LDSP- Linear Detection of Selection in Pooled sequence data

Stephens Lab

1 Introduction

Phase I - better estimate frequency using haplotypic information: The Intuition is that data at each SNP are binomial counts, which help estimate the frequency of a SNP in a pool. But by combining information across multiple corrected SNPs, you can improve the estimated frequency of the test SNP.

We demonstrate the feasibility of such an endeavor by a simple probability calculation:

Assume, the objective is to estimate the frequency of SNP t in the pool of haplotypes. Consider another SNP i (which we denote as S_i). For illustrative purposes, we suppose the two SNPS are perfectly correlated (e.g. $P(S_1 = 1, S_2 = 1) = 1$ & $P(S_1 = 0, S_2 = 0) = 1$)

one estimate: $P(S_t = 1) = \frac{n_1^1}{n_1}$
second estimate: $P(S_t = 1) = P(S_t = 1|S_i = 0)P(S_i = 0) + P(S_t = 1|S_i = 1)P(S_i = 1) = P(S_t = 1|S_i = 1)P(S_i = 1) = P(S_i = 1) = \frac{n_2^1}{n_2}$

where n_j^1 is the number of "1" allele reads at SNP j and n_j^0 is the number of "0" allele reads at SNP j and $n_j = n_j^1 + n_j^0$.

Therefore, effectively we have doubled our coverage. If instead of only one perfectly correlated SNP, we have a 1000 then sequencing only at 1x coverage will be like sequencing at 1000x coverage!

Phase II - detect selection using improved frequency estimate: To detect selection, we find sites that have had significant changes in their frequency compared to the founding population. We can do this by using a linear model which also allows us to model genetic drift with a normal error term.

2 Phase I

2.1 Prior from Li & Stephens

Consider one lineage for now.
Let $y = (y_1, y_2, ..., y_p)'$ denote the vector of allele frequencies in the study sample. Let $E[y_i] = \mu_i$ and the frequency of the test SNP be y_t and M denote the

$2mxp$ panel (i.e. $2m$ haplotypes and p SNPs). As in (Wen & Stephens, 2010), we assume

$$\vec{y}|M \sim N_p(\mu, \Sigma) \tag{1}$$

(Wen & Stephens, 2010) derived the estimates for μ and Σ from the haplotype copying model presented in (Li & Stephens, 2003).

$$\hat{\mu} = (1-\theta)f^{panel} + \frac{\theta}{2}1 \tag{2}$$

$$\hat{\Sigma} = (1-\theta)^2 S + \frac{\theta}{2}(1-\frac{\theta}{2})I \tag{3}$$

and S is obtained from Σ^{panel}, specifically,

$$S_{i,j} = \begin{cases} \Sigma^{panel}_{i,j} & i = j \\ e^{-\frac{-\rho_{i,j}}{2m}}\Sigma^{panel}_{i,j} & i \neq j \end{cases} \tag{4}$$

$\rho_{i,j} = -4Nc_{i,j}d_{i,j}$ where $d_{i,j}$ is the physical distance between markers i and j, N is the effective diploid population size, $c_{i,j}$ is the average rate of crossover per unit physical distance, per meiosis, between sites i and j (so that $c_{i,j}d_{i,j}$ is the genetic distance between sites i and j).
and,

$$\theta = \frac{(\sum_{i=1}^{2m-1}\frac{1}{i})^{-1}}{2m + (\sum_{i=1}^{2m-1}\frac{1}{i})^{-1}} \tag{5}$$

2.2 Data at SNP i

Let (n_i^0, n_i^1) denote the counts of "0" and "1" alleles at SNP i and $n_i = n_i^0 + n_i^1$. Then

$$n_i^1 \sim Bin(n_i, y_i) \dot{\sim} N(n_i y_i, n_i y_i(1-y_i))$$

where y_i is the true population frequency of the SNP i "1" allele.

$$\implies \hat{y_i}|y_i \ \sim \ N(y_i, \frac{y_i(1-y_i)}{n_i}) \tag{6}$$

where $\hat{y_i} = \frac{n_i^1}{n_i}$

Next we replace y_i by $\hat{y_i}$ in the variance for tractibility issues. Therefore,

$$\hat{y_i}|y_i \dot{\sim} N(y_i, \frac{\hat{y_i}(1-\hat{y_i})}{n_i}) \tag{7}$$

PUT HERE THE BETA DISTRIBUTION PART Letting $y_i^{true} = y_i$ and $y_i^{obs} = \hat{y_i}$,

$$\vec{y^{obs}}|\vec{y^{true}} \sim N_p(\vec{y^{true}},\ diag(\epsilon_1, ..., \epsilon_p)) \tag{8}$$

where $\epsilon_i = \frac{y_i^{obs}(1-y_i^{obs})}{n_i}$

2.3 Incorporating Dispersion

In the distribution of $\vec{y}$, we assumed that the panel and study individuals are from the sample population, and the parameters θ and ρ are estimated without error. Deviations from these assumptions will cause over-dispersion: the true allele frequencies will lie further from their expected values than the model predicts. To allow this, we modify equation 1 by introducing an over-dispersion parameter σ^2.

$$y^{\vec{true}}|M \sim N_p(\hat{\mu}, \sigma^2\hat{\Sigma}) \tag{9}$$

We estimate σ^2 by maximizing the multivariate normal likelihood:

$$y^{\vec{obs}}|M \sim N_p(\hat{\mu}, \sigma^2\hat{\Sigma} + diag(\epsilon_1, ..., \epsilon_p)) \tag{10}$$

To obtain the distribution for the true frequencies conditional on the observed data, we use Bayes theorem

$$P(y^{\vec{true}}|y^{\vec{obs}}, M) \propto P(y^{\vec{obs}}|y^{\vec{true}})P(y^{\vec{true}}|M)$$

Let

$$\bar{\Sigma} = \big(\frac{\hat{\Sigma}^{-1}}{\sigma^2} + diag(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p})\big)^{-1} \tag{11}$$

and,

$$\bar{\theta} = \bar{\Sigma}\ (\frac{\hat{\Sigma}^{-1}}{\sigma^2}\hat{\mu} + diag(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p})\ y^{\vec{obs}}) \tag{12}$$

Then since the normal is in the conjugate family,

$$y^{\vec{true}}|y^{\vec{obs}}, M \sim N_p(\bar{\theta}, \bar{\Sigma}) \tag{13}$$

Therefore a natural point estimate for $y^{\vec{true}}$ is $\bar{\theta}$.

2.4 Calculating the inverse of the covariance matrix when it is ill-conditioned

3 Phase II - estimating β

Let $f_{i,k,j}$ denote the frequency of the jth SNP in population i and replicate k. Then,

$$log(\frac{1 - f_{i,k,j}}{f_{i,k,j}}) = \mu_j + \beta_j g_i + \epsilon \tag{14}$$

where $\epsilon \sim N_(0, \sigma_d^2)$, σ_d^2 is the variance due to drift, μ_j is the frequency of the jth SNP in the founding population and

$$g_i = \begin{cases} -1 & i = 0 \\ 0 & i = 1 \\ 1 & i = 2 \end{cases}$$

The intuition here is that sites with large β coefficients are under selection.