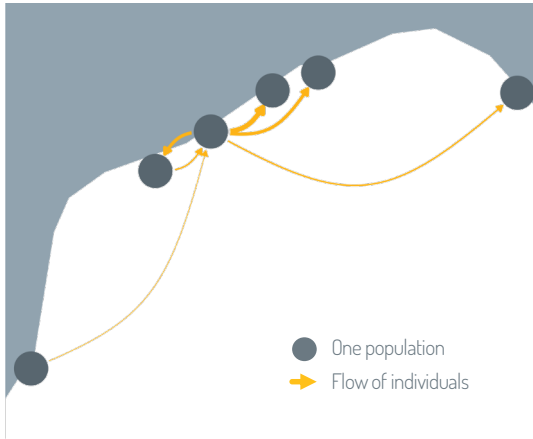




ASSESSING DIRECTIONAL POPULATION CONNECTIVITY FROM SNPS DATA: A BENCHMARK OF COMMONLY USED METHODS AND TOOLS ON SIMULATED DATASETS

Camille Sant, Didier Forcioli, Cécile Fauvelot, Jean-Olivier Irisson



POPULATION CONNECTIVITY

Exchange of individuals between geographically separated populations.

Physical tracking

Capture-mark-recapture

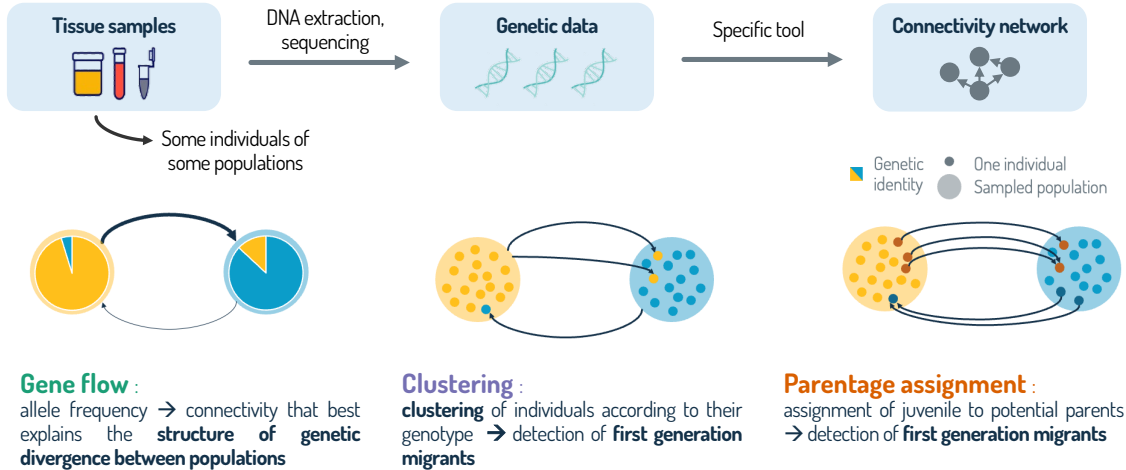
Diffusion models

Genetic

Usually consider
as empirical
... but also subject to bias,
especially due to sampling

Is the connectivity estimated by genetic approaches representative of reality?





Among the genetic approaches, which one is the most reliable?

*Is the connectivity estimated by genetic approaches representative of reality?
Among the genetic approaches, which one is the most reliable?*

Real connectivity network



Simulate
genetic data



Sub-sample
simulated
data



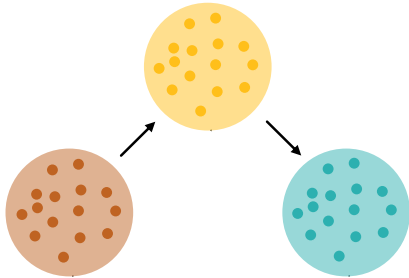
Apply the
available
tools



Networks found by tools



Compare to evaluate tools
accuracy



Demographic configuration:

- 3 populations
- Unidirectional flux

Simulation :

Tree of genome heritage :

- create diversity from a unique common ancestor model genome
→ msprime (Kelleher et al., 2016)
- realistic simulation of 10 000 generations
each reproduction events: migration, mutation, recombination
→ SLiM (Haller and Messer, 2019)

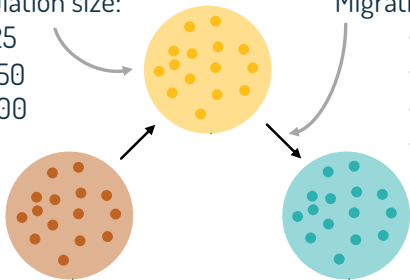
Genomes of the all individuals alive at the end of the simulation

Population size:

- 125
- 250
- 500

Migration rate:

- 1%
- 2%
- 4%
- 8%



Demographic configuration:

- 3 populations
- Unidirectional flux

Simulation :

Tree of genome heritage :

- create diversity from a unique common ancestor model genome
→ msprime (Kelleher et al., 2016)
- realistic simulation of 10 000 generations
each reproduction events: migration, mutation, recombination
→ SLiM (Haller and Messer, 2019)

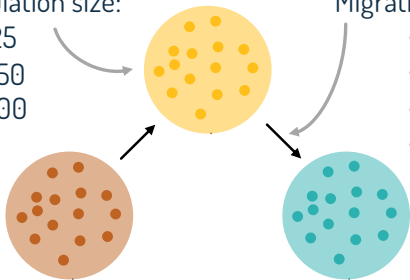
2 demographic variables, 12 demographic scenarios with 30 replicates

Population size:

- 125
- 250
- 500

Migration rate:

- 1%
- 2%
- 4%
- 8%



Sub-sampling

Individuals:

- 100%
- 30%
- 10%

SNPs:

- 1% (~6000 SNPs)
- 0.2% (~1500 SNPs)
- 0.1% (~800 SNPs)

Demographic configuration:

- 3 populations
- Unidirectional flux

Simulation :

Tree of genome heritage :

- create diversity from a unique common ancestor model genome
→ msprime (Kelleher et al., 2016)
- realistic simulation of 10 000 generations
each reproduction events: migration, mutation, recombination
→ SLiM (Haller and Messer, 2019)

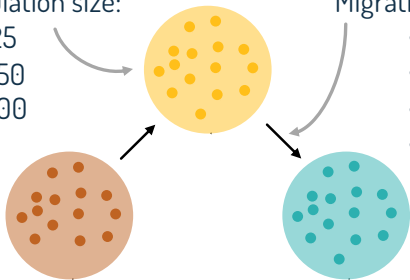
Sub-sampling both individuals and SNPs → more realistic and affordable sampling

Population size:

- 125
- 250
- 500

Migration rate:

- 1%
- 2%
- 4%
- 8%



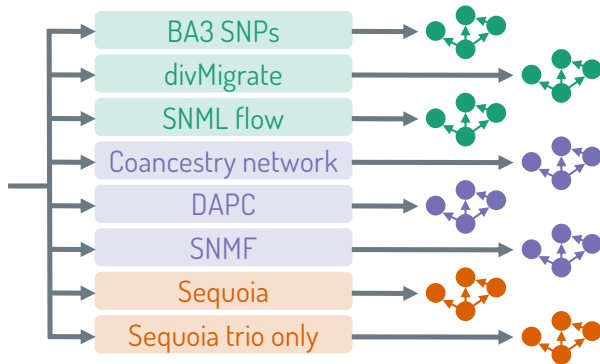
Sub-sampling

Individuals:

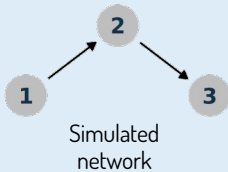
- 100%
- 30%
- 10%

SNPs:

- 1% (~6000 SNPs)
- 0.2% (~1500 SNPs)
- 0.1% (~800 SNPs)



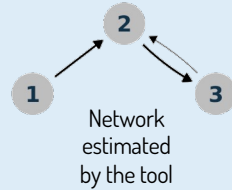
Apply commonly used tools to estimate connectivity on the simulated data → connectivity networks

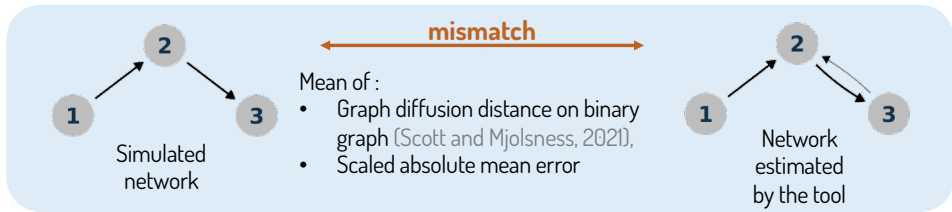


mismatch

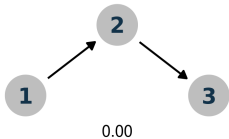
Mean of :

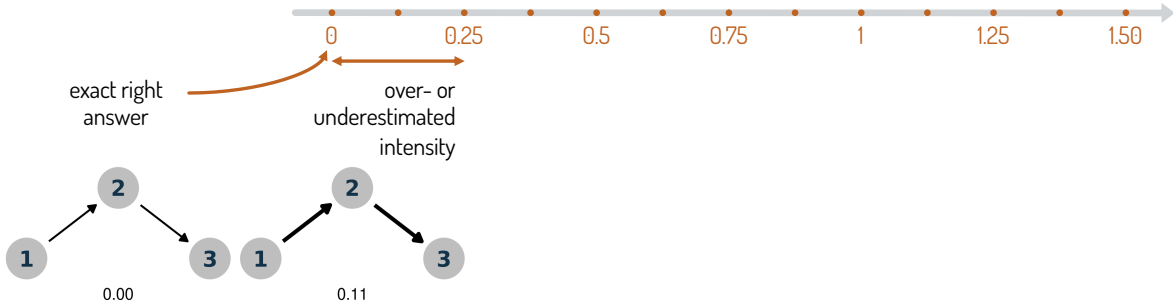
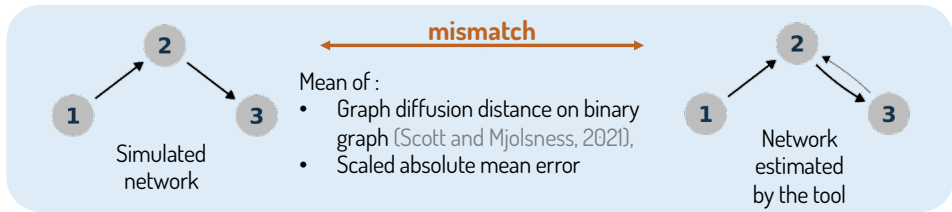
- Graph diffusion distance on binary graph (Scott and Mjolsness, 2021),
- Scaled absolute mean error

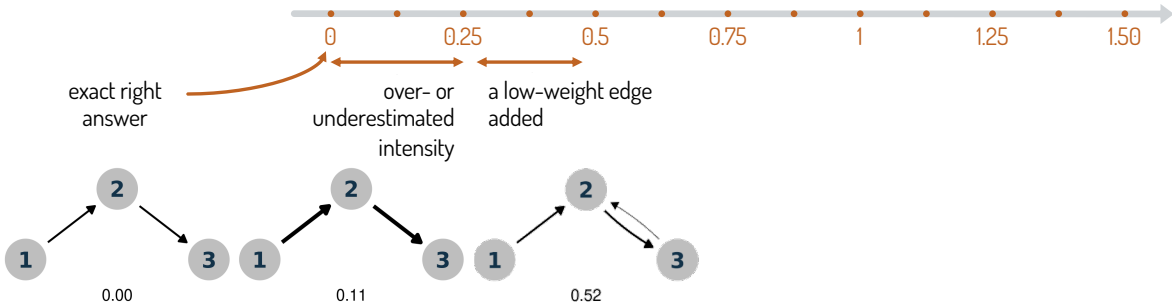
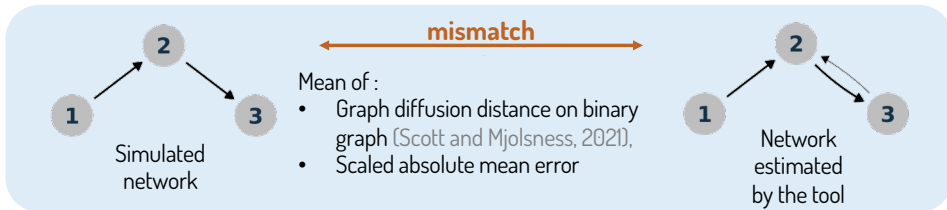


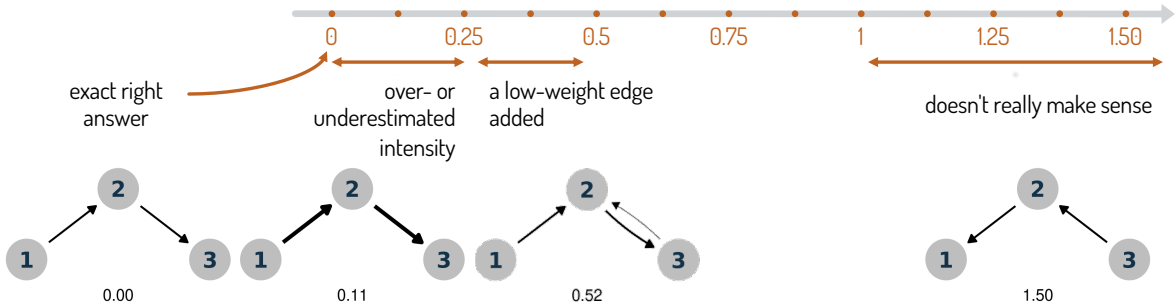
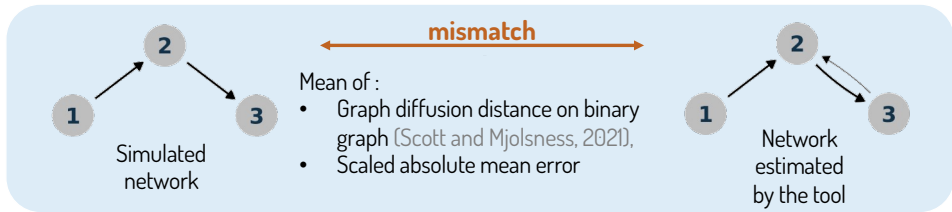


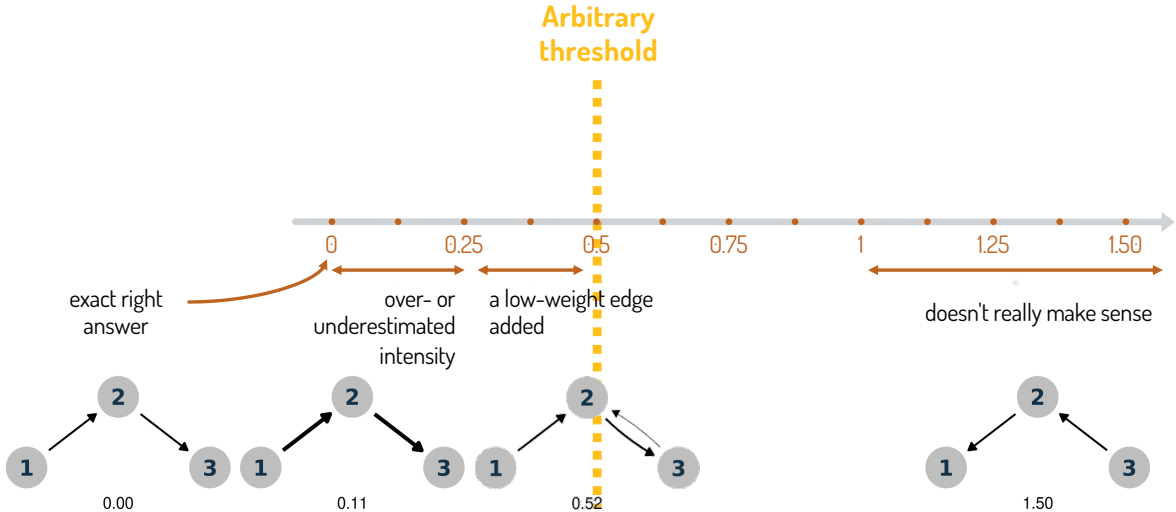
exact right
answer











Mismatch $< 0.5 \rightarrow$ good assessment

*Is the connectivity estimated by genetic approaches representative of reality?
Among the genetic approaches, which one is the most reliable?*

Real connectivity network



Simulate
genetic data

- population sizes
- migration rates

Sub-sample
simulated
data

- % sampled
individuals
- % sampled SNPs

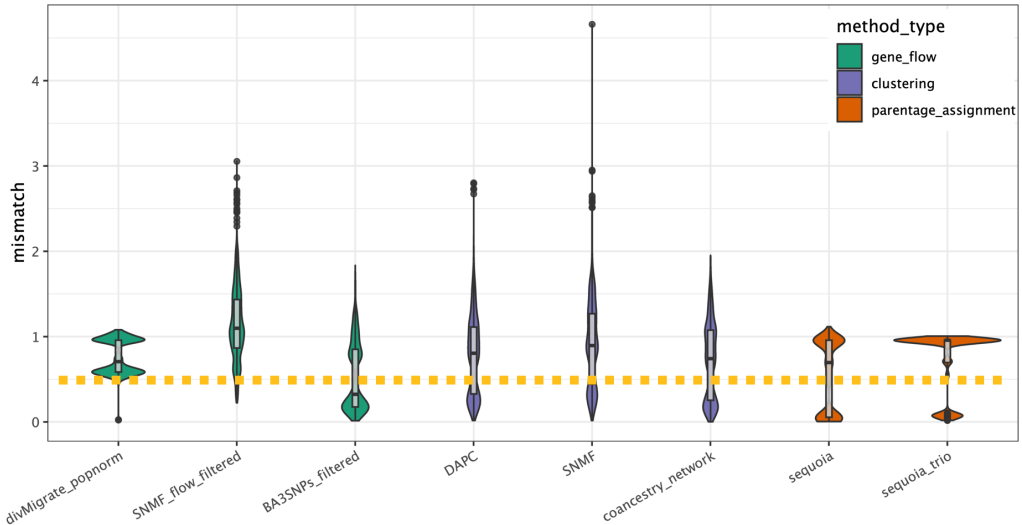
Apply the
available
tools

- tool

Networks found by tools



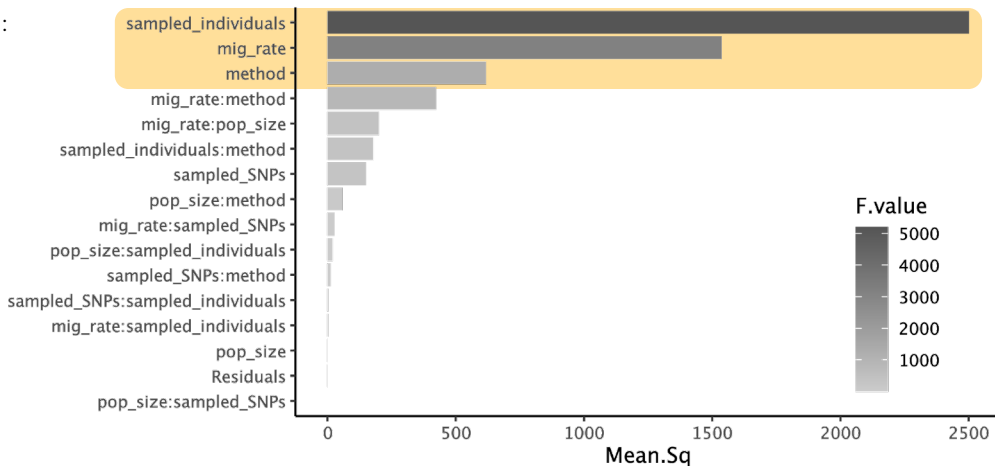
Compare to evaluate tools
accuracy



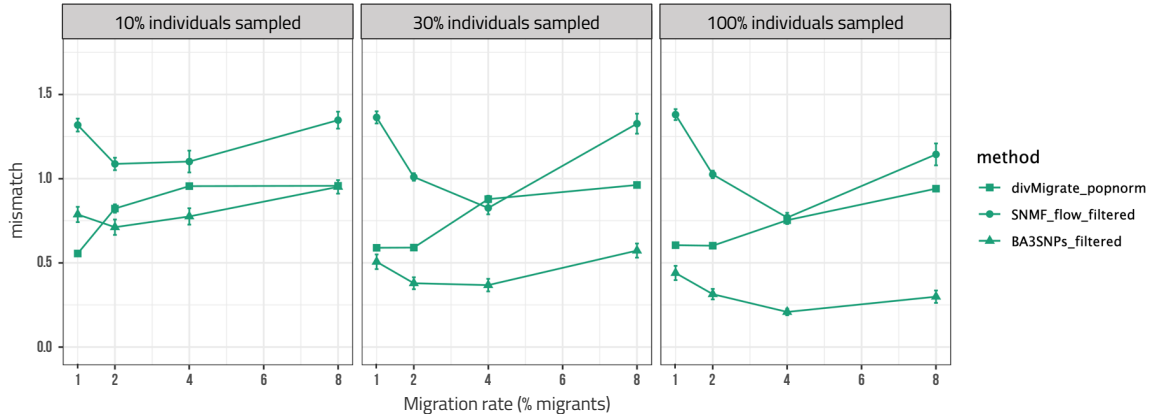
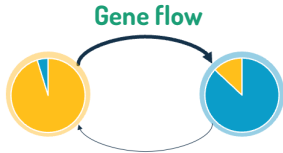
→ There is not a single method performing better than the others

Evaluate the impact of demographic and sampling variables on the mismatch value

ANOVA :

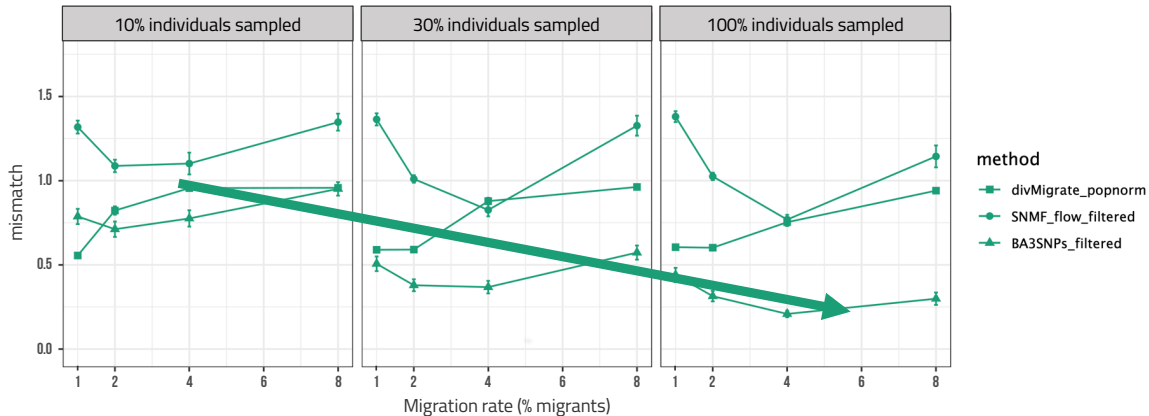


→ Mismatch value is highly impacted by the individual sampling rate and migration rate



Gene flow

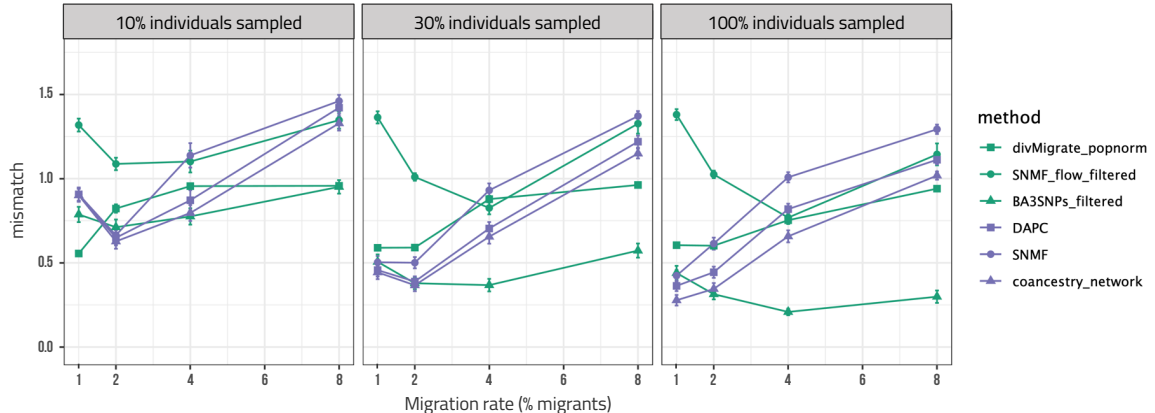
↗ sampling = ↘ error



Gene flow

↗ sampling = ↘ error

Clustering

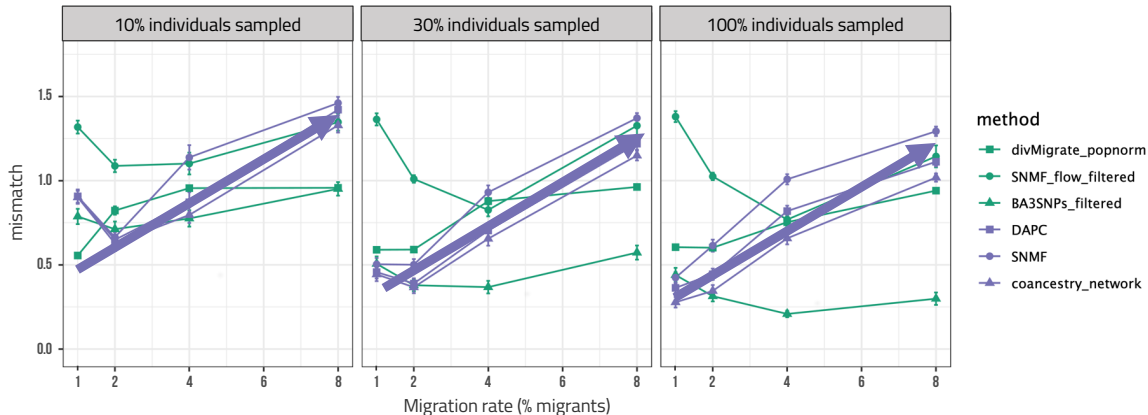


Gene flow

↗ sampling = ↘ error

Clustering

Good if low migration rate



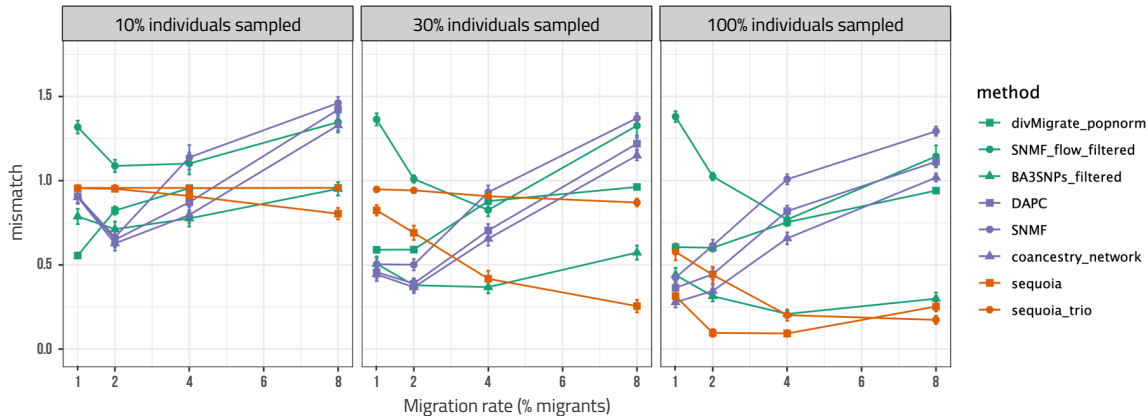
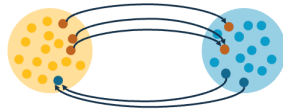
Gene flow

↗ sampling = ↘ error

Clustering

Good if low migration rate

Parentage assignment



Gene flow

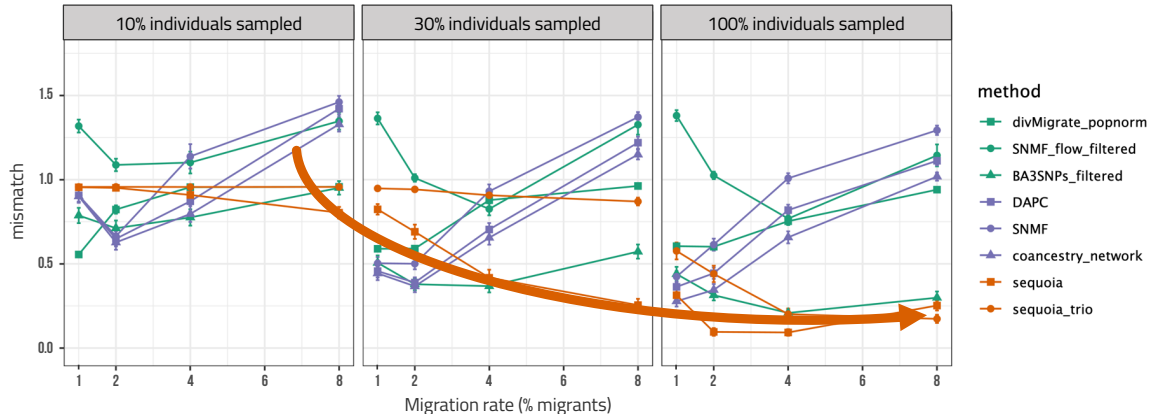
↗ sampling = ↘ error

Clustering

Good if low migration rate

Parentage assignment

Good if high sampling rate and migration rate



Gene flow

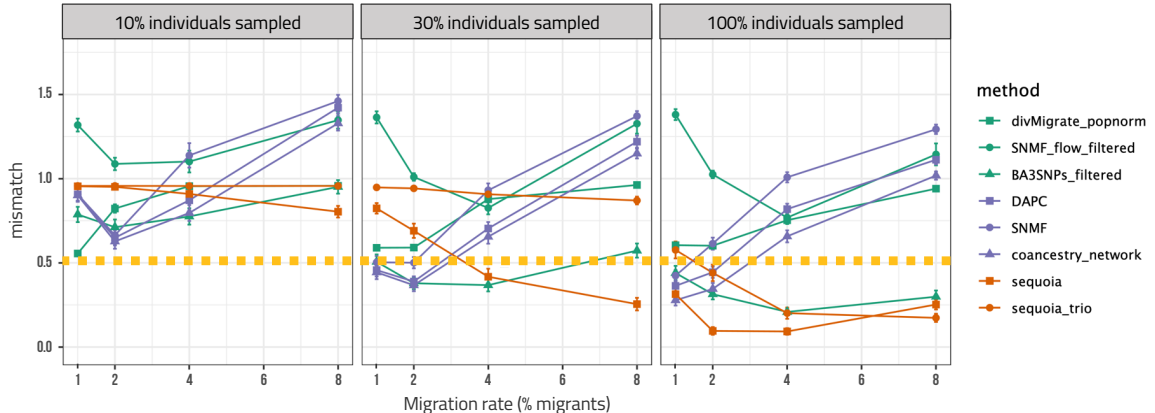
↗ sampling = ↘ error

Clustering

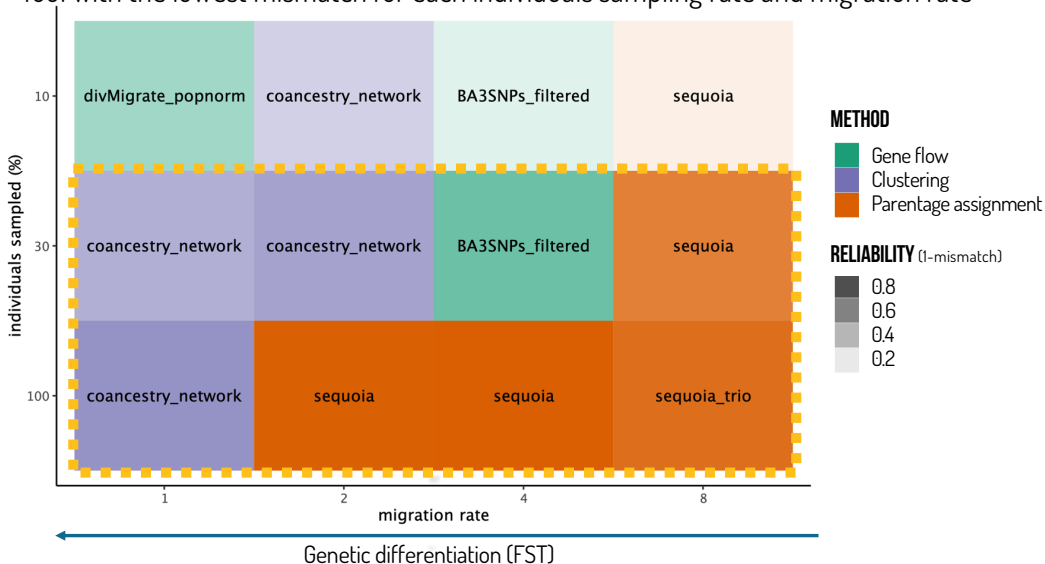
Good if low migration rate

Parentage assignment

Good if high sampling rate and migration rate



Tool with the lowest mismatch for each individuals sampling rate and migration rate



1. Estimating connectivity network between populations requires much more intensive sampling of individuals than conventional estimates of population structure

Population structure
estimate

Individuals sampling ++

Connectivity
network estimate

2. To estimate connectivity network, sampling more SNPs does not seem to compensate for low sampling rates of individuals, contrary to what has been demonstrated for population structure estimates

3. Guidelines :
 - 1 – Estimate population size and genetic differentiation using conventional approaches
 - 2 – **If enough individuals are sampled**, select an appropriate tool to access the connectivity network according to your situation

1. Estimating connectivity network between populations requires much more intensive sampling of individuals than conventional estimates of population structure

2. To estimate connectivity network, sampling more SNPs does not seem to compensate for low sampling rates of individuals, contrary to what has been demonstrated for population structure estimates



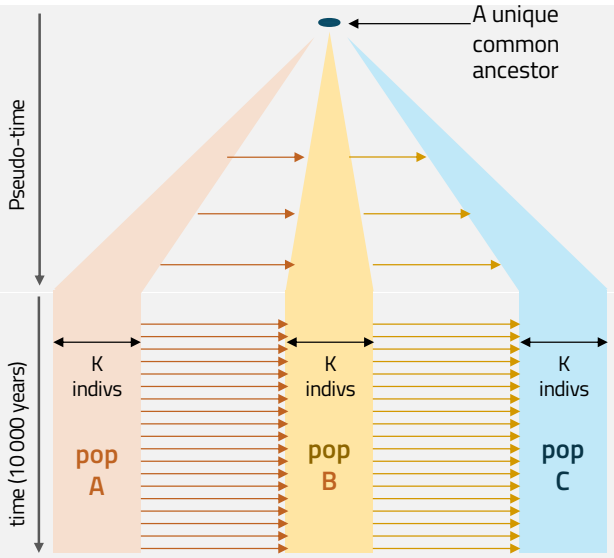
3. Guidelines :
1 – Estimate population size and genetic differentiation using conventional approaches
2 – **If enough individuals are sampled**, select an appropriate tool to access the connectivity network according to your situation

1. Estimating connectivity network between populations requires much more intensive sampling of individuals than conventional estimates of population structure
2. To estimate connectivity network, sampling more SNPs does not seem to compensate for low sampling rates of individuals, contrary to what has been demonstrated for population structure estimates
3. Guidelines :
 - 1 – Estimate population size and genetic differentiation using conventional approaches
 - 2 – **If enough individuals are sampled**, select an appropriate tool to access the connectivity network according to your situation

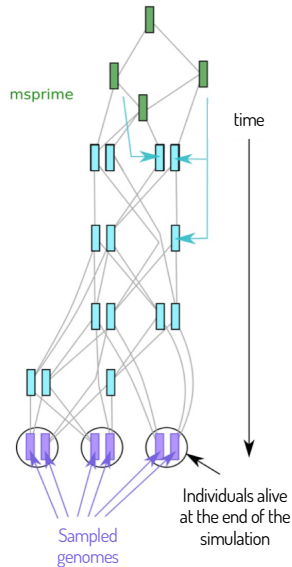


THANK YOU FOR YOUR ATTENTION

Demography



Genetic



At each reproduction:

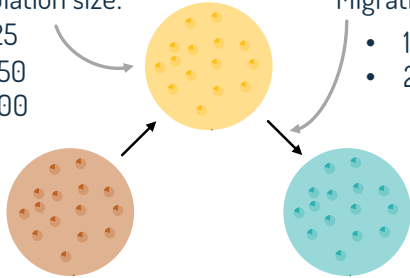
- migration
- reproduction
- mutation
- recombination

Tree of genome heritage starting from the common ancestor (the known reference).

→ We have access to the genomes of the final individuals

Population size:

- 125
- 250
- 500



Migration rate:

- 1% • 4%
- 2% • 8%

Sub-sampling

Individuals:

- 100%

SNPs:

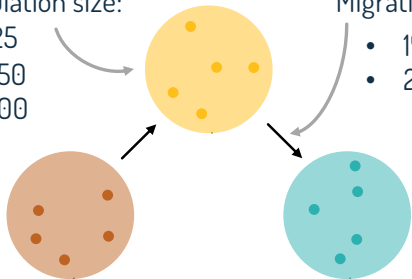
- 1% (~6000 SNPs)
- 0.2% (~1500 SNPs)
- 0.1% (~800 SNPs)

	SNPs										
	1	2	3	4	5	6	7	8	9	10	...
1	A	T	C	G	...						
2											
3											
4											
5											
6											
7											
8											
9											
10											
...											

Sub-sampling SNPs → more affordable sequencing

Population size:

- 125
- 250
- 500



Migration rate:

- 1% • 4%
- 2% • 8%

Sub-sampling

Individuals:

- 100%
- 30%
- 10%

SNPs:

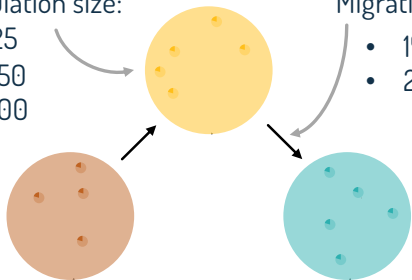
- 1% (~6000 SNPs)
- 0.2% (~1500 SNPs)
- 0.1% (~800 SNPs)

	SNPs										
	1	2	3	4	5	6	7	8	9	10	...
Individuals	1	A	T	C	G	...					
	2										
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	...										

Sub-sampling individuals → more realistic sampling

Population size:

- 125
- 250
- 500



Migration rate:

- 1% • 4%
- 2% • 8%

Sub-sampling

Individuals:

- 100%
- 30%
- 10%

SNPs:

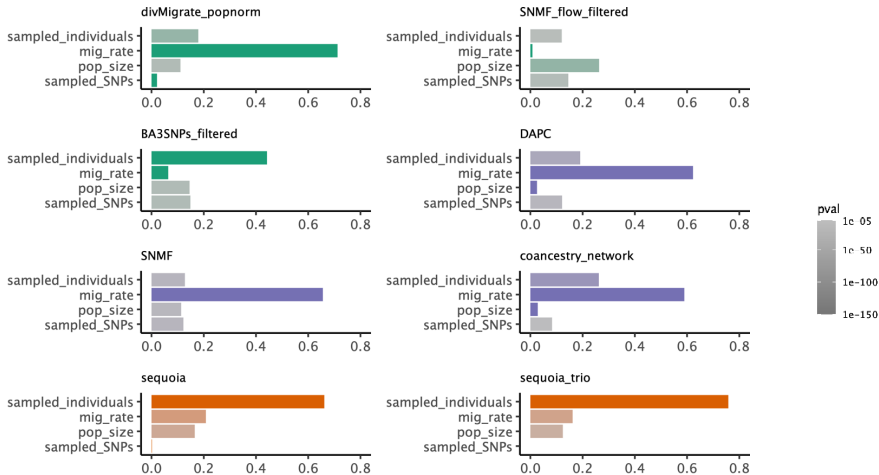
- 1% (~6000 SNPs)
- 0.2% (~1500 SNPs)
- 0.1% (~800 SNPs)

	SNPs										
	1	2	3	4	5	6	7	8	9	10	...
1	A	T	C	G	...						
2											
3											
4											
5											
6											
7											
8											
9											
10											
...											

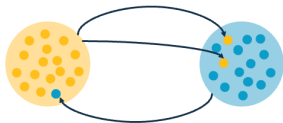
Sub-sampling both individuals and SNPs → more realistic and affordable sampling

Evaluate the impact of demographic and sampling variables on the mismatch value

GLM:
abs(coefficient estimate)



→ Individual sampling rate and migration rate are the variables most heavily impact the final error

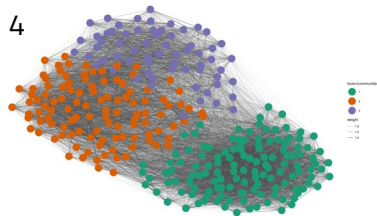
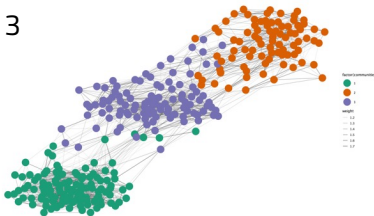
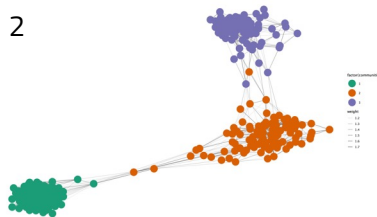
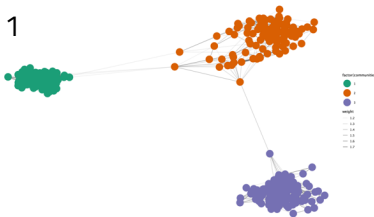


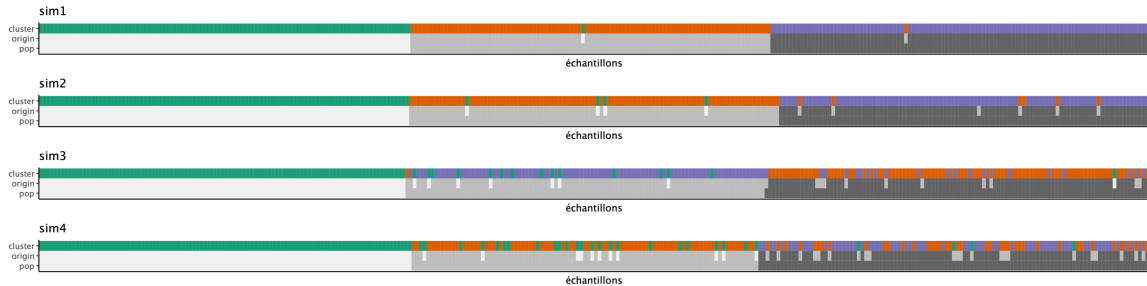
Assignment a une population:
clustering des individus en fonction de leur génotype → détection des **migrants de première génération**

⊖ Les populations fortement connectées ne sont pas différenciables.

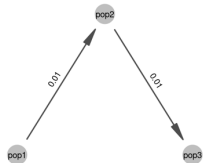
Network community detection

DAPC
 SNMF

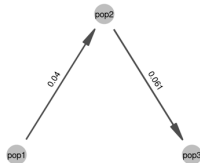




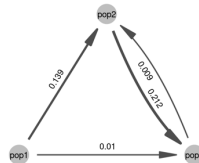
sim1
1%



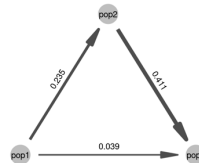
sim2
2%



sim3
4%



sim4
8%



P1: The more the weight of an edge changes, the more it impacts the distance value, $d(G0, G1) < d(G0, G1')$.

P2: A modification that qualitatively changes the flow (i.e. changes the source or sink character of a node or, even worse, leads to an unconnected graph) is more penalized than a modification of flow intensity, $d(G0, G2) < d(G0, G2') < d(G0, G2'')$.

P3: With equal graph size, the lower the density of the graph, the more impact a change has on the distance, $d(G0', G3) < d(G0, G2)$.

