Week-8
Admixture modelling: qpAdm

Aim: *Understanding admixture modelling, learning how to run qpAdm analysis*

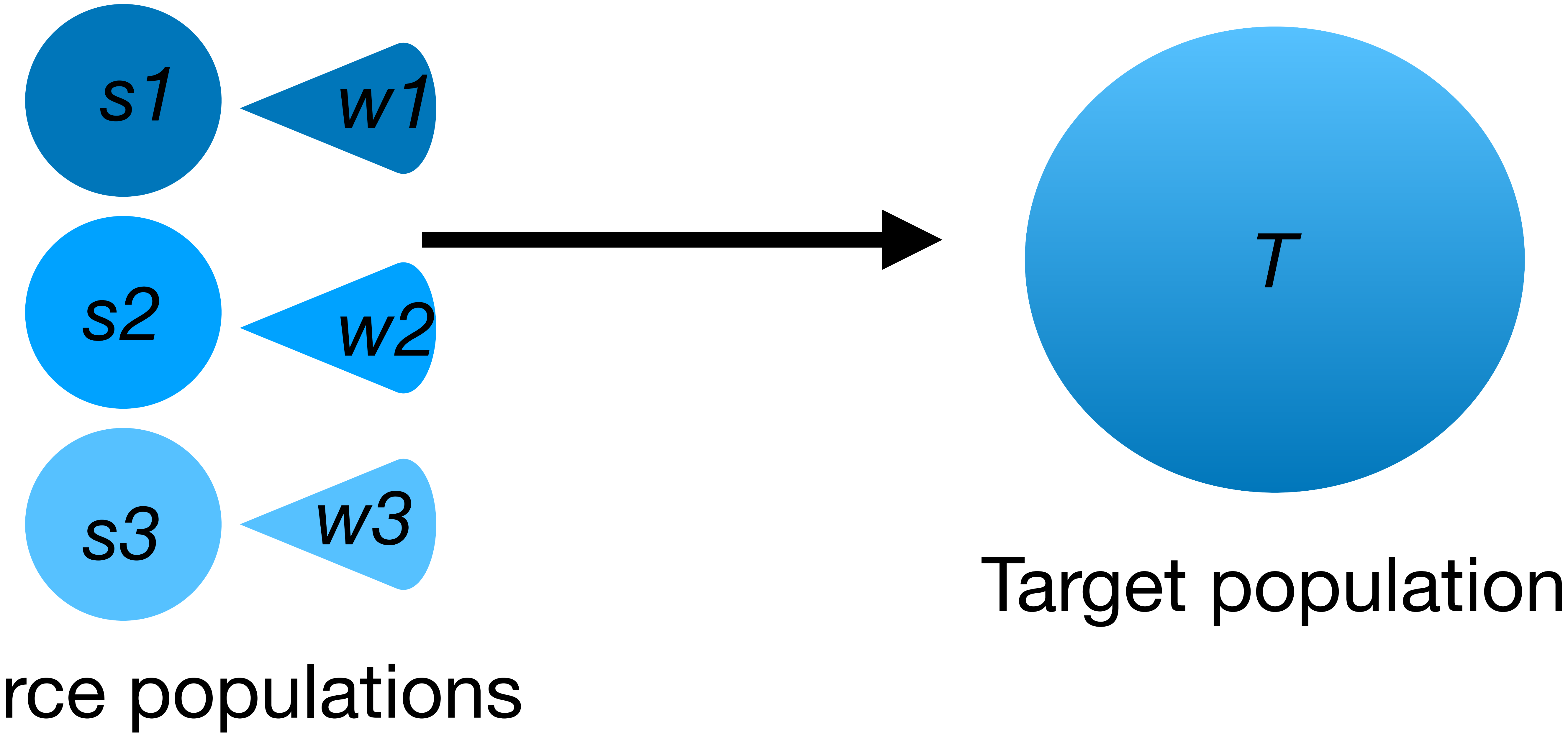
Hands-on: *Running qpAdm on an example dataset*

Reading suggestion: *Haak et al 2015 “Massive migration from the steppe was a source for Indo-European languages in Europe”*

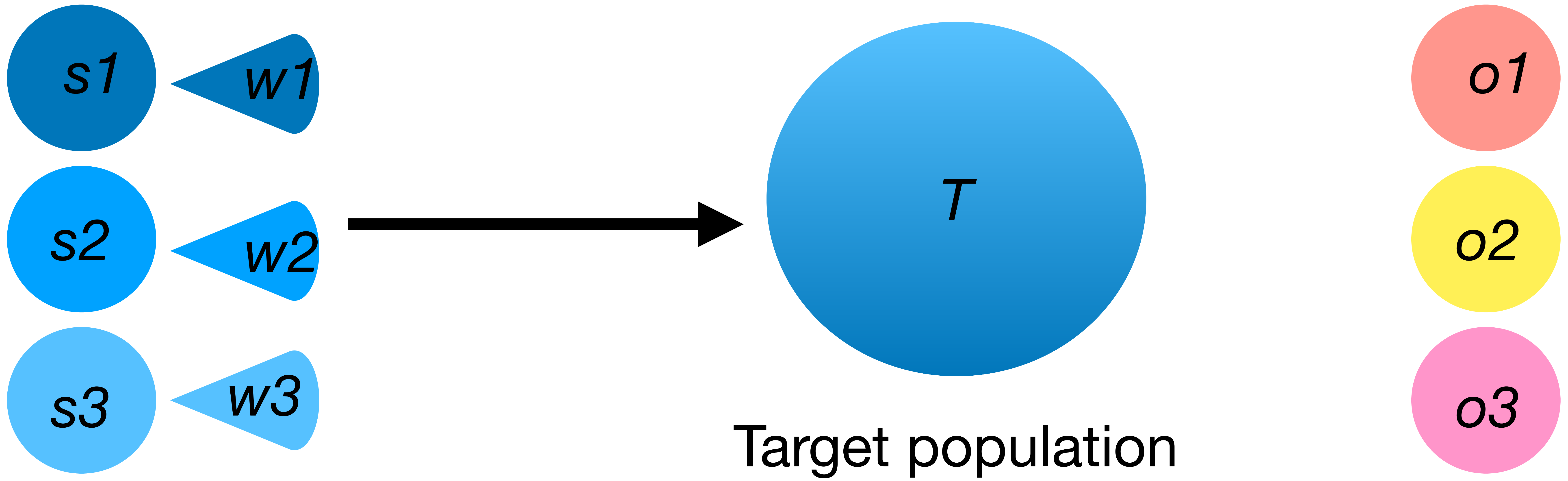
Admixture modelling: Estimating admixture proportions

F_4-statistics -> Phylogenetic hypotheses, relationships between populations

Admixture modelling -> estimating the contribution of a set of distinct gene pools to a target population / finding the proportions of mixture



If target population T is a mixture of source populations:



Target population

Source populations

w: ancestry proportion

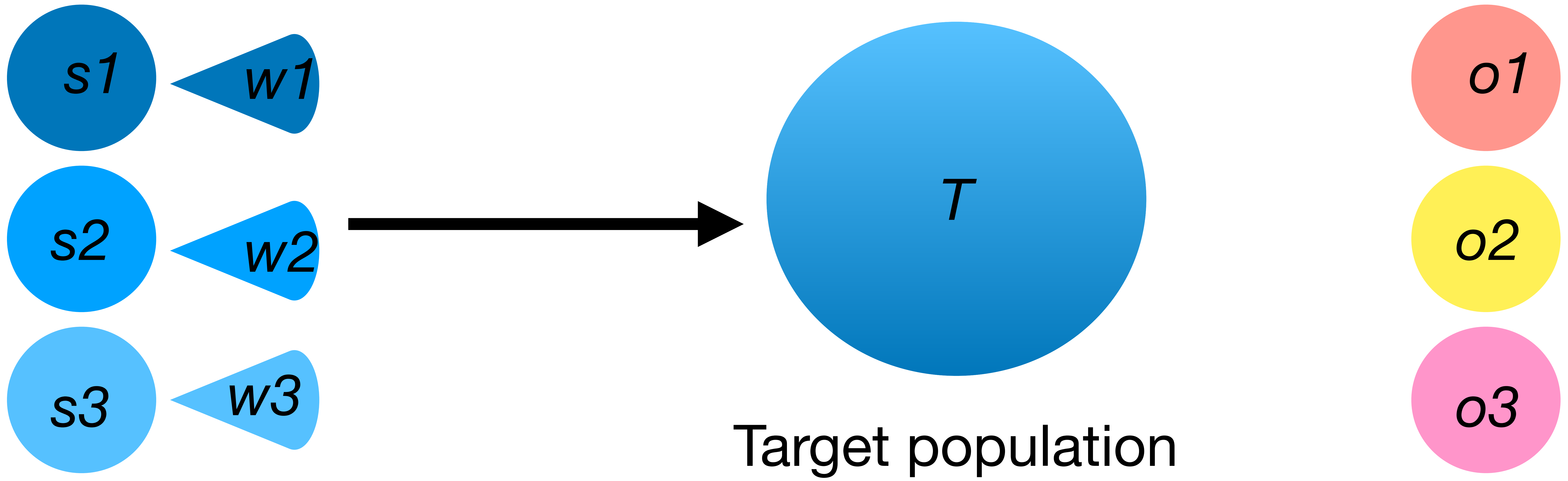
Outgroup populations

$$T = \sum_{i=1}^{n} w_i s_i$$

$$\sum_{i} w_i F_4(T, s_i, o_1, o_2) = F_4(T, T; o_1, o_2) = 0$$

no gene flow between S & T and Outgroups

Modelling using qpAdm:
We need a dataset, a parameter file, list of left populations, list of right populations



Target population

Source populations

w: ancestry proportion

Outgroup populations

Left populations

Right populations

Dataset - Data formats that can be used:

ANCESTRYMAP

EIGENSTRAT

PED

PACKEDPED

PACKEDANCESTRYMAP

In the server: You have your own datasets that you used for your projects, you can use them for this course.

Left population file: Target and sources

Target
Source 1
Source 2
Source 3
.
.
.
Source n

Right population file: Outgroups

Outgroup 1
Outgroup 2
Outgroup 3
Outgroup 4
.
.
.
Outgroup n

Parameter file:

genotypename:
snpname:
indivname:
popleft:
popright:

To identify an optimal model (Harney et al 2021)

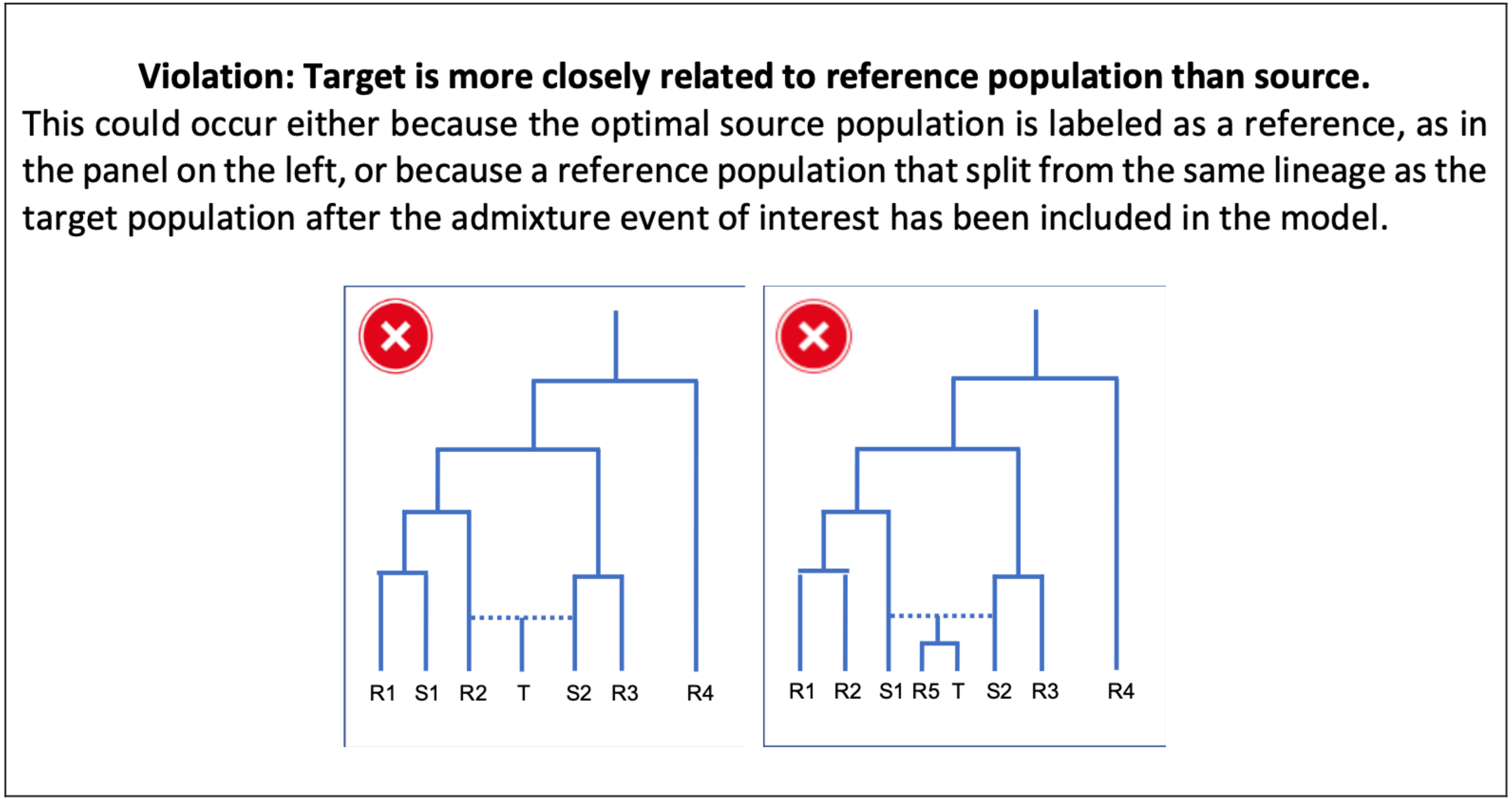
- Ensuring that the model includes right populations that are differentially related to the various source populations that are being used as potential left populations.

- Ensuring that models are directly comparable. It is not appropriate to compare two models that use entirely different sets of right populations. While it may not be possible to use identical sets of right populations for all models under consideration, the right population sets should be as similar as possible.

Models are considered acceptable if target and sources share a common ancestral lineage more recently than they share with outgrip populations

https://academic.oup.com/genetics/article/217/4/iyaa045/6070149

Some situations violating the results (from Harney et al 2021)

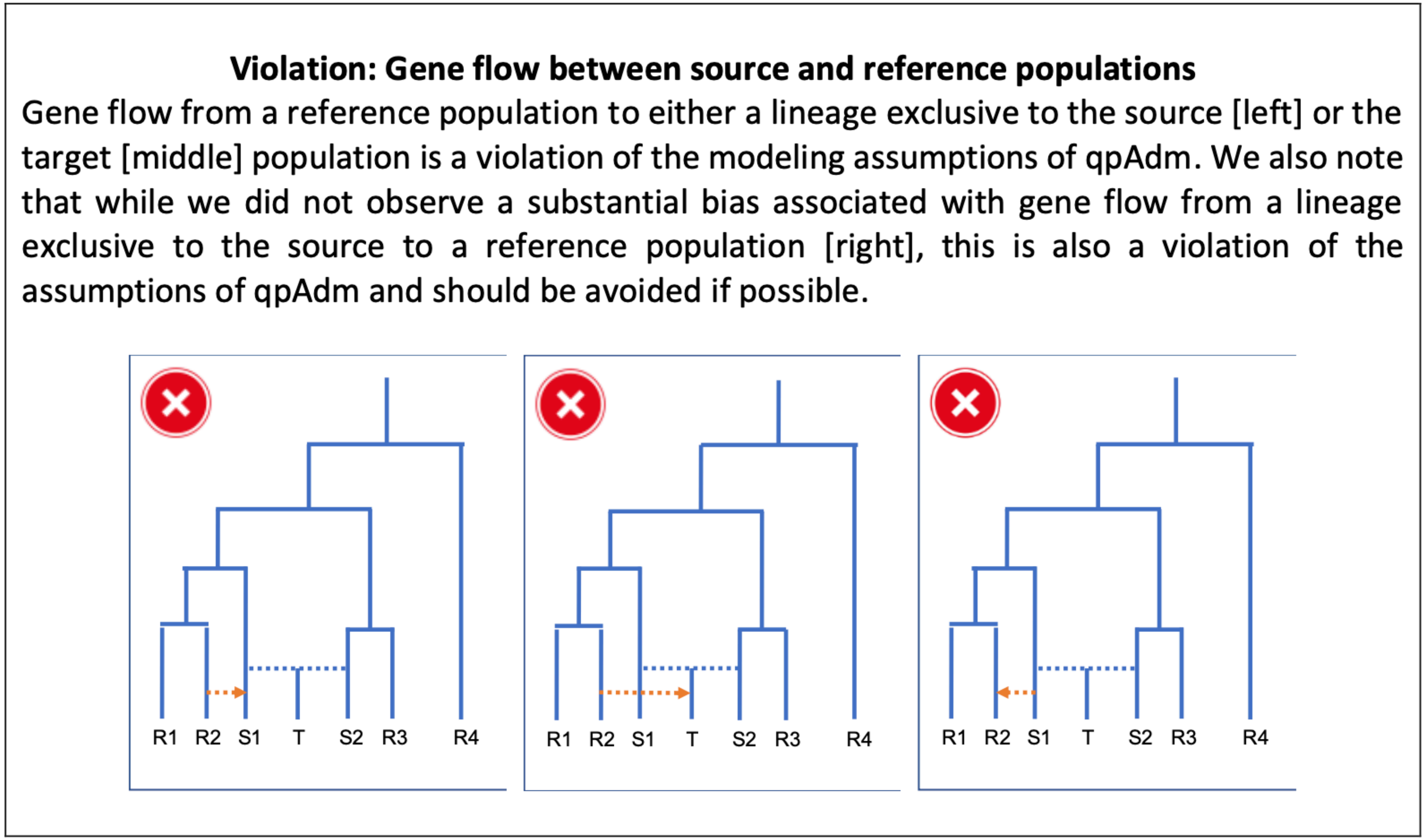


Violation: Target is more closely related to reference population than source.

This could occur either because the optimal source population is labeled as a reference, as in the panel on the left, or because a reference population that split from the same lineage as the target population after the admixture event of interest has been included in the model.

https://academic.oup.com/genetics/article/217/4/iyaa045/6070149

Some situations violating the results (from Harney et al 2021)

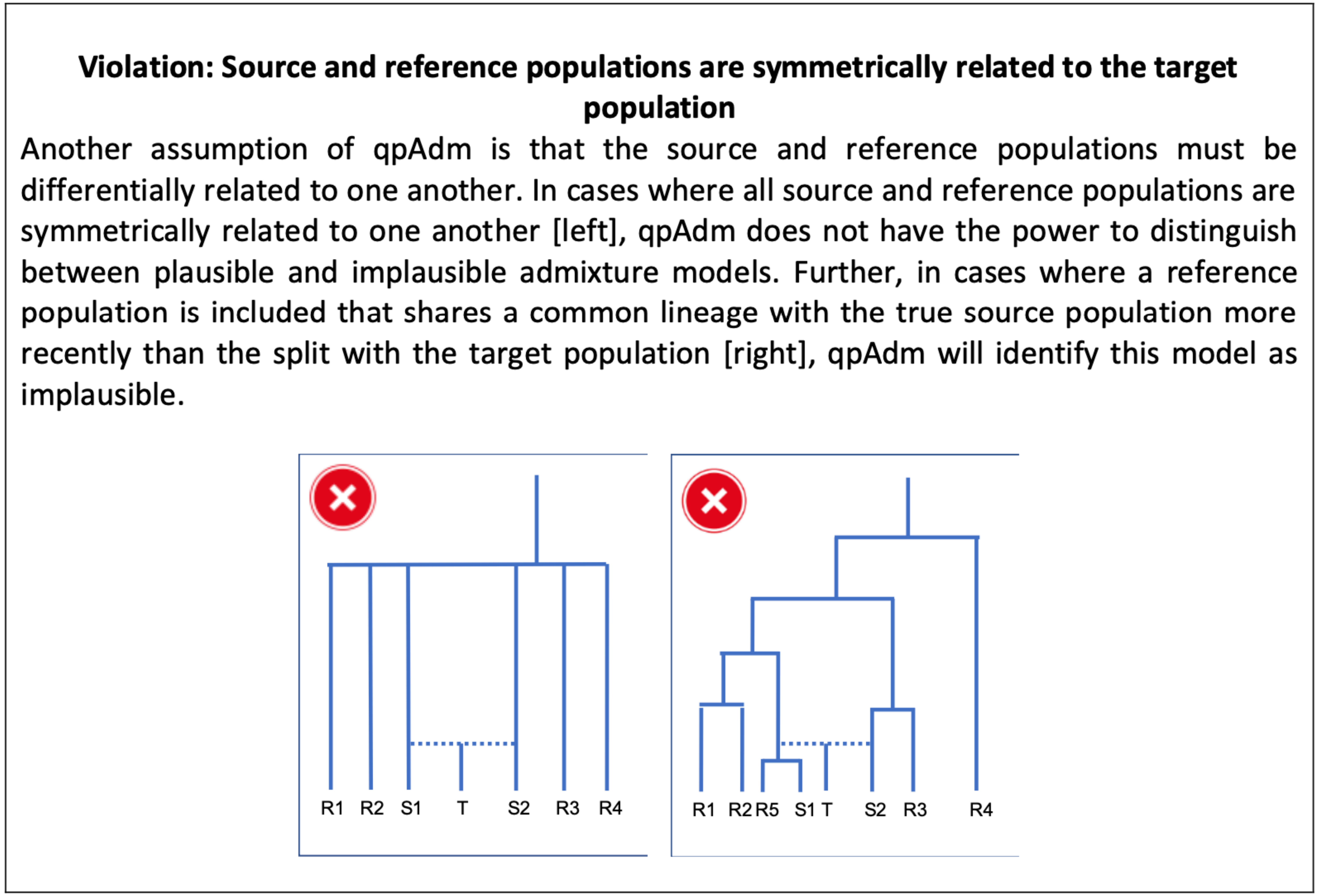


Violation: Gene flow between source and reference populations

Gene flow from a reference population to either a lineage exclusive to the source [left] or the target [middle] population is a violation of the modeling assumptions of qpAdm. We also note that while we did not observe a substantial bias associated with gene flow from a lineage exclusive to the source to a reference population [right], this is also a violation of the assumptions of qpAdm and should be avoided if possible.

https://academic.oup.com/genetics/article/217/4/iyaa045/6070149

Some situations violating the results (from Harney et al 2021)



Violation: Source and reference populations are symmetrically related to the target population

Another assumption of qpAdm is that the source and reference populations must be differentially related to one another. In cases where all source and reference populations are symmetrically related to one another [left], qpAdm does not have the power to distinguish between plausible and implausible admixture models. Further, in cases where a reference population is included that shares a common lineage with the true source population more recently than the split with the target population [right], qpAdm will identify this model as implausible.

https://academic.oup.com/genetics/article/217/4/iyaa045/6070149