ClassifyGxT:
Probabilistic classification of gene-by-treatment interactions on molecular count phenotypes

Will Valdar

Dept of Genetics

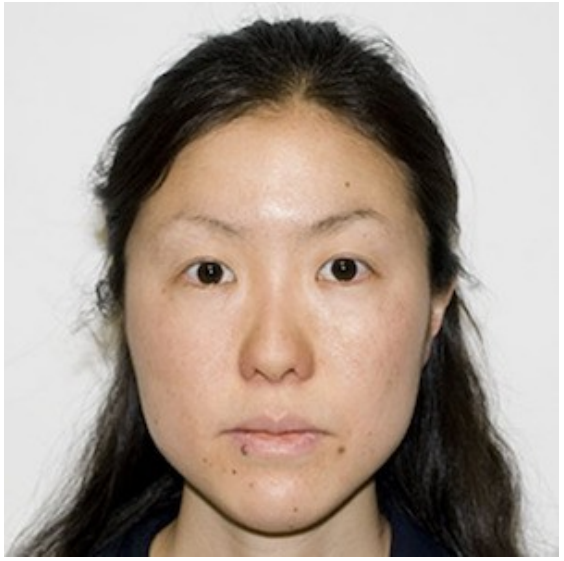
University of North Carolina at Chapel Hill

CTC-RG Oct 2, 2024

Overview

- Background, existing methods, motivation
- ClassifyGxT: classifying gene-by-treatment interactions
    - Probabilistic classification (Bayesian model selection)
    - Modeling molecular count phenotypes (Nonlinear regression)
- Applications to experimental data
- Summary

Acknowledgments



Yuriko Harigaya

Mike Love

in collaboration with

Jason Stein (UNC)

Brandon Le

Nana Matoba

Jordan Valone

<u>Funding</u>

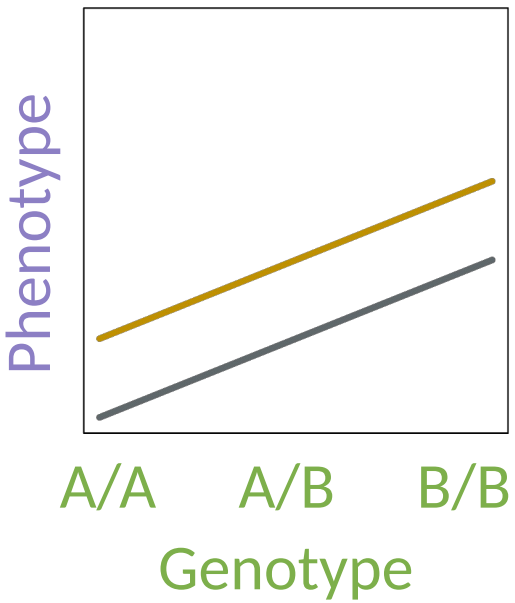
NIH R35-GM127000 (Valdar)

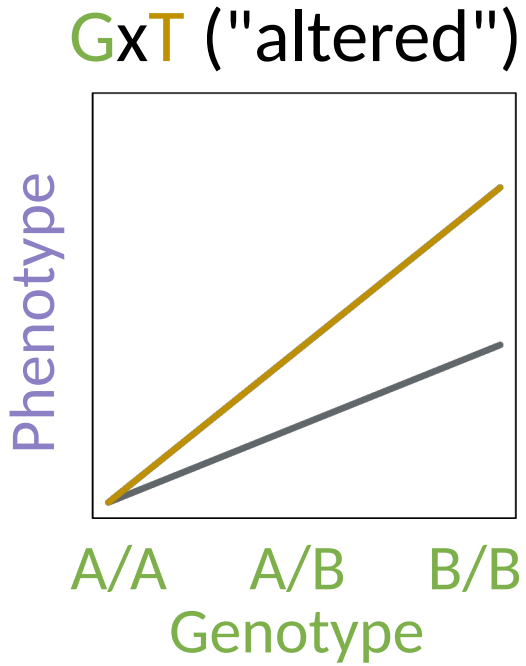
NIH R01-MH118349 (Stein, Love)

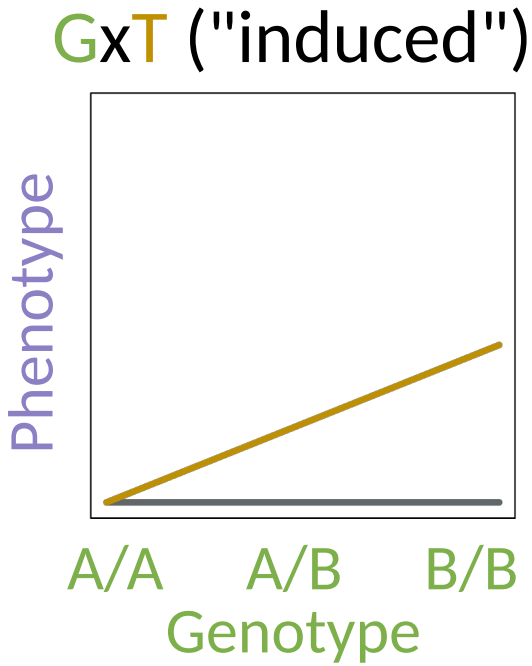
Background: gene-by-treatment (GxT) interactions

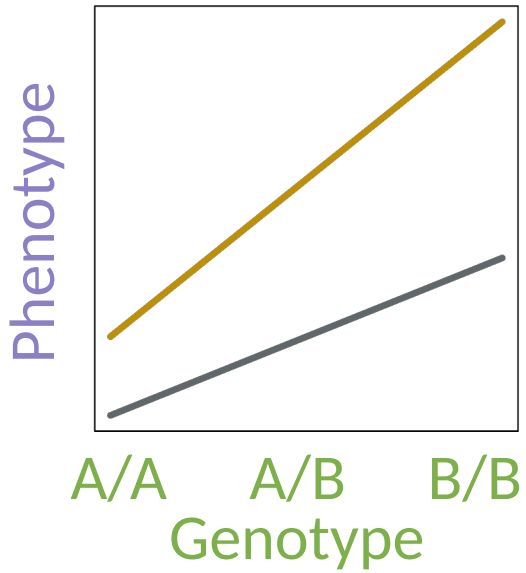
- Inter-individual variability in response to treatment
- Genotype, GxT interactions
- Informative for clinical decision making
- Current practice: presence/absence of GxT interaction
- We want to distinguish different types of GxT interaction

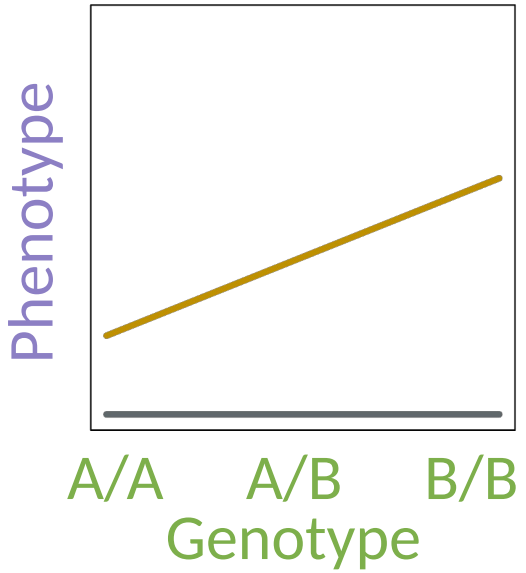
Background: types of GxT interactions







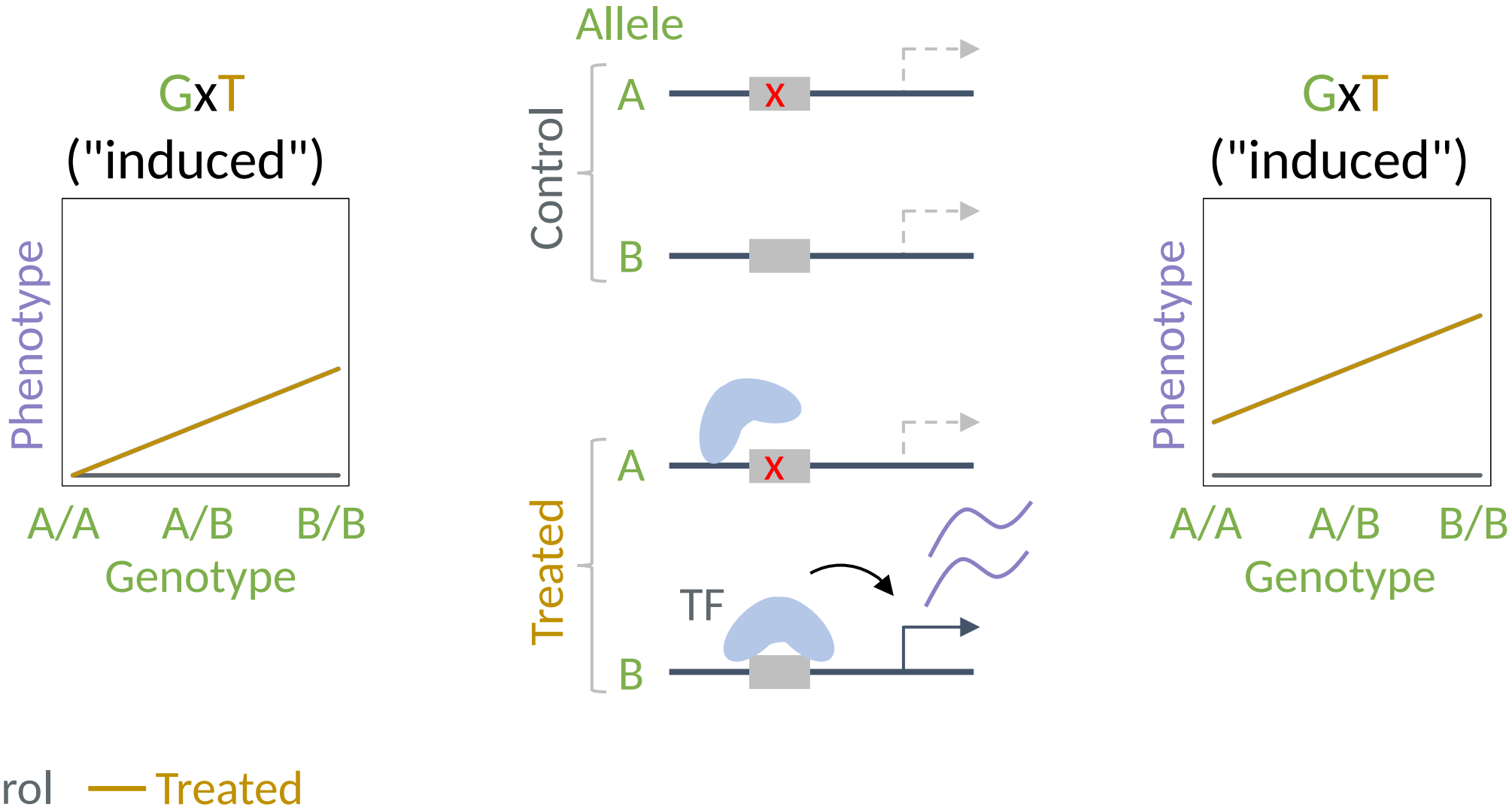




Control Treated

GxT classification can inform mechanisms

- Phenotype: molecular count data (e.g., gene expression)



GWAS/QTL mapping approaches to detect GxT

- stratified analysis

  control: phenotype = genotype + noise

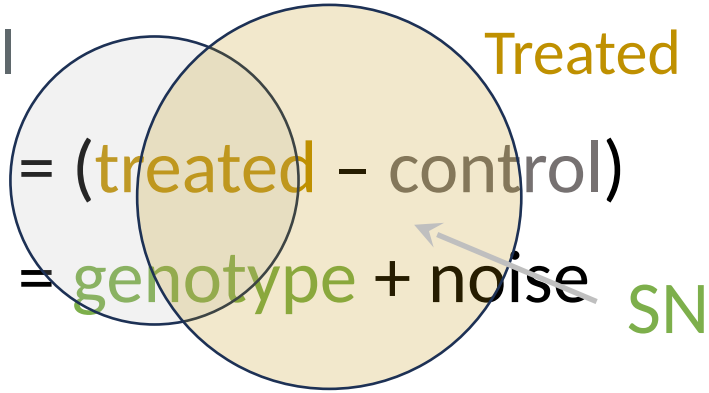
  treated: phenotype = genotype + noise

  - identifies induced loci
  - hard to rank importance

- "delta" approach

  Define: delta = (treated – control)

  Map using: delta = genotype + noise



SNPs w/ induced effect

  - simple and powerful
  - needs paired data
  - no GxT classification

- interaction model approach

  phenotype = genotype + treatment + genotype x treatment + noise

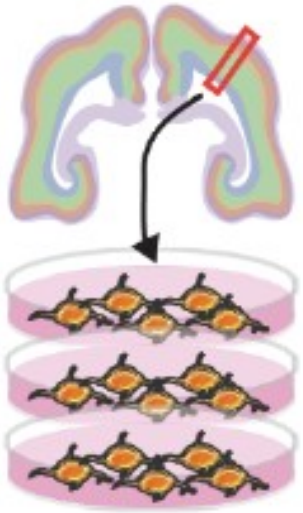
  interaction

  - flexible, no need for paired data
  - no GxT classification

A motivating dataset from Jason Stein lab

Jason Stein
(UNC)

Genotyped primary human
neural progenitor cells (hNPCs)
derived from fetal donors



+ Vehicle
(control)

+ Growth
stimulation
(treatment)

RNA-seq → "Interaction" approach

~100 gene-SNP pairs w/
significant GxT interactions

→ No principled way to prioritize SNPs with a specific type of GxT interaction

(Matoba, Le, Valone, *et al.*, 2024, in press)

Overview

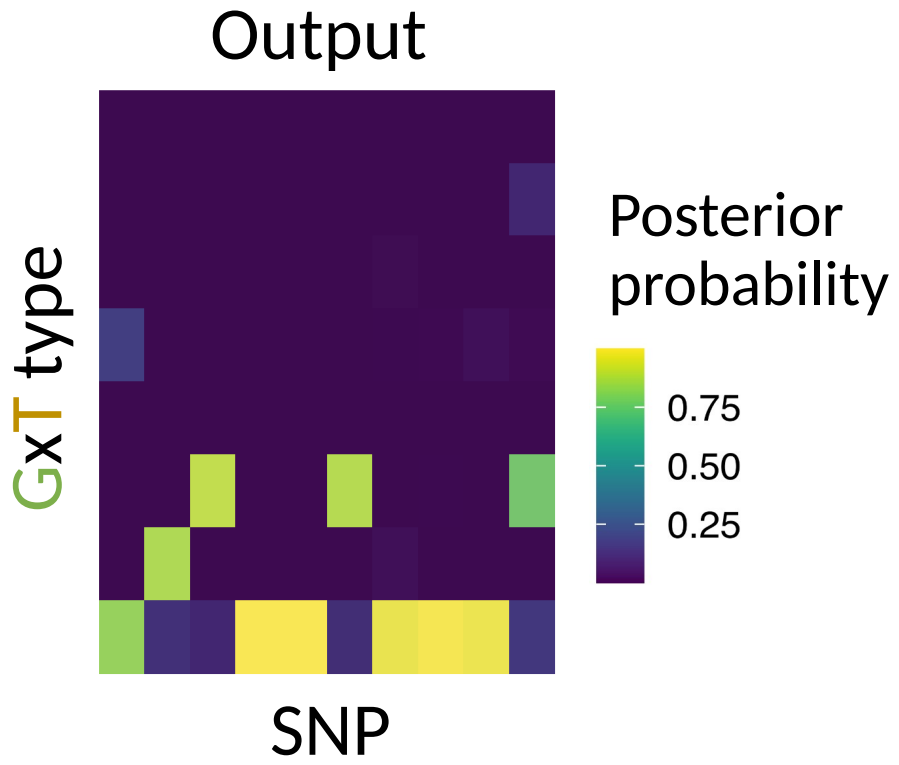
- Background, existing methods, motivation
- ClassifyGxT: classifying gene-by-treatment interactions
    - Probabilistic classification (Bayesian model selection)
    - Modeling molecular count phenotypes (Nonlinear regression)
- Applications to experimental data
- Summary

Goal: classifying GxT interactions

Input

A set of SNPs w/ GxT

- Individual genotype
- Individual phenotype

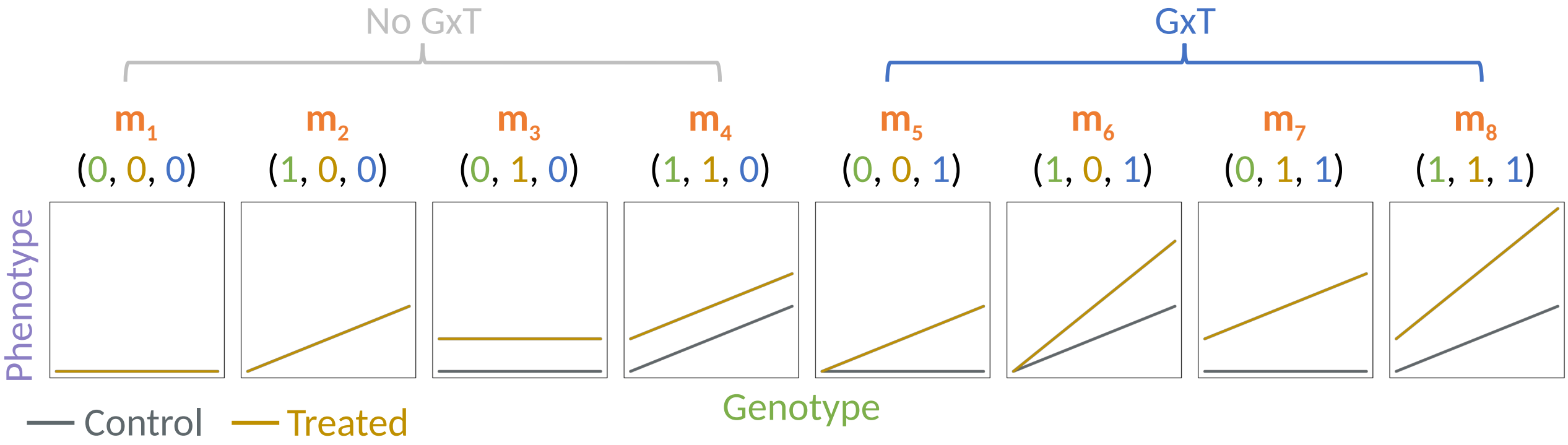


- Existing GxT classification methods require paired data (Barber *et al.*, 2010; Maranville *et al.*, 2011).

Approach: Bayesian model selection (BMS)

phenotype = genotype + treatment + genotype x treatment + noise

interaction



Modeling details for GxT linear model

linear model: $y_i = \beta_0 + \beta_g g_i + \beta_t t_i + \beta_{g\times t} g_i t_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$

model prior: $\Pr(\mathbf{m} = \mathbf{m}_j) = \frac{1}{8}$ for the j-th model $(j = 1, \cdots, 8)$

model posterior: $\Pr(\mathbf{m} = \mathbf{m}_j \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{m}_j)\Pr(\mathbf{m} = \mathbf{m}_j)}{\sum_{k=1}^{8} p(\mathbf{y} \mid \mathbf{m}_k)\Pr(\mathbf{m} = \mathbf{m}_k)}$

marginal likelihood: $p(\mathbf{y} \mid \mathbf{m}_j) = \int\int p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta} \mid \sigma^2) p(\sigma^2)\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\sigma^2$

effect priors:

$$\beta_0 \sim \mathrm{N}(0, \phi_0^2\sigma^2),\ \sigma^2 \sim \mathrm{IG}(\frac{\kappa}{2}, \frac{\lambda}{2})$$

$$\beta_g \sim \mathrm{N}(0, \phi_g^2\sigma^2),\ \beta_t \sim \mathrm{N}(0, \phi_t^2\sigma^2),\ \beta_{g\times t} \sim \mathrm{N}(0, \phi_{g\times t}^2\sigma^2)$$

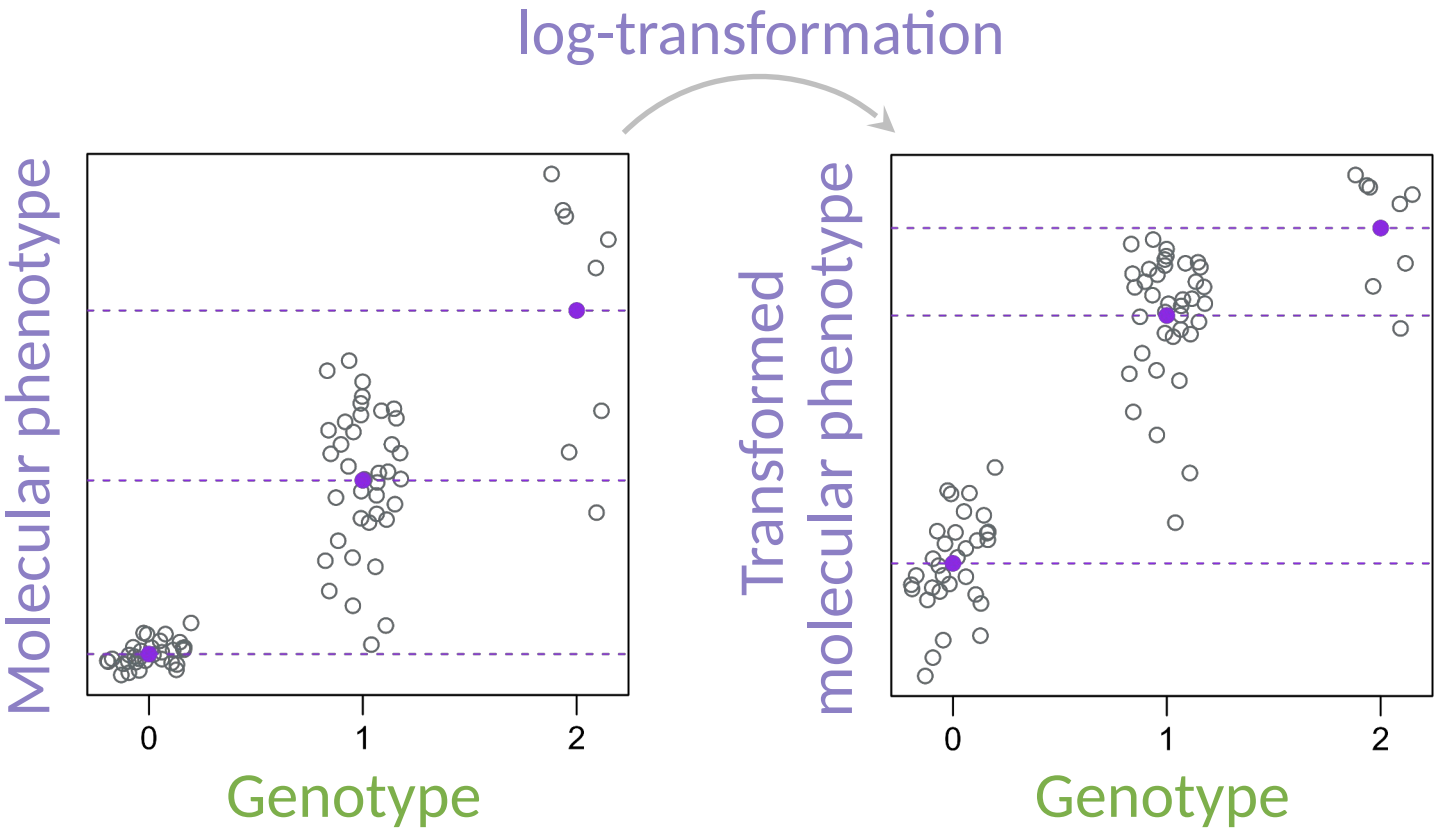
ϕ's, κ, and λ are hyperparameters.

Overview

- Background, existing methods, motivation
- ClassifyGxT: classifying gene-by-treatment interactions
    - Probabilistic classification (Bayesian model selection)
    - Modeling molecular count phenotypes (Nonlinear regression)
- Applications to experimental data
- Summary

Molecular count data and allelic additivity

- Molecular count data with respect to genotype is linear only on the original scale (Sun, 2012; Mohammadi *et al.*, 2017; Palowitch *et al.*, 2018).



- Non-linear on the log scale
- Linear assumption
  → Inaccurate effect size
  → Faulty classification?
- The aFC (Mohammadi et al) and ACME (Palowitch et al) methods account for this nonlinearity

Nonlinear model for GxT analysis on molecular phenotypes

- Extended aFC and ACME ("log-NL" for log-NonLinear)
  - For the *i*-th individual, the model is cast as

$$y_i = f(g_i, t_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),$$

where $f(\cdot)$ is a nonlinear function of g_i and t_i,

$$\begin{aligned} f(g_i, t_i) = \log \big(& (1 - \tfrac{g_i}{2})(1 - t_i) \exp(\beta_0) \\ & + (\tfrac{g_i}{2})(1 - t_i) \exp(\beta_0 + 2\beta_g) \\ & + (1 - \tfrac{g_i}{2})(t_i) \exp(\beta_0 + \beta_t) \\ & + (\tfrac{g_i}{2})(t_i) \exp(\beta_0 + 2\beta_g + \beta_t + 2\beta_{g\times t}) \big), \end{aligned}$$

- Not analytically tractable:
  - MCMC + bridge sampling
  - Maximum *a posteriori* (MAP) estimation + Laplace approximation

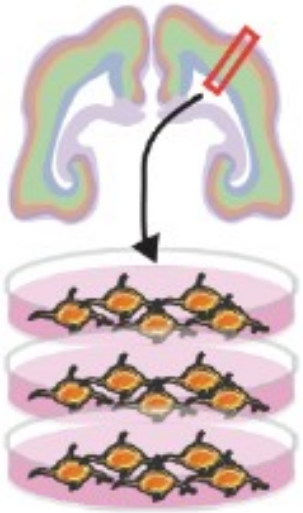
Overview

- Background, existing methods, motivation
- ClassifyGxT: classifying gene-by-treatment interactions
    - Probabilistic classification (Bayesian model selection)
    - Modeling molecular count phenotypes (Nonlinear regression)
- Applications to experimental data
- Summary

A motivating dataset from Jason Stein lab

Genotyped primary human neural progenitor cells (hNPCs) derived from fetal donors

Jason Stein (UNC)



+ Vehicle (control)

+ Growth stimulation (treatment)

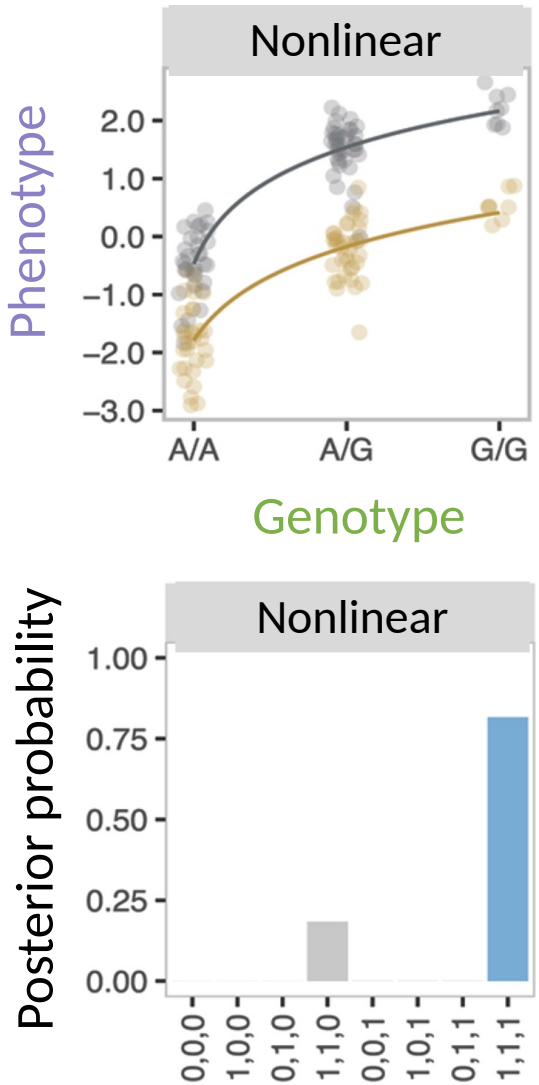
→ RNA-seq → "Interaction" approach

~100 gene-SNP pairs w/ significant GxT interactions

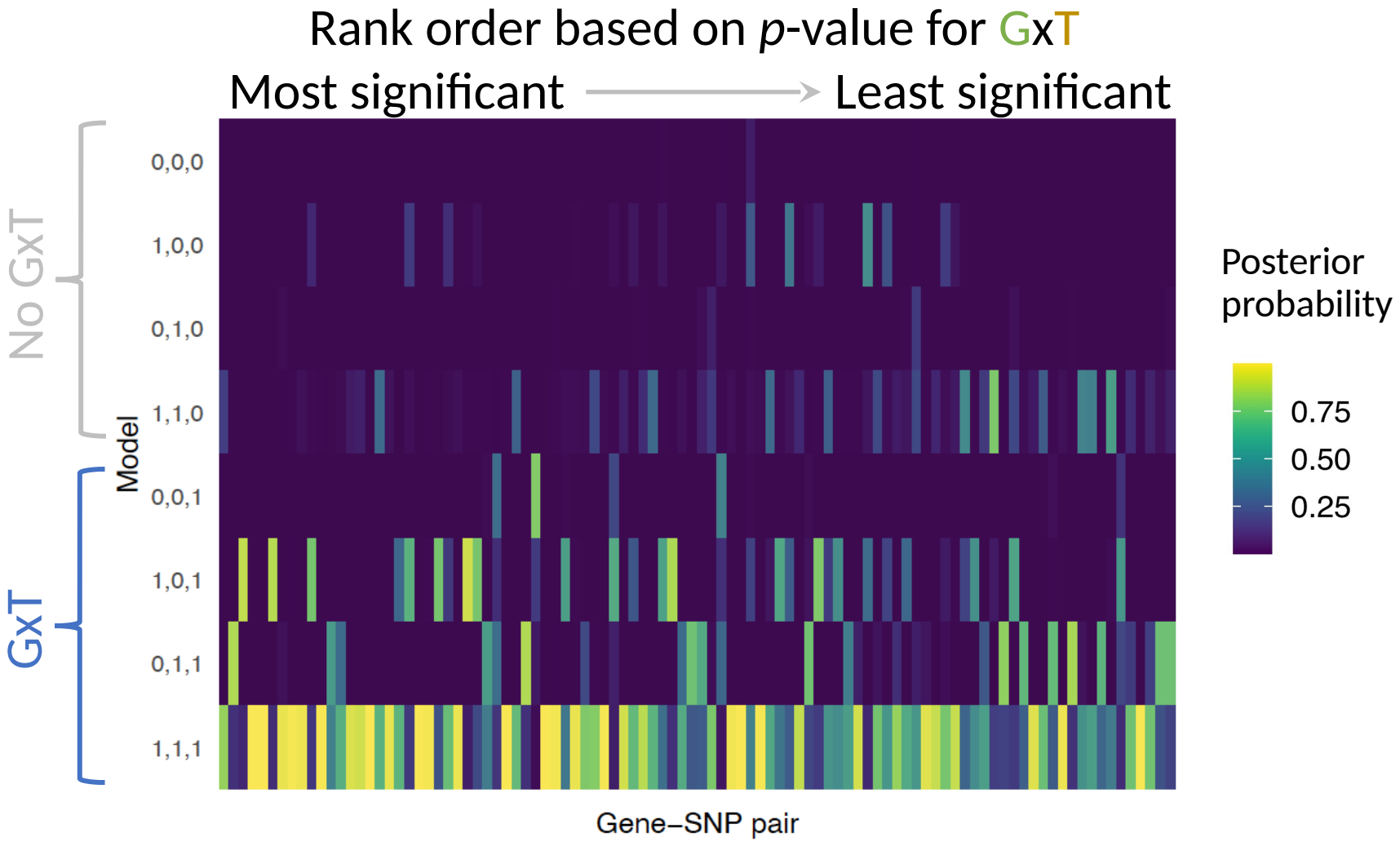
(Matoba, Le, Valone, *et al.*, 2024, in press)

ClassifyGxT captures evidence of different types of GxT

LINC02073 - rs7212610



All ~100 gene-SNP pairs



Summary

- Bayesian model selection framework for classifying GxT interactions on molecular count phenotypes (3 versions: log-NL, log-LM, RINT-LM)
- Provides more interpretable information about GxT interactions than the current practice
- non-linear modeling can help with molecular count phenotypes

GitHub

ClassifyGxT

software

bioRχiv

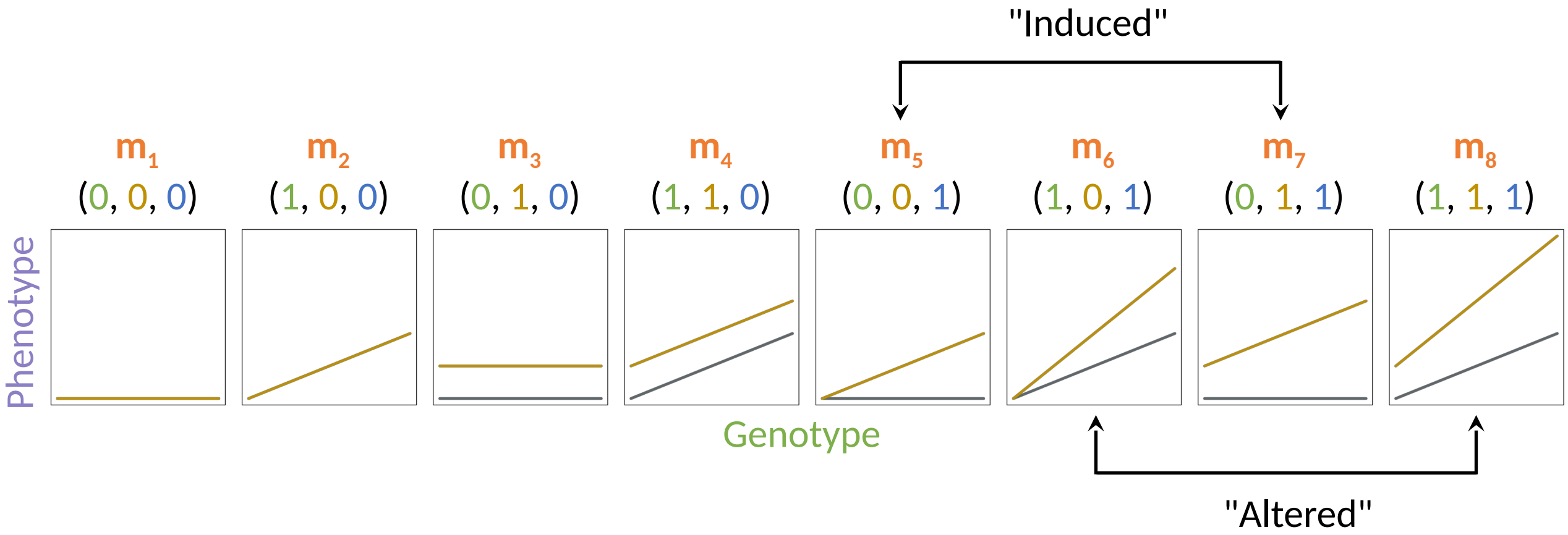
manuscript

pre-print

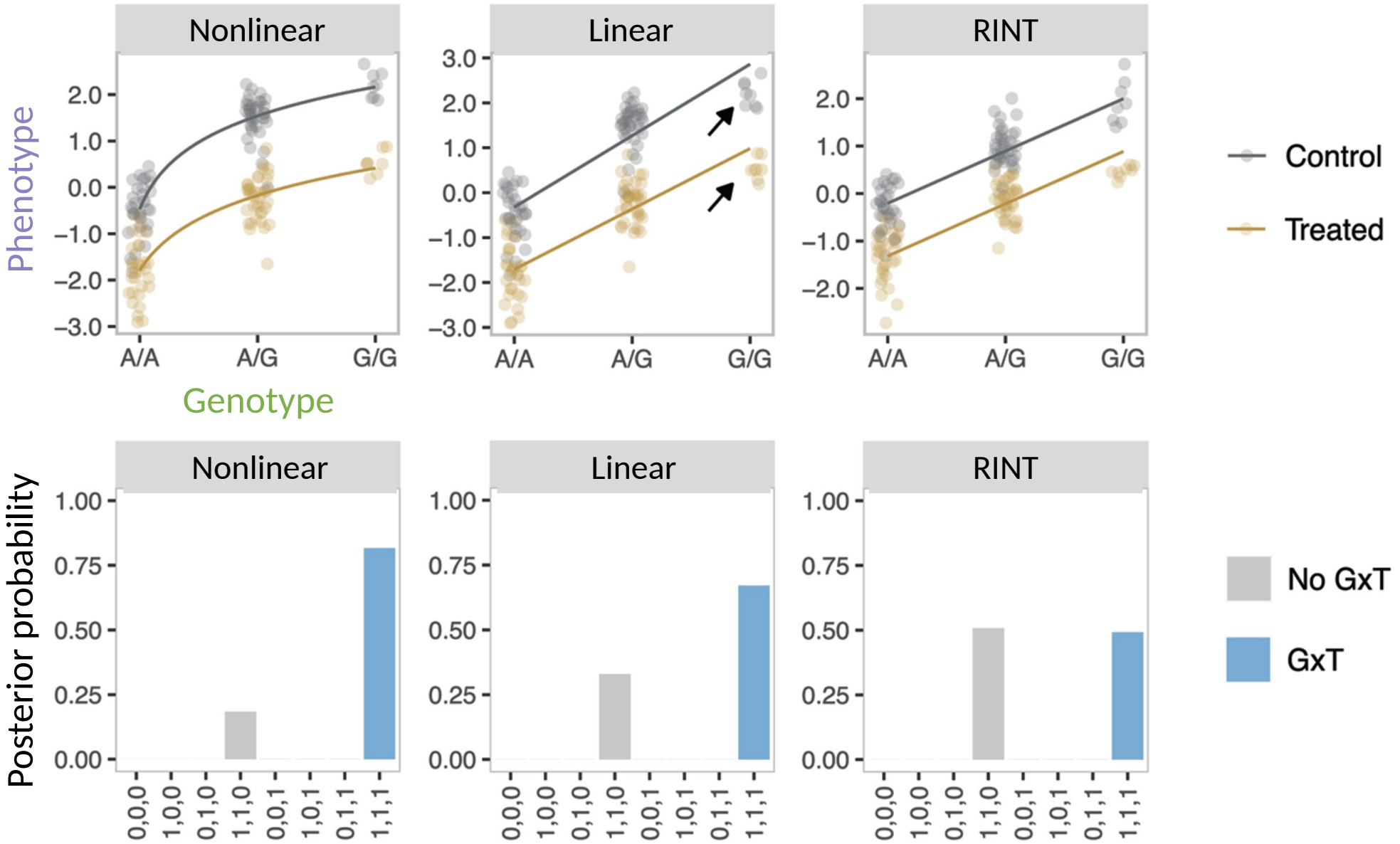
Or use links at https://valdarlab.unc.edu/software/

Extra slides

BMS: summarize posterior probability

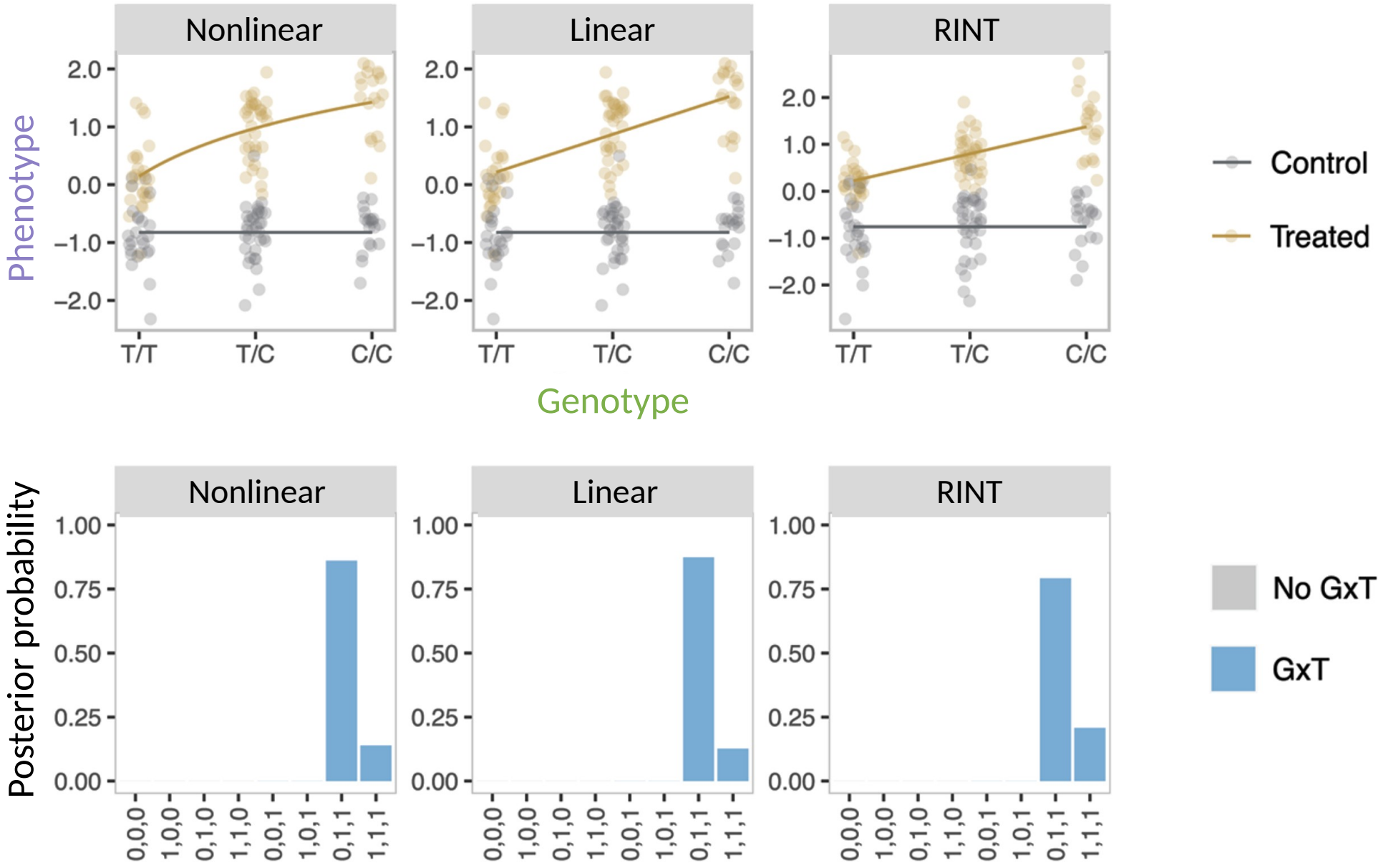


Nonlinear regression can be beneficial



LINC02073 - rs7212610

ClassifyGxT identifies treatment-induced genotype effect



TYR – rs10830237

Old slides

QTL mapping approaches to detect GxT

- stratified analysis

control: phenotype = genotype + noise

treated: phenotype = genotype + noise

Existing GxT detection method: "stratified" approach

- Based on QTL mapping
- Perform QTL mapping separately

For a given genetic variant (SNP) and the i-th subject ($i = 1, \cdots, N$):

$$y_i = \beta_0 + \beta_g g_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$$

y_i: phenotype; g_i: genotype

- Test H_0: $\beta_g = 0$
- Set a p-value threshold
- Get a list of significant SNPs

- No ranking

Existing GxT detection method: "delta" approach

- Based on QTL mapping
- Compute delta (Treated - Control)

For a given genetic variant (SNP) and the i-th subject $(i = 1, \cdots, N)$:

$$y_i = \beta_0 + \beta_g g_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$$

delta

genotype

- Test H_0: $\beta_g = 0$
- Set a p-value threshold
- Get a list of significant SNPs

- Requiring paired data
- No GxT classification

Existing GxT detection method: "interaction" approach

- Based on QTL mapping
- Combine the data

For a given genetic variant (SNP) and the i-th subject ($i = 1, \cdots, N$):

$$y_i = \beta_0 + \beta_g g_i + \beta_t t_i + \beta_{g\times t} g_i t_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$$

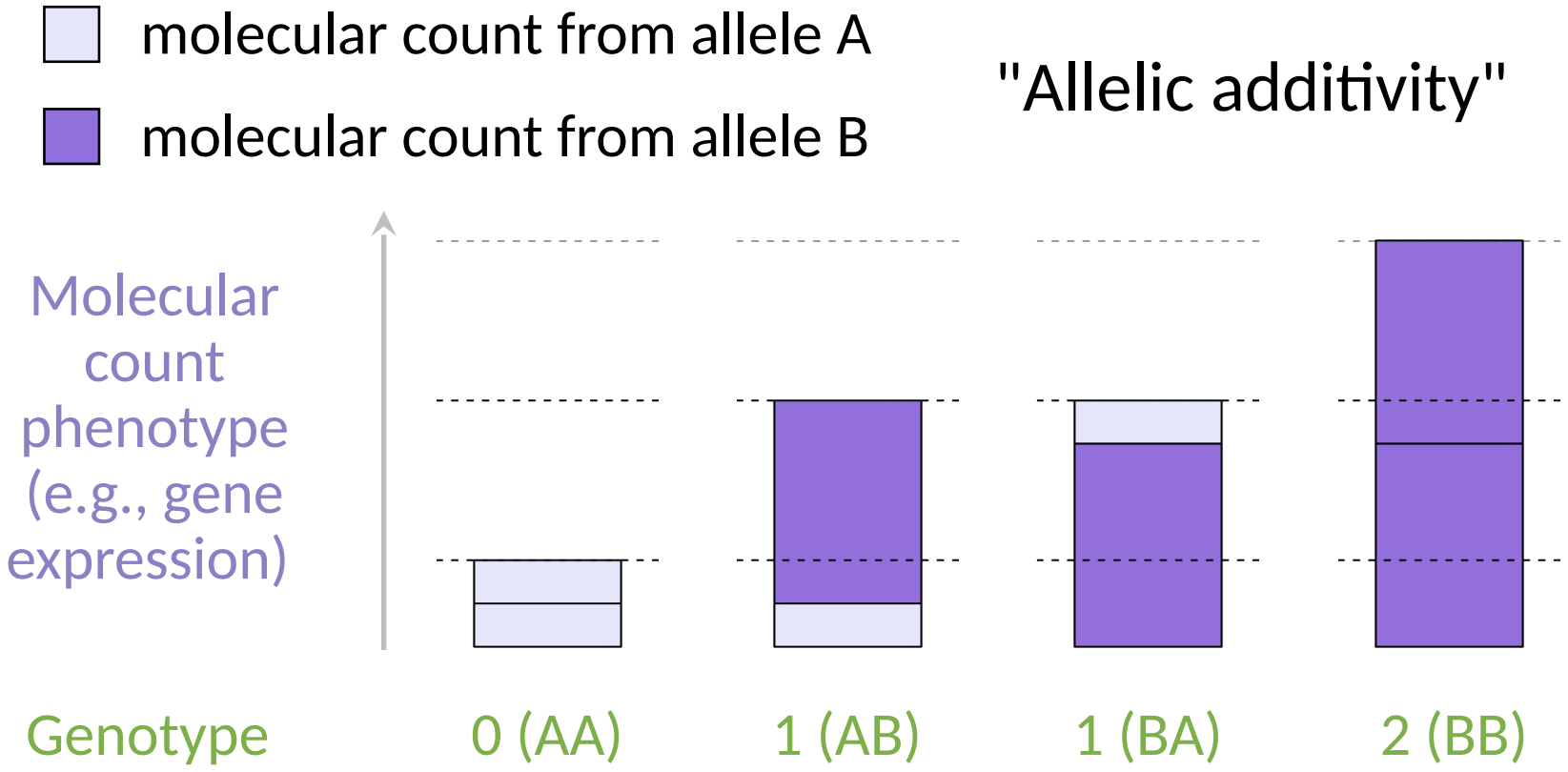
phenotype genotype treatment interaction

- Test H_0: $\beta_{g\times t} = 0$
- Set a p-value threshold
- Get a list of significant SNPs

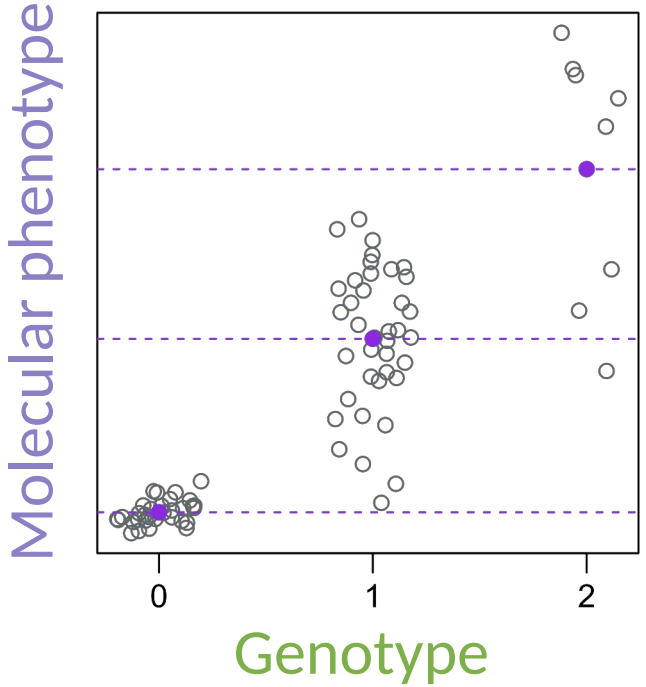
- Flexible (not requiring paired data)
- No GxT classification

Limitation of linear regression in modeling count data (1)

- Molecular count data with respect to genotype is linear only on the original scale (Sun, 2012; Mohammadi *et al.*, 2017; Palowitch *et al.*, 2018).



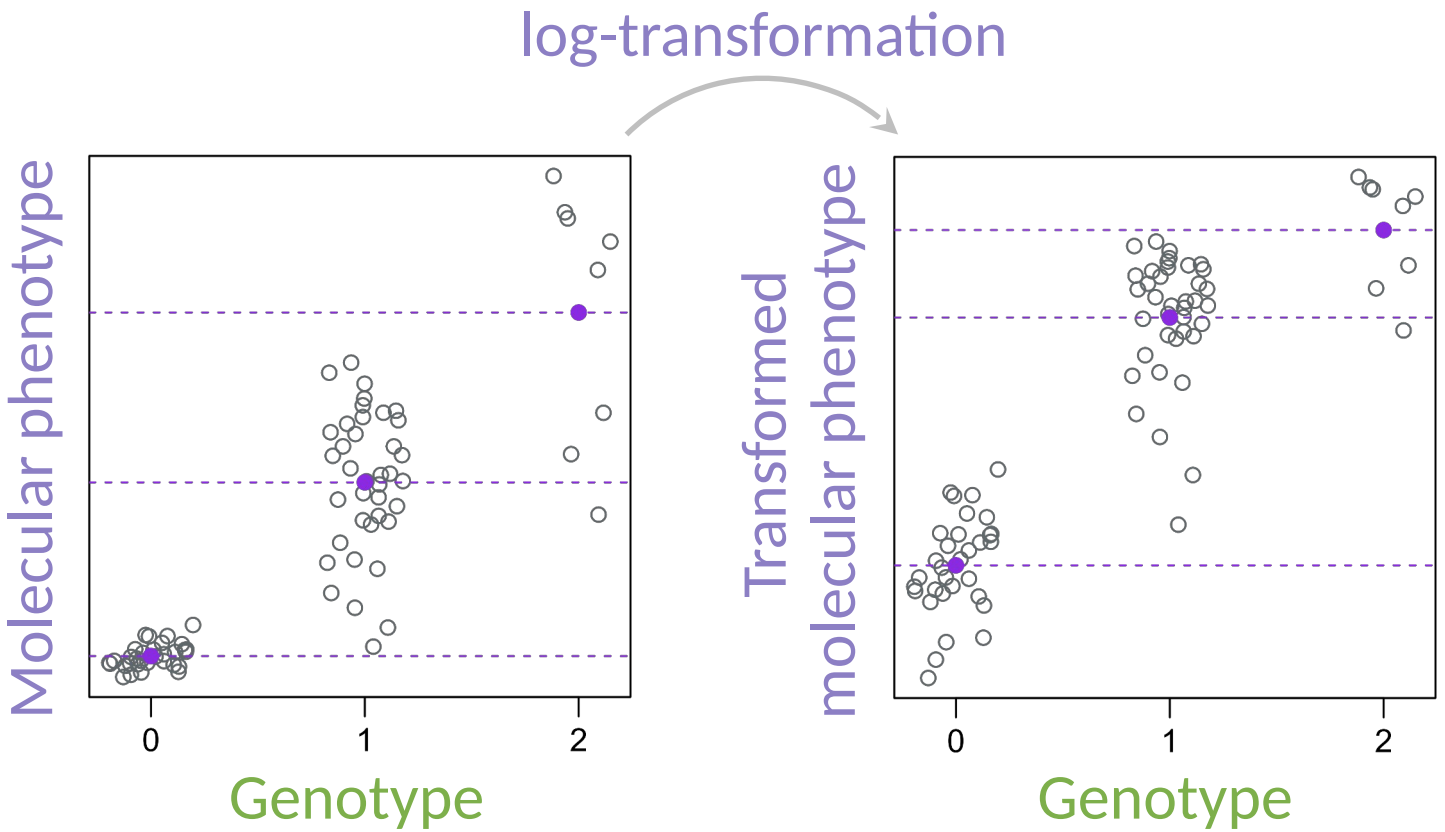
Reproduction of Sun, 2012; Mohammadi *et al.*, 2017



Data from Stein lab

- But, it exhibits variance heterogeneity.

Limitation of linear regression in modeling count data (2)



- Molecular count phenotype with respect to genotype is linear:
  - in the original scale
  - but not in the log scale [1-3].
- Linear assumption
  → Inaccurate effect size
- The aFC and ACME methods account for the nonlinearity [2,3].

(1: Sun, 2012; 2: Mohammadi *et al.*, 2017; 3: Palowitch *et al.*, 2018)