Joint inference of admixture and population history from the genetic covariance structure

Amika Sood[1,2] and **Alejandro Ochoa**[1,2]

[1]Duke Center for Statistical Genetics and Genomics, and [2]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA

Abstract

Inference of genetic admixture, or ancestry, is a fundamental analysis that could be greatly enhanced if estimates of the ancestral population history were more readily available. We focus on estimating the ancestral covariance structure, which reveals the differentiation of these ancestral populations and their pairwise relationships. Our previous work revealed a connection between these admixture parameters and the genetic covariance structure of the admixed individuals, given by straightforward linear algebra. In this work, we develop an innovative fast and accurate admixture inference algorithm based on fitting this linear algebra model. Our approach is made uniquely possible by estimating the covariance structure with popkin, which we recently developed and results in practically-unbiased estimates. We compare to several leading approaches, spanning a variety of likelihood and likelihood-free approaches, but which all have to estimate an enormous number of allele frequencies. In contrast, our approach marginalizes loci when estimating the covariance structure, resulting not only in less (but higher-quality) data to fit but also many fewer parameters to fit. This property of our model suggests that it may be more robust, since it has fewer degrees of freedom. Additionally, runtime is practically independent of the number of loci, so it is much faster than previous approaches when there are millions of loci. Our constrained non-convex optimization problem is tackled using a Memetic algorithm, which combines the Genetic algorithm with local optimization. We also consider estimating the ancestral covariance structure using the outputs of previous approaches. Our simulations show that our joint estimation approach results in more accurate estimates of the ancestral covariance matrix than estimates calculated from previous approaches, especially for larger numbers of ancestral populations. Our simulations also reveal the problem of non-unique solutions and we discuss our novel approach to regularize the problem. Lastly, we apply our estimator to several real human datasets and compare to estimates from leading approaches. In conclusion, we present a novel approach for estimating admixture that is fundamentally different from existing approaches, which shows promise in terms of its robustness and speed.

Unbiased kinship estimator: `popkin`

Genetic model. $x_{ij} \in \{0,1,2\}$: genotype of ind. j, biallelic SNP i, counting ref alelles. p_i: ancestral allele frequency. φ_{jk}: kinship coefficient. Genotype moments:

$$\mathrm{E}[x_{ij}] = 2p_i, \qquad \mathrm{Cov}(x_{ij}, x_{ik}) = 4p_i(1-p_i)\varphi_{jk}.$$

`popkin` is the only practically unbiased estimator for arbitrary population structures [1]. This application uses unsupervised version (no subpopulation labels):

$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \qquad \hat{A}_{\text{min}} = \min_{jk} A_{jk},$$

$$\hat{\varphi}_{jk}^{\text{new}} = 1 - \frac{A_{jk}}{\hat{A}_{\text{min}}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}.$$

Kinship to Coancestry conversion (model assumes no family structure [1, 2]):

$$\theta_{jk} = \begin{cases} 2\varphi_{jj} - 1 & \text{if} \quad j = k, \\ \varphi_{jk} & \text{if} \quad j \neq k. \end{cases}$$

Admixture: classic and covariance models

Standard admixture model [3] is a Bernoulli mixture model:

$$\mathrm{E}[\mathbf{X}/2] = \mathbf{\Pi} = \mathbf{P}\mathbf{Q}^\intercal.$$

Covariance of model (standard linear algebra):

$$\mathrm{Cov}(\mathbf{\Pi}) = \mathrm{Cov}(\mathbf{P}\mathbf{Q}^\intercal) = \mathbf{Q}\,\mathrm{Cov}(\mathbf{P})\mathbf{Q}^\intercal.$$

Parametrization from genetic model [4]:

$$\mathrm{Cov}(\mathbf{\Pi}) = \mathbf{V} \otimes \mathbf{\Theta}, \qquad \mathrm{Cov}(\mathbf{P}) = \mathbf{V} \otimes \mathbf{\Psi}.$$

Our final covariance model ($\mathbf{V}$ cancels out) [1, 2]:

$$\mathbf{\Theta} = \mathbf{Q}\mathbf{\Psi}\mathbf{Q}^\intercal.$$

Example simulation, input and output are shown in Fig. 1.

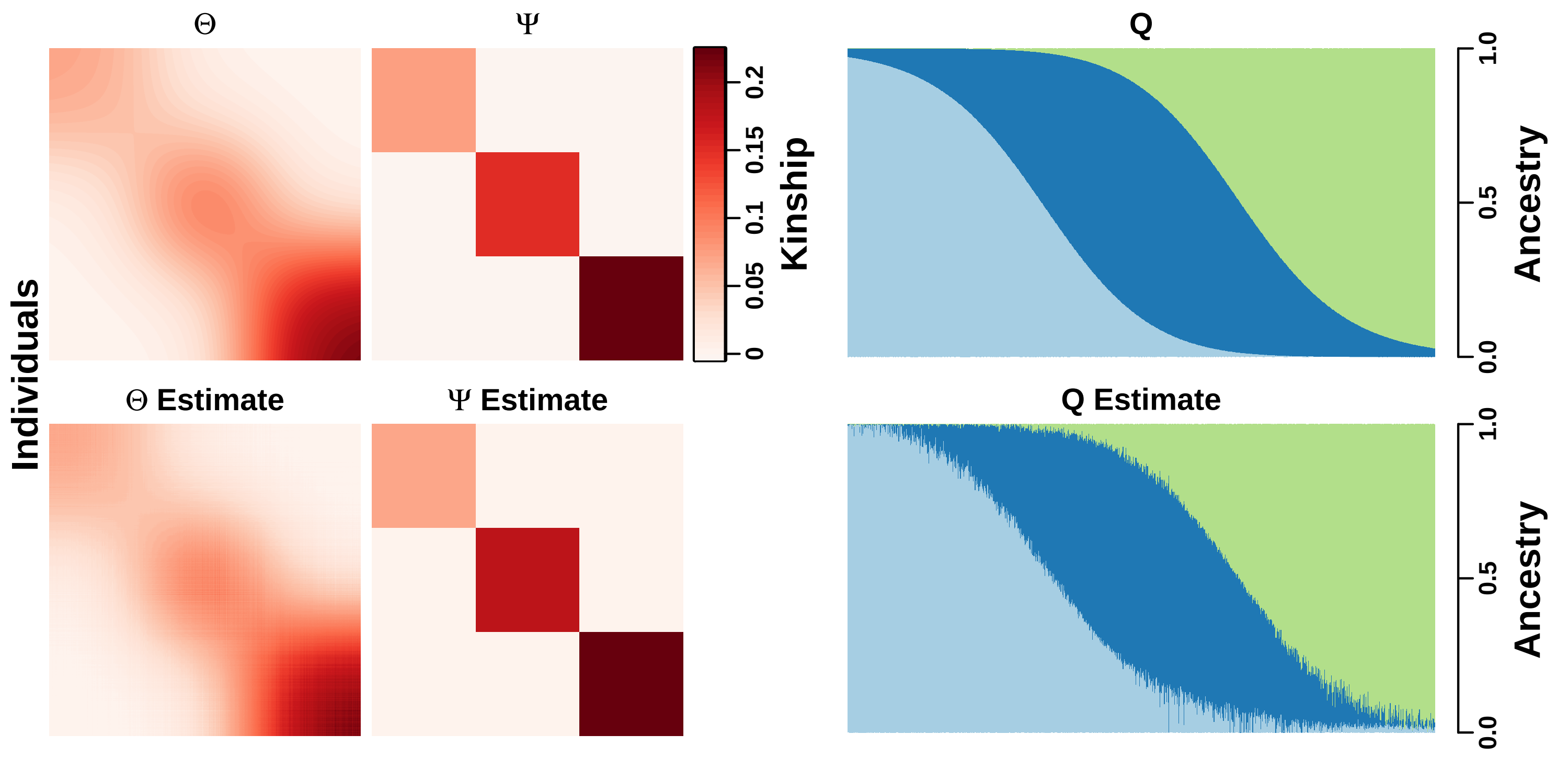

Figure 1:**Simulated data and our model fit.** $\mathbf{\Theta}, \mathbf{\Psi}, \mathbf{Q}$ are true parameters of simulation. $\mathbf{\Theta}$ estimate is obtained by `popkin`. $\mathbf{\Psi}, \mathbf{Q}$ estimates are from our Memetic algorithm infered solely from the $\mathbf{\Theta}$ estimate from `popkin`.

Our estimation approach

Objective function:

$$F = \left\|\hat{\mathbf{\Theta}} - \mathbf{Q}\mathbf{\Psi}\mathbf{Q}^\intercal\right\|^2 + \gamma\,\mathrm{tr}(\mathbf{\Psi}).$$

Penalize $\mathrm{tr}(\mathbf{\Psi})$ (minimize ancestral F_{ST}) to identify solution: Otherwise $\mathbf{Q}$ and $\mathbf{\Psi}$ are not uniquely defined, since objective depends on them solely through their product [5].

Constraints:

- All $\mathbf{Q}, \mathbf{\Psi}$ elements between 0 and 1.
- $\mathbf{Q}$ rows sum to one.

Non-convex problem: Gradient descent was not producing good solutions: suspected local minima or saddle points. The memetic algorithm overcomes these problems (Fig. 2).

DrAlexOchoa
ochoalab.github.io
alejandro.ochoa@duke.edu

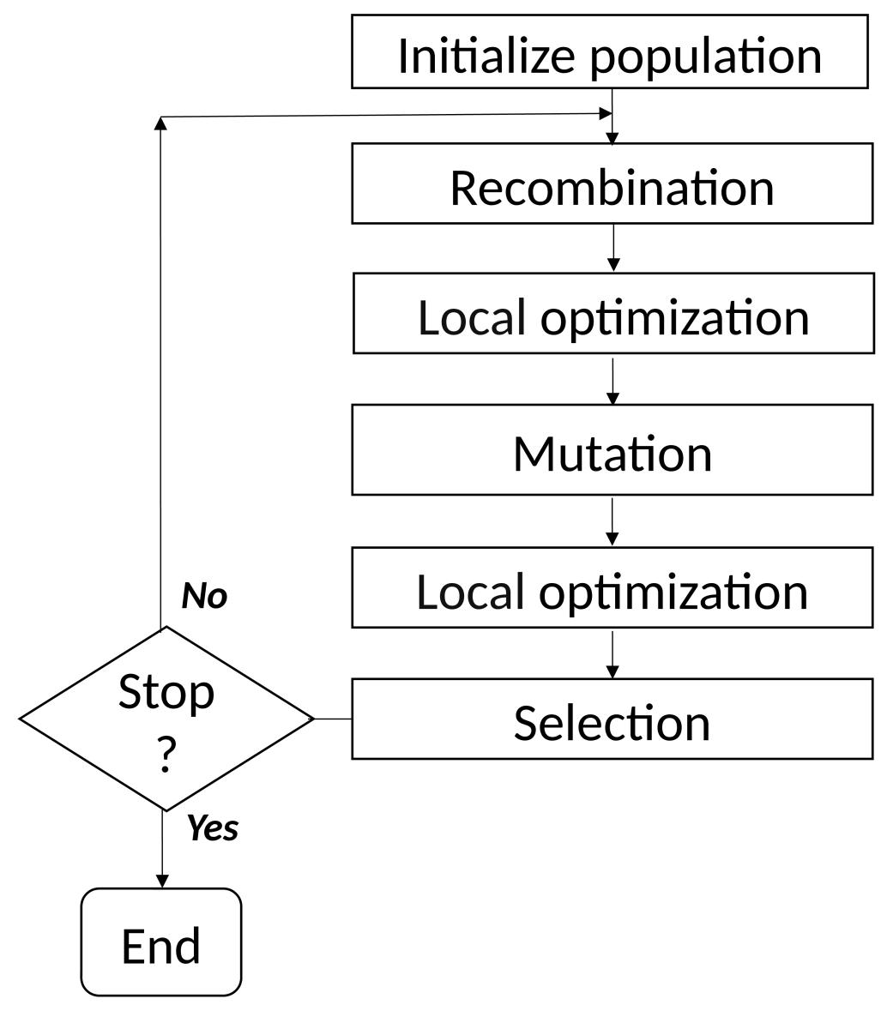
Initialize population
Recombination
Local optimization
Mutation
Local optimization
Selection
Stop ?
No
Yes
End

Figure 2:**Flowchart of proposed Memetic algorithm.** Memetic algorithms allow for efficient optimization of non-convex problems such as the present problem. Memetic combines the Genetic algorithm (Recombination, Mutation, and Selection) with local optimization steps.

Results and Conclusions

We compared various existing approaches in estimating $\mathbf{Q}$ and $\mathbf{\Psi}$ (latter from a popkin variant applied to allele frequency estimates $\mathbf{P}$ estimated by these existing approaches), and found our current estimator to be among the best (Fig. 3A-B). Our model quickly averages over loci with `popkin`, which reduces the overall runtime on very large datasets compared to existing approaches (Fig. 3C).

Overall, our new approach to admixture inference is accurate, provides ancestral population estimation (via $\mathbf{\Psi}$) and scales better for large datasets.

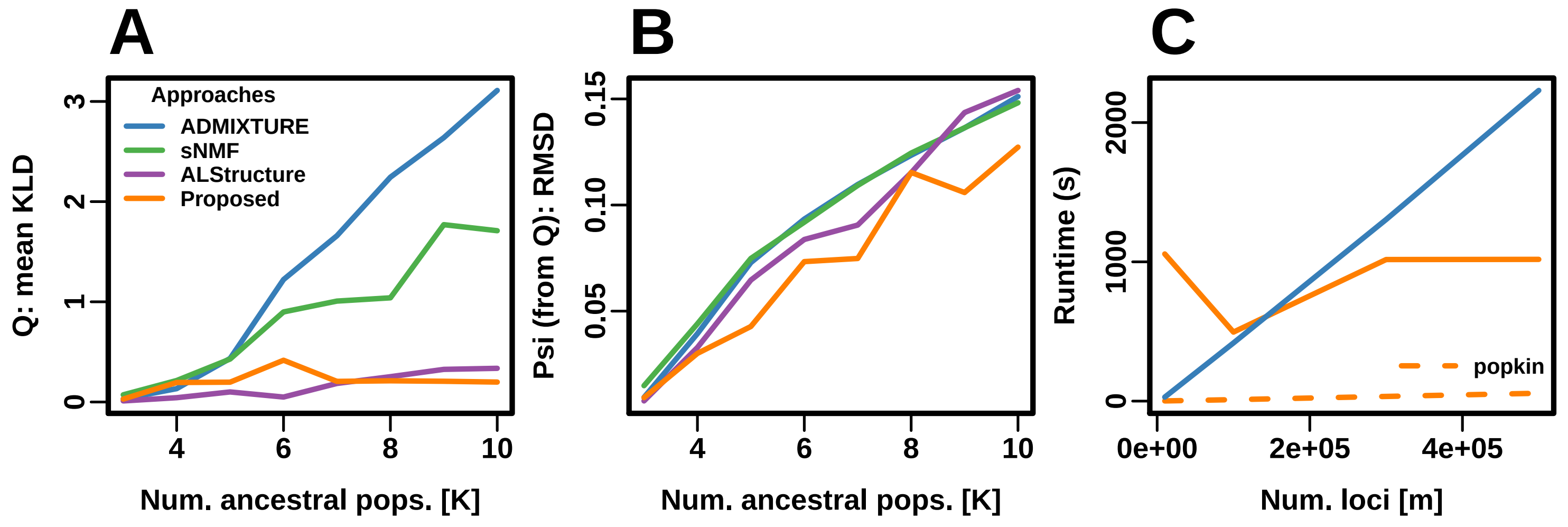

Figure 3:**Accuracy of Q, P and proposed Ψ estimates.** **A.** Error of $\mathbf{Q}$ estimates including our new proposal, which performs among the best. **B.** Error of $\mathbf{\Psi}$ estimated including our new proposal, which outperforms estimates from other approaches. **C.** Runtime of our approach is nearly independent of the number of loci m, whereas existing approaches run linearly with m. The `popkin` runtime is linear with m but negligible in comparison.

References

[1] Alejandro Ochoa and John D. Storey. "Estimating FST and kinship for arbitrary population structures". *PLoS Genet* 17(1) (2021), e1009241.

[2] Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". *bioRxiv* (2019).

[3] David H. Alexander, John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals". *Genome Res.* 19(9) (2009), pp. 1655–1664.

[4] B. S. Weir and W. G. Hill. "Estimating F-Statistics". *Annual Review of Genetics* 36(1) (2002), pp. 721–750.

[5] Irineo Cabreros and John D. Storey. "A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis". *Genetics* 212(4) (2019), pp. 1009–1029.