Estimating Association Effects of Copy Number Variants Using Penalized Regression with Lasso and Weighted Fusion Penalties

Yaqin Si[1], Wenbin Lu[1], Albert A. Tucci[1], Hui Wang[2], Yuhuan Cheng[1], Li-San Wang[2], Gerard Schellenberger[2], Wan-Ping Lee[2], Jung-Ying Tzeng[1]
1. North Carolina State University, Raleigh, NC, United States. 2. University of Pennsylvania, Philadelphia, PA, United States.

Goal

- **Copy number variants (CNVs)** are DNA gains or losses involving at least 50bp.
- We propose a statistical method for estimating CNV effects on phenotype.
- **Properties of the proposed methods:**
  - Incorporate **multiple CNV features** (see Figure 1)
    - Dosage (duplication vs. deletion, number of extra copies)
    - Length
  - Assess the joint effect of all CNV events within a genomic region
  - Avoid arbitrary CNV region definition
  - Allow CNV effects comparable across different studies

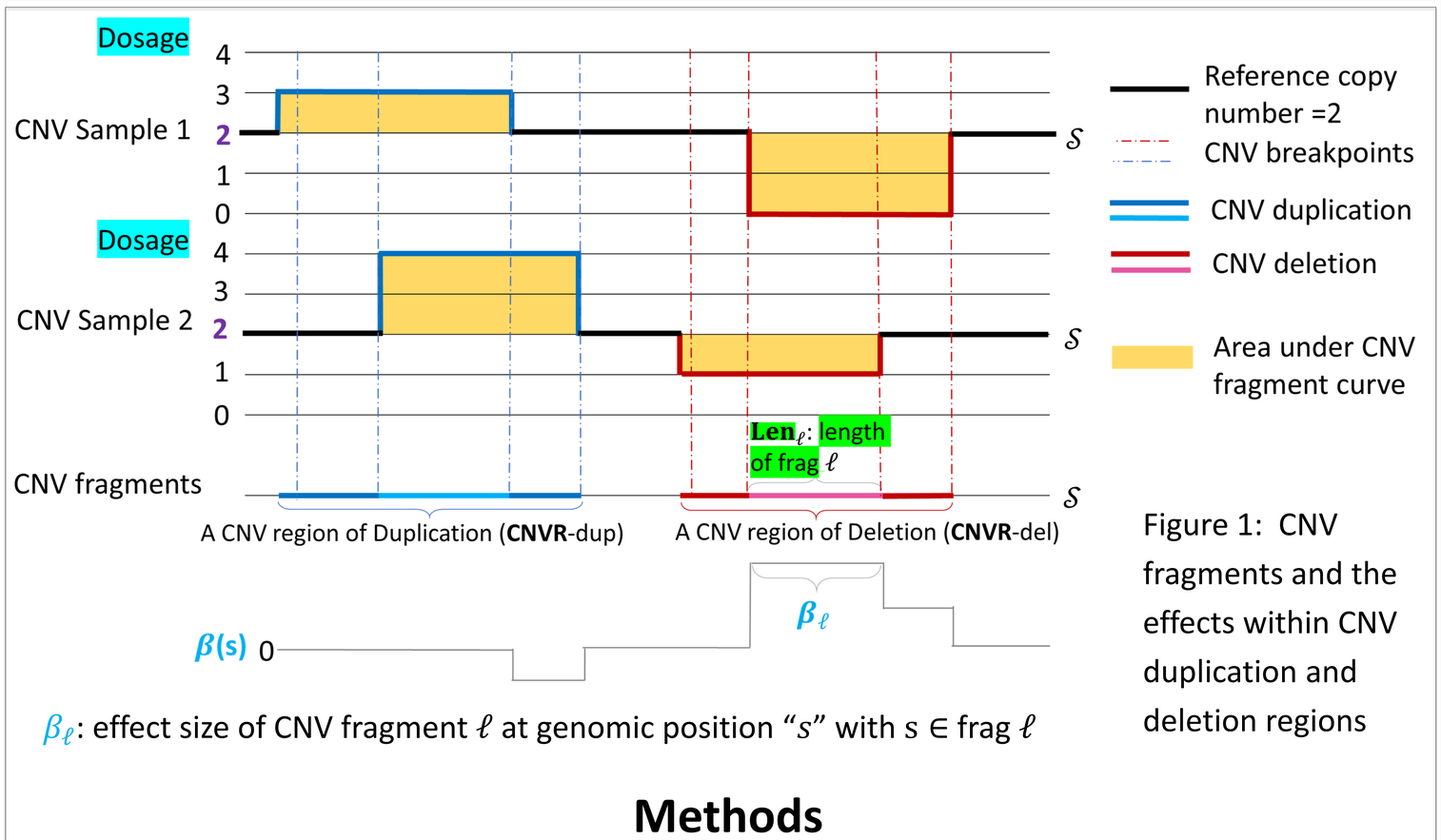

Figure 1: CNV fragments and the effects within CNV duplication and deletion regions

β_ℓ: effect size of CNV fragment ℓ at genomic position "s" with s ∈ frag ℓ

Methods

Proposed Model:

$$g(\mu_i) = \beta_0 + \int_s \beta(s) x_i^*(s) ds = \beta_0 + \sum_{\text{frag}\,\ell} \int_{s \in \text{frag}\,\ell} \beta(s)\, x_i^*(s) ds,$$

where $x_i^*(s) \equiv |\text{Dosage}_{is} - 2|$

- Dosage_{is}: subject i's dosage at position s
- 2: reference copy number
- $|\cdot|$: absolute value accommodates duplication and deletion

- Assume constant effect within a CNV fragment, i.e., $\beta(s) = \beta_\ell$ for s ∈ frag ℓ

  $\therefore \int_{s \in \text{frag}\,\ell} \beta(s) x_i^*(s) ds = \beta_\ell \int_{s \in \text{frag}\,\ell} x_i^*(s) ds \equiv \beta_\ell x_{i\ell}$

  where $x_{i\ell} = \int_{s \in \text{frag}\,\ell} x_i^*(s) ds$ = area under **CNV** curve (AUC) = $\text{Len}_{i\ell} \times |\text{Dosage}_{i\ell} - 2|$
- The proposed model can be simplified as

  $g(\mu_i) = \beta_0 + \sum_\ell \beta_\ell x_{\ell i}$ ← regress trait on $x \equiv$ "length"×"dosage"

Model Fitting:

- Proposed to impose two penalty terms:
  - Lasso penalty to deal with high dimensionalities and to select phenotype-associated CNVs, i.e., $\|\beta\|_1 < a$
  - Weighted Fusion penalty to encourage (not force) adjacent CNV fragments to have similar effect size, i.e., $\sum_{\ell=1}^{L} w_{\ell,\ell+1} (\beta_{\ell+1} - \beta_\ell)^2 < b$

Model interpretation

- Recall that $\beta(s) = \beta_\ell$ for s ∈ frag ℓ, i.e.,
  β_ℓ is CNV effect size at genomic position "s" with s in fragment ℓ
- In contrast, in the commonly considered model, $g(\mu_i) = \alpha_0 + \sum_\ell \alpha_\ell \cdot |\text{Dosage}_{i\ell} - 2|$,
  α_ℓ is the aggregated CNV effect across all positions in fragment ℓ (and hence would depend on the fragment length)
- If each study has its own CNV fragment definition, β_ℓ of the same genomic position is comparable across different studies but α_ℓ is not
- Can show that $\alpha_\ell = \beta_\ell \times \text{Length}_\ell$ (i.e., $\frac{\alpha_\ell}{\text{length}_\ell}$ is comparable across different studies)

Simulation and Real Data Analysis

Models to compare

Baseline 1: Ds_Lasso (X=|Dosage-2|)

Baseline 2: AUC_Lasso (X=AUC)

Model 1: AUC_Eql (Lasso + Fusion Equal weight)

Model 2: AUC_Cor (Lasso + Fusion Correlation-based weight)

Simulation scenarios and results:

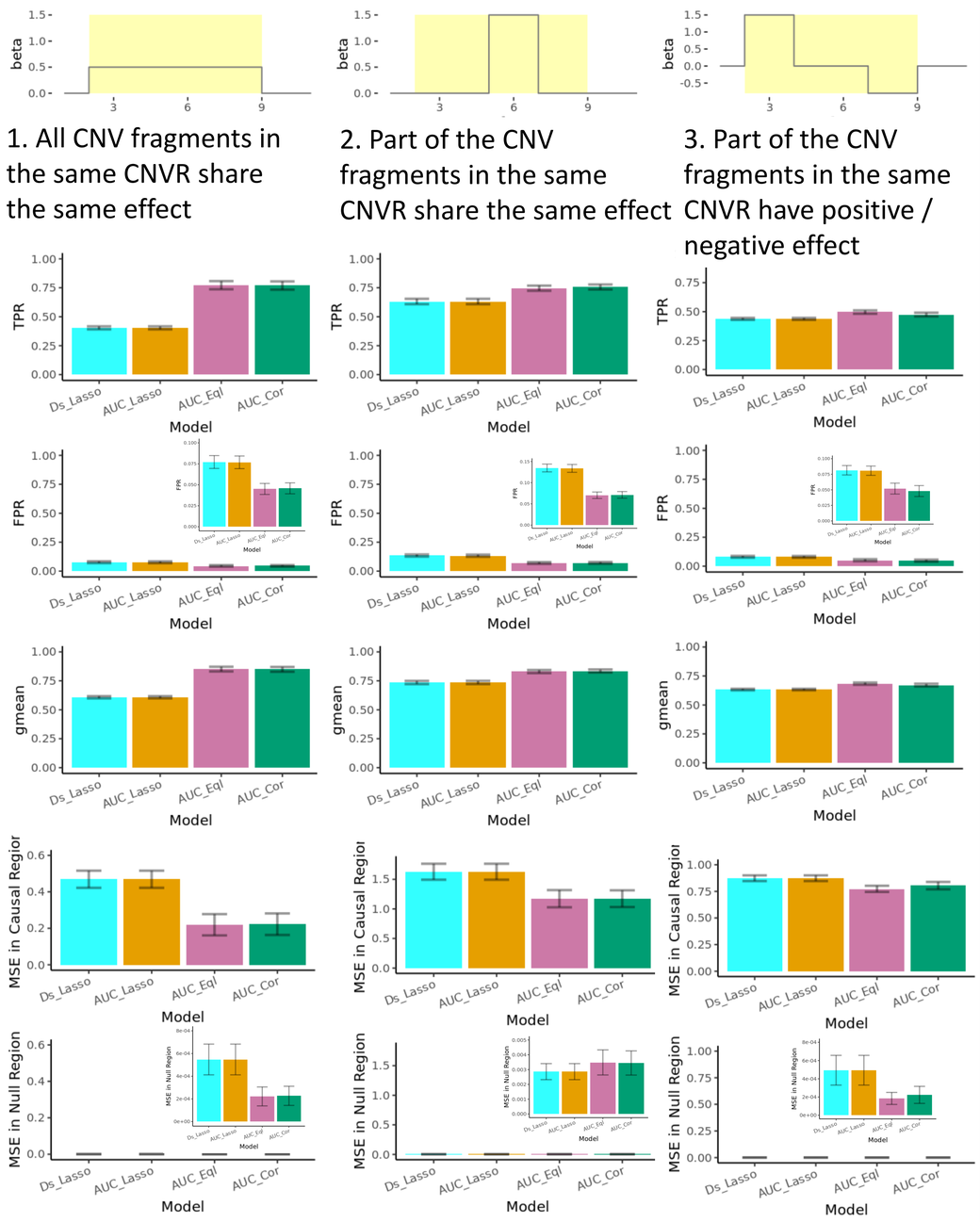

Figure 2: Simulation results under different effect scenarios

The proposed models show:

- **Better variable selection performance** (higher TPR, lower FPR, and higher G-measure) than the Lasso models.
- **Better effect size estimation performance** (Lower/equivalent MSE in Causal region and Null region) than Lasso models.

Real data application:

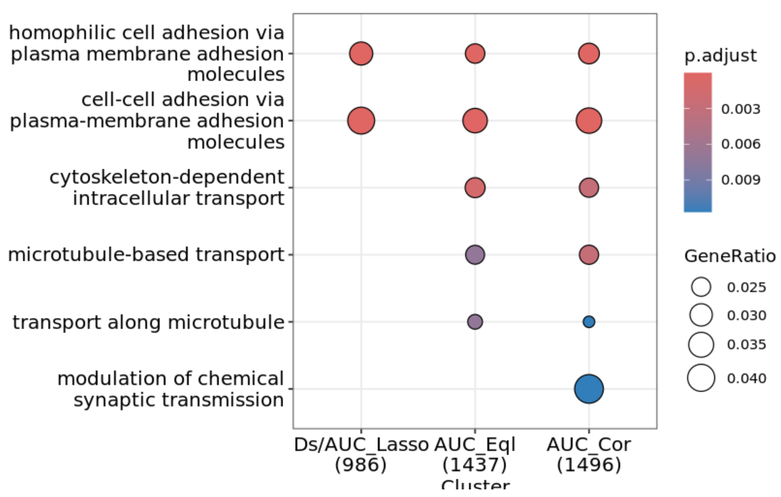

Figure 3: Gene ontology (GO) enrichment analysis for Alzheimer's Diseases associated genes identified by different models

The proposed models identify **additional enriched genes in** pathways related to **neuron structure and function**.

Future work: Apply to other phenotype-CNV association analysis, develop an R package for easy application