Supplementary Materials

emeraLD: Rapid LD Estimation with Massive Data Sets

Corbin Quick

- Here we describe subsampling techniques to approximate linkage disequilibrium (LD) between biallelic variants. We begin with the case where haplotype phase is known (genotypes take values 0 or 1), followed by the case where phase is unknown (genotypes take values 0, 1, or 2).
- We treat the sample correlation $r = (p_{jk} - p_j p_k)/s_j s_k$ as a parameter to be estimated by subsampling. Here, minor allele frequencies p_j and p_k (and standard deviations s_j and s_k) can be calculated efficiently and stored; because p_{jk} must be calculated for each pair of variants, we approximate to increase computational efficiency. For convenience, we treat allele frequencies as known constants.

Informed Subsampling with Phased Genotypes

- Here, we describe a subsampling approach to approximate the sample correlation $r = (p_{jk} - p_j p_k)/s_j s_k$ using phased genotypes. Consider the estimator $\tilde{r}(\ell, \Delta) = [\tilde{p}_{jk}(\ell, \Delta) - p_j p_k]/s_j s_k$, where

$$\tilde{p}_{jk}(\ell, \Delta) = \begin{cases} \frac{p_j}{\ell} \sum_{i=1}^{\ell} \tilde{G}_{ik}^{(j)} & \Delta = 1 \\ \frac{p_k}{\ell} \sum_{i=1}^{\ell} \tilde{G}_{ij}^{(k)} & \Delta = 0 \end{cases}$$

- for $\Delta \in \{0, 1\}$ and where each $\tilde{G}_{ik}^{(j)}$ (or $\tilde{G}_{ij}^{(k)}$) is independently sampled from the subset of haplotypes with $G_{ij} = 1$ (or $G_{ik} = 1$).
- Clearly $\tilde{r}(\ell, \Delta)$ is an unbiased estimator for r, and has empirical variance

$$\text{var}_n[\tilde{r}(\ell, \Delta)] = \frac{p_{jk}}{\ell s_j^2 s_k^2} \left[\Delta p_j^2 (p_j - p_{jk}) + (1 - \Delta) p_k^2 (p_k - p_{jk}) \right].$$

- Therefore, given that we sample ℓ minor allele carriers of either variant j or variant k, the optimal estimator $\tilde{r}_\ell$ is given by taking $\Delta = I(p_j \leq p_k)$. Intuitively, carriers of the rarer allele are more informative for estimating the size of the intersection.
- Letting ρ denote the true LD value in the population, the MSE of the approximate estimator is

$$\text{MSE}(\tilde{r}_\ell) \coloneqq \mathbb{E}[(\tilde{r}_\ell - \rho)^2] = \mathbb{E}[(r - \rho)^2] + \mathbb{E}[(\tilde{r}_\ell - r)^2],$$

- so for $p_j \leq p_k$ (WLOG) we have $\text{MSE}(\tilde{r}_\ell) - \text{MSE}(r)$= $(p_j - p_{jk}) p_{jk} / \ell s_j^2 s_k^2$.
- The variance of the estimator is maximized with respect to p_{jk} when $p_{jk} = p_j/2$, and maximized with respect to p_j when $p_j = 1/2$ (because $1/2 \geq s_k \geq s_j \geq p_j$). It follows that $\text{MSE}(\tilde{r}_\ell) - \text{MSE}(r) \leq 1/\ell$.

Informed Subsampling with Unphased Genotypes

- Here, we describe a subsampling approach to approximate the sample correlation $r = c_{jk}/s_j s_k$ using unphased genotypes. We define the sample covariance between variants j and k as $c_{jk} = \frac{1}{n} \sum_{i=1}^{n} G_{ij} G_{ik} - 4 p_j p_k$, and we can write

$$\frac{1}{n} \sum_{i=1}^{n} G_{ij} G_{ik} = p_{k,1} \hat{\mathbb{E}}(G_j | G_k = 1) + 2 p_{k,2} \hat{\mathbb{E}}(G_j | G_k = 2)$$

- where $p_{k,m}$ is the proportion of individuals with genotype m at variant k, and $\hat{\mathbb{E}}(G_j | G_k = m)$ is the mean genotype at variant j among individuals with genotype m at variant k in the overall sample of n individuals.

- Define the approximate estimator

$$\tilde{c}_{jk}(\ell_1, \ell_2) = p_{k,1}\tilde{\mathbb{E}}_{\ell_1}(G_j|G_k = 1) + 2p_{k,2}\tilde{\mathbb{E}}_{\ell_2}(G_j|G_k = 2) - 4p_jp_k,$$

- where $\tilde{\mathbb{E}}_\ell(G_j|G_k = m)$ is estimated by sampling ℓ genotypes from individuals with genotype m at variant k. The approximate estimator is unbiased and has empirical variance

$$\text{var}_n[\tilde{c}_{jk}(\ell_1, \ell_2)] = \frac{p_{k,1}^2}{\ell_1}\text{var}_n(G_j|G_k = 1) + \frac{4p_{k,2}^2}{\ell_2}\text{var}_n(G_j|G_k = 2).$$

- Supposing that variants j and k are independent (which maximizes the variability of the estimator),

$$\text{var}_n[\tilde{c}_{jk}(\ell_1, \ell_2)] = \left(\frac{p_{k,1}^2}{\ell_1} + \frac{4p_{k,2}^2}{\ell_2}\right)s_j^2,$$

- which is minimized by choosing $\ell_1 : \ell_2$ in proportion to $p_{k,1} : 2p_{k,2}$, or in other words oversampling homozygotes by a factor of 2.

- We can now define the Minimax optimal approximate estimator $\tilde{c}_{jk}^\ell = \tilde{c}_{jk}(\ell_1^*, \ell_2^*)$, where

$$\ell_1^* = \frac{2p_{k,2}}{2p_{k,2} + p_{k,1}}\ell \quad \text{and} \quad \ell_2^* = \frac{p_{k,1}}{2p_{k,2} + p_{k,1}}\ell.$$

- Therefore, the optimal approximate estimator has $\text{var}_n(\tilde{c}_{jk}^\ell) \leq 4p_k^2s_j^2/\ell$ (note that $2p_k = p_{k,1} + 2p_{k,2}$), and letting $\tilde{r}_\ell = \tilde{c}_{jk}^\ell/s_js_k$, we have

$$\text{MSE}(\tilde{r}_\ell) - \text{MSE}(r) = \text{var}_n(\tilde{r}_\ell) \leq \frac{4p_k^2}{\ell s_k^2} \leq \frac{2}{\ell}.$$

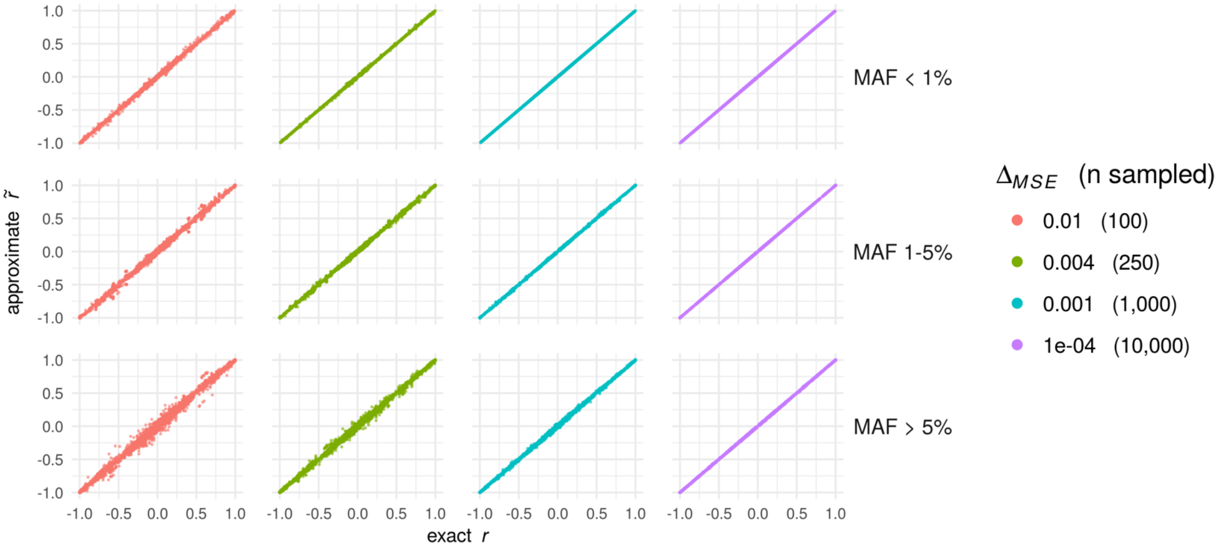
- Here, we have not assumed Hardy-Weinberg Equilibrium (HWE) for either variant. Supposing that both variants are in HWE, we can write $\mathbb{E}(G_jG_k) = 2p_{jk}(1 + p_j + p_k - p_{jk}) + 2(p_k - p_{jk})(p_j - p_{jk})$, and because p_{jk} is the only unknown parameter, the most efficient subsampling estimator would use as many minor-allele homozygotes as possible before sampling any heterozygotes. We avoid this assumption to ensure that estimates are robust.

Time Complexity of Approximation by Informed Subsampling

- By subsampling ℓ individuals or haplotypes whenever $\min(MAC_j, MAC_k) > \ell$, we are guaranteed at most ℓ operations for each pair of variants.

- For computational efficiency, we sample subsets of minor-allele carriers once for each variant as genotype data are processed.

Supplementary Figure 1: Approximate vs. Exact LD Estimates



- Here, we show approximate vs. exact LD estimates from the Haplotype Reference Consortium. The number of minor-allele carriers sampled ℓ is equal to $1/\Delta_{MSE}$, where Δ_{MSE} is the maximum MSE induced by approximation