1 LDA and LDA score

In population genetics, linkage disequilibrium (LD) is defined as the non-random association of alleles at different loci in a given population [1]. With the popular "chromosome painting" technique [2], we can obtain the ancestry probabilities of each individual as each locus. In this section, we propose an ancestry linkage disequilibrium (LDA) approach to measure the association of alleles from painting data.

1.1 Definition

Let $A(i,j,k)$ denote the probability of the kth ancestry ($k = 1,\ldots,K$) at the jth SNP ($j = 1,\ldots,J$) of a chromosome for the ith individual ($i = 1,\ldots,N$).

We define the distance between SNP l and m as the average L_2 norm between ancestries at those SNPs. Specifically we compute the L_2 norm for the ith genome as

$$D_i(l,m) = \|A(i,l,\cdot) - A(i,m,\cdot)\|_2 = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(A(i,l,k) - A(i,m,k))^2}. \tag{1}$$

Then we calculate the distance between SNP l and m by averaging $D_i(l,m)$ ($i = 1,\ldots,N$).

$$D(l,m) = \frac{1}{N}\sum_{i=1}^{N} D_i(l,m). \tag{2}$$

We define $D^\star(l,m)$ as the theoretical distance between SNP l and m if there were no linkage disequilibrium of ancestry (LDA) between them. $D^\star(l,m)$ is estimated by

$$D^\star(l,m) \approx \frac{1}{N}\sum_{i=1}^{N}\|A(i^\star,l,\cdot) - A(i,m,\cdot)\|_2, \tag{3}$$

where $i^\star \in (1,\ldots,N)$ are re-sampled without replacement at SNP l. Using the empirical distribution of ancestry probabilities accounts for variability in both the average ancestry and its distribution across SNPs. Ancestry assignment can be very precise in regions of the genome where our reference panel matches our data, and uncertain in others where we only have distant relatives of the underlying populations.

The LDA between SNP l and m is a similarity, defined in terms of the negative distance $-D(l,m)$ normalised by the expected value $D^\star(l,m)$ under no LD, as:

$$LDA(l,m) = \frac{D^\star(l,m) - D(l,m)}{D^\star(l,m)}. \tag{4}$$

LDA therefore takes an expected value 0 when haplotypes are randomly assigned at different SNPs, and positive values when the ancestries of haplotypes are correlated.

LDA is a pairwise quantity. To arrive at a per-SNP property, we define the LDA score (LDAS) of SNP j as the total LDA of this SNP with the rest of the genome, i.e. the integral of the LDA for that SNP. Because this quantity decreases to zero as we move away from the target SNP, this is in practice computed within an XcM-window (we use $X = 5$ as LDA is approximately zero outside this region in our data) on both sides of the SNP. Note that we measure this quantity in terms of the genetic distance, and therefore LDAS is measuring the length of Ancestry-specific haplotypes compared to individual-level recombination rates.

As a technical note, when the SNPs approach either end of the chromosome, they no longer have a complete window, which results in a smaller LDAS. This would be appropriate for measuring total ancestry correlations, but to make LDAS useful for detecting anomalous SNPs, we use the $LDAS$ of the symmetric side of the SNP to estimate the LDAS within the non-existent window.

$$LDAS(j;X) = \begin{cases} \int_{gd(j)-X}^{gd(j)+X} LDA(j,l)\,dgd & \text{if } X \le gd(j) \le tg - X, \\ \int_{0}^{gd(j)+X} LDA(j,l)\,dgd + \int_{2gd(j)}^{gd(j)+X} LDA(j,l)\,dgd & \text{if } gd(j) < X, \\ \int_{gd(j)-X}^{tg} LDA(j,l)\,dgd + \int_{gd(j)-X}^{2gd(j)-tg} LDA(j,l)\,dgd & \text{if } gd(j) > tg - X. \end{cases} \tag{5}$$

where $gd(l)$ is the genetic distance (i.e. position in cM) of SNP l, and tg is the total genetic distance of a chromosome. In computation, we also assume the LDA on either end of the chromosome equals to the LDA of the SNP closest to the end: $LDA(j, gd = 0) = LDA(j, l_{mingd})$ and $LDA(j, gd = td) = LDA(j, l_{maxgd})$, where gd is the genetic distance, l_{mingd} and l_{maxgd} are the index of the SNP with the smallest and largest genetic distance, respectively.

The integral $\int_{gd(j)-X}^{gd(j)+X} LDA(j,l)\,dgd$ is computed assuming linear interpolation of the LDA score between adjacent SNPs.

1.2 Quality control

One major source of bias in the estimate of LDAS is due to sparse SNP sampling, as the LDA score is calculated by summing the space under piecewise linear functions. To handle this without making further distributional assumptions, we propose a quality control method.

An upper bound and lower bound of the estimates of LDAS can be obtained by replacing the linear interpolation in Eq.5 with a step function. In detail, we take the larger and smaller LDA value of two neighbouring SNPs, respectively, as the fixed LDA in the genetic distance between the two SNPs in the integral over the XcM-window on both sides of the SNP. Specifically:

$$LDA_{upper}(j,l) = \max\{LDA(j,l), LDA(j,l+1)\}$$

and

$$LDA_{lower}(j,l) = \min\{LDA(j,l), LDA(j,l+1)\},$$

which can be substituted into Eq.5 to obtain an upper and lower-bound respectively of the LDAS of SNP j: $LDAS_{upper}(j;X)$ and $LDAS_{lower}(j;X)$. When computing $LDAS_{lower}(j;X)$, we assume $LDA(j, gd = 0) = LDA(j, gd = td) = 0$ for conservative estimation.

Intuitively, the maximum possible error of LDAS of SNP j is

$$LDAS_{error}(j;X) = LDAS_{upper}(j;X) - LDAS_{lower}(j;X). \tag{6}$$

Hence, we could remove SNPs with $LDAS_{error} \geq \delta$ where δ is a specified threshold (we used $\delta = 0.5$).

However, the above method cannot totally address the sparsity of the painting data. In practice, the pairwise LDA shrinks to almost 0 when the closest SNPs are more than 2cM away. We therefore remove SNPs in very sparse regions based on their 2cM windows: SNP j is removed if at least one of $n_m(j) < \theta$ for $m = 0.5, 1, 1.5, 2$, where $n_m(j)$ is the number of SNPs that is $(m - 0.5, m]$cM away from SNP j and θ is a specified threshold (we used $\theta = 10$).

In conclusion, we suggest the hybrid of $LDAS_{error} < \delta$ and $n_m \geq \theta$(m=0, 0.5, 1.0, 1.5) as the quality control of SNPs, which alleviates the bias estimates due to sparsity of the painting data and therefore avoids extreme LDA scores.

References

[1] Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

[2] Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS genetics*, 8(1):e1002453, 2012.