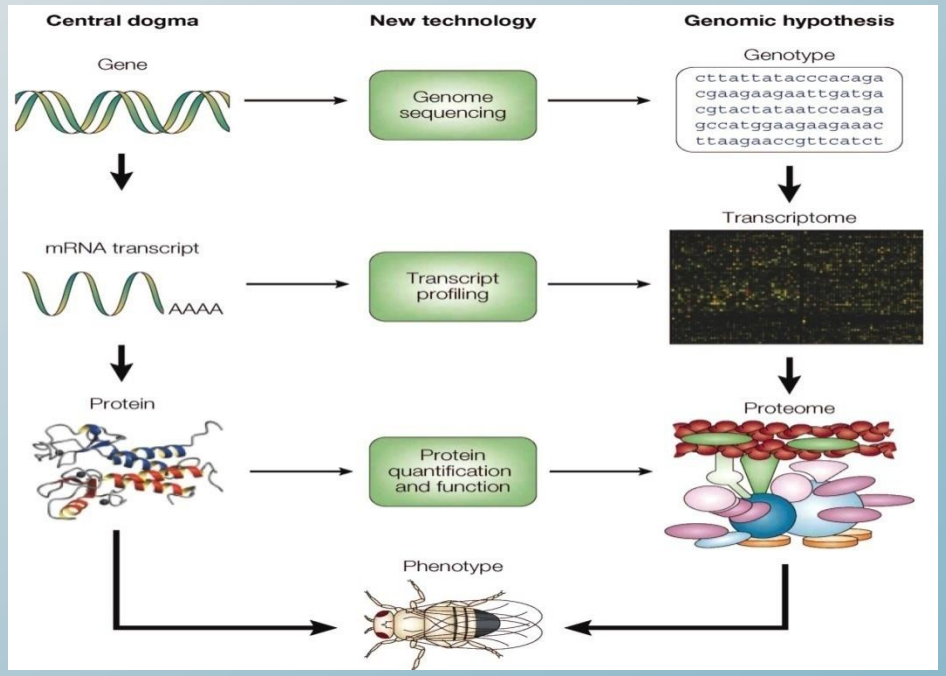
- **Chromosome to DNS**



- **From Genome to Genes**

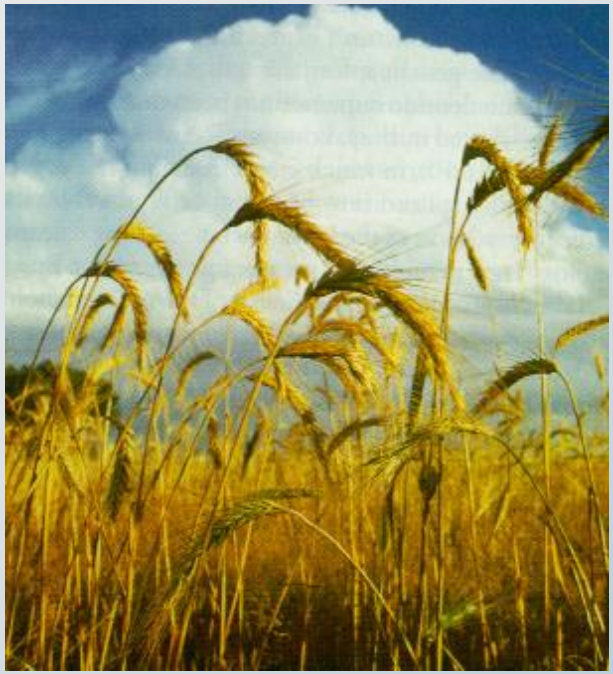


- **From Gene to Phenotype**



Partition of Phenotypic Variation

Phenotype $y = \mu + E + G + GE + e$



G

GE

E

Genetic Effect: A, D, I

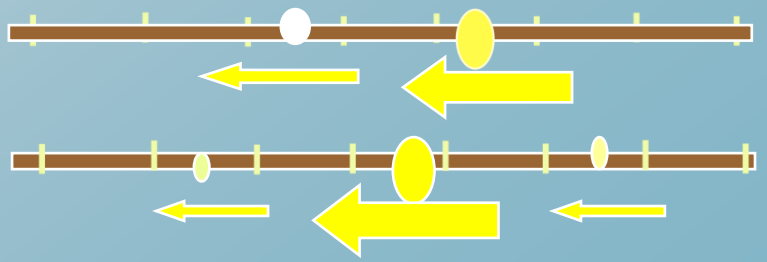
GE: AE, DE, IE

Macro Env, Micro Env

Genetic Effects

QTL Position & Effects

Genome



Partitioning of Phenotypic Variation

$$y = \mu + E + G + GE + e$$

- **Gene Effects (G):**
  Addtive (A), Dominance (D)
  Gene-Gene Interaction (AA, AD, DD)
- **Environment Effects (E):**
  Location, Weather, Treatment
- **Gene-Environment Interaction (GE)**
  AE, DE, AAE, ADE, DDE
- **Random Error (*e*)**

Partitioning of Phenotypic Variation

$$y = \mu + E + G + GE + e$$

- **Classical quantitative genetics**

$$G = A + D + (AA + AD + DD)$$

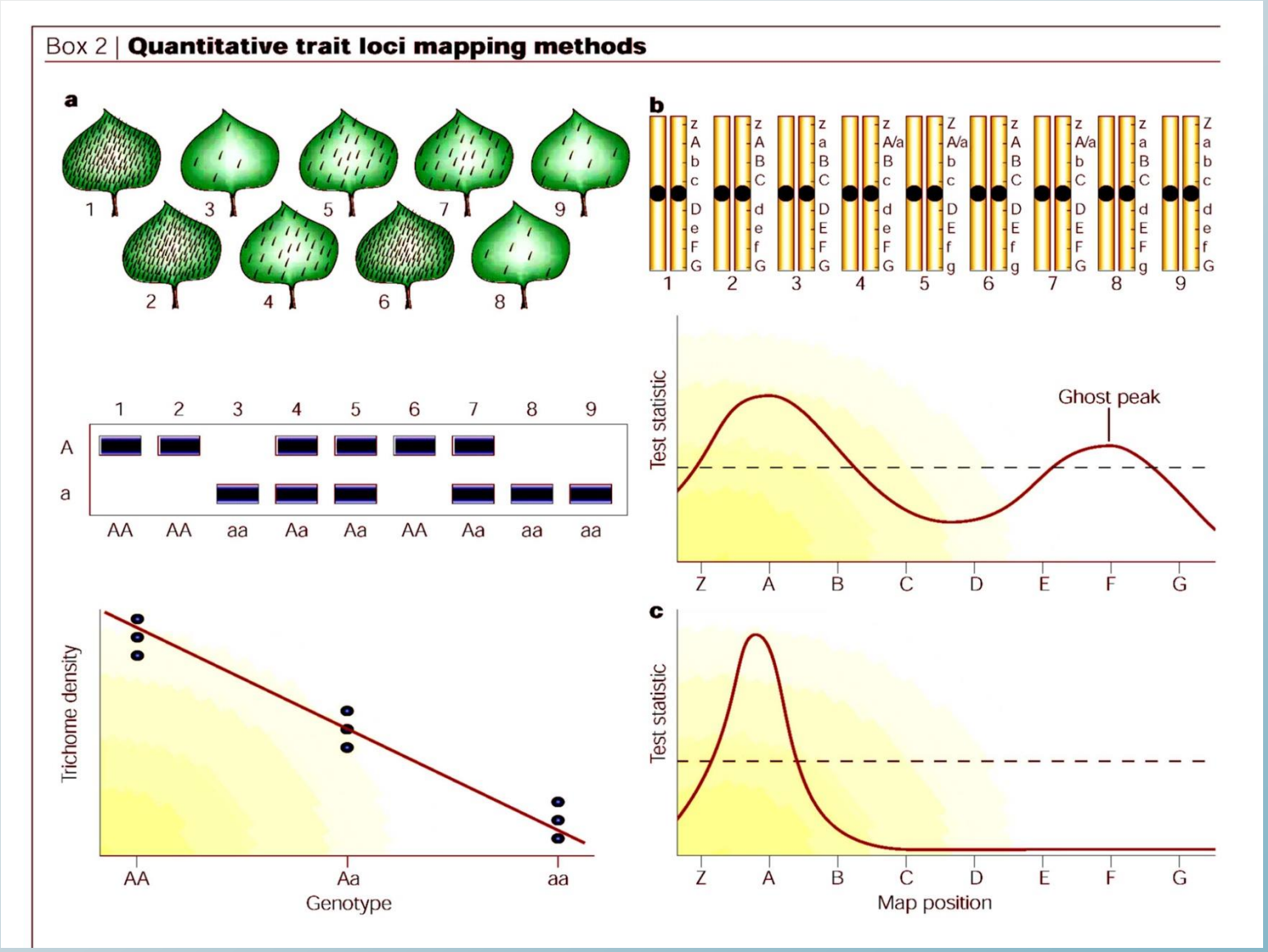
$$GE = AE + DE + (AAE + ADE + DDE)$$

- **Molecule quantitative genetics**

$$G = \sum a + \sum d + (\sum\sum aa + \sum\sum ad + + \sum\sum dd)$$

$$GE = \sum ae + \sum de + (\sum\sum aae + \sum\sum ade + + \sum\sum dde)$$

Method of Mapping QTL



NATURE REVIEWS 2001

Presentation of QTL

Multiple QTL

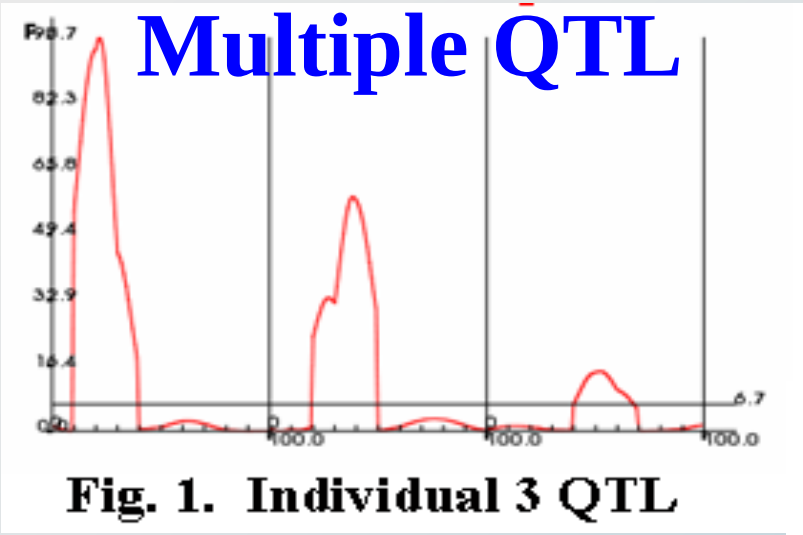


Fig. 1. Individual 3 QTL

QTL × Env

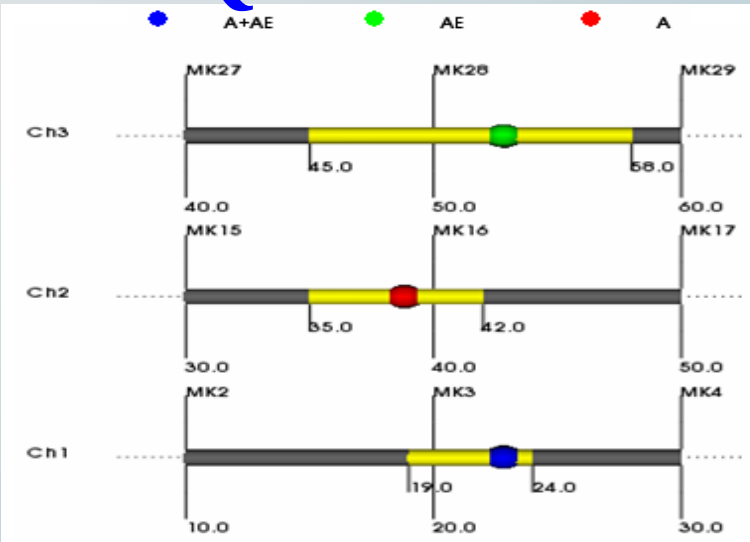


Fig. 2. Differential expression of 3 QTL

QTL Network

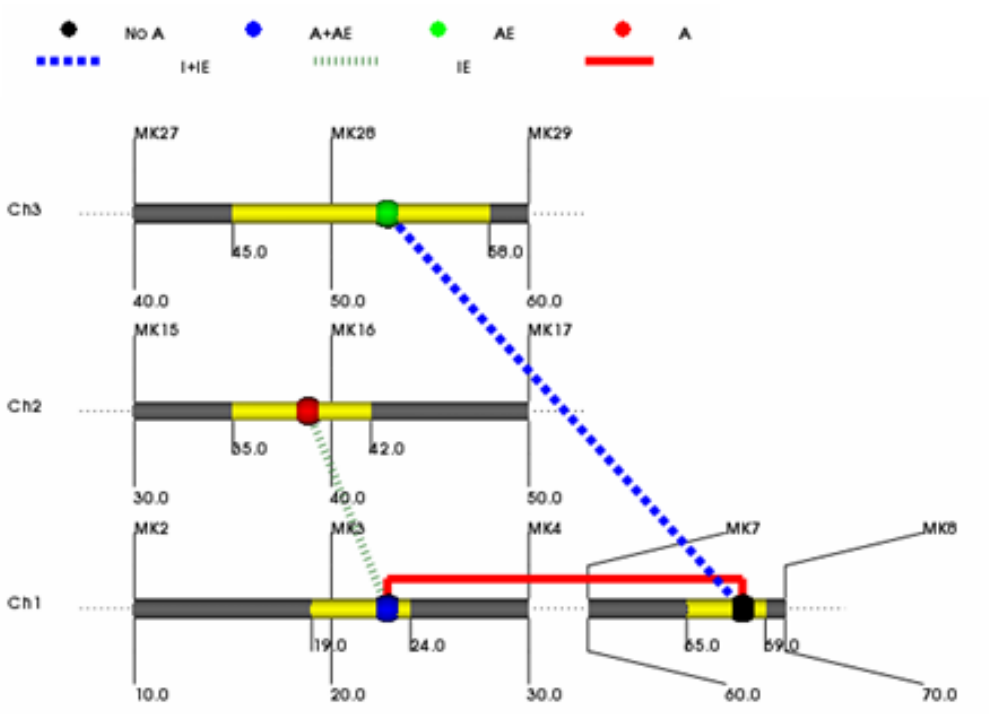
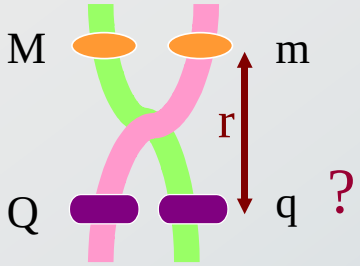


Fig. 3. Differential expression of QTL network

Populations for Mapping QTL



$(\frac{MQ}{MQ})$ $P_1 \times P_2$ $(\frac{mq}{mq})$

$\left(\begin{array}{cccc} \frac{MQ}{MQ} & \frac{Mq}{Mq} & \frac{mQ}{mQ} & \frac{mq}{mq} \\ \frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2} \end{array}\right)$

$(\frac{MQ}{mq})$ F_1 → Haploid → DH

$P_1 \times F_1$ $P_2 \times F_1$

$\left(\begin{array}{cccc} MQ & Mq & mQ & mq \\ \frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2} \end{array}\right)$

BC_1

$\left(\begin{array}{cccc} \frac{MQ}{MQ} & \frac{MQ}{Mq} & \frac{MQ}{mQ} & \frac{MQ}{mq} \\ \frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2} \end{array}\right)$ ⊗

F_2

BC_2

$\left(\begin{array}{cccc} \frac{mq}{MQ} & \frac{mq}{Mq} & \frac{mq}{mQ} & \frac{mq}{mq} \\ \frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2} \end{array}\right)$ ⊗

$\left(\begin{array}{ccccccccc} \frac{MQ}{MQ} & \frac{MQ}{Mq} & \frac{Mq}{Mq} & \frac{MQ}{mQ} & \frac{MQ}{mq}/\frac{Mq}{mQ} & \frac{Mq}{mq} & \frac{mQ}{mQ} & \frac{mQ}{mq} & \frac{mq}{mq} \\ \frac{(1-r)^2}{4} & \frac{r(1-r)}{2} & \frac{r^2}{2} & \frac{r(1-r)}{2} & \frac{r^2+(1-r)^2}{2} & \frac{r(1-r)}{2} & \frac{r^2}{2} & \frac{r(1-r)}{2} & \frac{(1-r)^2}{4} \end{array}\right)$

Experimental Design
(Observations n = 250×12×3=9,000)

- **Mapping Population**

  250 DH lines
- **Environments**

  4 Location, 3 Years
- **Replications**

  3 Blocks

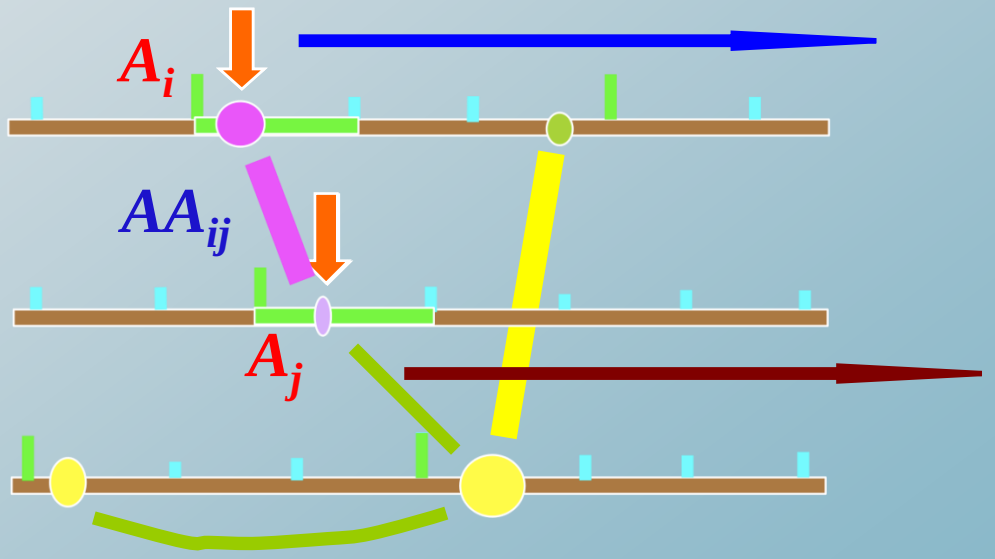
12

Mixed-model-based Composite Interval Mapping

(Wang & Zhu et al. 1999, TAG, V99)

Mapping QTL with A+AA and QE Interaction (DH, RIL)

$$y = \mu + A_i + A_j + AA_{ij} + E + A_iE + A_jE + AA_{ij}E + G_M + G_{MM} + G_ME + G_{MM}E + \varepsilon$$



Mixed-Linear Model Approaches

$$\mathbf{y} = \mathbf{Xb} + \mathbf{U}_E\mathbf{e}_E + \mathbf{U}_{QE}\mathbf{e}_{QE}$$
$$+\mathbf{U}_M\mathbf{e}_M + \mathbf{U}_{MM}\mathbf{e}_{MM} + \mathbf{U}_{ME}\mathbf{e}_{ME} + \mathbf{U}_{MME}\mathbf{e}_{MME} + \mathbf{e}_\varepsilon$$
$$= \mathbf{Xb} + \sum \mathbf{U}_u\mathbf{e}_u$$
$$\sim N(\mathbf{Xb}, \underset{(n\times n)}{\mathbf{V}} = \sum_{u=1}^{m}\sigma_u^2\mathbf{U}_u\mathbf{R}_u\mathbf{U}_u^T)$$

The Likelihood Function for QTL Mapping Model

$$L(\mathbf{b},\mathbf{V}) = (2\pi)^{-\frac{n}{2}}\left|\mathbf{V}\right|^{-\frac{1}{2}}\exp[-\tfrac{1}{2}(\mathbf{y}-\mathbf{Xb})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})]$$

Estimation of QTL Main Effects

$$\ddot{\mathbf{b}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

Prediction of QTL-Environment Interaction Effects

$$\ddot{\mathbf{e}}_u = \sigma_u^2\mathbf{U}_u^T[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^+\mathbf{X}^T\mathbf{V}^{-1}]\mathbf{y}$$

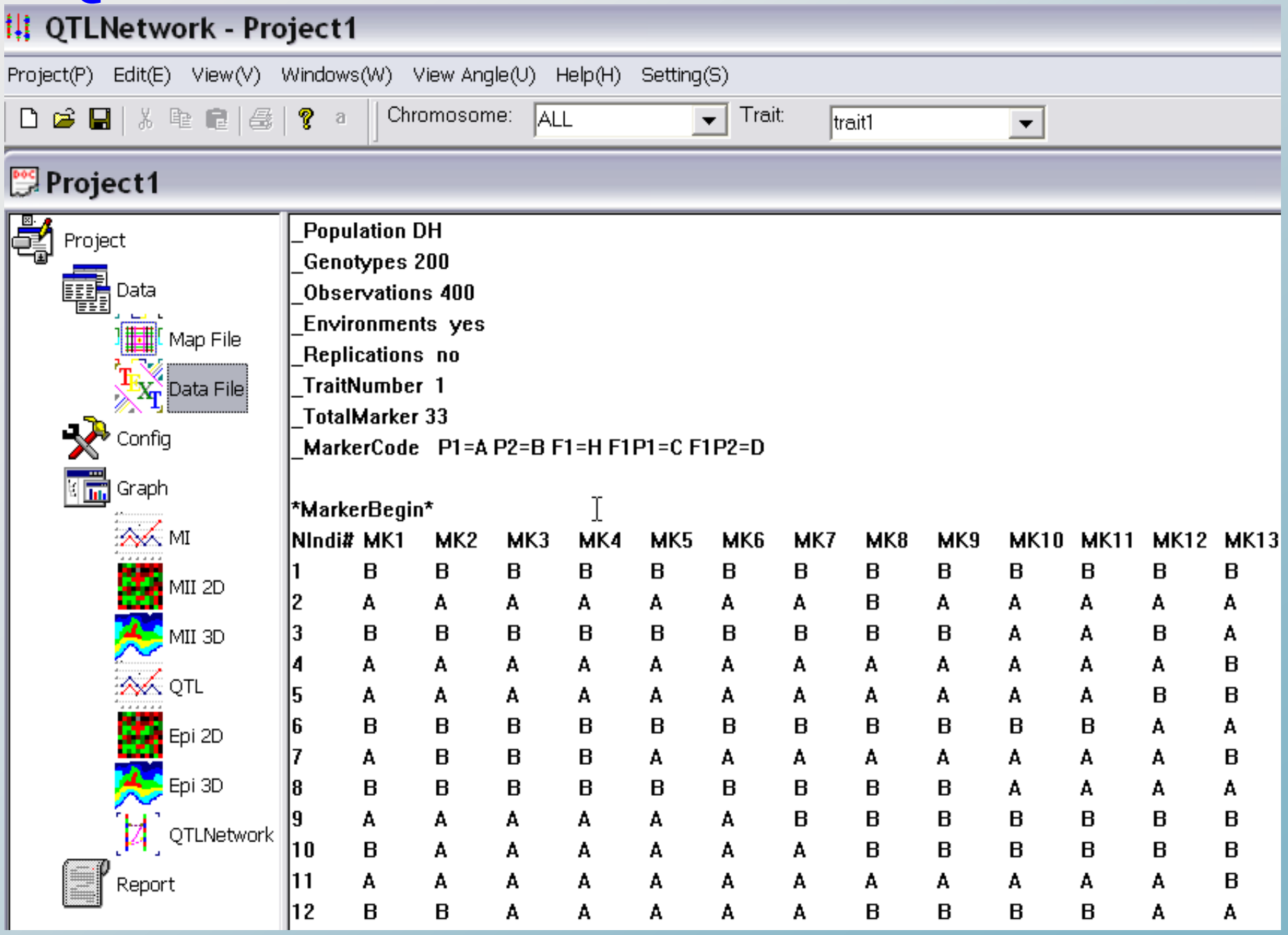
Challenges in Computation for Inverses of Big Matrix $\mathbf{V}_{(9000\times 9000)}$

- Rice Map Distance = 2031 cM
- Step of QTL Searching= 1 cM
- Steps of Two-dimession Search

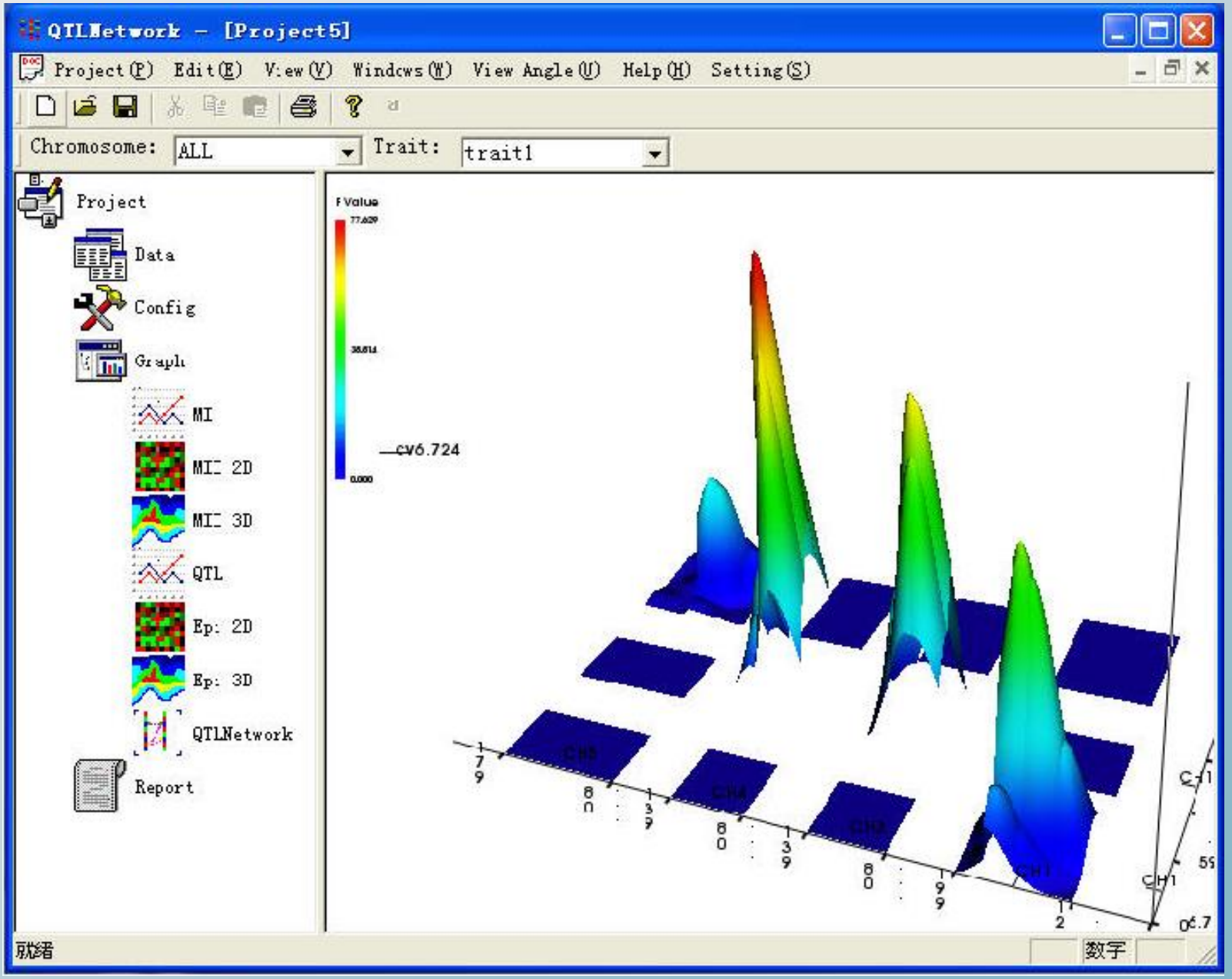
  = 2031× 2030/2 = 2.06 Millions

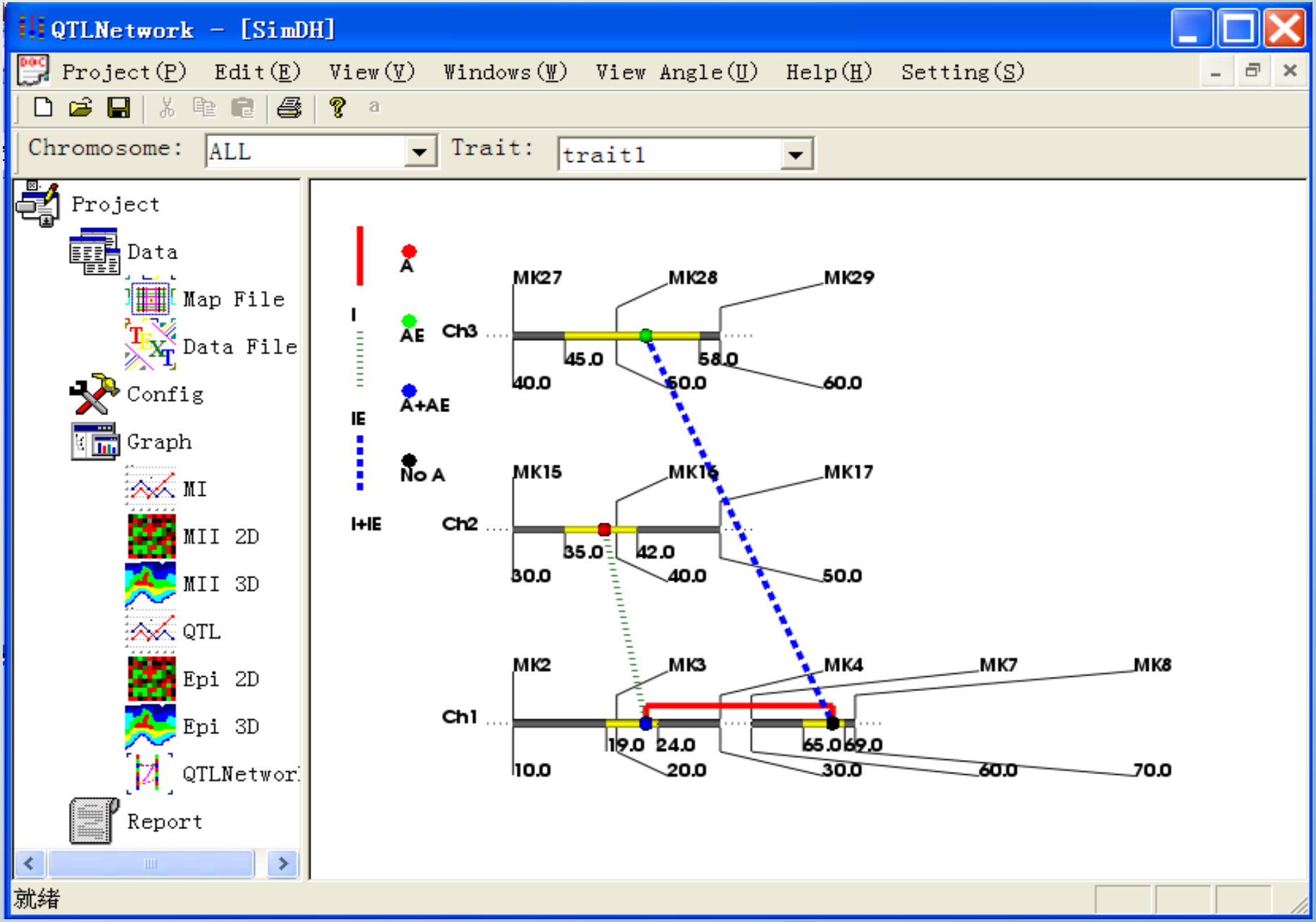
- Detecting QTLs for 10 Traits Needs to Calculate 20.6 Millions Inverses of V

QTLNetwork version 2.0

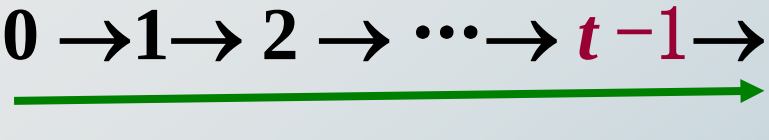


QTLNetwork 2.0




QTLNetwork - [SimDH]
Project(P) Edit(E) View(V) Windows(W) View Angle(U) Help(H) Setting(S)
Chromosome: ALL
Trait: trait1
Project
Data
Map File
Data File
Config
Graph
MI
MII 2D
MII 3D
QTL
Epi 2D
Epi 3D
QTLNetwor
Report
A
AE
A+AE
No A
I
IE
I+IE
Ch3
MK27
MK28
MK29
40.0
45.0
50.0
58.0
60.0
Ch2
MK15
MK16
MK17
30.0
35.0
40.0
42.0
50.0
Ch1
MK2
MK3
MK4
MK7
MK8
10.0
19.0
20.0
24.0
30.0
65.0
69.0
60.0
70.0
就绪

Mapping Developmental QTL for Complex Traits



Unconditional Model for Phenotypic Value at Time t

$$y(t) = \mu(t) + G_Q(t) + E(t) + G_QE(t)$$

$$+ G_M(t) + G_ME(t) + \varepsilon(t)$$

Analyzing Q & QE Effects from Time 0→t

Conditional Model for Phenotypic Value at Time t

$$y(t|t\text{-}1) = \mu(t|t\text{-}1) + G_Q(t|t\text{-}1) + E(t|t\text{-}1)$$

$$+ G_QE(t|t\text{-}1) + G_M(t|t\text{-}1) + G_ME(t|t\text{-}1) + \varepsilon(t|t\text{-}1)$$

Analyzing Net Q & QE Effects From Time t -1 → t

Table 2. Chromosomal regions and estimated genetic effects of QTLs for plant height (cm) at different stages in two environments. (Yan, *et al.* 1998, Genetics, V150)

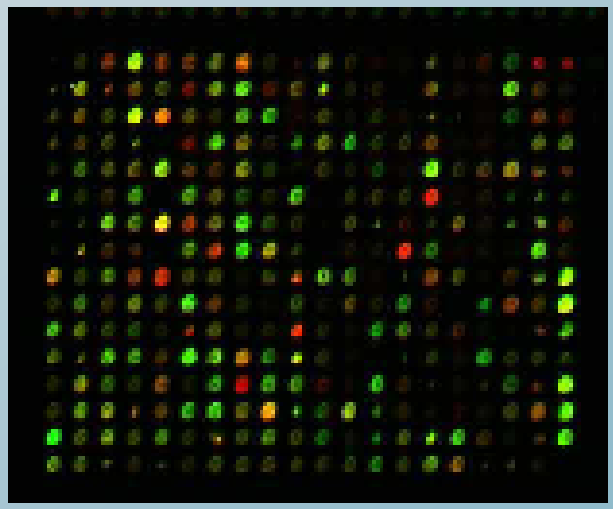
QTL	Maker Interval	Dist-ance (cM)	Days	Main effect		GE in Hangzhou		GE in Hainan	
				t	$t\|t-1$	t	$t\|t-1$	t	$t\|t-1$
Ph1	RZ730-RZ801	33.1	10	-2.14	-2.14			-1.08	-1.08
			20	-2.47				-0.95	
			30	-3.6				-1.45	-0.52
			40	-4.6				-3	-0.52
			50	-5.06			-1.68	-3.12	1.05
			60	-9.54	-0.55	-3.09	-1.7	-2.64	1.1
			70	-12.57		-4.39		-3.51	
			80	-15.01		-4.47		-4.35	
			90	-16.98		-4.07		-4.36	
Ph2	Amy1A/C-RG95	12.8	10	1.03	1.03			0.63	0.63
			20	1.51				0.76	
			30	1.93				0.93	
			40	2.91		0.99		1.25	
			50	2.77	0.92		-0.74	1.05	-0.74
			60	4.74		2.55			
			70	6.07		2.67			
			80	7.73		2.42		1.93	
			90	7		1.62		1.69	

Challenges in Presentation of G-G and G-E During Developmental Stages

- For G-G interaction presentation, there needs two-dimension display
- When Genes express differently across times and spaces, there needs four-dimension display or three-dimension dynamic display

Experimental Design for Microarray Testing

- **Array Design：**
  - **14112 Genes (84 x 68)/Array**
- **Treatment Design（3 Way Factors）：**
  - **Factor G: 14112 Genes**
  - **Factor C：8 Cancer Cells**
  - **Factor M：7 Medicine**
- **Replicates: 3**



Combining Analysis

$$y_{ijkl} = G_l + A_i + C_j + M_k + GA_{li} + GC_{lj} + GM_{lk} + \varepsilon_{ijkl}$$

Challenges in Mixed-Model Approaches for Array Analysis

$$\mathbf{y} = \mathbf{X}\mathbf{b}_G + \mathbf{U}_A\mathbf{e}_A + \mathbf{U}_C\mathbf{e}_C + \mathbf{U}_M\mathbf{e}_M$$
$$+\mathbf{U}_{GA}\mathbf{e}_{GA} + \mathbf{U}_{GC}\mathbf{e}_{GC} + \mathbf{U}_{GM}\mathbf{e}_{GM} + \mathbf{e}_\varepsilon$$
$$= \mathbf{X}\mathbf{b}_G + \sum_{u=1}^{7}\mathbf{U}_u\mathbf{e}_u \sim N(\mathbf{X}\mathbf{b}_G, \underset{(n\times n)}{\mathbf{V}} = \sum_{u=1}^{7}\sigma_u^2\mathbf{U}_u\mathbf{U}_u^T)$$

Prediction of Random Effects

$$\hat{\mathbf{e}}_{u(1)} = \mathbf{U}_u^T\mathbf{V}_{(1)}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$
$$= \mathbf{U}_u^T[\mathbf{V}_{(1)}^{-1} - \mathbf{V}_{(1)}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}_{(1)}^{-1}\mathbf{X})^+\mathbf{X}^T\mathbf{V}_{(1)}^{-1}]\mathbf{Q}_{(1)}\mathbf{y}$$

Experiment Size

n = 14112 ×8×7×3 = 2.37 Millions

A Two-step Strategy for Detecting Differential Gene Expression of cDNA Microarray Data

(Lu, Zhu, &, Liu, 2005, Current Genetics. 47: 121–131)

Choosing a subset of potential genes with differential expression

$$y_{ijk(l)} = \mu_{(l)} + A_{i(l)} + V_{j(l)} + D_{k(l)} + \gamma_{ijk(l)}$$

Combining analysis of multiple genes

$$y_{ijkl} = G_l + A_i + C_j + M_k + GA_{li} + GC_{lj} + GM_{lk} + \varepsilon_{ijkl}$$

Experiment Size Reduced to ≈ 300

n = 300×8×7×3 ≈ 50,400

Thank You!