

Evaluation of ASAP performance in relation to the number of PCs

The number of principal components (PCs) might affect PANE accuracy. To explore this, we compared the error rates of our tool when using different numbers of principal components.

For such a task, we used the simulated set with no missing data. Briefly, we simulated 20 unadmixed and 16 admixed populations (Supp. Table 1), with admixture events involving two or three source populations with split times spanning from 75 KYA to 9 KYA. The contribution from minor sources ranged between 5% and 40%, providing a variety of scenarios to evaluate the performance of ASAP in different contexts. These simulations also covered a wide range of genetic relationships, with pairwise F_{st} values from 0.01 to 0.23, including bottleneck scenarios.

We focused on six populations, covering a wide range of separation time between the two sources that contributed to their admixture event. Then, we calculated the variance explained by each PC, using N PCs, with $N = \{2, 4, 6 \dots 20\}$. For each run of PANE, we calculated the error as the absolute value of the difference between the estimated and observed minor ancestry component. Therefore, values close to 0 are high accuracy indicators, and negative and positive values represent an underestimation or overestimation, respectively.

In detail, we observed the lowest error for N larger than 8 when the true sources were used as reference groups (Fig. ST1A). However, when sister groups of true sources were used ('proxy source'), the accuracy decreased for N larger than 10, with the highest error observed for Pop 2 and 14 (Fig. ST1B).

Given this pattern, we opted to run PANE with 10 PCs to balance minimizing error and maintaining accuracy, ensuring reliable inference across the different population scenarios.

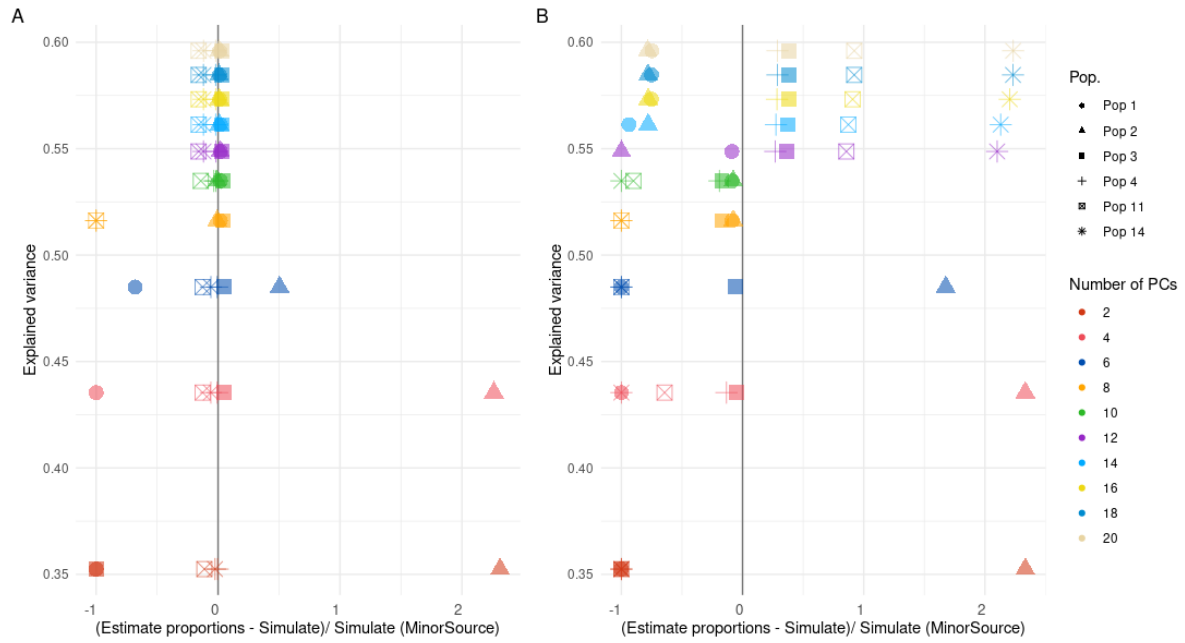


Fig. ST1: Comparison of the PANE performance using different numbers of PCs. Each symbol represents a different population. The y-axis shows the variance explained based on the number of PCs considered (indicated by the different colors). On the x-axis, it is shown how much the result obtained for the minor source differs from the expected one, also in relation to the weight of the contribution itself.

We applied PANE on the average PCs of each population, using the true sources as sources in A) and their sister groups ('proxy sources') in B).