

1 LDA and LDA score

1.1 Definition

Let $A(i, j, k)$ denote the probability of the k th ancestry ($k = 1, \dots, K$) at the j th SNP ($j = 1, \dots, J$) of a chromosome for the i th haplotype ($i = 1, \dots, N$).

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 i th genome as

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

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

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

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

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

where $i^* \in (1, \dots, N)$ 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 $D^*(l, m)$ under no LD, as:

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

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 score for that SNP. Because this quantity decreases to zero as we move away from the target SNP, this is in practice computed within an X cM-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.

$$LDAS(j; X) = \begin{cases} \int_{gd(j)-X}^{gd(j)+X} LDA(j, l) dgd & \text{if } X \leq gd(j) \leq tg - X, \\ \int_0^{gd(j)+X} LDA(j, l) dgd + \int_{2gd(j)}^{gd(j)+X} LDA(j, l) dgd & \text{if } gd(j) < X, \\ \int_{gd(j)-X}^{tg} LDA(j, l) dgd + \int_{gd(j)-X}^{2gd(j)-tg} LDA(j, l) dgd & \text{if } gd(j) > tg - X. \end{cases} \quad (5)$$

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 X cM-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

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

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.

2 Simulation

In the evolution, different alleles may be selected by different ancestries to survive diseases. When these ancestries admix, a combination of increased population density and exposure to various diseases lead to strong balancing selection for these haplotypes. In this section, we investigate how balancing haplotype selection affects the LDA score.

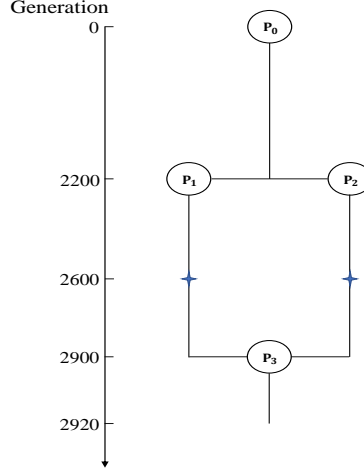


Figure 1: The evolutionary pattern of populations in the simulation.

Fig.1 shows the evolutionary pattern of populations in the simulation. An ancient population P_0 evolves for 2200 generations before splitting into two sub-populations P_1 and P_2 . After evolving 400 generations, we added a mutation at a different locus in P_1 and P_2 . Both added mutations are then positively selected in the following 300 generations, after which they merge to P_3 . In P_3 , both added mutations experience strong positive selection for 20 generations. Finally, we sample 1000 individuals from P_3 to compute their ancestry proportions of P_1 and P_2 using the "chromosome painting technique, and calculate the LDA score of the simulated chromosome positions.

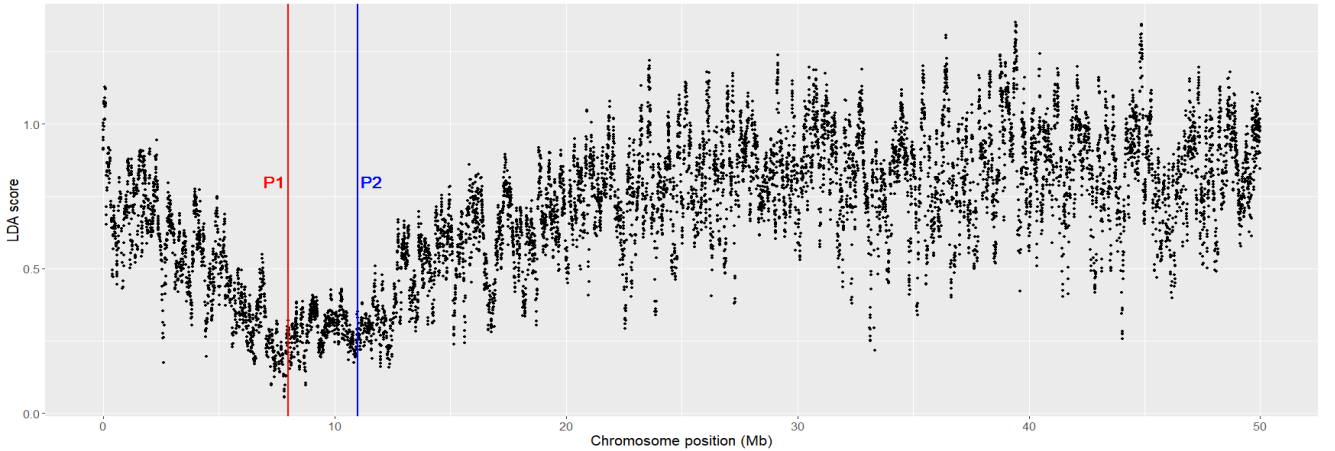


Figure 2: LDA score plot of the simulated chromosome positions. The red and blue lines are the positions of the selected allele in P_1 and P_2 , respectively.

Fig.2 is the LDA score of the simulated chromosome positions under a single large sample. The selected allele in P_1 is at 8Mb (red line) while that in P_2 is at 11Mb (blue line). Clearly the LDA score is much lower around the selected regions compared with the unselected regions.

For the rigor of the simulation, we simulated 100 times with a smaller sample. In detail, we reduced the mutation rate and the total genome distance, i.e. 20Mb total genome distance with selected mutations added in 5Mb and 8Mb for P_1 and P_2 respectively. Then We compared the average LDA score in the selected (defined from 3Mb to 10Mb) and unselected regions over 100 simulations. Furthermore, the impact of evolving generations in P_3 on LDA score is also investigated in the simulation.

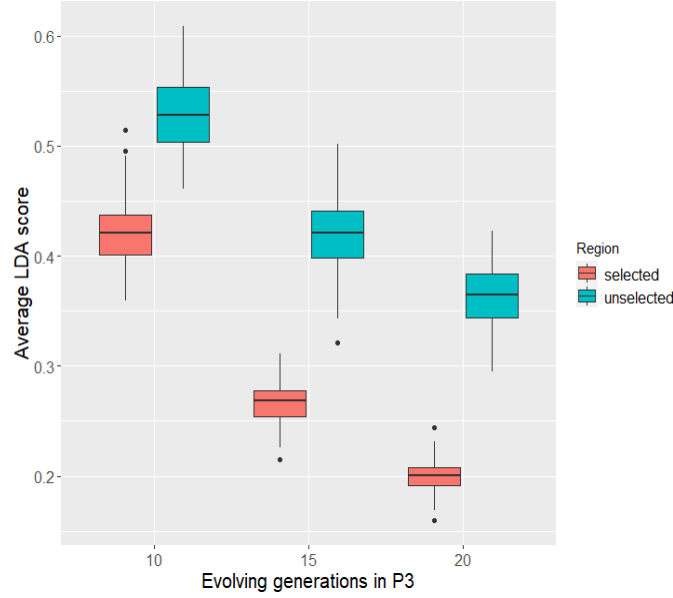


Figure 3: The average LDA score of the selected and unselected regions with different number of evolving generations in P_3 over 100 simulations.

The simulation results in Fig.3 show an evident reduction of LDA score in the selected regions. Furthermore, with the increment of generations in P_3 , more recombination events break haplotypes, which creates many shorter LDA blocks and thus reduces the LDA scores.

In conclusion, we demonstrate through simulation studies that balancing haplotype selection and longer generations in mixed populations are associated with low LDA scores.