LD-SEQ: Increasing the effective coverage of SEQuencing experiments using Linkage Dis-equilibrium

Hussein Al-Asadi, Matthew Stephens

1 Introduction

Calculate effective coverage 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 estimate.

2 The Prior

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

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

(Wen & Stephens, 2010) derive estimates for μ and Σ:

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

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

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,

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

and,

$$\Sigma^{panel}_{i,j} = \begin{cases} f^{panel}_i(1-f^{panel}_i) & i = j \\ f^{panel}_{ij} - f^{panel}_i f^{panel}_j & i \neq j \end{cases} \tag{6}$$

where f^{panel}_{ij} is the panel frequency of the haplotype "1-1" consisting of loci i and loci j

3 The likelihood for pooled sequencing experiments

Let (n^0_i, n^1_i) denote the counts of 0 and 1 alleles at SNP i, $n_i = n^0_i + n^1_i$, and y^{true}_i is the population frequency of the SNP i "1" allele.

$$n^1_i|y^{true}_i \sim Bin(n_i, y^{true}_i) \dot{\sim} N(n_i y_i, n_i y^{true}_i(1-y^{true}_i))$$

$$\implies \frac{n^1_i}{n_i}|y^{true}_i \dot{\sim} N(y^{true}_i, \frac{y^{true}_i(1-y^{true}_i)}{n_i}) \tag{7}$$

let $y_i^{obs} = \frac{n_i^1}{n_i}$ and replace y_i^{true} with y_i^{obs} in the variance (a common simplification). Therefore our equation becomes,

$$y_i^{obs}|y_i^{true} \dot{\sim} N(y_i^{true}, \frac{y_i^{obs}(1-y_i^{obs})}{n_i}) \tag{8}$$

$y_1^{obs}|y_1^{true}, y_2^{obs}|y_2^{true}, \ldots, y_p^{obs}|y_p^{true}$ are independent therefore we can write,

$$y^{obs}|y^{true} \dot{\sim} N_p(y^{true},\ diag(\epsilon_1, ..., \epsilon_p)) \tag{9}$$

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

4 Avoiding $y_i^{obs} = 0$

If the coverage is low, then a frequency estimate can be zero (i.e. $\frac{n_i^1}{n_i} = 0$) which will introduce complications when we must invert matrices. Therefore we make the following modification,

$$y_i^{obs} = \frac{n_i^1 + \frac{1}{2}}{n_i + 1} \tag{10}$$

which has nice Bayesian properties.

5 The likelihood for ancient dna sequencing experiments

5.1 The Likelihood under the full model

Instead of sequencing from a pool, we (1) sample an individual from the population and (2) conduct a sequencing experiment on that individual. We repeat this process for all individuals. Let X_k denote the genotype of the kth individual (which is unobserved).

$$X_{i,j} \sim Binomial(2, y_i^{true})$$

$$n_{i,j}^1 | X_{i,j} \sim Bin(n_{i,j}, \frac{X_{i,j}}{2})$$

Integrating out X_k, we see that the likelihood $f(y_i^{true})$ (up to a constant) is,

$$\sum_{j=1}^{k} \log\left((y_i^{true})^2 0^{n_{i,j}^0} + (1-y_i^{true})^2 0^{n_{i,j}^1} + y_i^{true}(1-y_i^{true}) 2^{1-n_{i,j}}\right)$$

We can approximate the likelihood above with a normal likelihood using the laplace approximation (see appendix), yielding

$$P(D|y_i^{true}) \propto N(u_i, \epsilon_i^2)$$

$$u_i = y_i^{true}$$

$$\epsilon_i^2 \approx \frac{-1}{f''(\hat{y}_i)}$$

where $y_i^{obs} = \hat{y}_i$ is the MLE of the likelihood which must be computed numerically. However, $f''(p)$ can be computed analytically as,

$$\sum_{i=1}^{k} \left(\frac{2\ 0^{n_{i,j}^0} + 2\ 0^{n_{i,j}^1} - 2^{2-n_{i,j}}}{p^2 0^{n_{i,j}^0} + (1-p)^2 0^{n_i} + p(1-p)2^{1-n_{i,j}}} - \frac{(2p0^{n_{i,j}^0} - 2(1-p)0^{n_{i,j}^1} + (1-p)2^{1-n_{i,j}} - p2^{1-n_{i,j}})^2}{(p^2 0^{n_{i,j}^0} + (1-p)^2 0^{n_{i,j}^1} + p(1-p)2^{1-n_{i,j}})^2}\right)$$

As before,

$$P(y^{obs}|y^{true}) \dot{\sim} N_p(y^{true},\ diag(\epsilon_1^2, ..., \epsilon_p^2)) \tag{11}$$

5.2 Using an approximate likelihood with similar MSE

Let our estimate of y_i^{true} be,

$$y_i^{obs} = \frac{\sum_{j=1}^{k} n_{i,j}^1}{\sum_{j=1}^{k} n_{i,j}} \tag{12}$$

and let, $\epsilon_i = y_i^{obs}(1 - y_i^{obs})\frac{\sum_{j=1}^{k} n_{i,j} + (n_{i,j})^2}{2(\sum_{j=1}^{k} n_{i,j})^2}$ (which can be drived using using the law of total variance), then by CLT $P(y^{obs}|y^{true}) \dot{\sim} N_p(y^{true},\ diag(\epsilon_1^2, ..., \epsilon_p^2))$ as before.

6 The Posterior

In the distribution of y^{true}, 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.

$$y^{true}|M \sim N_p(\hat{\mu}, \sigma^2\hat{\Sigma}) \tag{13}$$

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

$$y^{obs}|M \sim N_p(\hat{\mu}, \sigma^2\hat{\Sigma} + diag(\epsilon_1, ..., \epsilon_p)) \tag{14}$$

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,

$$\bar{\Sigma} = \big(\frac{\hat{\Sigma}^{-1}}{\sigma^2} + diag(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p})\big)^{-1} \tag{15}$$

and,

$$\bar{\mu} = \bar{\Sigma}\ (\frac{\hat{\Sigma}^{-1}}{\sigma^2}\hat{\mu} + diag(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p})\ y^{obs}) \tag{16}$$

since the normal is in the conjugate family,

$$y^{true}|y^{obs}, M \sim N_p(\bar{\mu}, \bar{\Sigma}) \tag{17}$$

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

Note: we can also calculate effective coverage here (just simply use the reverse mapping of bin to normal approximation)

6.1 Avoiding prior mean bias

Note: for simplicity we now assume the overdispersion parameter $\sigma^2 = 1$

As mentioned above, we assume the the panel and sample individuals are drawn from the same population. This is the never the case in reality but in some applications, the frequencies of alleles of interest have changed significantly but the correlation structure has changed slightly (i.e. very little recombination between nearby SNPs). Therefore we would just like to use the information from SNP correlations. $L(y_i^{true})$ will do the job.

$$L(y_i^{true}) = P(y^{obs}|y_i^{true}, M) \propto \frac{P(y_i^{true}|y^{obs}, M)}{P(y_i^{true}|M)} \tag{18}$$

We showed above,

$$y_i^{true}|y^{obs}, M \sim N(\bar{\mu_i}, \bar{\Sigma}_{ii}) \tag{19}$$

and,

$$y_i^{true}|M \sim N(\hat{\mu_i}, \hat{\Sigma_{ii}}) \tag{20}$$

Thus,

$$P(y^{obs}|y_i^{true}, M) \propto e^{\frac{-(y_i^{true} - \bar{\mu_i})^2}{2\bar{\Sigma}_{ii}} + \frac{(y_i^{true} - \hat{\mu_i})^2}{2\hat{\Sigma}_{ii}}} \tag{21}$$

Completing the square, we can see

$$\frac{-(y_i^{true}-\bar{\mu_i})^2}{2\bar{\Sigma}_{ii}}+\frac{(y_i^{true}-\hat{\mu_i})^2}{2\hat{\Sigma}_{ii}}=\frac{1}{2}(\frac{1}{\hat{\Sigma}_{ii}}-\frac{1}{\bar{\Sigma}_{ii}})(y_i^{true}-\frac{\hat{\mu_i}\bar{\Sigma}_{ii}-\bar{\mu_i}\hat{\Sigma}_{ii}}{\bar{\Sigma}_{ii}-\hat{\Sigma}_{ii}})^2+K \tag{22}$$

Let,

$$\tilde{\mu_i}=\frac{\hat{\mu_i}\bar{\Sigma}_{ii}-\bar{\mu_i}\hat{\Sigma}_{ii}}{\bar{\Sigma}_{ii}-\hat{\Sigma}_{ii}} \tag{23}$$

and,

$$\tilde{\sigma}_i^2=\frac{1}{-\frac{1}{\hat{\Sigma}_{ii}}+\frac{1}{\bar{\Sigma}_{ii}}} \tag{24}$$

$$L(y_i^{true})\propto e^{-\left(y_i^{true}-\tilde{\mu_i}\right)^2\big/2\tilde{\sigma}_i^2} \tag{25}$$

with $\tilde{\mu_i}$ being the MLE of y_i^{true}.

7 Computing the Effective Coverage

To calculate effective coverage (n_e) and the effective proportion (p_e), we approximate the normal likelihood with a binomial likeihood (which can also be justified by using the Laplace approximation (see appendix).

7.0.1 Simply taking the reverse mapping of the well known binomial to normal transformation

$$p_e=\tilde{\mu_i} \tag{26}$$

and,

$$\frac{p_e(1-p_e)}{n_e}=\tilde{\sigma}_i^2 \tag{27}$$

Then,

$$n_e=\frac{\tilde{\mu_i}(1-\tilde{\mu_i})}{\tilde{\sigma}_i^2} \tag{28}$$

8 Appendix

8.1 Using the Taylor expansion

Let,

$$f(p)=logl(p)=log(p^{n_1}(1-p)^{n-n_1}) \tag{29}$$

Taking the taylor expansion of $f(p)$ around its maximum ($\hat{p}$) we get,

$$f(p)\approx f(\hat{p})+\frac{(p-\hat{p})^2}{2}f''(\hat{p}) \tag{30}$$

Therefore,

$$e^{f(p)}=p^{n_1}(1-p)^{n-n_1}\approx Ce^{\frac{-(p-\hat{p})^2}{2\frac{-1}{f''(\hat{p})}}} \tag{31}$$

working back from equation 25 let

$$\tilde{\mu_i}=\hat{p} \tag{32}$$

and,

$$\tilde{\sigma}_i^2=\frac{-1}{f''(\hat{p})}=\frac{(1-\hat{p})\hat{p}^2}{n_1} \tag{33}$$

where $\hat{p}=\frac{n_1}{n}$. Solving the equations for n:

$$n=\frac{\tilde{\mu_i}(1-\tilde{\mu_i})}{\tilde{\sigma}_i^2} \tag{34}$$

which is the same as above!

8.2 Computational Issues

8.2.1 Calculating the inverse of the covariance matrix

$\hat{\Sigma}$ is singular when SNPs are perfectly correlated, to fix this we... (look at notes)

We modify 15 which is now probably non-singular?

$$\bar{\Sigma} = \left(\frac{\hat{\Sigma}^{+}}{\sigma^2} + diag\left(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p}\right)\right)^{-1} \tag{35}$$