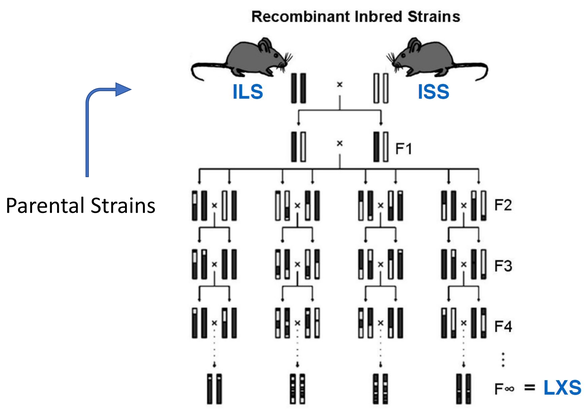
The Role of Covariates in Estimating Model-based Heritability Score for High-throughput Sequencing Data.

Pratyaydipta Rudra[1], **Julius Olaifa**[1], **Joshua Habiger**[1]
[1]Department of Statistics , Oklahoma State University.

Goals of the study

- Heritability (h^2) is the amount of a trait's variation that may be attributed to genotypic variation.
- Heritability analysis of intermediate traits such as gene-expression obtained from high-throughput sequencing data is challenging due to the non-normality of the data.
- A framework to estimate and test heritability for such data was proposed by Rudra et al [1, 2]. But the proposed model did not allow covariates.
- Here, we propose a model that can accommodate covariates.

Motivating experiment



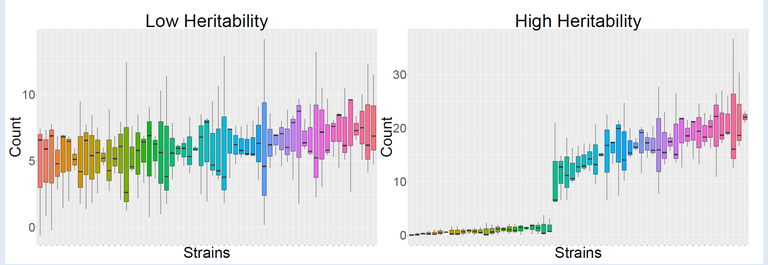
LXS mouse panel

- miRNA and mRNA expression data collected from 59 strains of mice from the LXS mouse panel. [3]
- Observations within a strain are assumed to be genetically identical.
- Mice were randomly assigned to either ethanol group (treatment) or saline group (control).

A statistical framework to estimate h^2

- Use a GLMM to model the data:
  - Random effects for the strains.
  - Data distributed as negative binomial or some other distribution that can account for overdispersion.
  - $g(\mu_{ij}) = \beta_{0i} + b_{si},\ b_{si} \sim N(0, \sigma_b^2)$

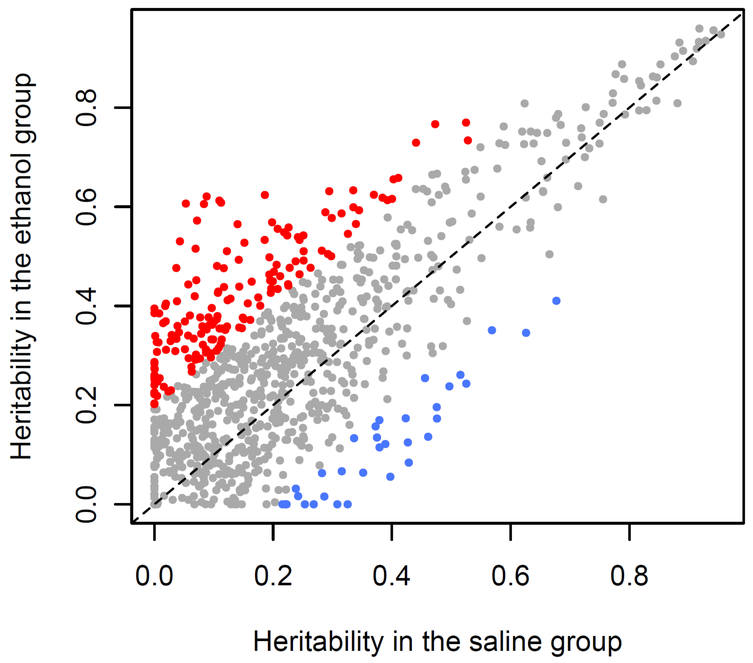
$$V(y_{ij}) = \underbrace{E(V(y_{ij}|strain_{si}))}_{\text{Within Strain}} + \underbrace{V(E(y_{ij}|strain_{si}))}_{\text{Between Strain}}$$



- Use variance partition coefficient (VPC) [4] to estimate heritability.
  - $VPC_i = \dfrac{E\left(V\left(y_{ij}|b_{si}\right)\right)}{E\left(V\left(y_{ij}|b_{si}\right)\right) + V\left(E\left(y_{ij}|b_{si}\right)\right)}$

Presence of covariate

- Treatment is a covariate (categorical) that has not been accounted for by the previous model.
- For categorical covariates, separate heritability estimation can be carried out for each group.



The red and blue points show the genes with difference in $h^2 > 0.2$

Questions:
- Can we gain anything by analyzing the two groups simultaneously by using the group label as a covariate?
- Can we test if the difference in heritability (for a certain gene) is statistically significant across the groups?

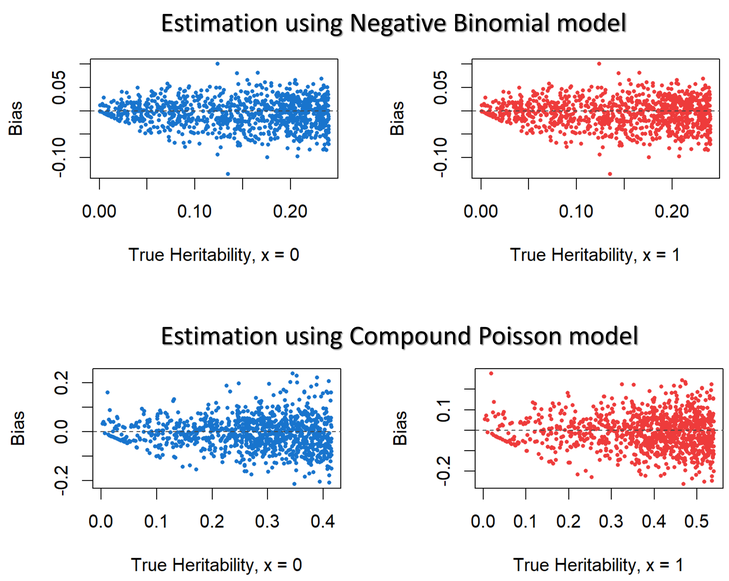
Proposed model

- For a covariate x that we want to include in the model, using the random slope model:
  $$g(\mu_{ij}) = \alpha_i + \beta_i\, x_{ij} + b_{si},\quad b_{si} \sim N(0, \sigma_b^2)$$
- One can also use a random slope model:
  $$g(\mu_{ij}) = \alpha_i + \beta_i\, x_{ij} + b_{0si} + b_{1si},\quad \boldsymbol{b_{si}} \sim N_2(\mathbf{0}, \Sigma_b)$$
- To test for differential heritability,
  $$H_0: \beta_i = 0\ vs\ H_1: \beta_i \neq 0$$
  - Likelihood ratio test has been developed.
- Can be extended for more than one covariate.

Questions:
- Will the variance partition valid after including a stochastic covariate in the model?
  - Yes!
- How to test differential and significant heritability simultaneously?

Simulation Study

- Data for 1000 genes were generated using the negative binomial or compound Poisson model [1].
- 50 strains, with 3 observations within each strain.
- x is a binary covariate.



- Bias is within expected range for the random intercept model.
- Performance is best when the data generation model is same as the model fitted, and when σ_b^2 is not too small.

Future work

- We are currently conducting extensive simulations to answers questions such as the followings.

Questions:
- What is the relative performance of the proposed models?
- What is the effect of different choices and mis-specifications of the covariance matrix Σ_b of the random effects?
- What are the relative advantages of these models compared to a corresponding linear mixed model after transforming the data using a variance stabilizing transformation? [1]

- Analyze read data using the proposed models.
  - Categorical covariates.
  - Continuous covariates.

References

[1] Rudra, P., et al. (2017). Model based heritability scores for high-throughput sequencing data. *BMC bioinformatics.*
[2] R Package HeritSeq, CRAN.
[3] Radcliffe, R. A., et al (2020). Systems genetics analysis of the LXS recombinant inbred mouse strains: Genetic and molecular insights into acute ethanol tolerance. *Plos one.*
[4] Goldstein, H., et al. (2002). Partitioning variation in multilevel models. Underst Stat: Stat Issues Psychol Educ Soc Sci. .

Contact: prudra@okstate.edu