

# 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  $y = (y_1, y_2, \dots, y_p)'$  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

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

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

(Wen & Stephens, 2010) derived the estimates for  $\mu$  and  $\Sigma$  from the haplotype copying model presented in (Li & Stephens, 2003).

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

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

and  $S$  is obtained from  $\Sigma^{panel}$ , specifically,

$$S_{i,j} = \begin{cases} \Sigma_{i,j}^{panel} & i = j \\ e^{-\frac{\rho_{i,j}}{2m}} \Sigma_{i,j}^{panel} & i \neq j \end{cases} \quad (4)$$

$\rho_{i,j} = -4Nc_{i,j}d_{i,j}$  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,

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

## 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

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

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

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

where  $\hat{y}_i = \frac{n_i^1}{n_i}$

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

$$\hat{y}_i | y_i \sim N(y_i, \frac{\hat{y}_i(1 - \hat{y}_i)}{n_i}) \quad (7)$$

But, if the coverage is low, then the estimate the frequency of a SNP can be 0 (i.e.  $\frac{n_i^1}{n_i} = 0$ ). We make the following modification which has a Bayesian justification of assuming a prior of  $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ .

$$y_i^{obs} = \frac{n_i^1 + \frac{1}{2}}{n_i + 1} \quad (8)$$

Letting  $y_i^{true} = y_i$ ,

$$y_i^{obs} | y_i^{true} \sim N_p(y_i^{true}, \text{diag}(\epsilon_1, \dots, \epsilon_p)) \quad (9)$$

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  $\theta$  and  $\rho$  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  $\sigma^2$ .

$$y^{true}|M \sim N_p(\hat{\mu}, \sigma^2 \hat{\Sigma}) \quad (10)$$

We estimate  $\sigma^2$  by maximizing the multivariate normal likelihood:

$$y^{obs}|M \sim N_p(\hat{\mu}, \sigma^2 \hat{\Sigma} + \text{diag}(\epsilon_1, \dots, \epsilon_p)) \quad (11)$$

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

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

Let

$$\bar{\Sigma} = \left( \frac{\hat{\Sigma}^{-1}}{\sigma^2} + \text{diag}\left(\frac{1}{\epsilon_1}, \dots, \frac{1}{\epsilon_p}\right) \right)^{-1} \quad (12)$$

and,

$$\bar{\theta} = \bar{\Sigma} \left( \frac{\hat{\Sigma}^{-1}}{\sigma^2} \hat{\mu} + \text{diag}\left(\frac{1}{\epsilon_1}, \dots, \frac{1}{\epsilon_p}\right) y^{obs} \right) \quad (13)$$

Then since the normal is in the conjugate family,

$$y^{true}|y^{obs}, M \sim N_p(\bar{\theta}, \bar{\Sigma}) \quad (14)$$

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

### 2.4 Calculating the inverse of the covariance matrix

When  $\hat{\Sigma}$  is singular, we decompose using SVD and calculate the pseudo-inverse,

$$\hat{\Sigma} = U \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} V^T \quad (15)$$

where the pseudo-inverse is,

$$\hat{\Sigma}^+ = V \begin{bmatrix} S^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T \quad (16)$$

We modify 12 which is now probably non-singular??

$$\bar{\Sigma} = \left( \frac{\hat{\Sigma}^+}{\sigma^2} + \text{diag}\left(\frac{1}{\epsilon_1}, \dots, \frac{1}{\epsilon_p}\right) \right)^{-1} \quad (17)$$

### 3 Phase II - estimating $\beta$

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

$$\log\left(\frac{1 - f_{i,k,j}}{f_{i,k,j}}\right) = \mu_j + \beta_j g_i + \epsilon \quad (18)$$

where  $\epsilon \sim N(0, \sigma_d^2)$ ,  $\sigma_d^2$  is the variance due to drift,  $\mu_j$  is the frequency of the  $j$ th 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  $\beta$  coefficients are under selection.