A Fully Bayesian Hierarchical Modeling Strategy for Identifying Gene Expression Heterosis using Parallel Computing with Graphics Processing Units (GPUs)

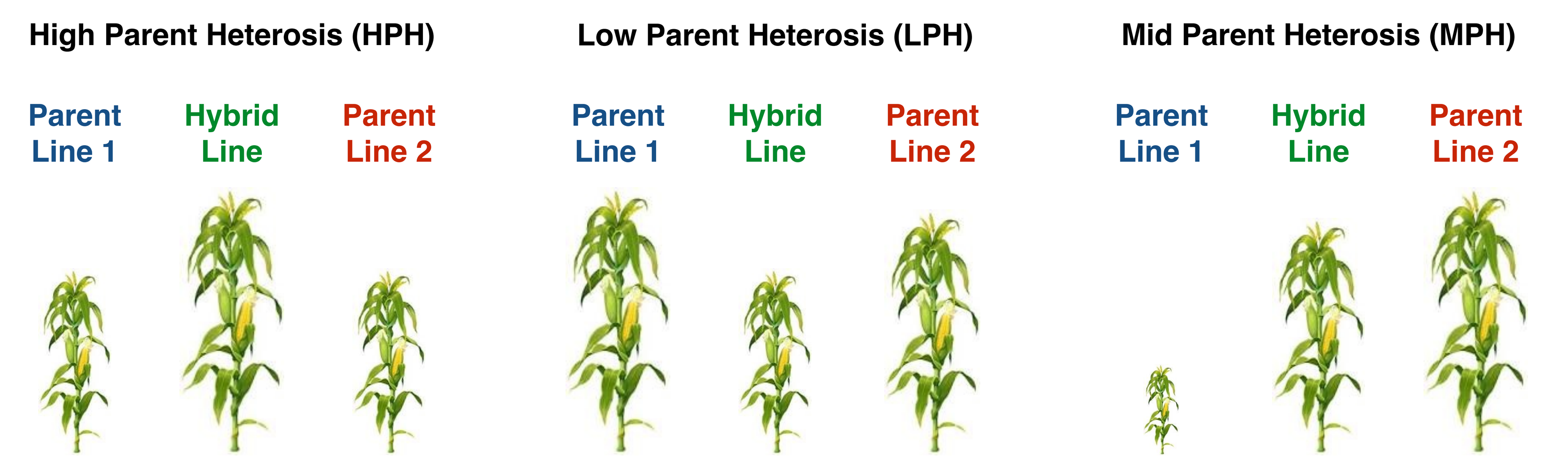
Will Landau and Dr. Jarad Niemi

Iowa State University, Ames, IA, USA

Abstract

Heterosis, or hybrid vigor, occurs when the mean trait value of offspring is more extreme than that of either parent. Well before Darwin first described heterosis in 1876, people used it for practical purposes. Within the last century, heterosis has been used to improve many crop species for food, feed, and fuel industries. Despite intensive study and successful utilization of heterosis, the basic molecular genetic mechanisms responsible for heterosis remain unclear. To learn about these mechanisms, researchers have begun to measure the expression levels of thousands of genes in parental lines and their hybrid offspring. The expression level of each gene can be viewed as a trait alongside more traditional traits like plant height, grain yield, and drought tolerance. This approach presents challenges, such as the simultaneous analysis of tens of thousands of gene expression traits. The main focus of this presentation is a fully Bayesian hierarchical modeling strategy for modeling count-based expression data from next-generation RNA sequencing. Also featured are the high performance computing methods that make this method tractable.

Phenotypic Heterosis



Gene Expression Heterosis in RNA Sequencing Data

		Parent Line 1			Hybrid Line		Parent Line 2	
	Gene 1	100	225	0	70	279	300	106
	Gene 2	0	1	1	50	501	2	1
HPH	Gene 3	3	4	2	700	900	0	0
LPH	Gene 4	893	400	760	5	5	1000	513
	...	...	...	...	...	...	...	...
MPH	Gene 34897	10	13	6	819	761	902	912

Notation

- $y_{g,n}$: observed count (gene g, library n).
- $\eta(g,n)$: expression effect.
- Parameterize $\eta(g,n)$:

$$\eta(g,n) = \begin{cases} \phi_g - \alpha_g & \text{if library } n \text{ came from parent 1} \\ \phi_g + \delta_g & \text{if library } n \text{ came from the hybrid} \\ \phi_g + \alpha_g & \text{if library } n \text{ came from parent 2} \end{cases}$$

- ϕ_g: parental mean expression effect
- α_g: half parental difference
- δ_g: difference between hybrid and parental mean

The Hierarchical Model

$$y_{g,n} \overset{\text{ind}}{\sim} \text{Poisson}(\exp(\rho_n + \epsilon_{g,n} + \eta(g,n)))$$

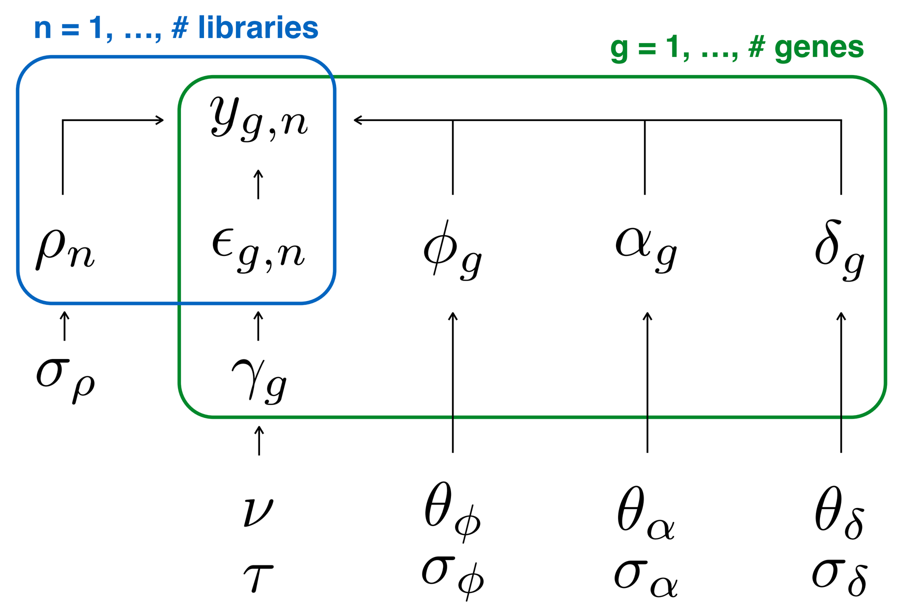
$$\rho_n \overset{\text{ind}}{\sim} \text{N}(0, \sigma_\rho^2)$$
$$\sigma_\rho \sim \text{U}(0, s_\rho)$$

$$\epsilon_{g,n} \overset{\text{ind}}{\sim} \text{N}(0, \gamma_g^2)$$
$$\gamma_g^2 \overset{\text{ind}}{\sim} \text{Inv-Gamma}\left(\text{shape} = \frac{\nu}{2},\ \text{scale} = \frac{\nu\tau^2}{2}\right)$$
$$\nu \sim \text{U}(0, d)$$
$$\tau^2 \sim \text{Gamma}(\text{shape} = a, \text{rate} = b)$$

$$\phi_g \overset{\text{ind}}{\sim} \text{N}(\theta_\phi, \sigma_\phi^2)$$
$$\theta_\phi \sim \text{N}(0, c_\phi^2)$$
$$\sigma_\phi \sim \text{U}(0, s_\phi)$$
$$\alpha_g \overset{\text{ind}}{\sim} \text{N}(\theta_\alpha, \sigma_\alpha^2)$$
$$\theta_\alpha \sim \text{N}(0, c_\alpha^2)$$
$$\sigma_\alpha \sim \text{U}(0, s_\alpha)$$
$$\delta_g \overset{\text{ind}}{\sim} \text{N}(\theta_\delta, \sigma_\delta^2)$$
$$\theta_\delta \sim \text{N}(0, c_\delta^2)$$
$$\sigma_\delta \sim \text{U}(0, s_\delta)$$

- Greek letters are parameters,
- Roman letters are assumed constant.

Directed Acyclic Graph: Opportunities for Parallelism

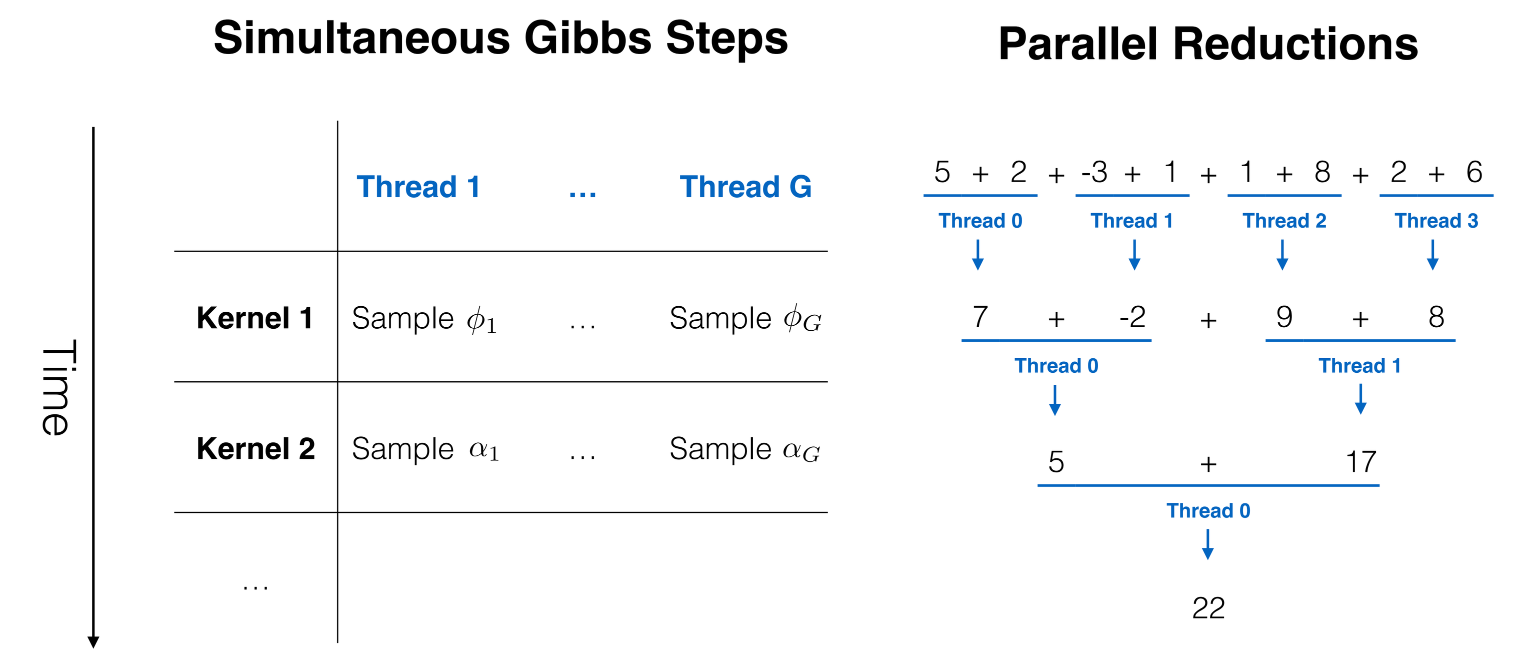


Fitting the Model: slice sampling within Gibbs

- Iteratively sample parameters from their full conditional posterior distributions.

Parameters	Full Conditional Posterior Distributions	Parallel Computing
$\epsilon_{g,n}$, ϕ_g, α_g, δ_g	Approximate (slice sampling)	Simultaneous Gibbs steps
ρ_n, ν	Approximate (slice sampling)	Parallel reductions
θ_ϕ, θ_α, θ_δ	Normal distributions	Parallel reductions
τ^2	Gamma distribution	Parallel reductions
γ_g^2	Inverse gamma distributions	Parallel reductions
σ_ρ^2, σ_ϕ^2, σ_α^2, σ_δ^2	Inverse gamma distributions truncated above	Parallel reductions

Parallel Computing: CUDA Graphics Processing Units



Inference

- Using samples $\phi_g^{(m)}$, $\alpha_g^{(m)}$, and $\delta_g^{(m)}$ $(m = 1, \ldots, M)$ from the appropriate posterior predictive distributions, we can calculate the posterior probabilities that gene g has...

Heterosis	HPH	LPH	MPH
Probability	$\frac{1}{M}\sum_{m=1}^{M} I\left(\delta_g^{(m)} > \left\lvert\alpha_g^{(m)}\right\rvert\right)$	$\frac{1}{M}\sum_{m=1}^{M} I\left(\delta_g^{(m)} < -\left\lvert\alpha_g^{(m)}\right\rvert\right)$	$\frac{1}{M}\sum_{m=1}^{M} I\left(\left\lvert\delta_g^{(m)}\right\rvert > \epsilon\right)$

- $I(\cdot)$ is the indicator function, and $\epsilon > 0$ is an appropriate threshold for mid parent heterosis.

Acknowledgements

This research was supported by National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health and the joint National Science Foundation / NIGMS Mathematical Biology Program under award number R01GM109458. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation. Drs. Dan Nettleton and Peng Liu of the Iowa State University Department of Statistics helped oversee and direct this work.