Assessing the Performance of qpAdm:
A Statistical Tool for Studying Population Admixture

Éadaoin Harney[1,2], John Wakeley[1]

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA
[2]The Max Planck-Harvard Research Center for the Archaeoscience of the Ancient Mediterranean, Cambridge, MA, 02138, USA and Jena, D-07745, Germany

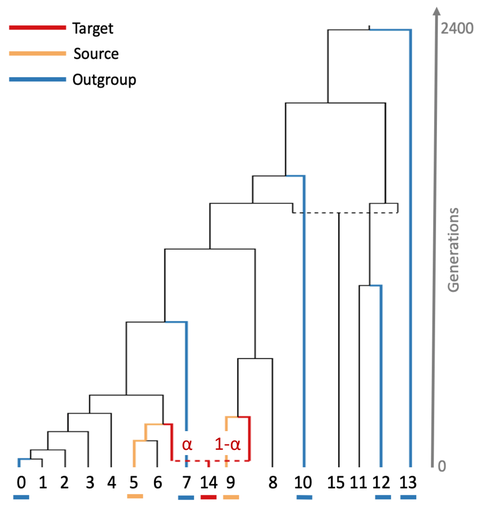
STUDY AIMS

qpAdm is a statistical tool for modeling the ancestry of admixed populations. qpAdm assesses the plausibility of admixture models and estimates admixture proportions, without requiring users to specify the underlying relationships of all the populations included in the model. Although qpAdm is theoretically described in Haak et al (2015), relatively little has been done to assess its performance using data with a known population history. We performed a simulation-based study to assess the behavior of qpAdm under various scenarios in order better understand its performance and to identify areas of potential weakness.

DATA GENERATION

We simulated genome-wide data using msprime (Kelleher, 2016), with the following specifications:

- 22 chromosomes
- Mutation rate: 2.0×10^{-8}
- Recombination rate: 1.0×10^{-8}
- Effective population sizes: 2.5×10^{4} – 8.0×10^{5}
- Sample 10 diploid individuals per population
- ~30 million SNPs generated per simulation



The simulated tree (left) involves a recent admixture event in the history of population 14. We therefore study the ancestry of this population, using the following **standard model**:

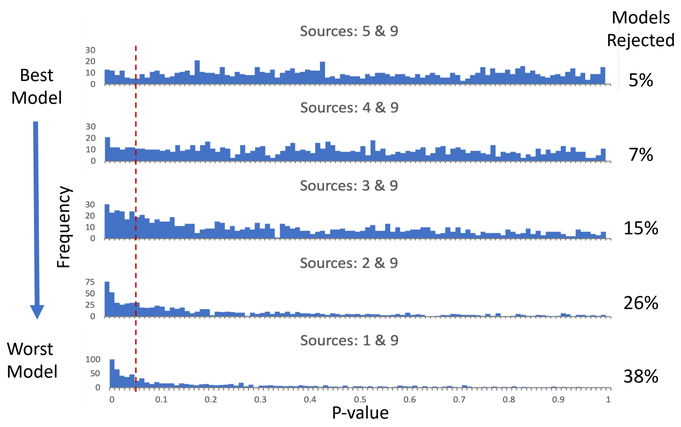
- Target: 14
- Sources: 5 & 9
- Outgroups*: 13, 12, 10, 7 & 0

*In qpAdm, the term outgroup (also commonly called "right" populations), is a misnomer. For qpAdm models, the optimal set of outgroup populations should include populations that are differentially related to the source and target populations (e.g. populations 0 and 7).

HOW OFTEN DOES qpAdm REJECT THE CORRECT MODEL?

EXPECTATION:

For a correct model, the distribution of p-values produced by qpAdm is expected to be uniform, therefore using a threshold of $p > 0.05$ to evaluate models, we expect to reject ~5% of correct models.



RESULT:

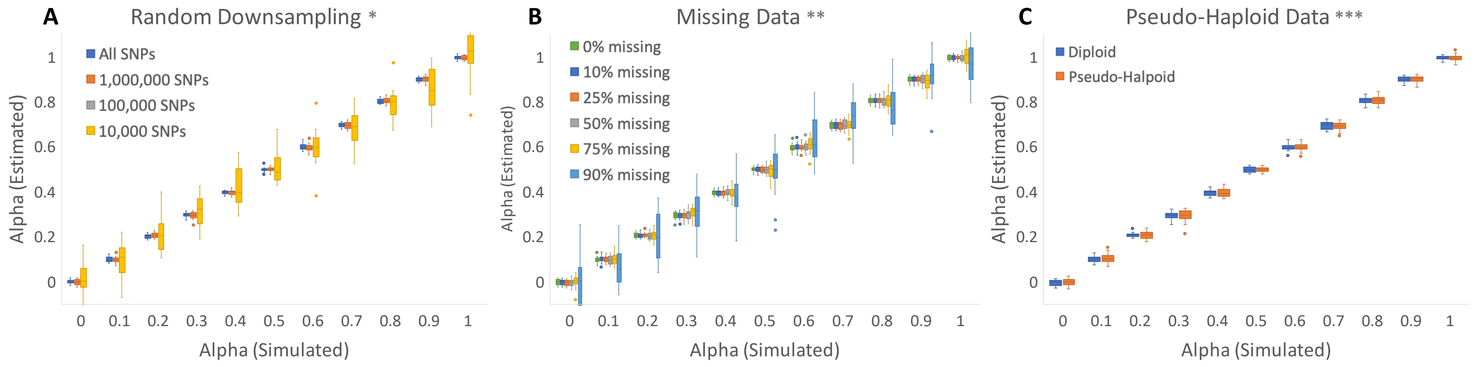
Running 1,000 replicates of the standard model (α=0.5), we find that qpAdm p-values are uniformly distributed when optimal sources are chosen. This distribution is biased towards 0 when non-optimal sources are used.

We confirm that by using a threshold of $p > 0.05$, qpAdm rejects correct models in 5% of cases.

HOW MUCH DATA DOES qpAdm REQUIRE TO ESTIMATE ADMIXTURE PROPORTIONS?

RESULT:

Simulating 20 replicates of the standard model, with varying admixture proportions (α=0-1), we find that qpAdm accurately estimates admixture proportions, even when the number of SNPs analyzed is limited (A), there are high rates of missing data (B), or the data is pseudo-haploid (C).

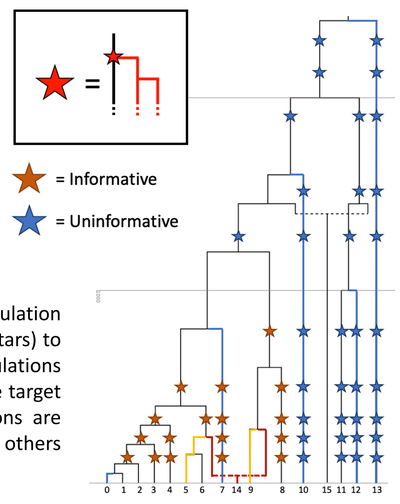


*The indicated number of SNPs were randomly selected from the full dataset (~30 million SNPs) for analysis.
**For each individual, genotypes at the indicated proportion of sites were designated as missing from the down-sampled 1,000,000 SNPs dataset.
***A random allele at each site was selected from the diploid genotype to represent the pseudo-haploid genotype. Data was generated from the down-sampled, 1,000,000 SNP dataset, with 25% missingness.

…THERE ARE TOO MANY OUTGROUP POPULATIONS?

PROBLEM:

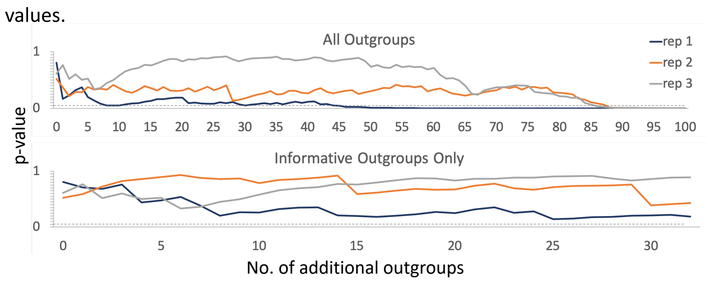
A theoretical upper limit on the number of outgroup populations that can be included in a qpAdm model is predicted to exist. However, this upper limit has never been formally studied.



APPROACH:

We added additional population branching events (represented by stars) to the simulated tree (right). Populations that are differentially related to the target (14) and source (5 + 9) populations are considered **informative**, while all others are considered **uninformative**.

We sequentially ran qpAdm on the standard model, randomly adding one new population to the outgroup set and observed the change in associated p-values.



RESULTS:

We find that with the addition of a large number of population outgroups, p-values begin to approach zero. This decline is not driven by the inclusion of particular informative outgroups, but appears to be dependent on the total number of outgroups.

WHAT HAPPENS IF…

…THE POPULATION HISTORY INVOLVES CONTINUOUS GENE FLOW?

PROBLEM:

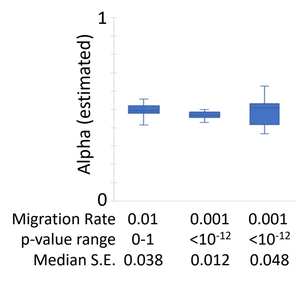
qpAdm assumes that gene flow occurs via pulses of admixture. However, gene flow may also occur as the result of continuous migration. How qpAdm interprets continuous gene flow has not been formally studied.

APPROACH:

We simulated data that involves continuous gene flow between neighboring populations (right), following a 6-way population split 1000 generations ago.

We consider the p-values and admixture proportion estimates produced by qpAdm, under the following model:

- Target=2
- Sources=1 & 3
- Outgroups = 0, 4, 6, 7 & 8



Migration Rate	0.01	0.001	0.001
p-value range	0-1	$<10^{-12}$	$<10^{-12}$
Median S.E.	0.038	0.012	0.048

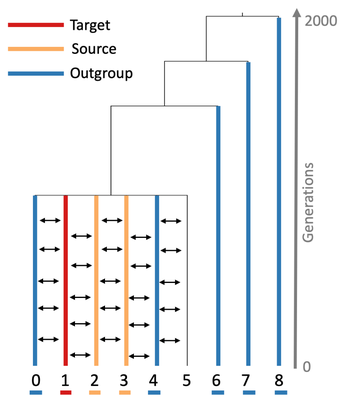
RESULTS:

We find that qpAdm does not distinguish between admixture and continuous gene flow, particularly in cases where the migration rate is high.

…THE TARGET POPULATION FALLS IN AN EXTREME POSITION ALONG A CLINE?

PROBLEM:

qpAdm often produces models with admixture proportions that fall outside the range of 0-1. This has been interpreted as evidence of a genetic cline of ancestry, in which the target population falls in a more extreme position than either of the proposed sources (Lazaridis et al, 2017). We explore this question using a model with continuous gene flow.



Migration Rate	0.01	0.001	0.001
p-value range	0-1	$<10^{-12}$	$<10^{-12}$
Median S.E.	0.038	0.012	0.048

APPROACH:

Using the same model of continuous gene flow as the panel on the left, we consider the p-values and admixture proportion estimates produced by qpAdm, under the following model:

- Target=1
- Sources=2 & 3
- Outgroups = 0, 4, 6, 7 & 8

RESULTS:

We confirm that in cases where the target population falls in a more extreme position along a genetic cline than either of the source populations, admixture proportions outside the range of 0-1 are produced.

REFERENCES

- Haak W, et al. "Massive migration from the steppe was a source for Indo-European languages in Europe." *Nature* 522.7555 (2015): 207.
- Kelleher J, Etheridge AM, and McVean G (2016), Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes, *PLoS Comput Biol* 12(5): e1004842.
- Lazaridis I, et al. "Genetic origins of the Minoans and Mycenaeans." *Nature* 548.7666 (2017): 214.