Week 2 Drosophila Work

Overview

The purpose of my work this week is to build simulated datasets similar to those used by Pallares et al. The purpose of these simulated datasets is to be able to test the power and Type I error of both the model prosposed by Pallares et al. as well as the interaction model proposed by Rebecca / Arbel. Ideally, we aim to show that the interaction model is able to effectively identify types of genetic effects (e.g. amplification) that the origin model is unable to.

Dataset Simulation

As a reference for the dataset simulations, I used the simulation setup that Pallares et al. uses to compare the CMH and beta-binomial approaches for detecting genetic signal. I made some small modifications to this setup (e.g. allowing for different effects in the control and high sugar environments) that I felt would be useful in evaluating the interaction model. Specifically, we assume that n individuals are sampled throughout the experiment ($\frac{1}{2}$ at T_0, $\frac{1}{4}$ in C at T_N, $\frac{1}{4}$ in HS at T_N). For each individual, l loci are sampled. Let π_{ij} be the genotype of individual i at locus j (1 for homozygous alternative, 0.5 for heterozygous, 0 otherwise). Let y_{ij} be the number of alternative allele reads of individual i at locus j, and let r_j be the total number of allele reads at locus j (assume this is the same for all individuals). Finally, let $\pi_{\cdot j}^{(c)}$ be the average alternative allele frequency for all individuals at locus j in condition c, where $c \in \{T_0, C, HS\}$.

Then, we follow the following procedure for generating a simulated dataset:

(1)

(a) Generate $\pi_{\cdot 1}^{(T_0)}, ..., \pi_{\cdot l}^{(T_0)} \sim \hat{F}_{maf}$ independently, where $\hat{F}_{maf}$ is the estimated distribution of minor allele frequencies from the data.

(b) Generate the change in minor allele frequencies for each site j and each condition c, $\Delta_j^{(c)}$ according to a mixture distribution. For more details on the possible parameters of this mixtures please see the code below. But, in summary, this mixture distribution can put probability of 5 different cases. (1) Both conditions null, (2) HS alt, C null, (3) C alt, HS null, (4) HS and C alt, same magnitude, (5) HS and C alt, HS effect is amplified. Once these change values are calculated, use them to calculate the minor allele frequencies for each condition at each site at T_N

(2) For each locus j, generate genotypes for each individual, $\pi_{1j}, ..., \pi_{1n}$ according to the corresponding minor allele frequency for each site and condition.

(3)