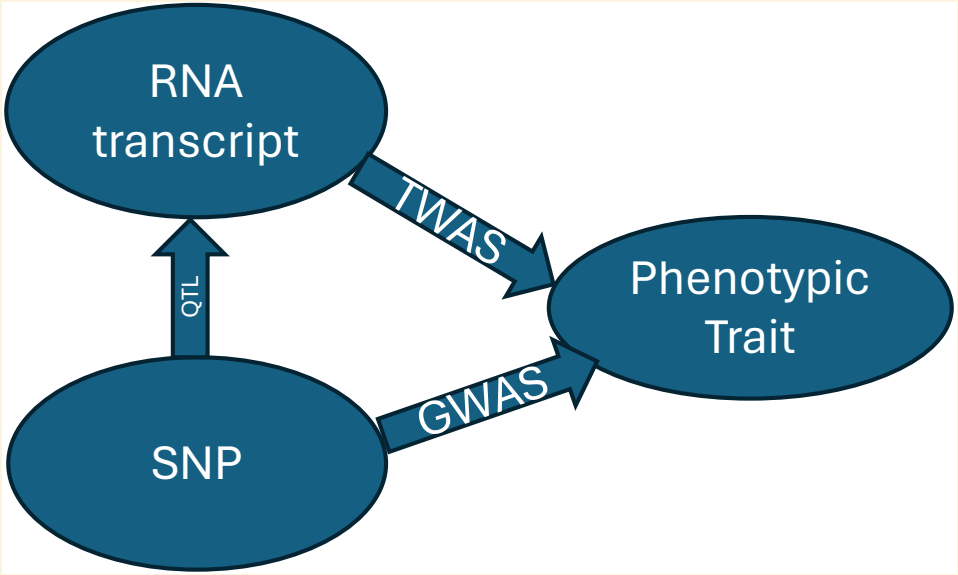
Combined GWAS and TWAS using SVEN

Somak Dutta and Vivekananda Roy

Iowa State University

AG2PI Workshop #26
May 17, 2024

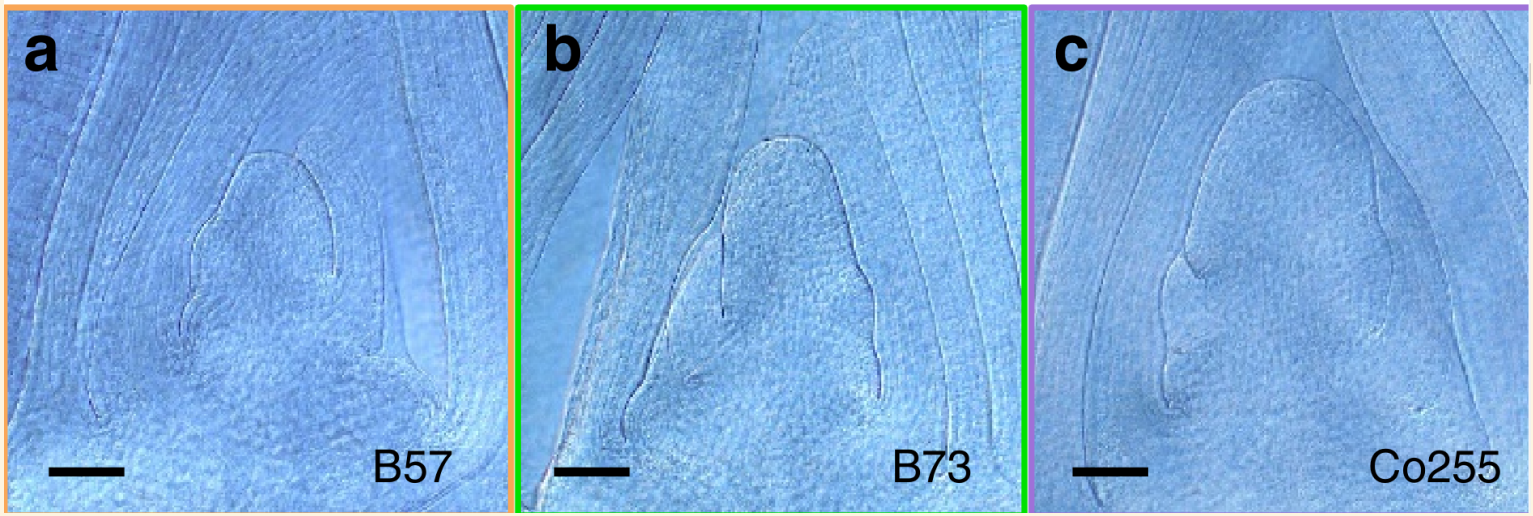
GWAS vs TWAS



- Genome-wide association studies (GWAS) use SNPs (and other types of genetic markers) to discover genes associated with phenotypic variation in traits of interest.
- Transcriptome-wide association studies (TWAS) can identify correlations between gene expression levels and phenotypic variation in these same traits.
- Although, GWAS and TWAS results can be combined using the Fisher's Combined Test, the statistical power to detect causal genes could be increased by combining GWAS and TWAS through a single model.

Diversity in the maize SAM

- The maize shoot apical meristem (SAM) comprises a small pool of stem cells that generate all above-ground organs.
- $n = 360$ genotypes, $p_1 = 1,279,929$ SNPs, $p_2 = 39,035$ gene expressions (RNA transcript).
- For the SAM data, the goal is to identify candidate genes associated with SAM morphometric variation.



Variation in maize SAM morphology. Examples of small (a), intermediate (b) and large (c) SAM lines (Figure taken from Leiboff et al., 2015)

Notations

- y: BLUEs or BLUPs of genotypes (e.g., logarithm of SAM volumes).
- X: The $n \times p_1$ covariate matrix of SNPs without intercept.
- Z: The $n \times p_2$ covariate matrix of gene expressions without intercept.
- Multiple linear regression model: $y = \mu_0 1_n + X\beta + Z\alpha + \epsilon$, where $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$.
- β, α: Vectors of regression coefficients, σ^2: error variance.
- The **goal** is to identify the best subset of predictors, among all possible subsets of predictors.
- Multilocus methods can consider multiple genetic variants simultaneously and account for population structure and relatedness between individuals and thus have better potential to identify important genetic variants.

Model setup

- Linear hierarchical regression model:

$$
\begin{aligned}
y|\beta,\beta_0,\sigma^2,\gamma,\delta &\sim \mathcal{N}_n\left(1_n\mu_0 + X_\gamma\beta_\gamma + Z_\delta\alpha_\delta, \sigma^2 I\right),\\
\beta_j|\mu_0,\sigma^2,\gamma &\overset{\text{ind}}{\sim} \mathcal{N}\left(0, \frac{\gamma_j}{\lambda_1}\sigma^2\right) \text{ for } j = 1,\ldots,p_1,\\
\alpha_j|\mu_0,\sigma^2,\delta &\overset{\text{ind}}{\sim} \mathcal{N}\left(0, \frac{\delta_j}{\lambda_2}\sigma^2\right) \text{ for } j = 1,\ldots,p_2,\\
\left(\mu_0,\sigma^2\right)|\gamma,\delta &\sim f\left(\mu_0,\sigma^2\right) \propto 1/\sigma^2,\\
\gamma|w_1 &\sim f(\gamma|w_1) = w_1^{|\gamma|}(1-w_1)^{p_1-|\gamma|},\\
\delta|w_2 &\sim f(\delta|w_2) = w_2^{|\delta|}(1-w_2)^{p_2-|\delta|},
\end{aligned}
$$

where

- γ is a $p_1 \times 1$ vector of latent binary variable indicating the inclusion or exclusion of SNPs: i.e., $\gamma_j = 1$ if jth SNP is *important*.
- δ is a $p_2 \times 1$ vector of latent binary variable indicating the inclusion or exclusion of genes, i.e., $\delta_j = 1$ if jth gene is *important*.
- $|\gamma| = \sum_{j=1}^{p_1} \gamma_j$ is the size of a model γ,
- $|\delta| = \sum_{j=1}^{p_2} \delta_j$ is the size of a model δ and
- $\lambda_2, \lambda_1 > 0$ and $w_1, w_2 \in (0,1)$ are assumed to be known non-random functions of n, p_1 and p_2.

Model explained

Suppose $p_1 = p_2 = 2$, then there are 16 possible models:

$$
\begin{array}{lll}
\gamma = (0,0), \delta = (0,0): & Y = \mu_0 1 + & +\epsilon \\
\gamma = (1,0), \delta = (0,0): & Y = \mu_0 1 + \beta_1 X_1 & +\epsilon \\
\gamma = (0,1), \delta = (0,0): & Y = \mu_0 1 + \beta_2 X_2 & +\epsilon \\
\gamma = (1,1), \delta = (0,0): & Y = \mu_0 1 + \beta_1 X_1 + \beta_2 X_2 & +\epsilon \\
\gamma = (0,0), \delta = (1,0): & Y = \mu_0 1 + \alpha_1 Z_1 & +\epsilon \\
\vdots & \vdots & \vdots \\
\gamma = (1,1), \delta = (1,1): & Y = \mu_0 1 + \beta_1 X_1 + \beta_2 X_2 + \alpha_1 Z_1 + \alpha_2 Z_2 & +\epsilon
\end{array}
$$

- Posterior distribution of (γ, δ): Analytical integration w.r.t $(\mu_0, \beta_\gamma, \alpha_\delta, \sigma^2)$ leads to the marginal posterior distribution of (γ, δ): $f(\gamma, \delta | y)$.

- Large models tend to over-fit but are penalized by the prior inclusion probabilities w_1 and w_2. Under certain assumptions,

$$\lambda_i \sim \frac{n}{p_i^2}, \quad w_i \sim \frac{\sqrt{n}}{p_i}, \quad (i = 1, 2)$$

guarantee model selection consistency.

Selection of variables with embedded screening (SVEN)

- Given the data, SVEN finds out the combinations of SNPs and genes that have *high posterior probabilities.*
- Using stochastic search algorithms, SVEN produces
  - **MAP** model: the model with the highest posterior probability,
  - **WAM**: weighted average model.
  - **MIP**: marginal inclusion probabilities of each marker.

Example

Synthetic dataset generated from the model: $n = 60$ lines

$$Y = 0.5 + \mathbf{0.1X_1} + 0X_2 + 0Z_1 + \mathbf{0.4Z_2} + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

We'll vary σ. Note that the SE of genotypic BLUEs decrease with increasing number of reps. [X_1, X_2, Z_1, Z_2 are generated from $N(0, 1)$ distributions.]

	γ_1	γ_2	δ_1	δ_2	$\sigma = 1.4$	$\sigma = 0.5$	$\sigma = 0.1$
1	·	·	·	·	0.60	·	·
2	1	·	·	·	·	·	·
3	·	1	·	·	0.01	·	·
4	1	1	·	·	·	·	·
5	·	·	1	·	·	·	·
6	1	·	1	·	·	·	·
7	·	1	1	·	·	·	·
8	1	1	1	·	·	·	·
9	·	·	·	1	0.38	0.97	·
10	1	·	·	1	·	0.01	0.98
11	·	1	·	1	·	0.01	·
12	1	1	·	1	·	·	0.01
13	·	·	1	1	·	0.01	·
14	1	·	1	1	·	·	0.01
15	·	1	1	1	·	·	·
16	1	1	1	1	·	·	·
Total					1	1	1

Duplicated SNPs or GEs can hurt discoveries

Same model as before. Fix $\sigma = 0.5$

$$Y = 0.5 + \mathbf{0.1X_1} + 0X_2 + \mathbf{0.4Z_2} + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

but now $Z_1 \equiv Z_2$. We study the effect of shrinkage. Assume $\lambda_1 = \lambda_2 = \lambda$

	γ_1	γ_2	δ_1	δ_2	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$
1	·	·	·	·	·	·	·
2	1	·	·	·	·	·	·
3	·	1	·	·	·	·	·
4	1	1	·	·	·	·	·
5	·	·	1	·	·	·	·
6	1	·	1	·	0.44	0.39	0.01
7	·	1	1	·	·	·	·
8	1	1	1	·	·	0.01	·
9	·	·	·	1	·	·	·
10	1	·	·	1	0.44	0.39	0.01
11	·	1	·	1	·	·	·
12	1	1	·	1	·	0.01	·
13	·	·	1	1	·	·	·
14	1	·	1	1	0.11	0.20	0.90
15	·	1	1	1	·	·	·
16	1	1	1	1	·	0.01	0.07
Total					1	1	1

The total probability gets divided into *equivalent* models. Same things happen when Z_1 and Z_2 are very highly correlated.

Comments

- If the trait is complex, it is advised to keep λ_1 and λ_2 high (e.g., 0.1 – 1) (Li et al., 2023).
- It is advised to thin the SNPs to reduce the effect of slow LD decay.
- Quality control of the SNP and gene data are crucial but are beyond the scope of this workshop.
- Note, you can also use the SVEN to run only GWAS or only TWAS or only QTL mapping.

References:

- Leiboff et al. (2015). Genetic control of morphometric diversity in the maize shoot apical meristem. *Nature Communications*, 6:8974.
- Li et al. (2013). Model based screening embedded Bayesian variable selection for ultra-high dimensional settings. *Journal of Computational and Graphical Statistics*, 32(1), 61-73.

Demonstration

We use the SAM panel (Leiboff et al, 2015): $n = 360$ maize lines.

- $p_1 = 10,000$ SNPs are selected.
  - SNPs with MAF $< 5\%$ are dropped.
  - Major alleles are coded as 0, minor allele coded as 1.
  - Missing values are replaced by the probability of being a minor allele.
  - Stored in a sparse matrix format (R-package Matrix).

- $p_2 = 2000$ genes are selected. For each gene:
  - gene expressions less than 5%-ile are set to zero.
  - gene expressions bigger than 95%-ile are set to 2.
  - remaining values are linearly interpolated to [0,2]

- Phenotypic values y generated synthetically.

Data and code available at:
`https://faculty.sites.iastate.edu/somakd/workshops`