Representing long-range genetic similarity on a background of spatially heterogeneous IBD

Vivaswat Shastry, John Novembre

SMBE 2024
PUERTO VALLARTA

Acknowledgements

<u>Labs</u>:

- Berg
- Novembre
- Steinrücken

<u>Community</u>:

Program in Computational Biology

Biological Sciences Division

Outline

- Introduction + Motivation

- Spatially heterogeneous models of isolation-by-distance (IBD)

- Our model for long-range gene flow events (`FEEMSmix`)

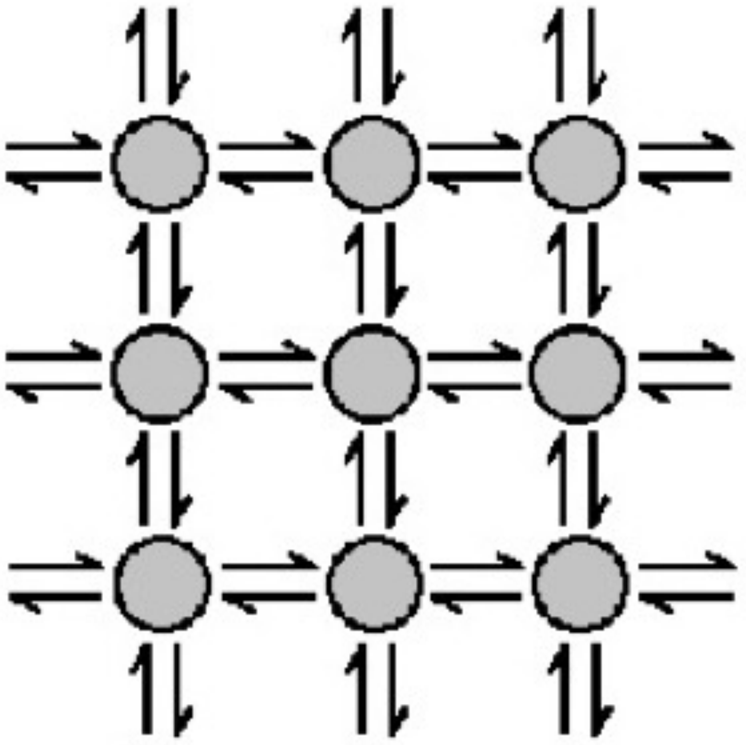
- Results:
    1. Simulations
    2. North American grey wolves
    3. Afro-Eurasian panel of humans

Outline

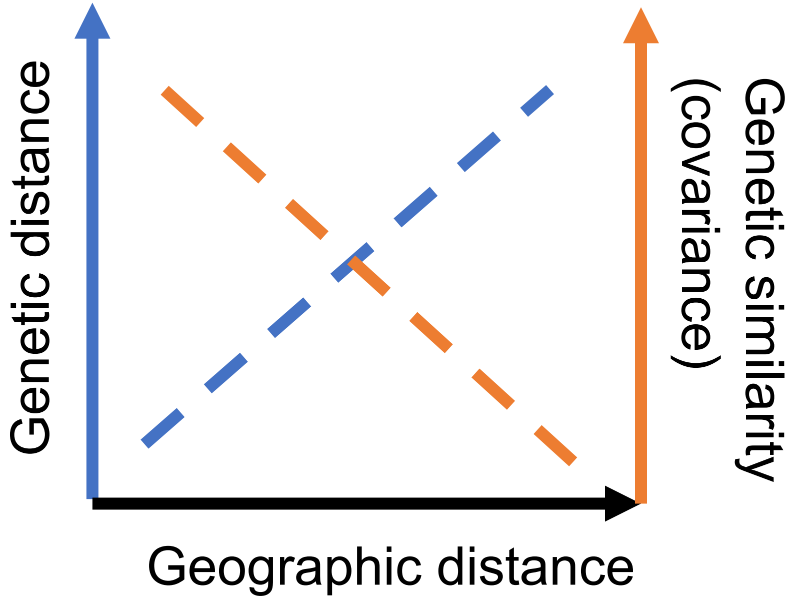
- Introduction + Motivation
- Spatially heterogeneous models of isolation-by-distance (IBD)
- Our model for long-range gene flow events (`FEEMSmix`)
- Results:
  1. Simulations
  2. North American grey wolves
  3. Afro-Eurasian panel of humans

Major mode of gene flow in most species: IBD

"…accumulation of local genetic differences under geographically restricted dispersal."



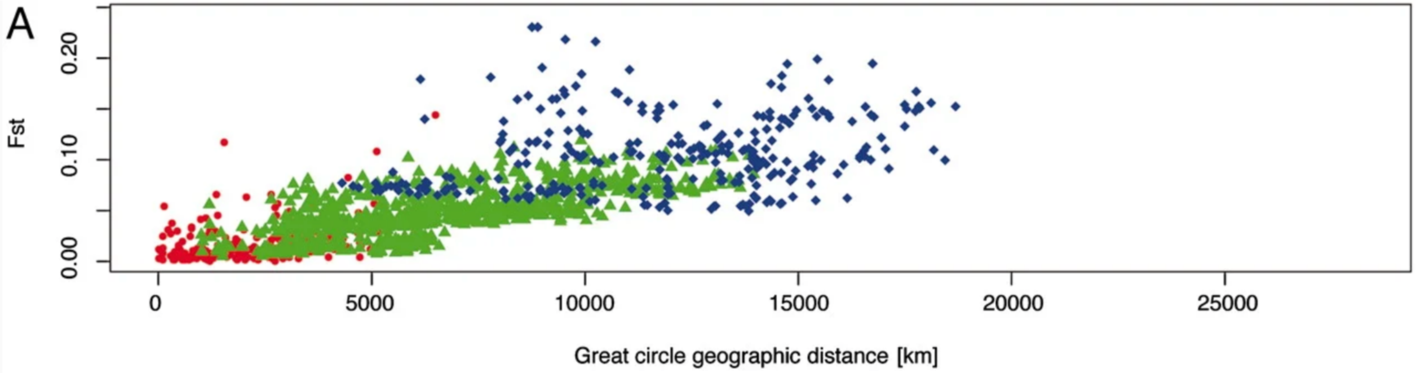
Pattern of IBD



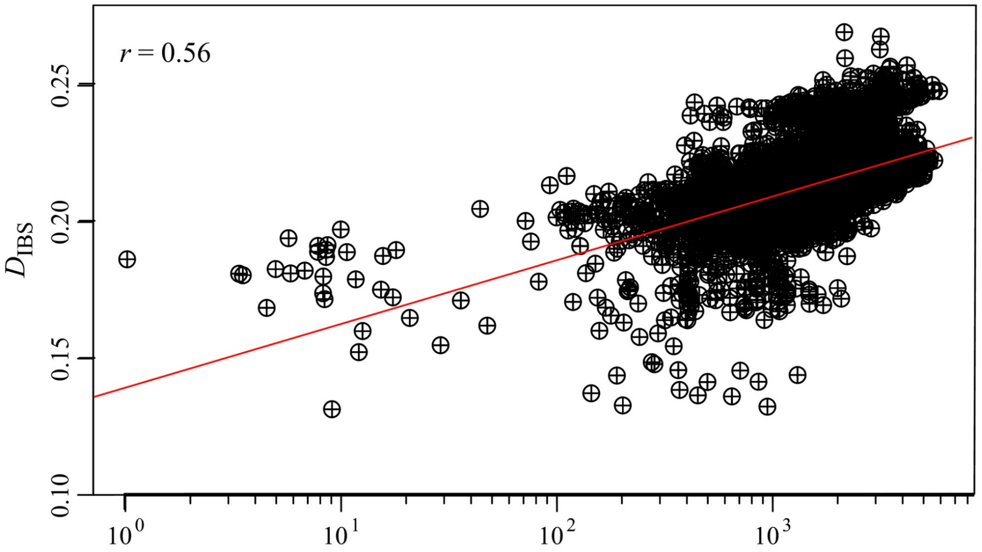
(initially formulated by Wright 1943)

A few examples of IBD in different species

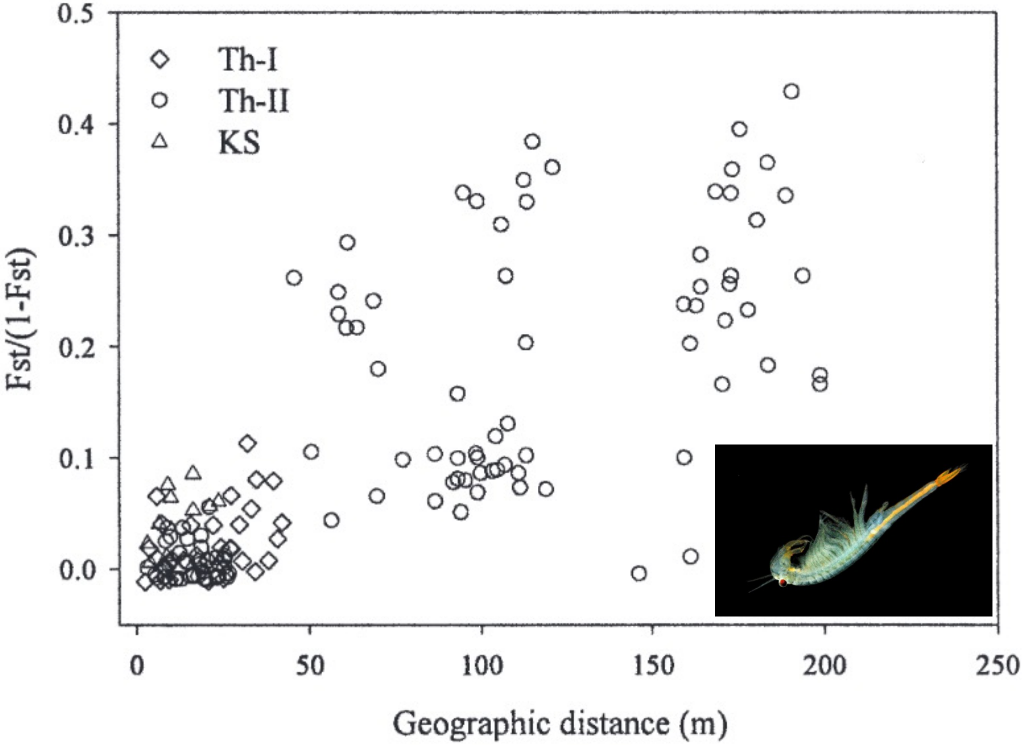
Humans (Ramachandran *et al* 2005)

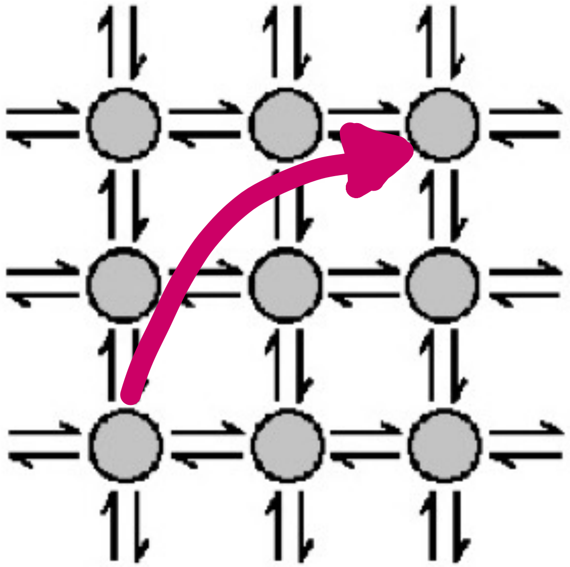


Grey wolves (Schweizer *et al* 2016)



Fairy shrimp (Hulsmans *et al* 2007)





Long-range gene flow events can cause deviations from this pattern of IBD…

Long-range gene flow events can cause deviations from this pattern of IBD…

…but so can a host of other genetic processes like fluctuating population size & assortative mating and other ecological processes like spatially heterogeneous landscapes & habitat fragmentation

Long-range gene flow can be caused by:

- Human-mediated translocations

Wolves in Yellowstone

Long-range gene flow can be caused by:

- Human-mediated translocations

- Extreme weather events

Tornado carrying a cow

Long-range gene flow can be caused by:

- Human-mediated translocations
- Extreme weather events
- Natural migration for greener pastures

Outline

- Introduction + Motivation

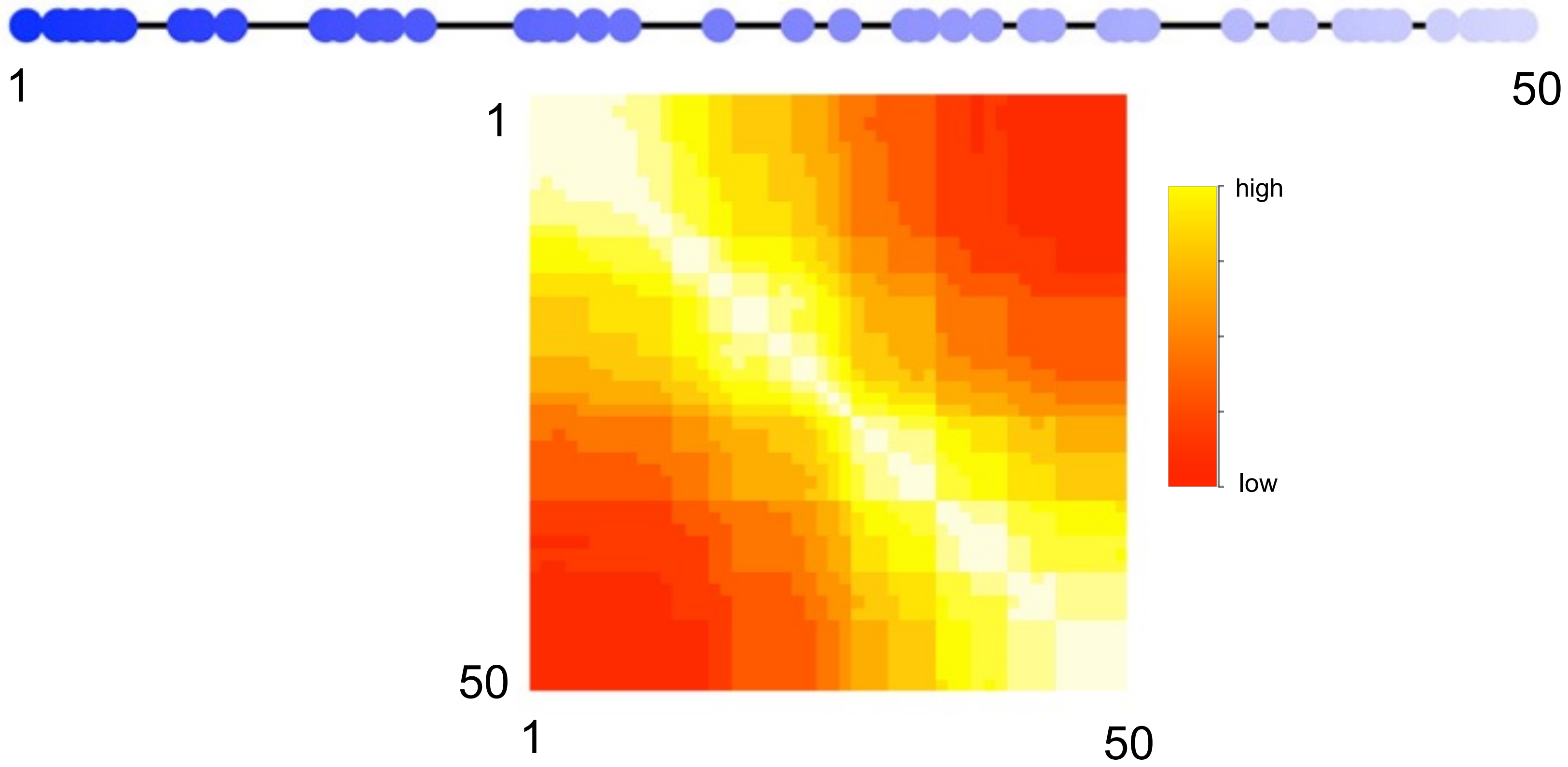
- Spatially heterogeneous models of isolation-by-distance (IBD)

- Our model for long-range gene flow events (`FEEMSmix`)

- Results:
  1. Simulations
  2. North American grey wolves
  3. Afro-Eurasian panel of humans

Spatial models of genetic variation

- Typically, revolves around modeling the **observed covariance matrix** of genotypes with an **expected covariance matrix**

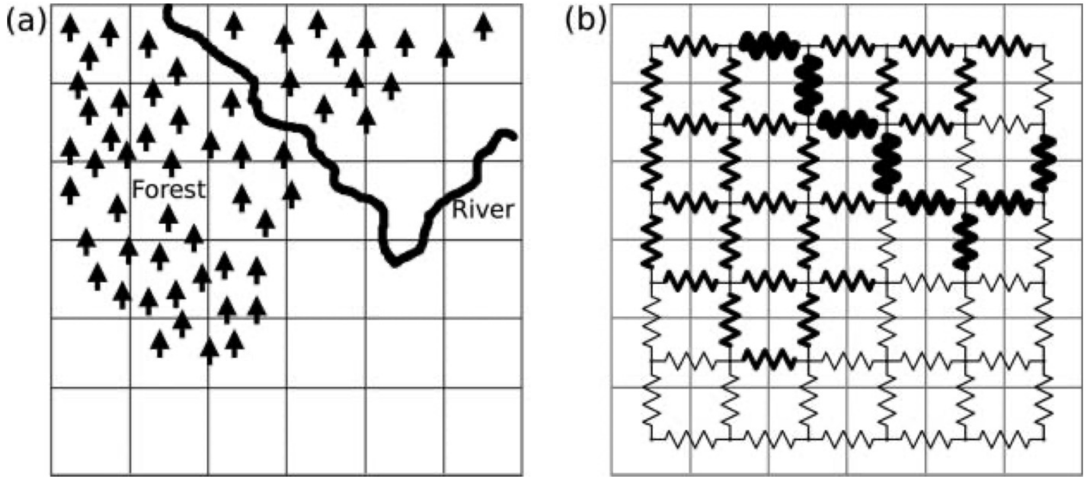


Constructing the expected covariance matrix

1. Parametric variogram (`SpaceMix`)

2. Deep neural networks (`disperseNN` & `Locator`)

3. **Pairwise coalescent times (approximated in `EEMS/FEEMS` using circuit theory)**

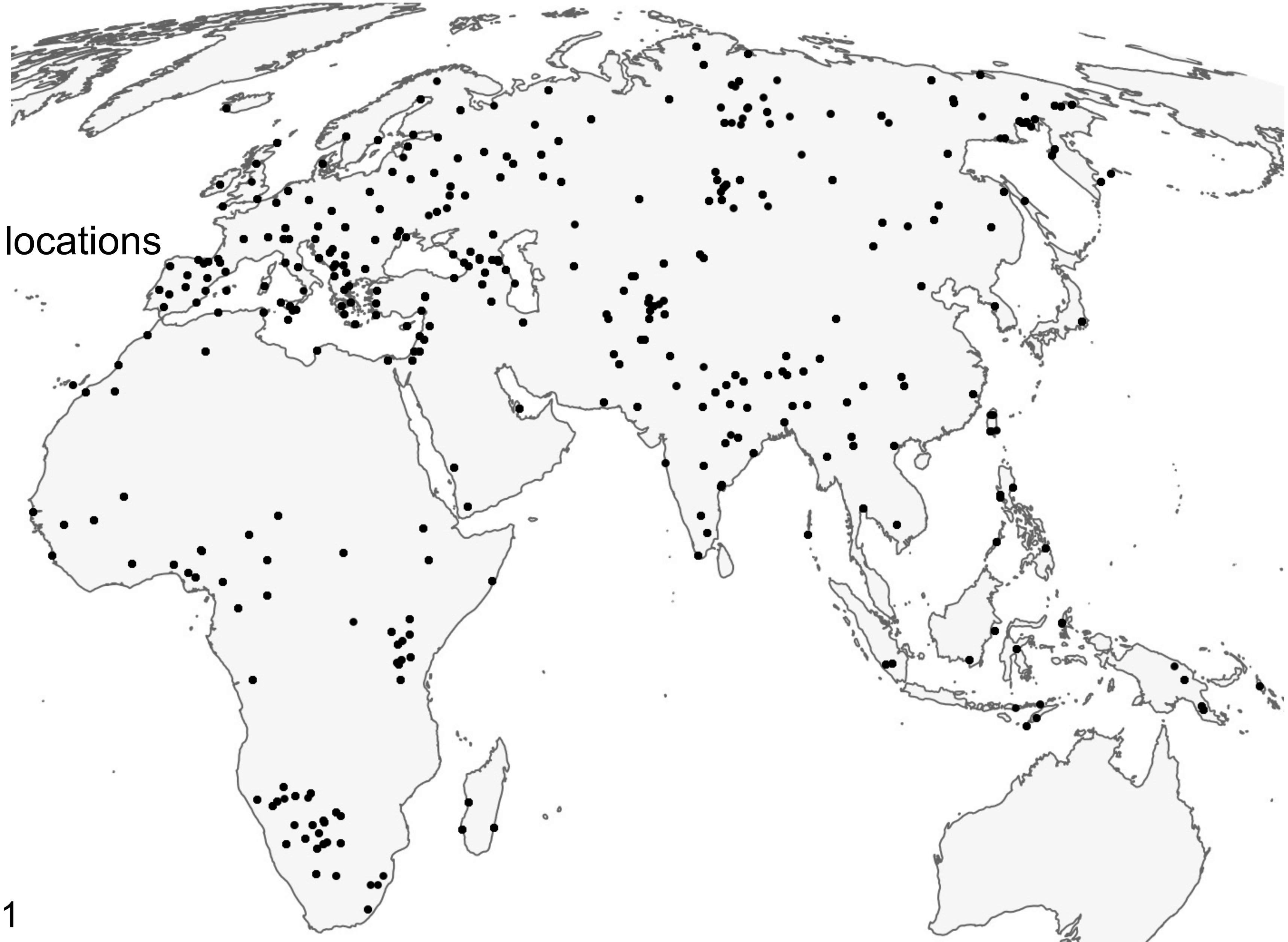


Fast
Estimation of
Effective
Migration
Surfaces

1. Bradburd *et al* 2016
2. Smith *et al* 2020
   Battey *et al* 2020
3. Hanks & Hooten 2013
   Petkova *et al* 2016
   Lundgren & Ralph 2017
   Marcus, Ha *et al* 2021

Brief overview of FEEMS

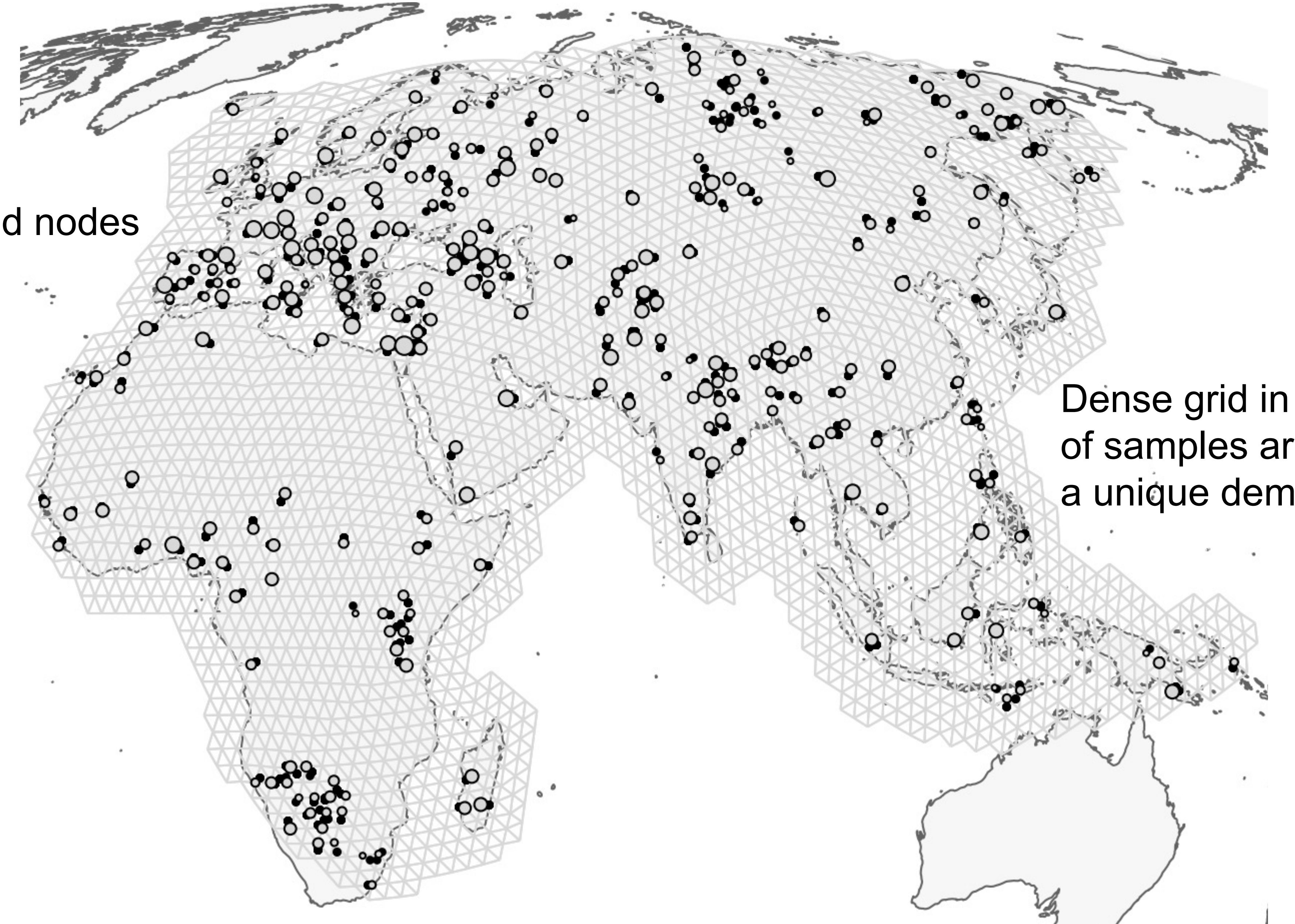
Sample locations



Marcus, Ha *et al* 2021

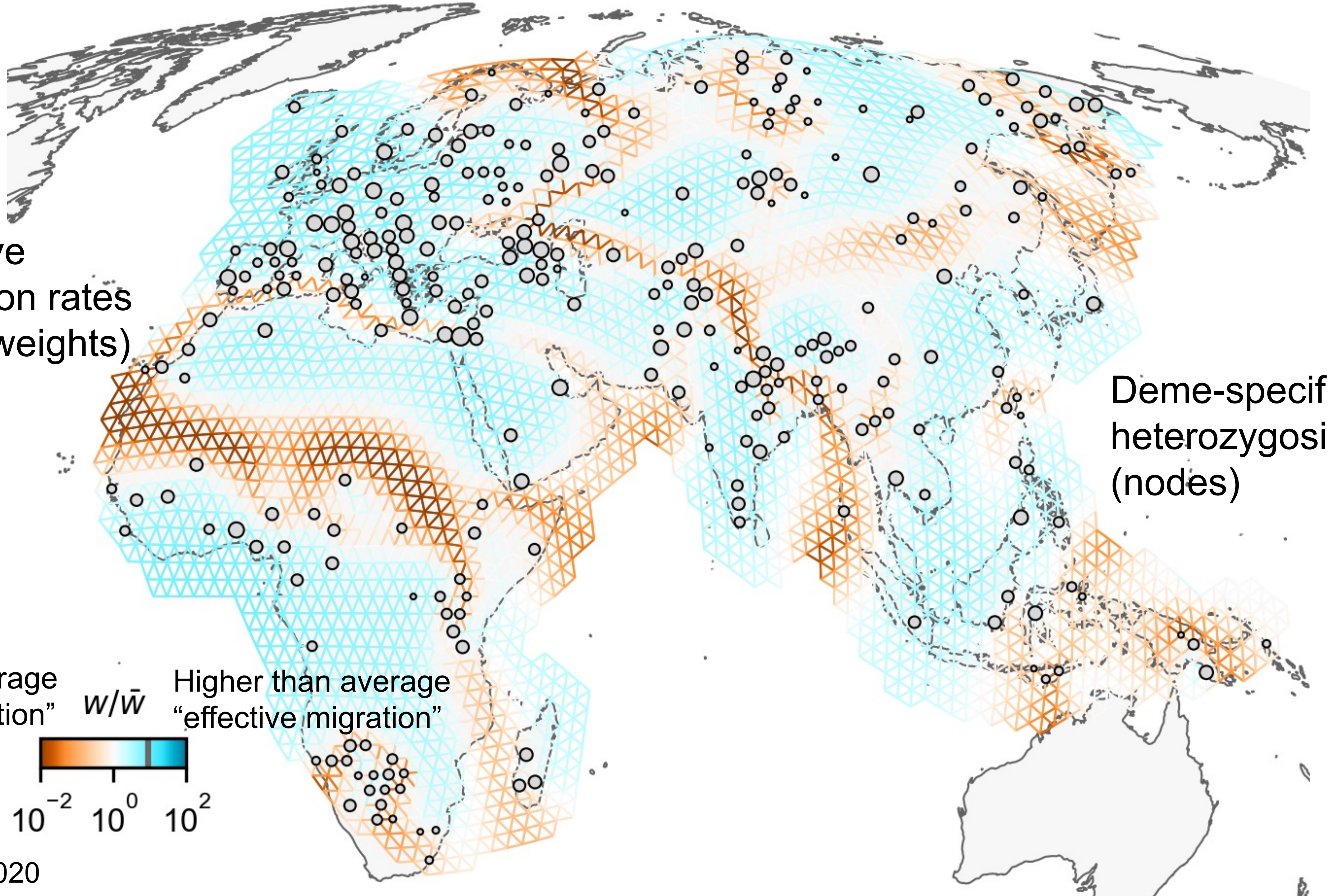
Graph construction & spatial assignment

Observed nodes (demes)



Dense grid in which ~80% of samples are assigned to a unique deme

Penalized likelihood estimation



Effective migration rates (edge weights)

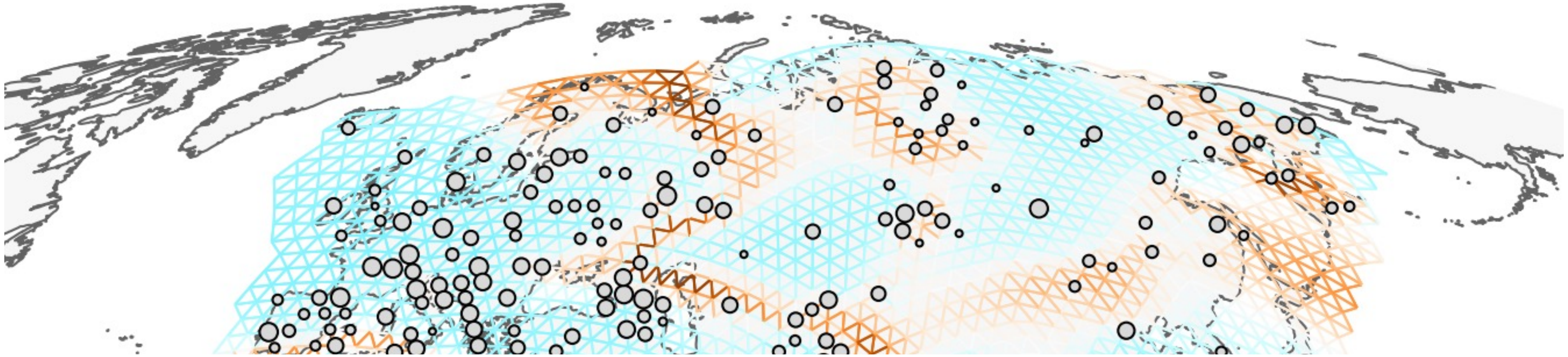
Deme-specific heterozygosity (nodes)

Lower than average “effective migration”

Higher than average “effective migration”

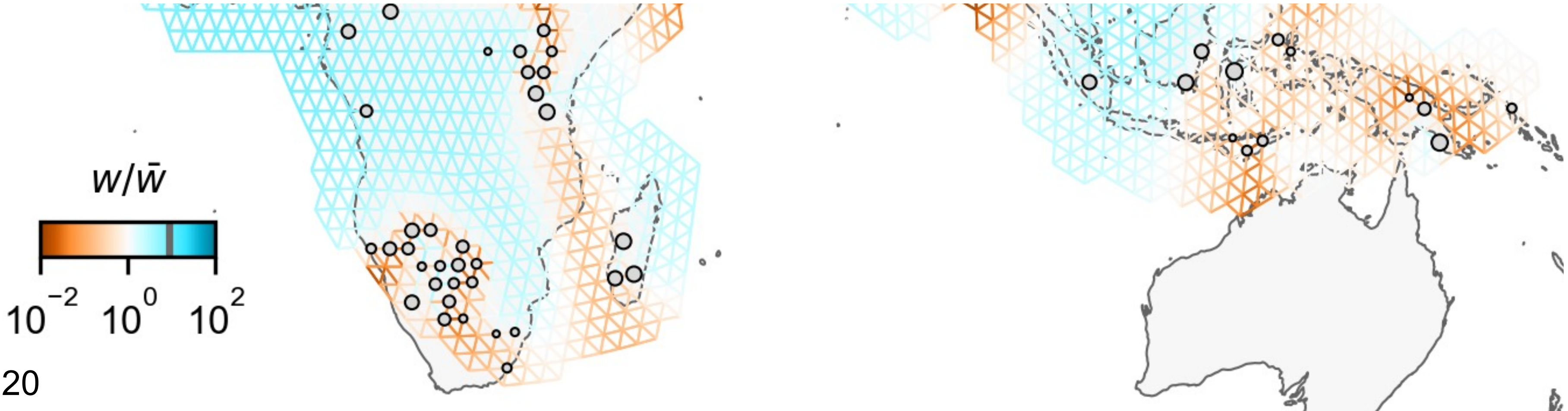
Data from Peter *et al* 2020

Penalized likelihood estimation



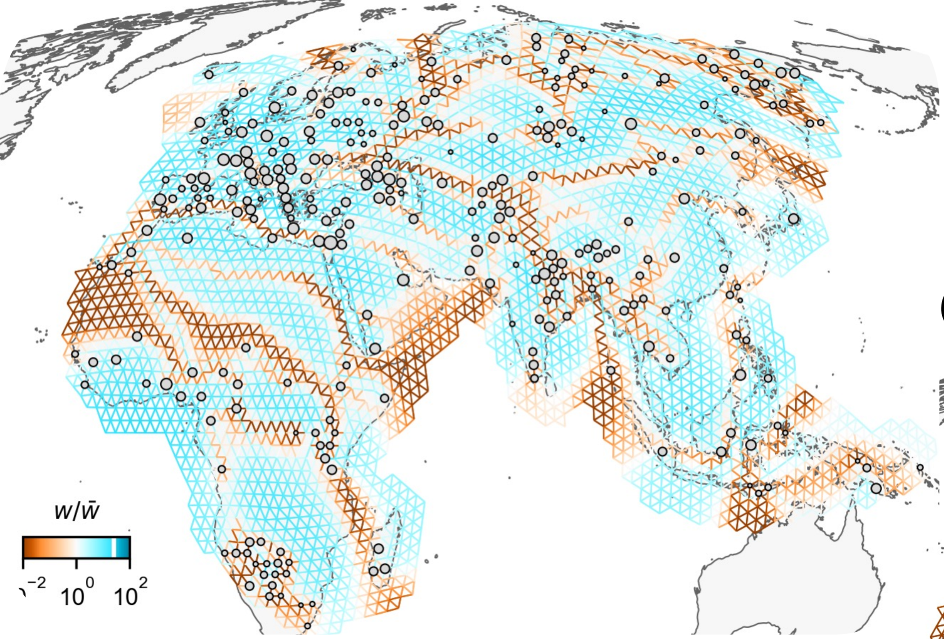
Assumptions:

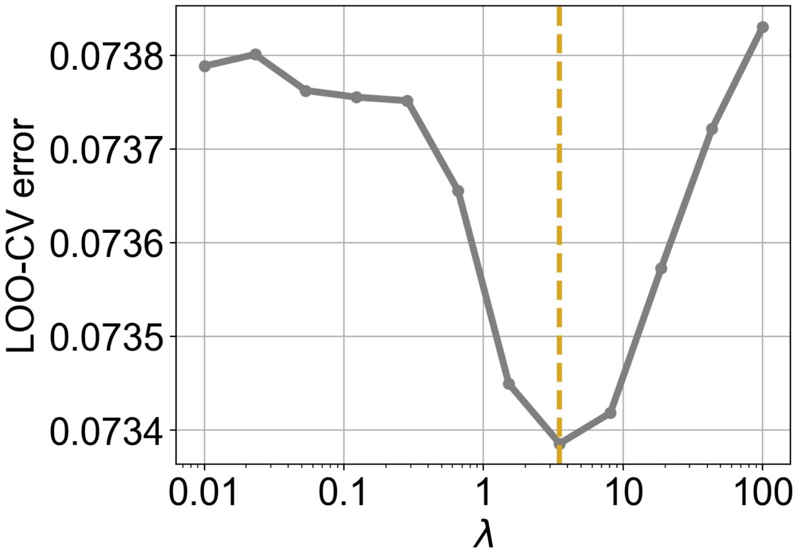
- •Allele frequency in a deme is normally distributed (unlinked SNPs)
- •Individuals are exchangeable within a deme
- •**Symmetric, time-stationary migration rates**



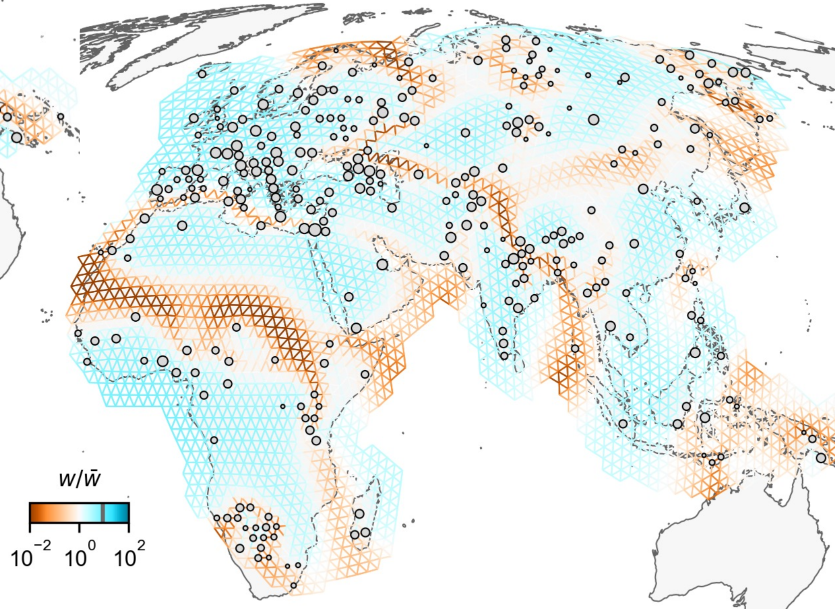
Data from Peter *et al* 2020

Lower λ ➔ discontinuous maps

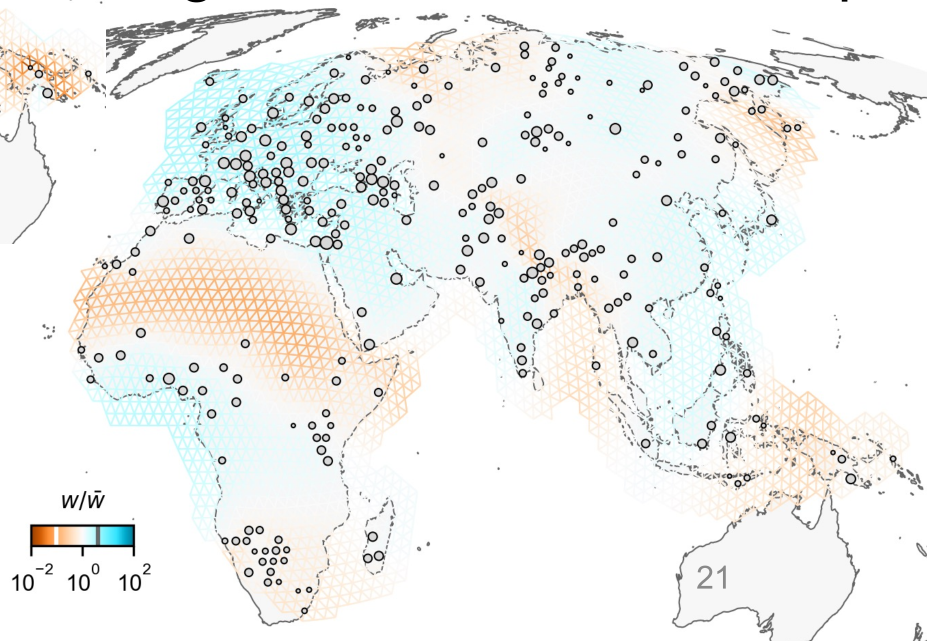




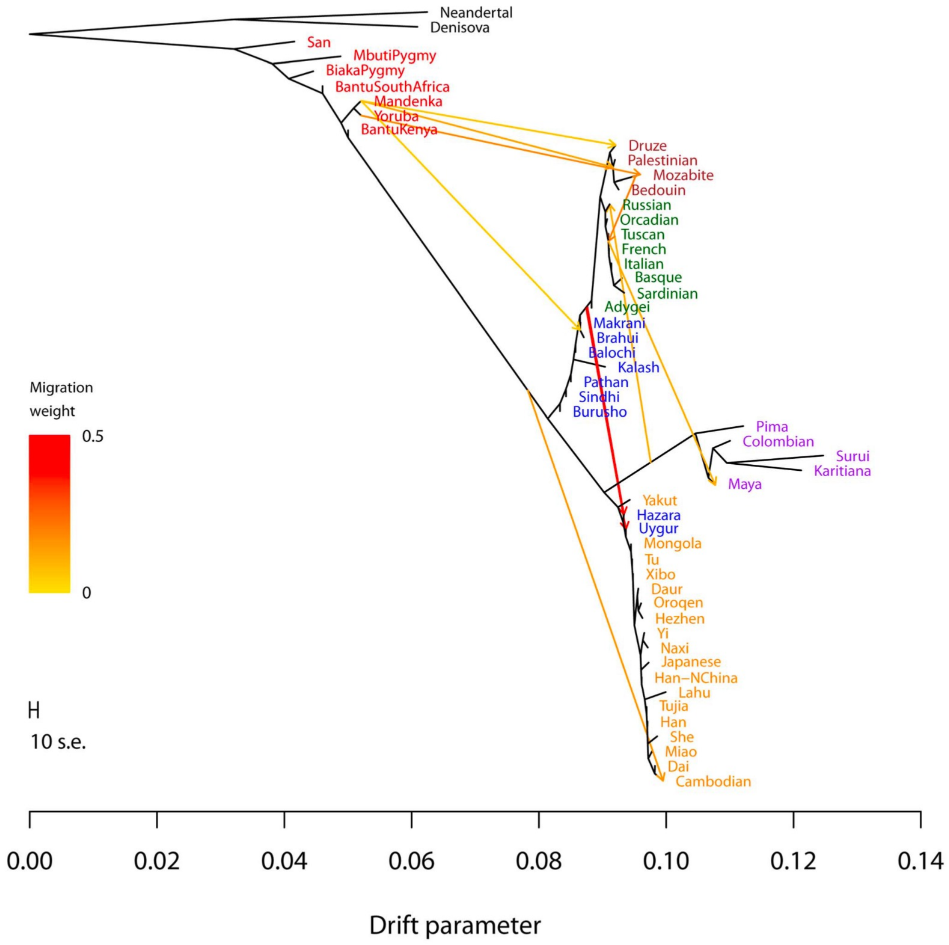
Optimal λ (chosen by CV)



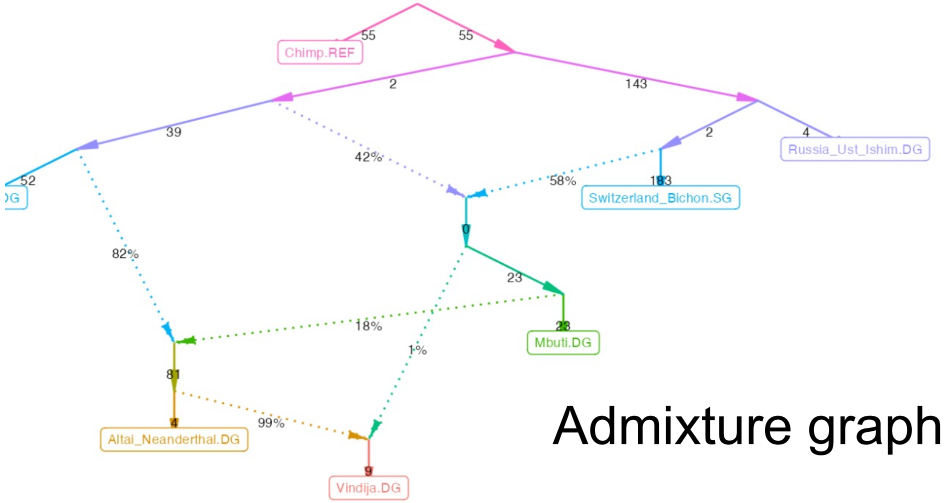
Higher λ ➔ smoother maps



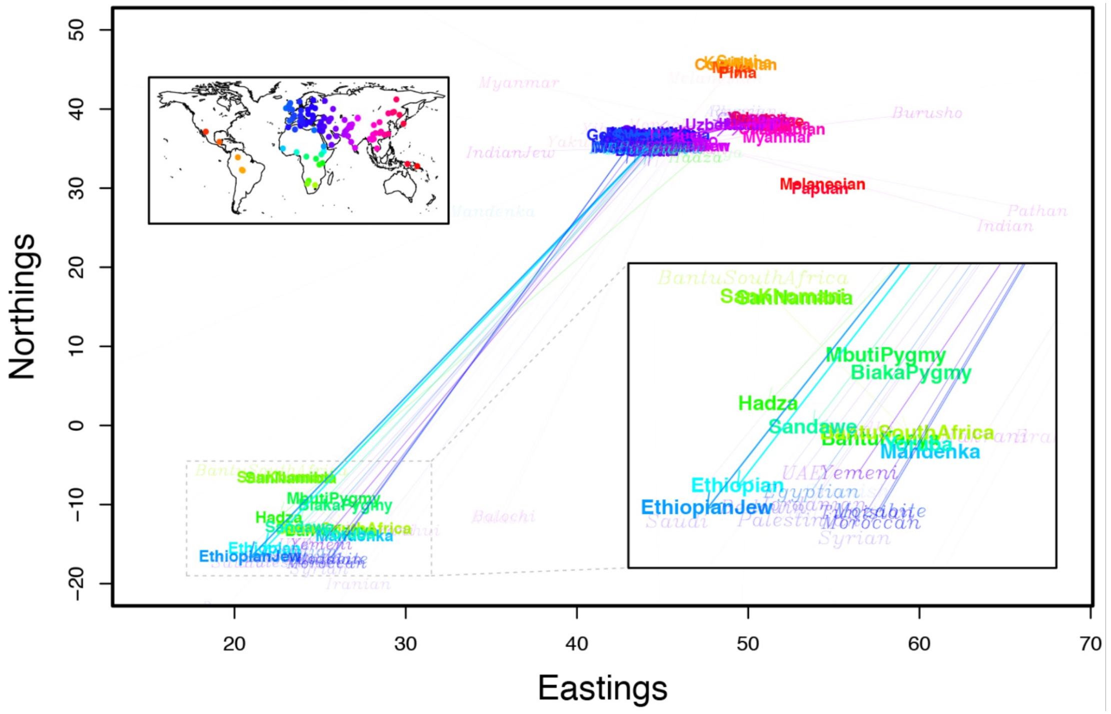
Previous work in estimating “long-range” gene flow events



`TreeMix` (Pickrell & Pritchard 2012)



Admixture graphs (Lipson *et al* 2013)



`SpaceMix` (Bradburd *et al* 2016)

Outline

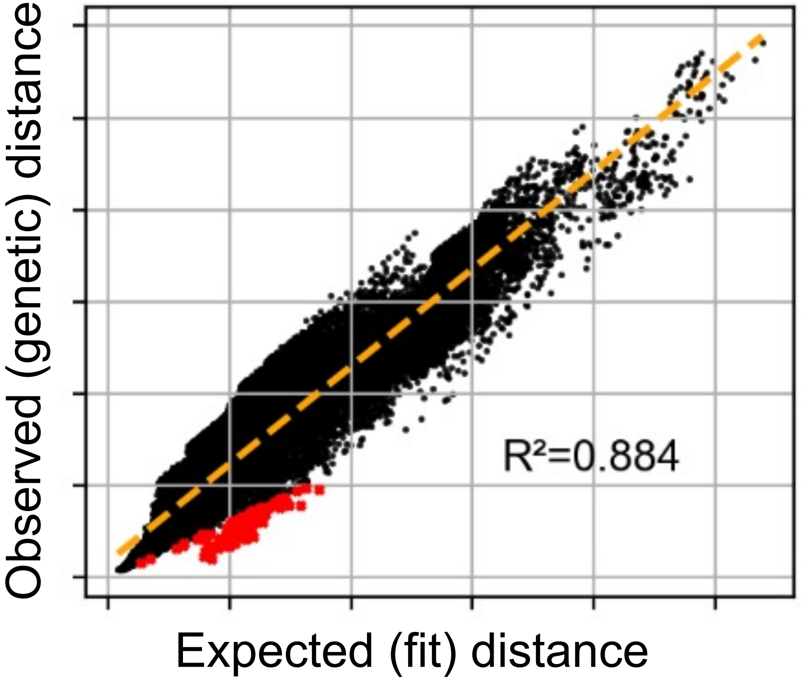
- Introduction + Motivation

- Spatially heterogeneous models of isolation-by-distance (IBD)

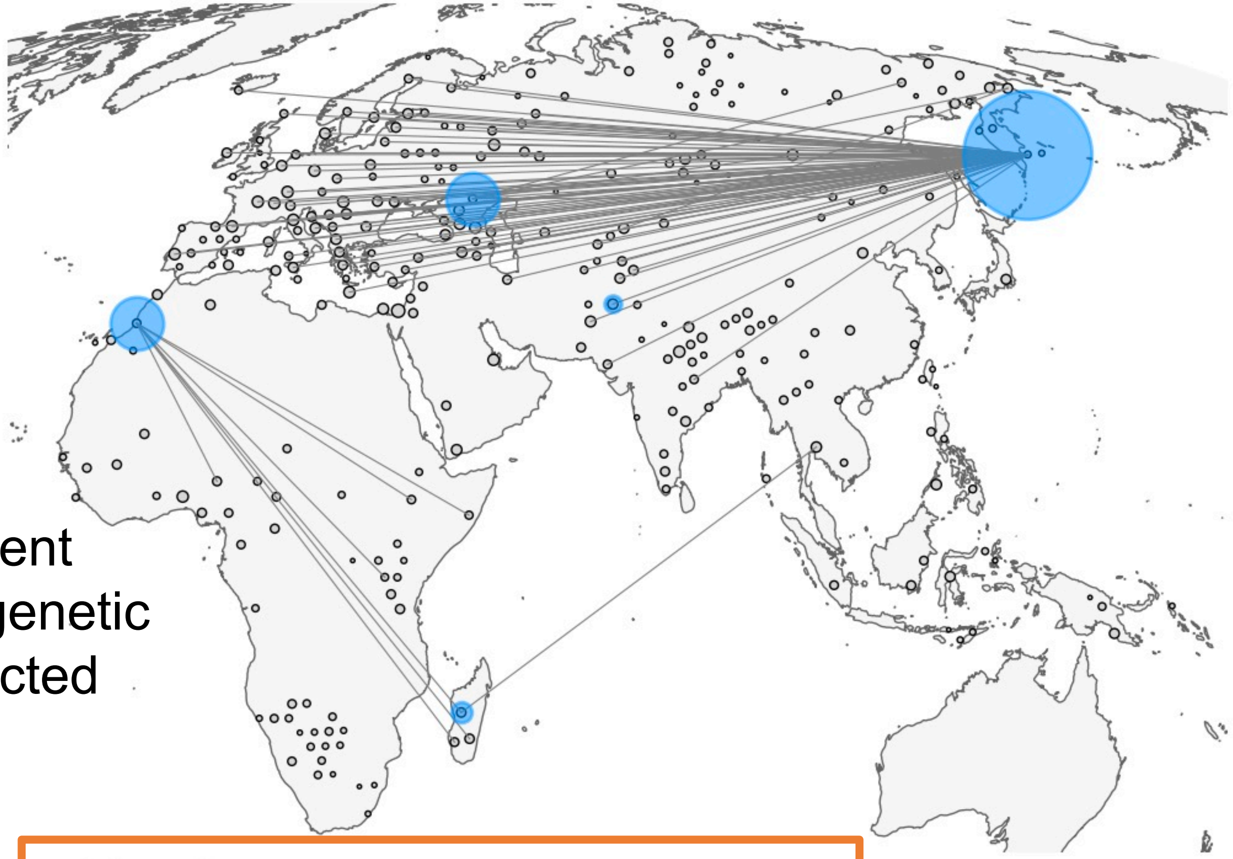
- Our model for long-range gene flow events (`FEEMSmix`)

- Empirical results:
  1. North American grey wolves
  2. Afro-Eurasian panel of humans

FEEMSmix workflow



(red outliers represent pairs with smaller genetic distance than expected under the model)



$$T'_{ss} = T_{ss}$$
$$T'_{sd} = cT_{sd} + (1 - c)T_{ss}$$
$$T'_{dd} = (1 - c)^2 T_{dd} + 2c(1 - c)T_{sd} + c^2 T_{ss}$$

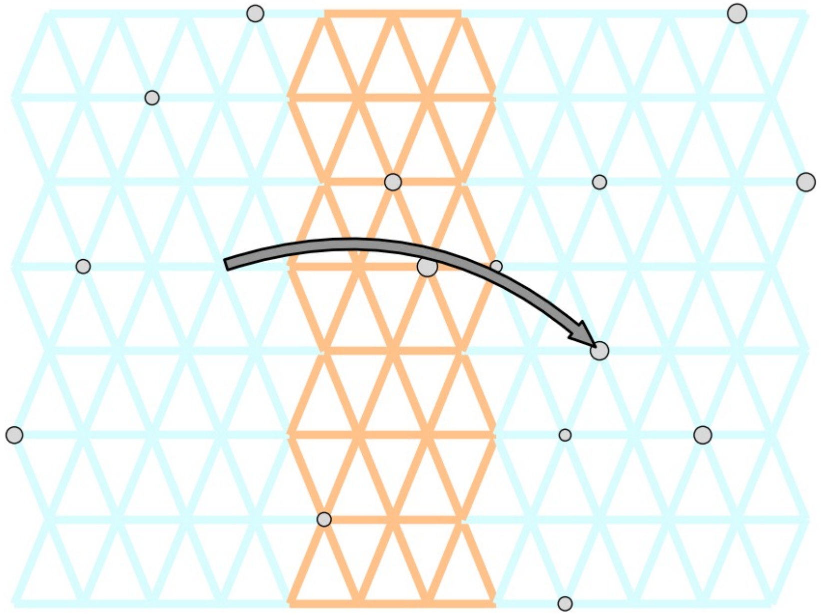
Outline

- Introduction + Motivation
- Spatially heterogeneous models of isolation-by-distance (IBD)
- Our model for long-range gene flow events (`FEEMSmix`)
- Results:
  1. Simulations
  2. North American grey wolves
  3. Afro-Eurasian panel of humans

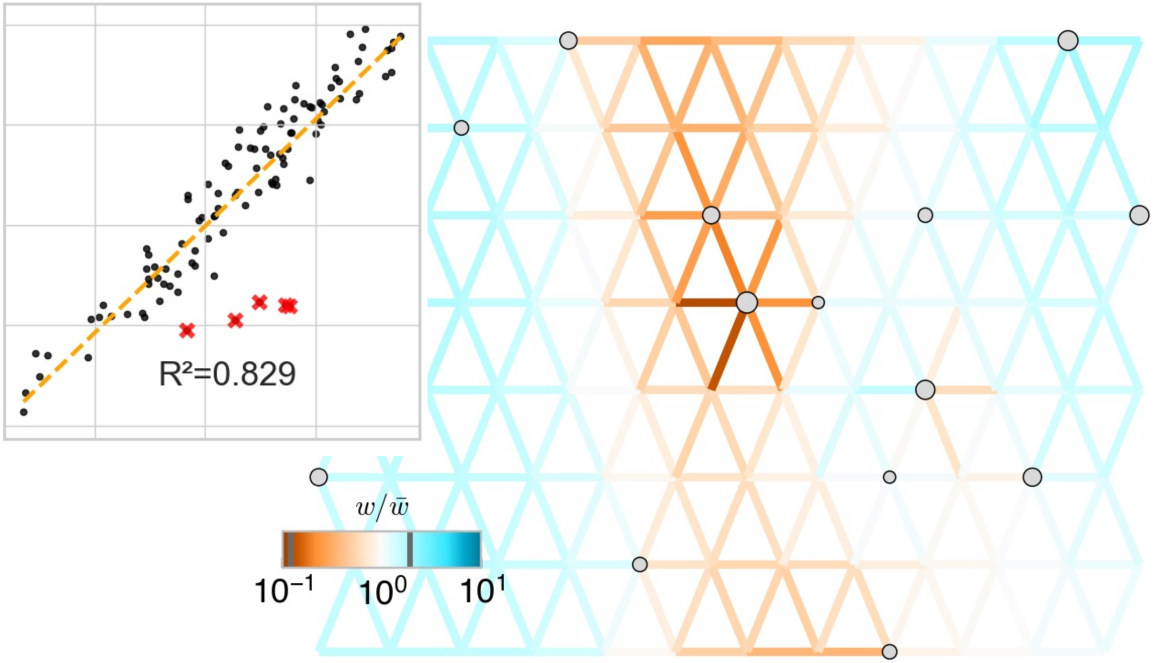
Brief simulation results

- 8x12 grid (only 15% sampled)
- 1-10 samples/deme
- 1,000 SNPs
- Corridor *m* is 10x barrier *m*
- Varying population size *N* across grid

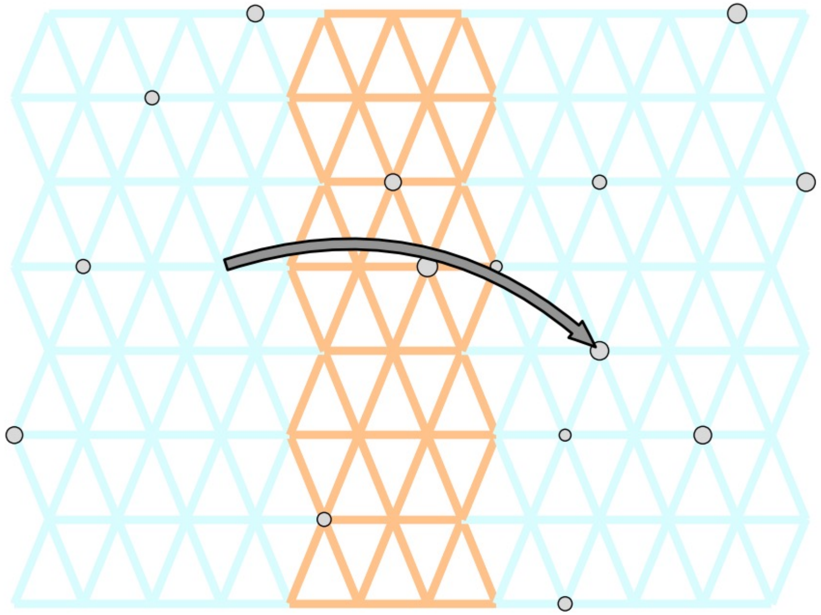
F_{ST} across barrier ≈ 0.1



Simulated truth
(instantaneous pulse of *c* = 0.5)

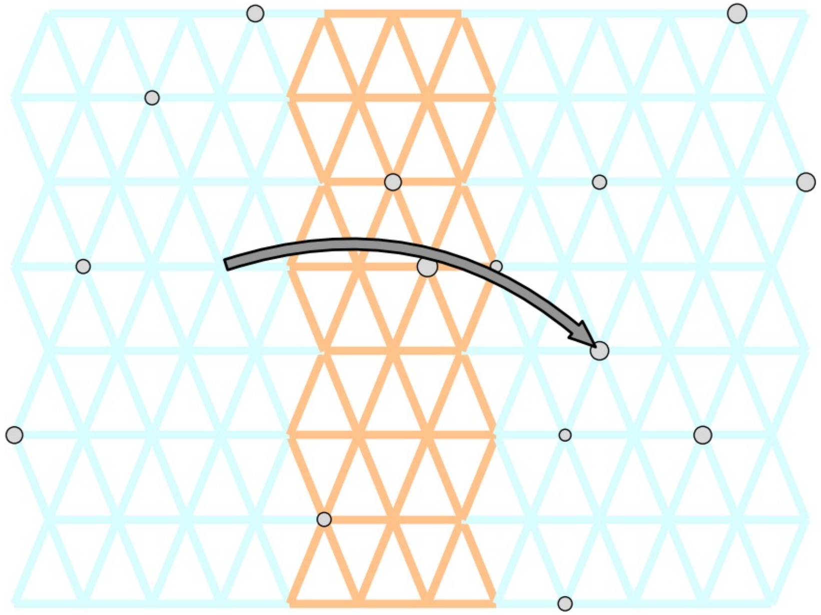


Initial FEEMS fit

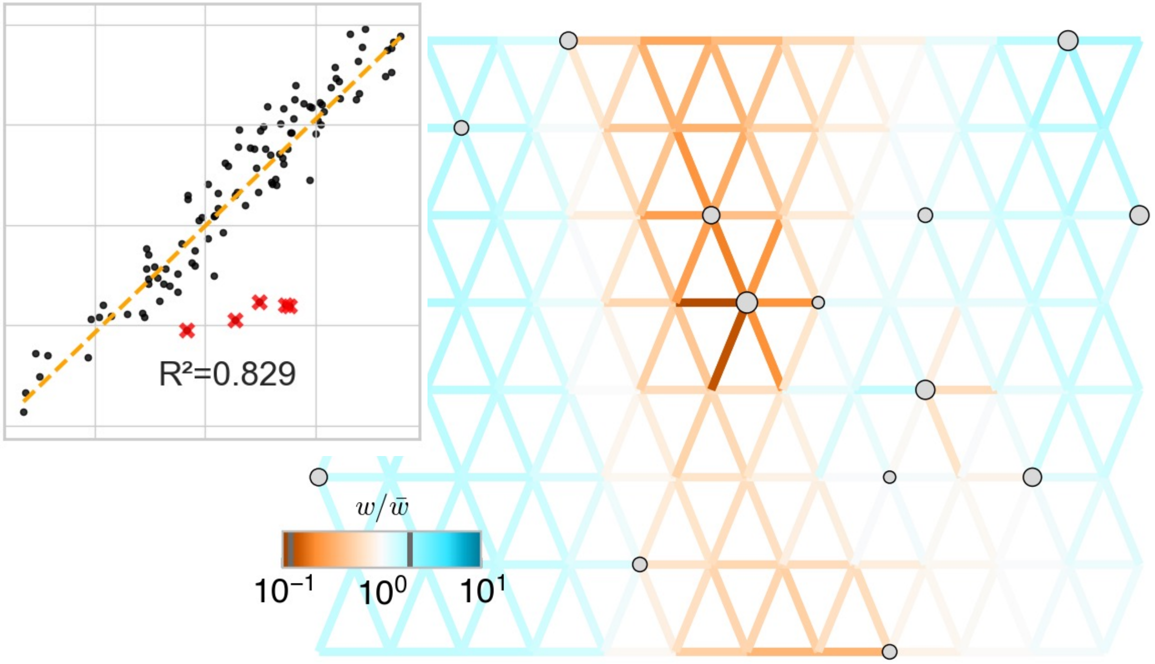


Simulated truth

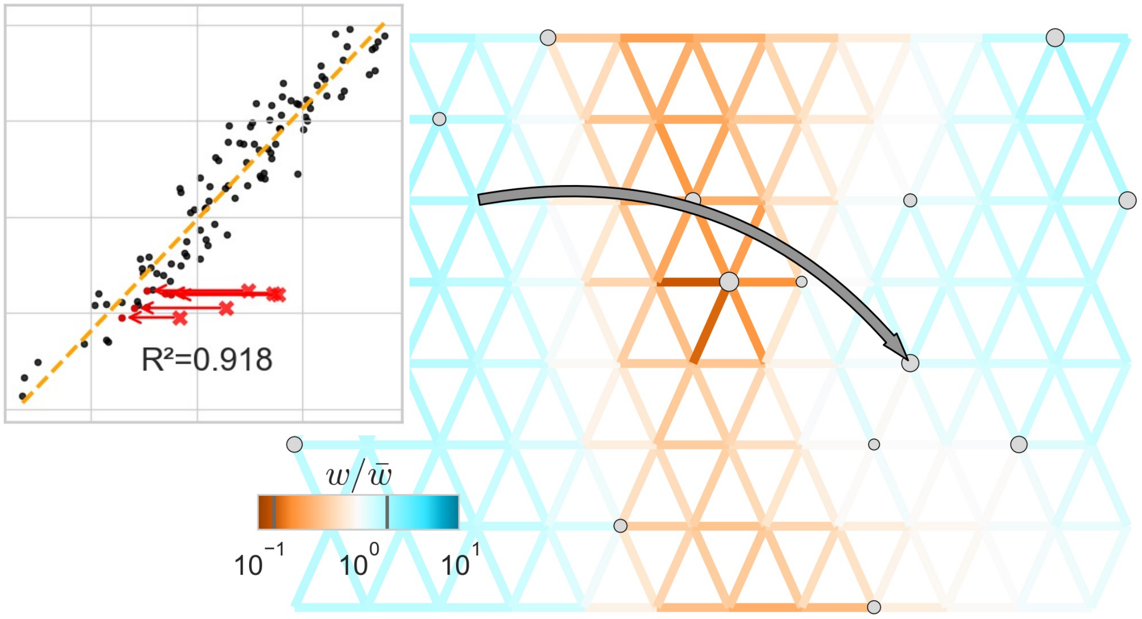
F_{ST} across barrier ≈ 0.1



Simulated truth

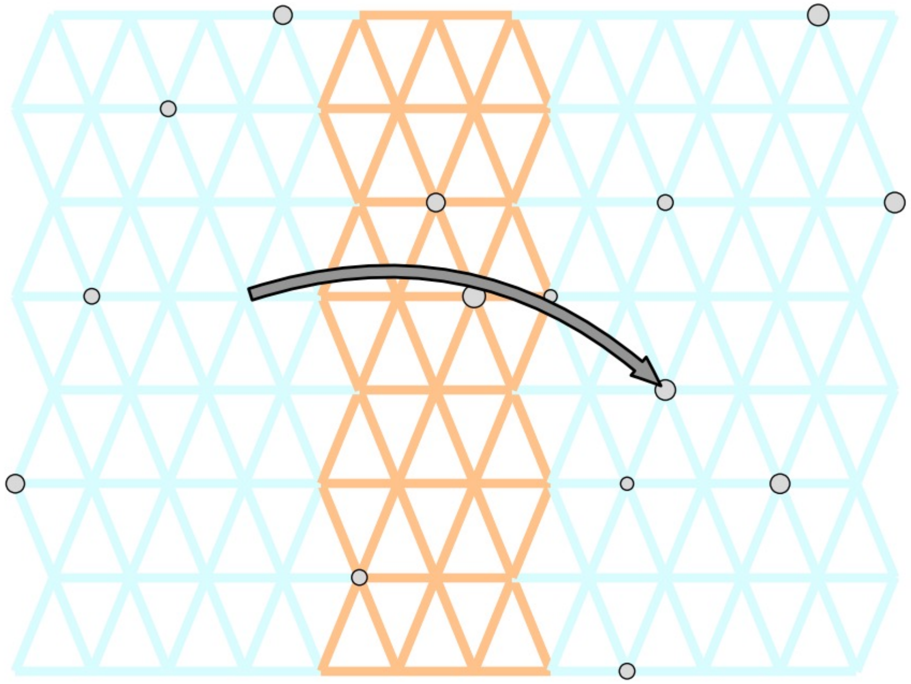


Initial `FEEMS` fit

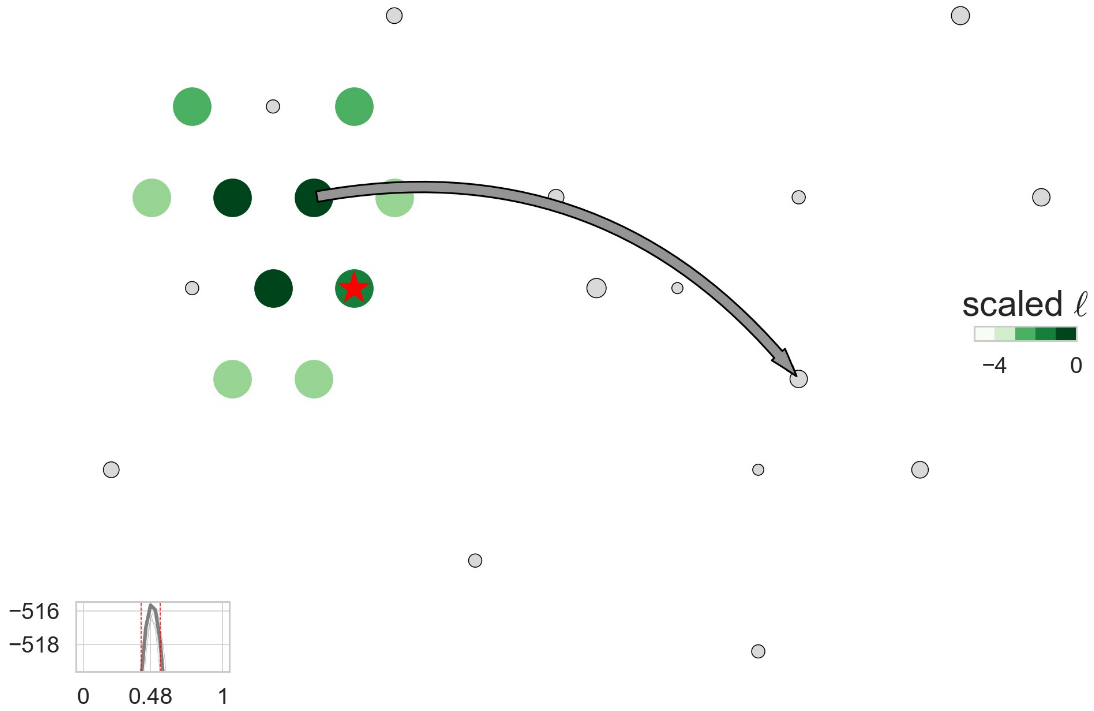


`FEEMSmix` fit

True source is within <u>two</u> log-likelihood units of MLE source



Simulated truth
(c = 0.5)



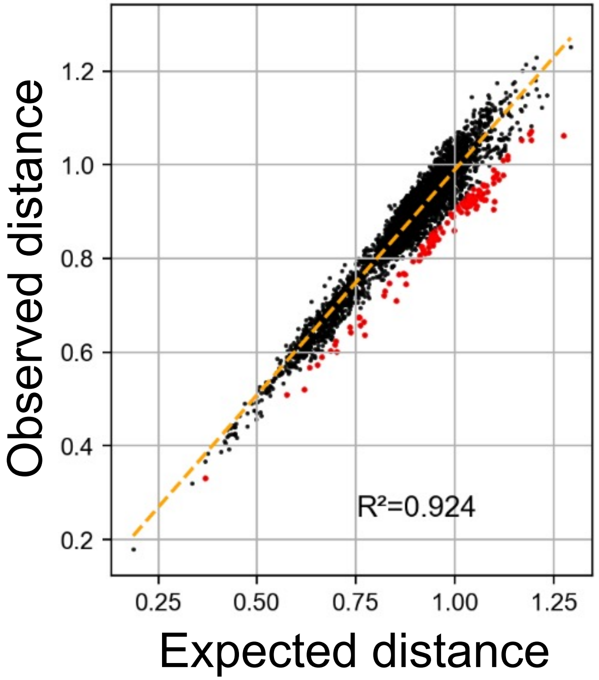
Estimated `FEEMSmix` contour
(est. c = 0.48)

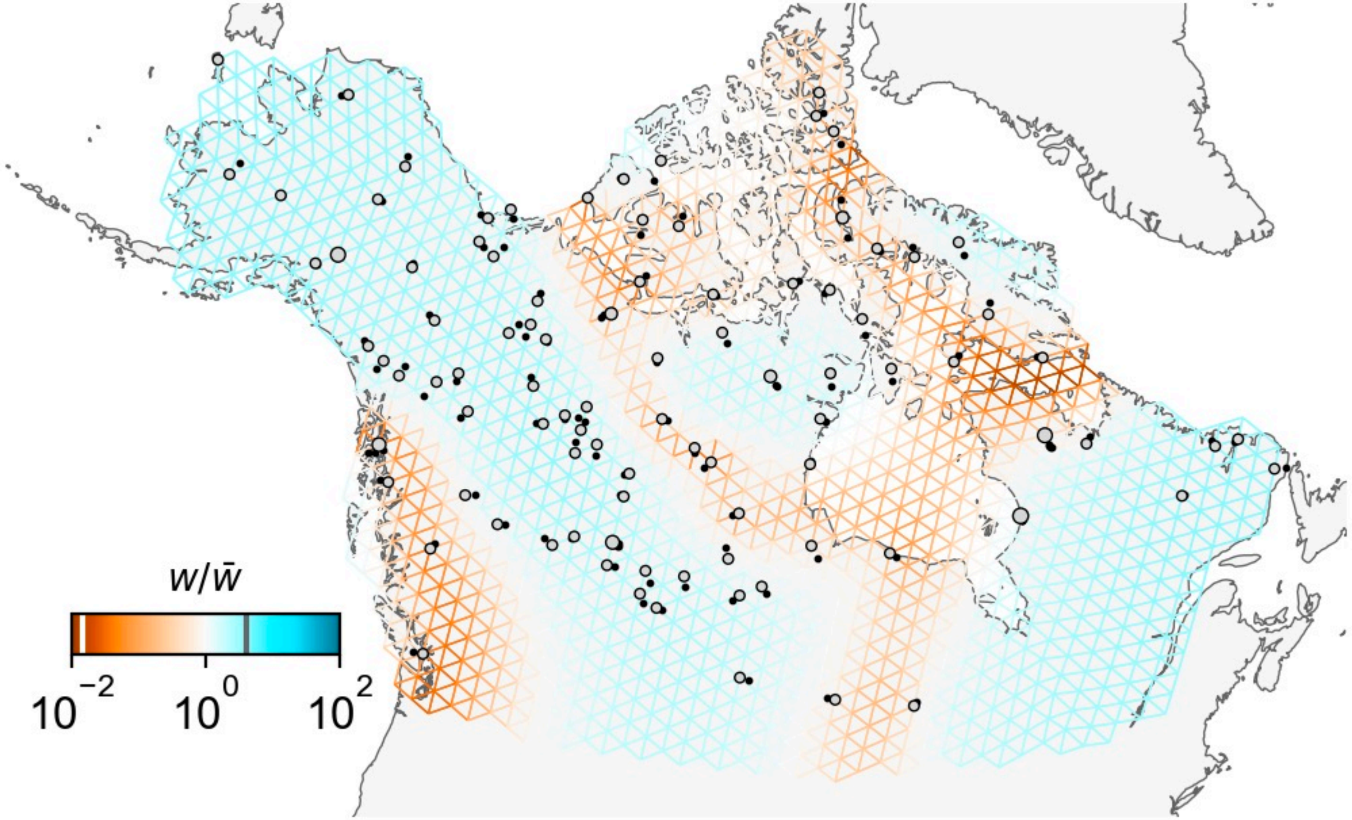
Outline

- Introduction + Motivation
- Spatially heterogeneous models of isolation-by-distance (IBD)
- Our model for long-range gene flow events (`FEEMSmix`)
- Results:
    1. Simulations
    2. North American grey wolves
    3. Afro-Eurasian panel of humans

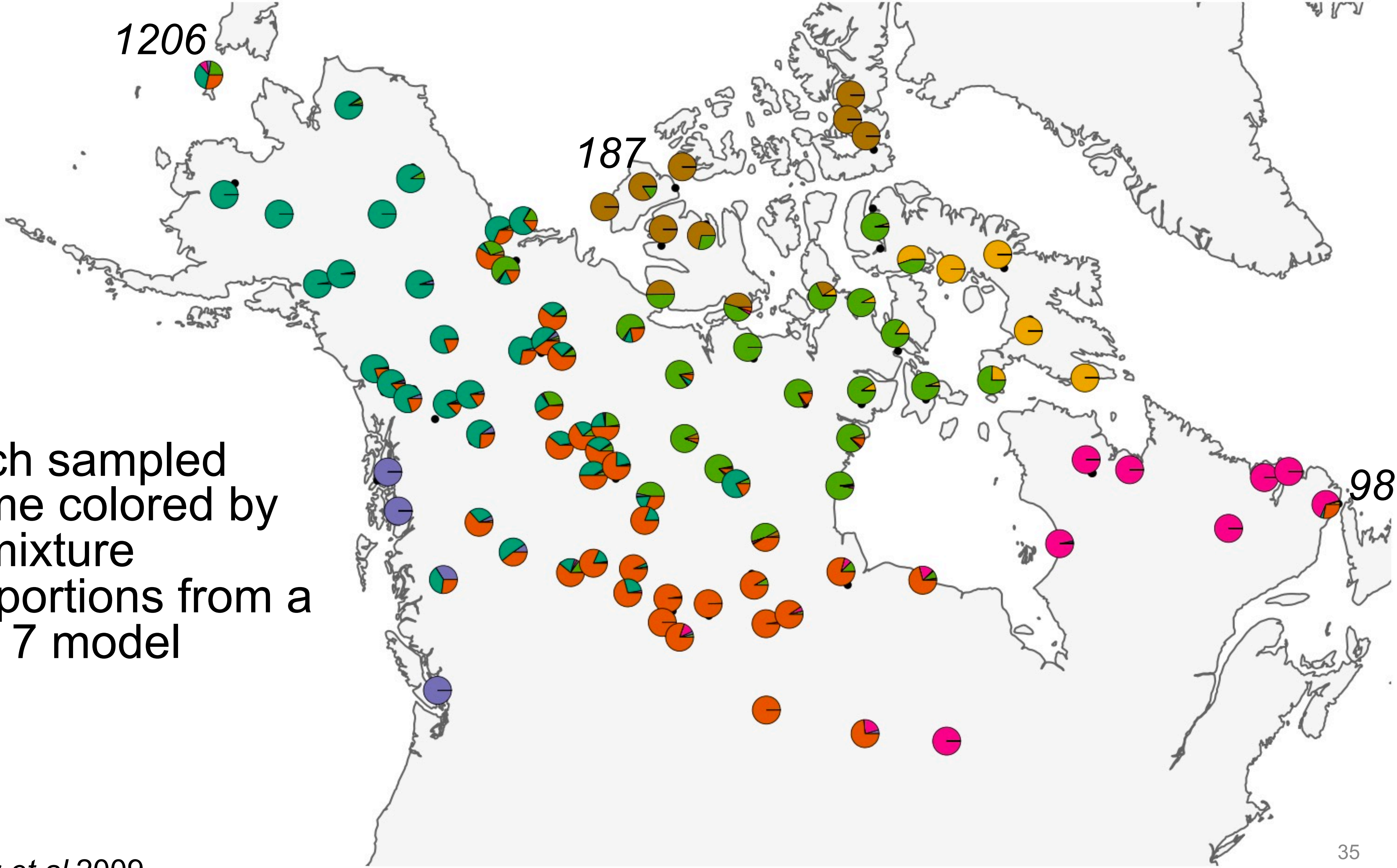
North American grey wolves

- 111 samples
- 17.8k SNPs
- 94 demes (~1 sample/deme)



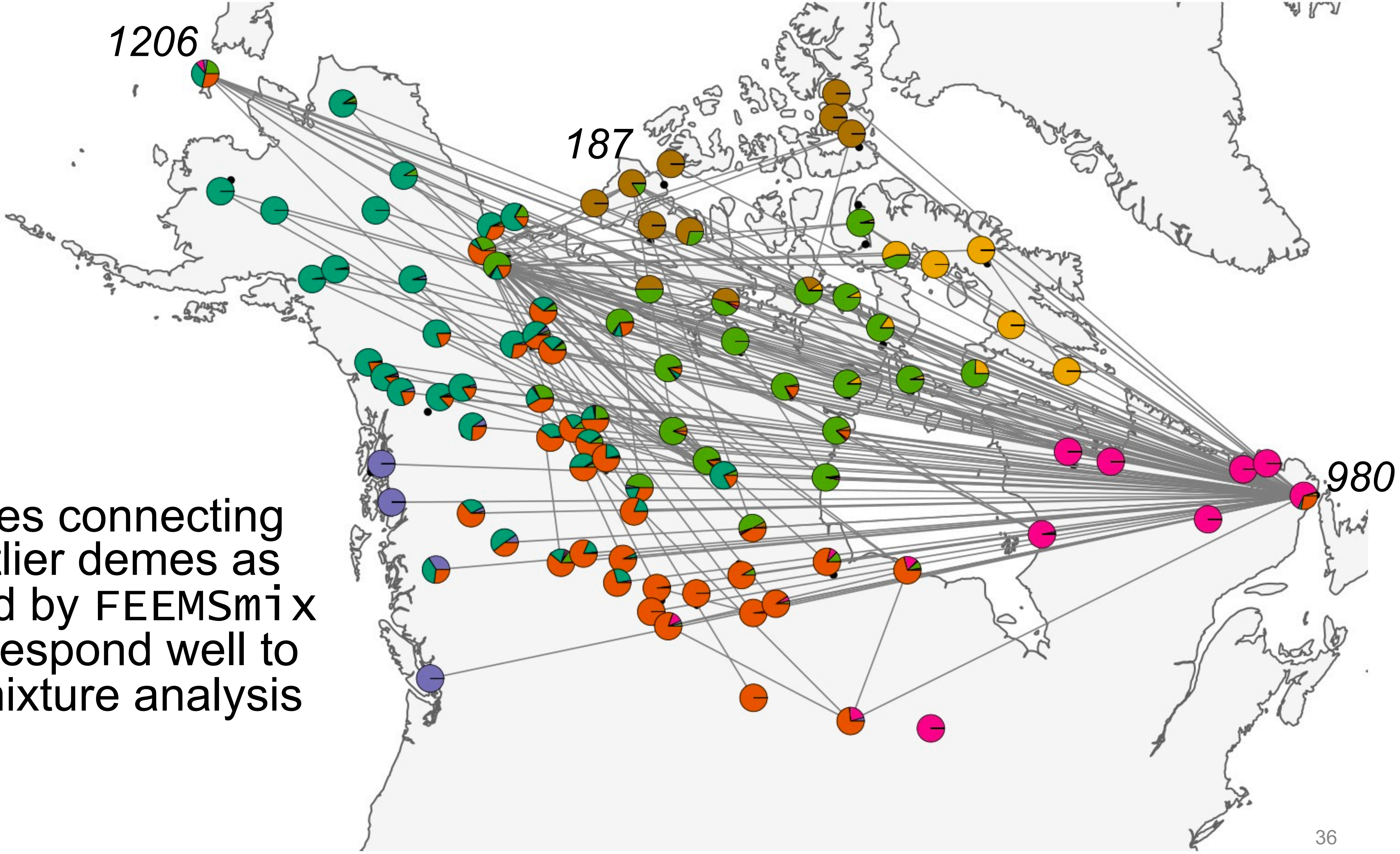


Schweizer *et al* 2016



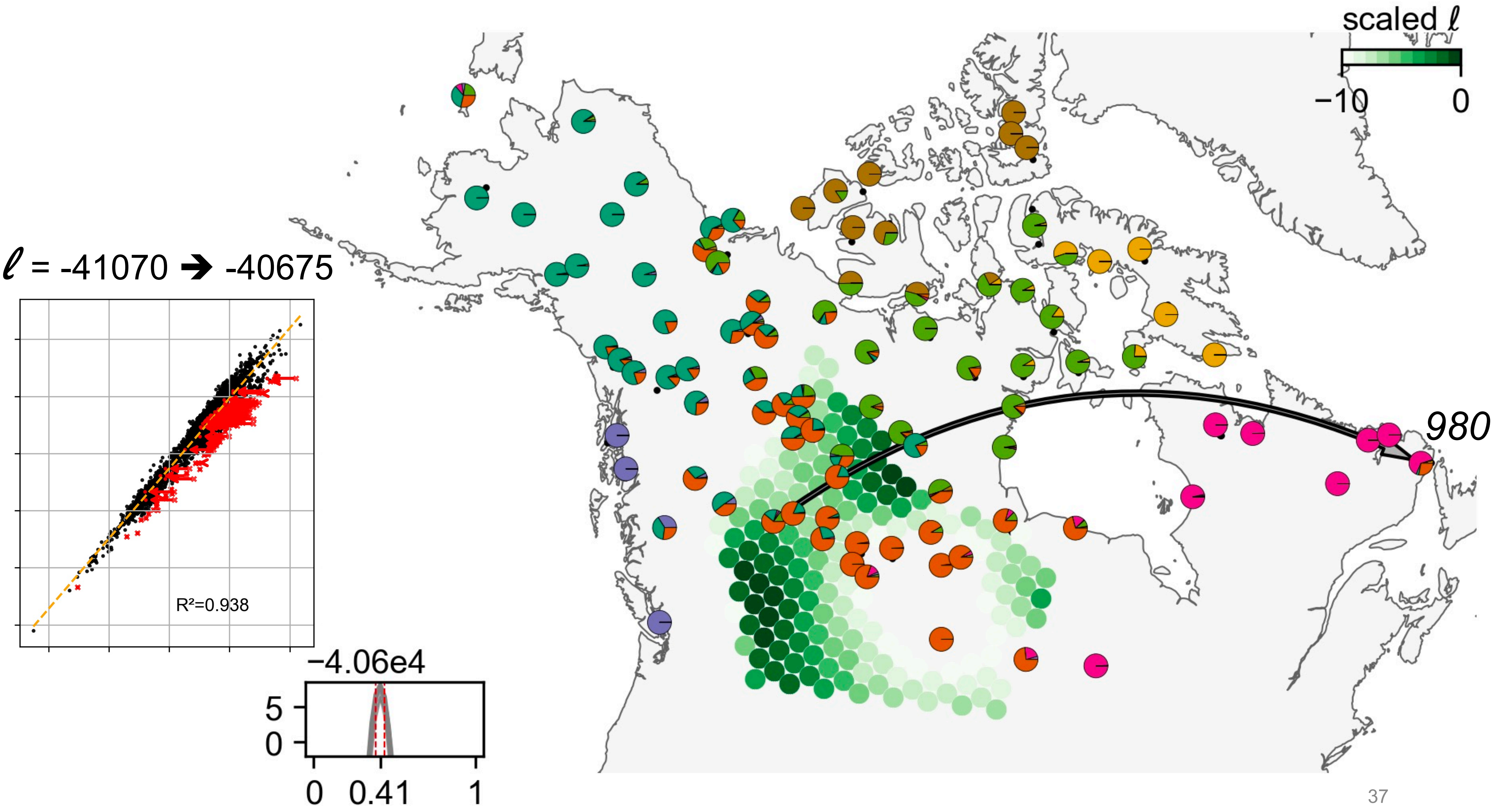
Each sampled deme colored by admixture proportions from a *K* = 7 model

Alexander *et al* 2009

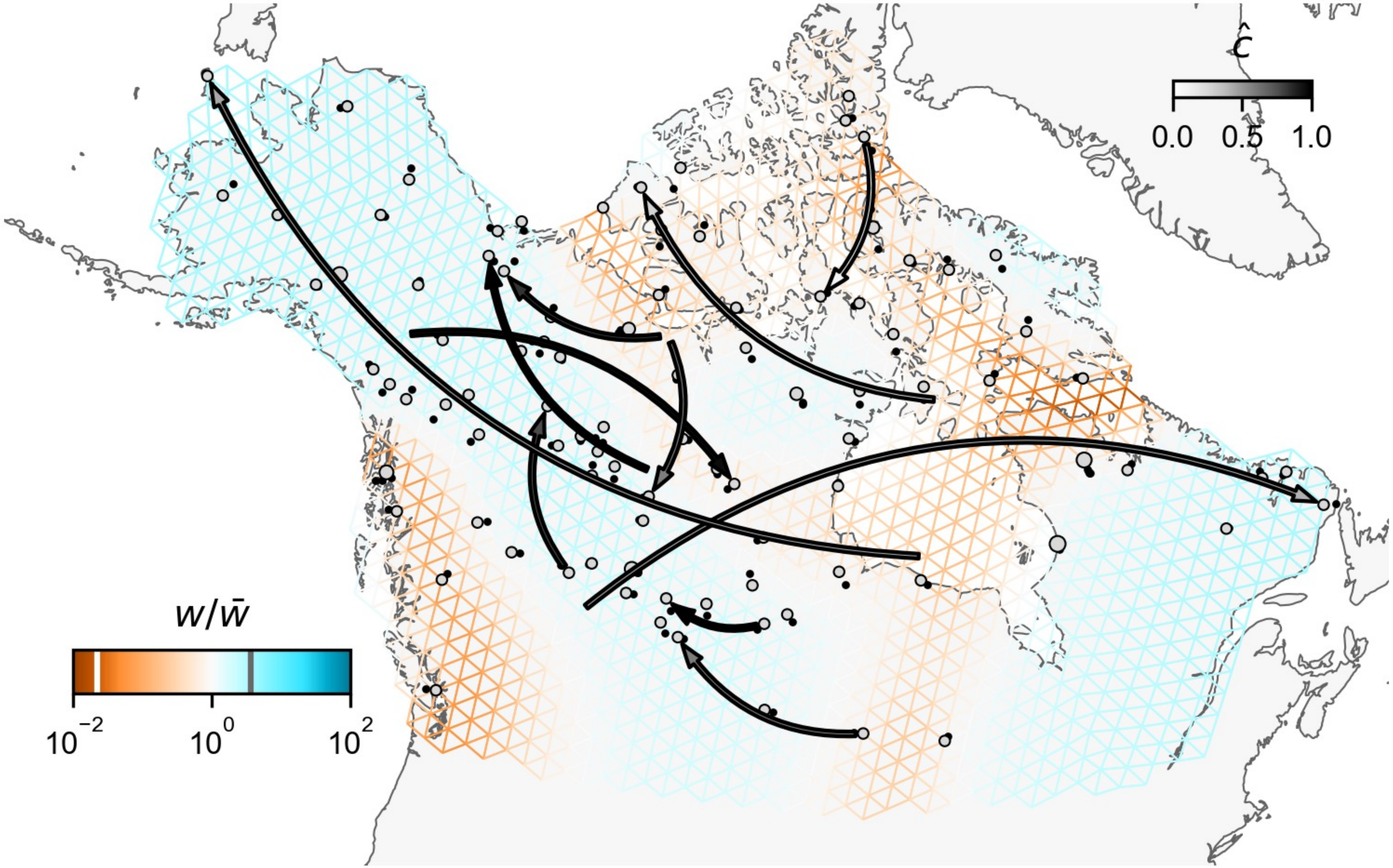


Lines connecting outlier demes as found by `FEEMSmix` correspond well to admixture analysis

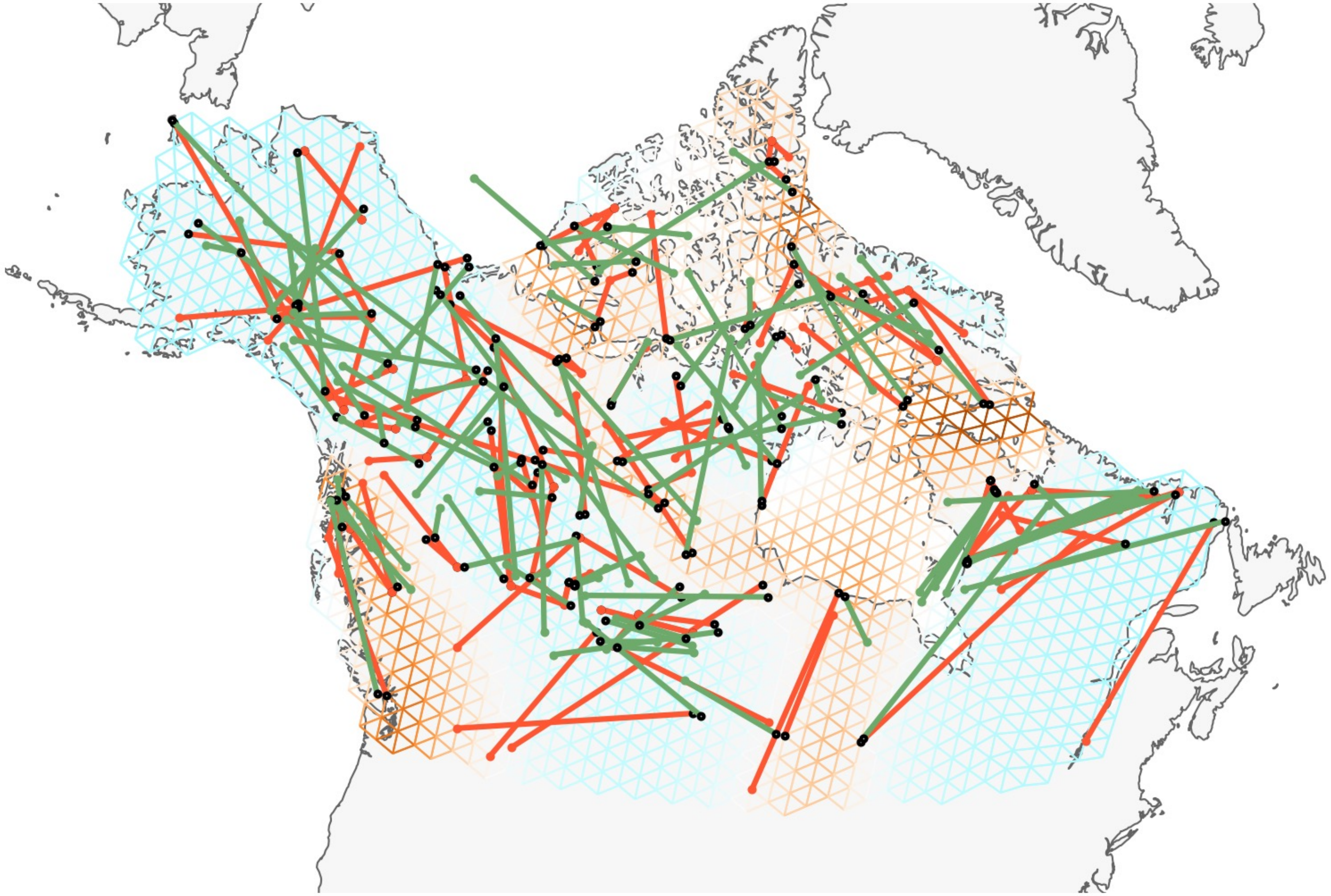
ℓ = -41070 ➔ -40675



The full picture: wolves move around a lot

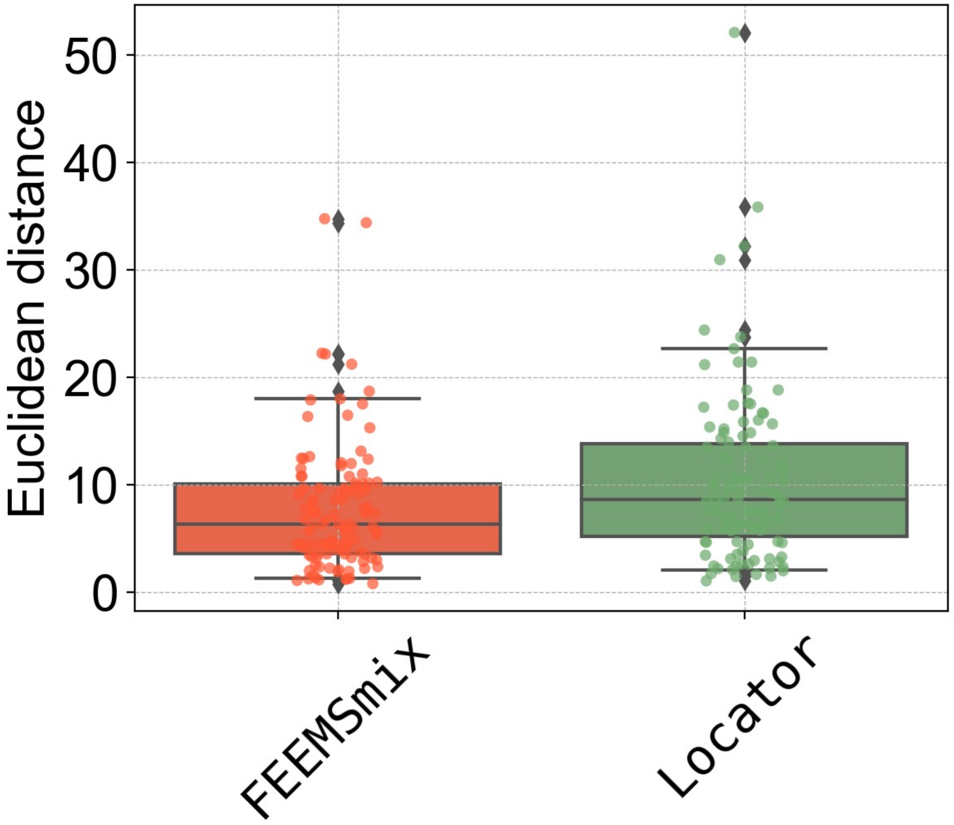


Bonus: spatial prediction!



Median error

~6.4 km ~8.6 km



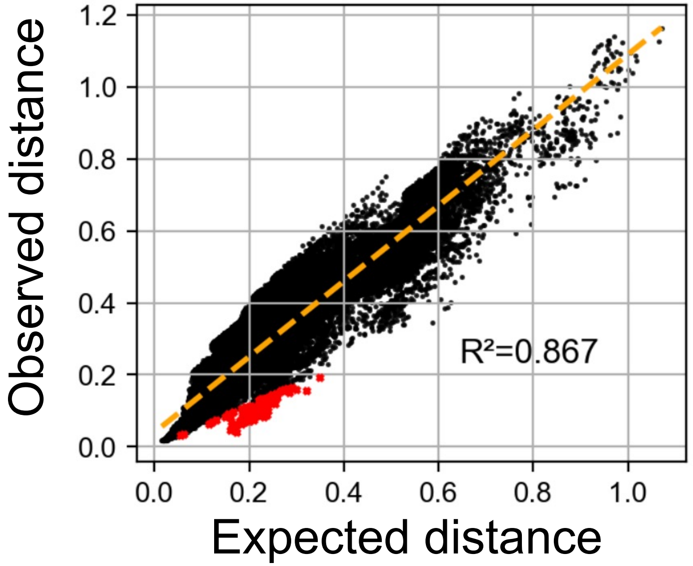
Battey *et al* 2020

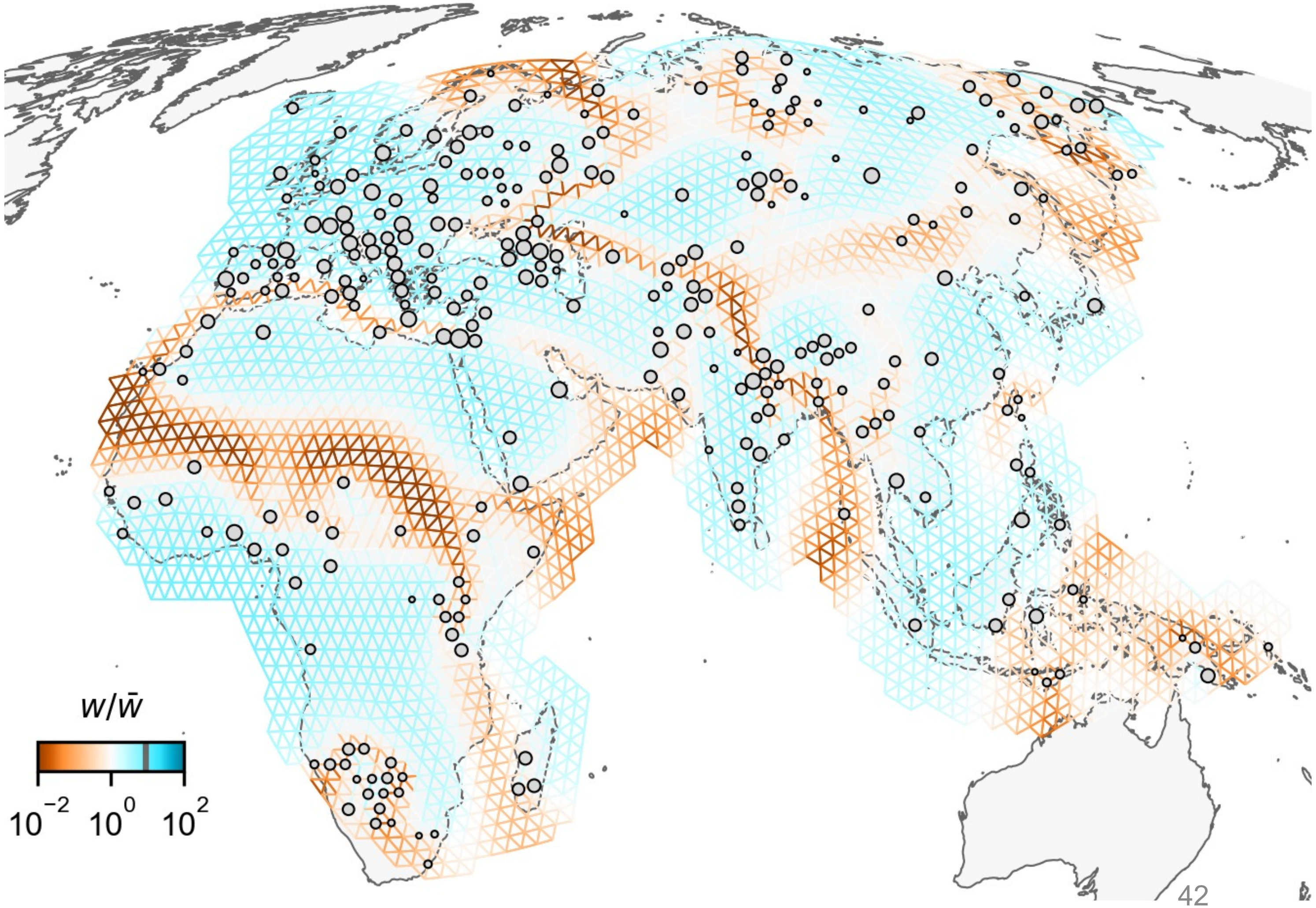
Outline

- Introduction + Motivation
- Spatially heterogeneous models of isolation-by-distance (IBD)
- Our model for long-range gene flow events (FEEMSmix)
- Results:
  1. Simulations
  2. North American grey wolves
  3. Afro-Eurasian panel of humans

Afro-Eurasian panel of humans

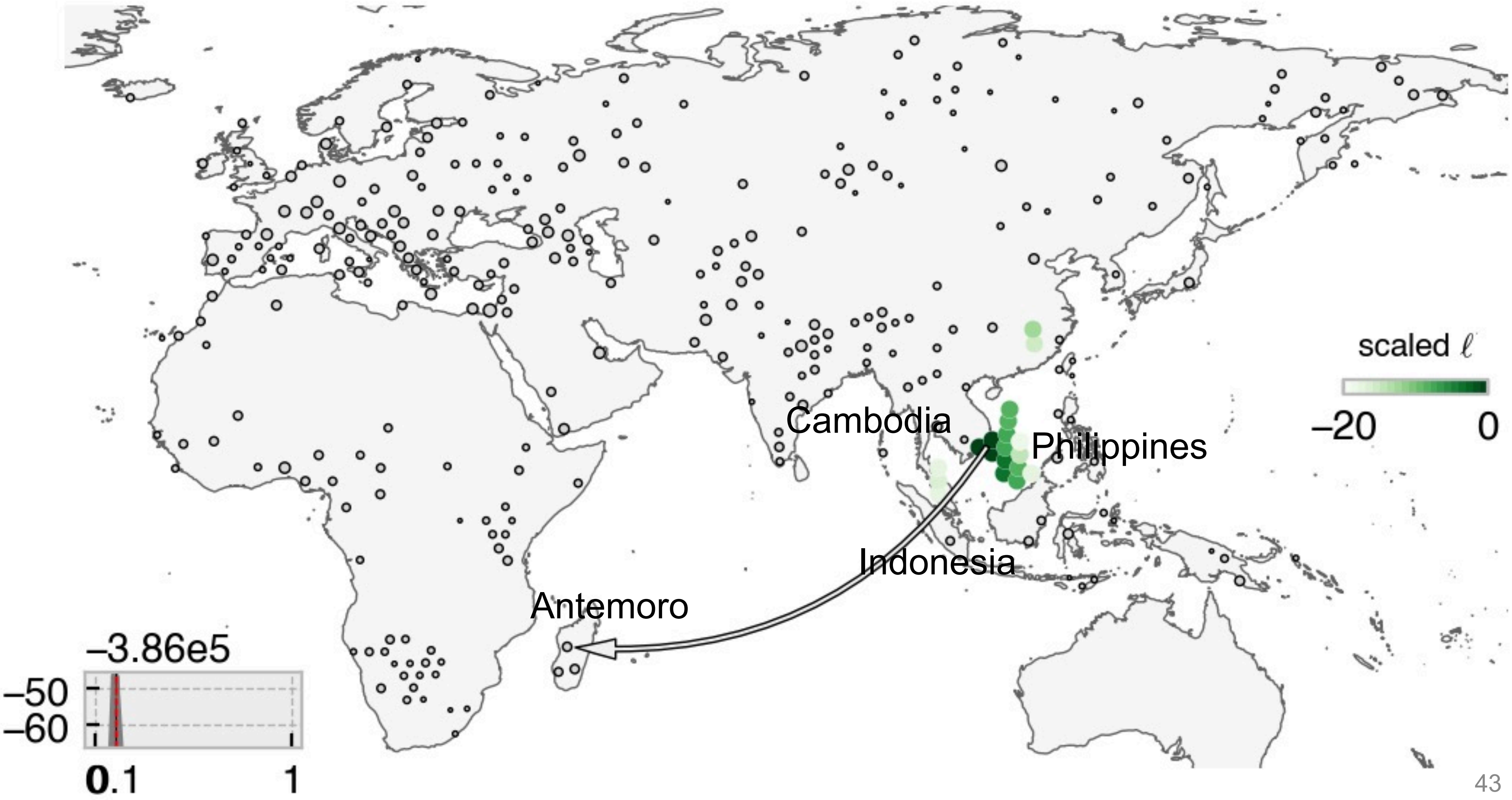
- 4,700 samples
- 20k SNPs
- 297 demes





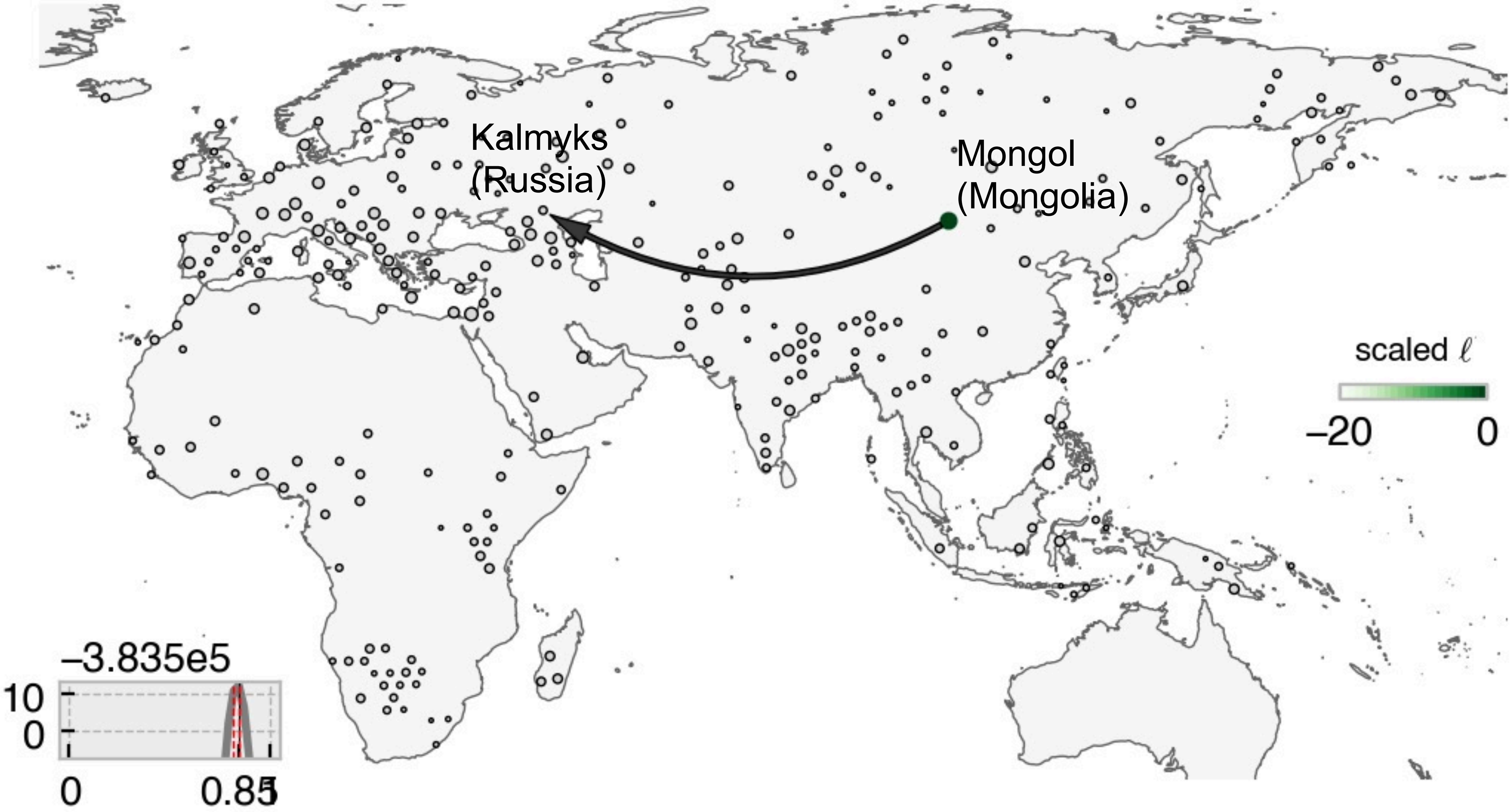
Peter *et al* 2020

Southeast Asian ancestry source detected in Malagasy

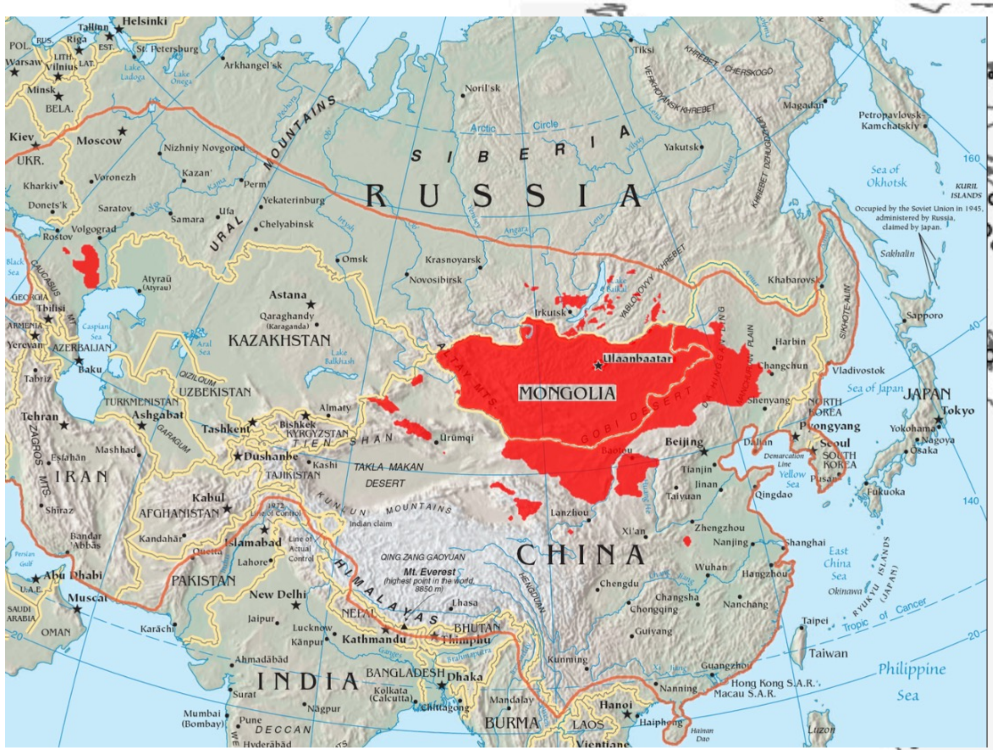


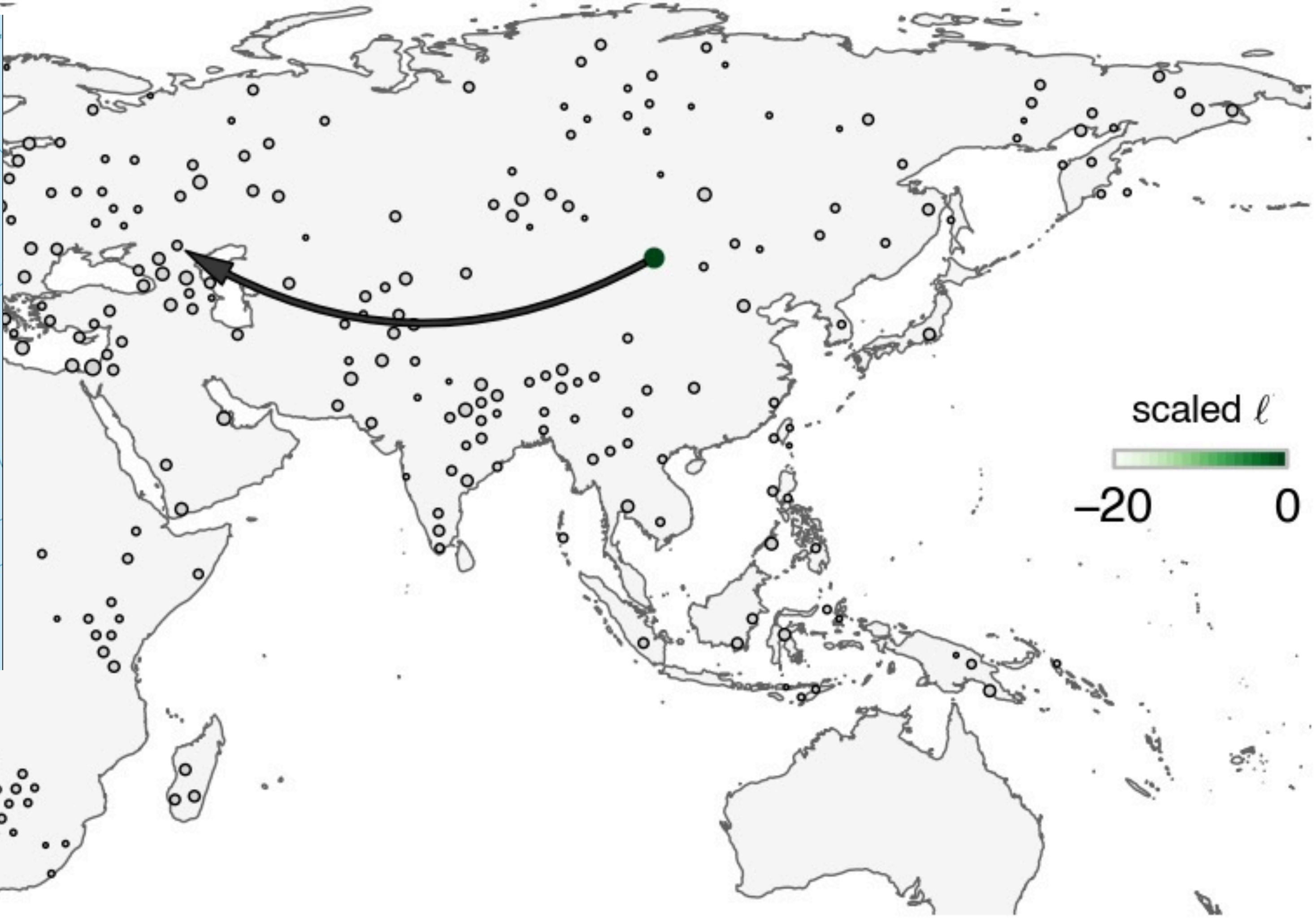
Pierron *et al* 2017

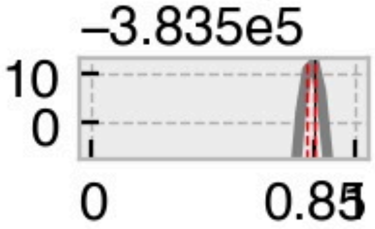
Kalmyks: only Mongolic-speaking people living in Europe



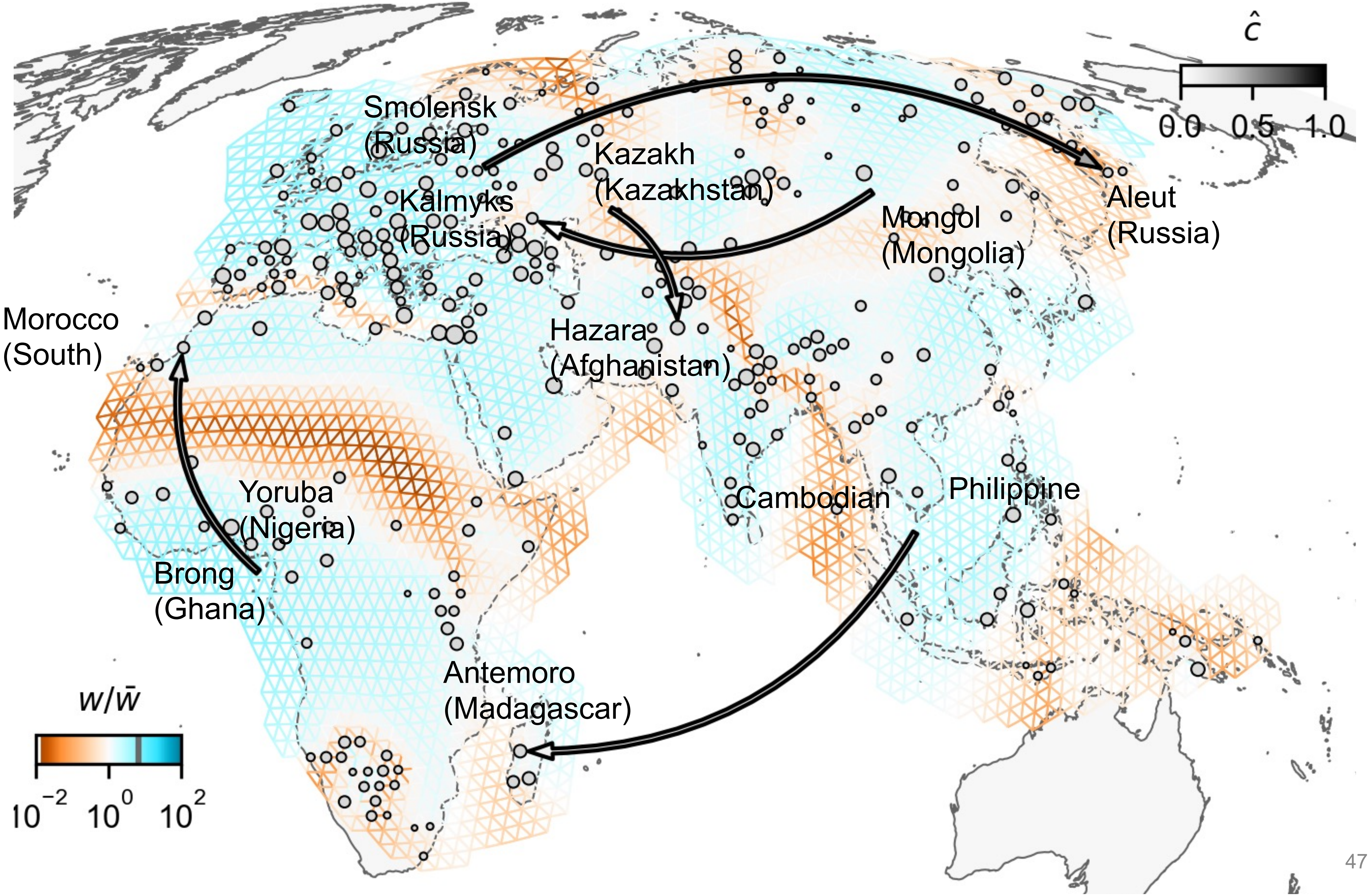
Kalmyks: only Mongolic-speaking people living in Europe







Erdeniev 1985


$\hat{c}$
0.0
0.5
1.0
Smolensk (Russia)
Kazakh (Kazakhstan)
Kalmyks (Russia)
Aleut (Russia)
Mongol (Mongolia)
Morocco (South)
Hazara (Afghanistan)
Yoruba (Nigeria)
Brong (Ghana)
Cambodian
Philippine
Antemoro (Madagascar)
$w/\bar{w}$
10^{-2}
10^{0}
10^{2}

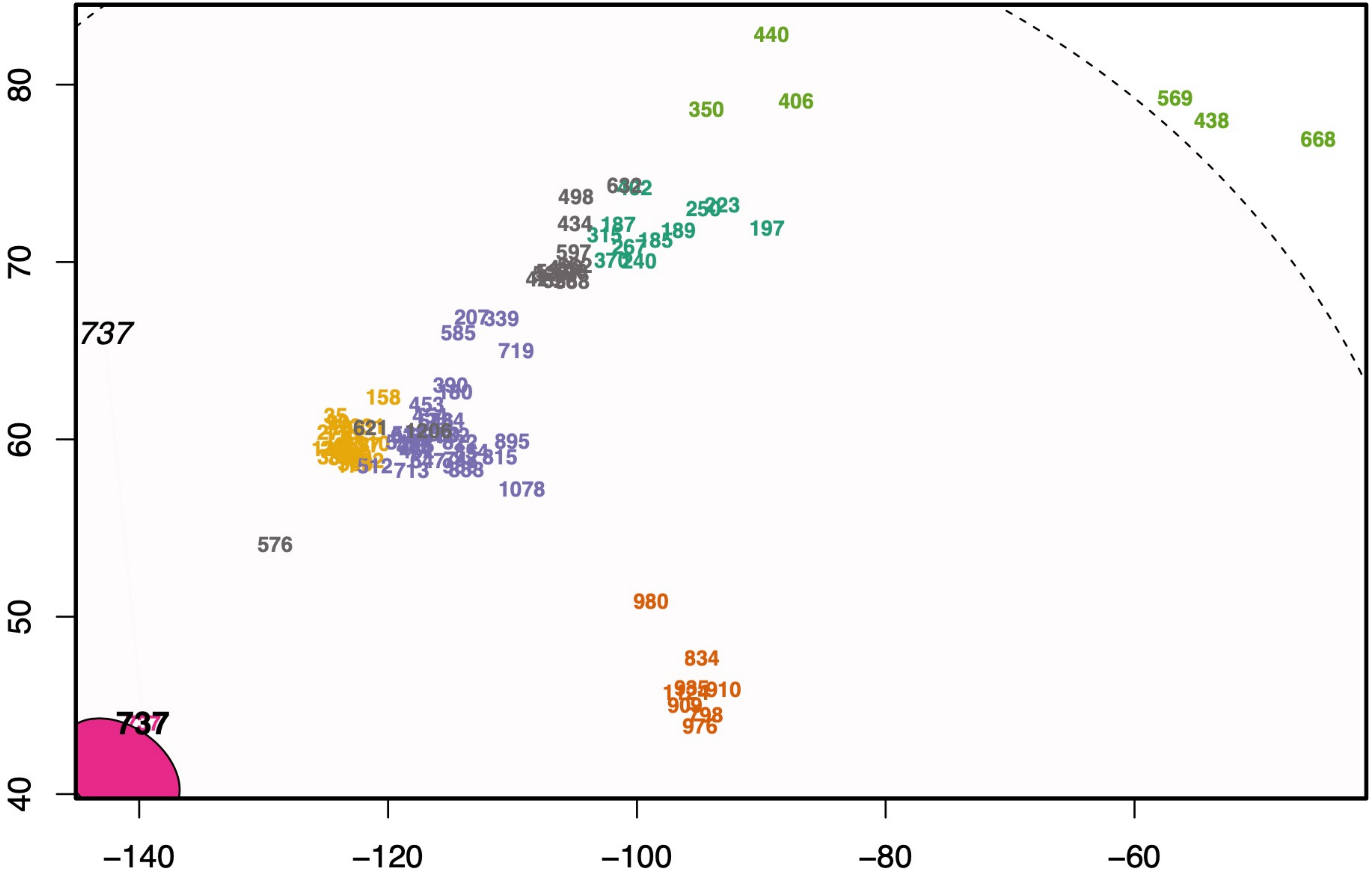
Watch this space for updates coming *very* soon!

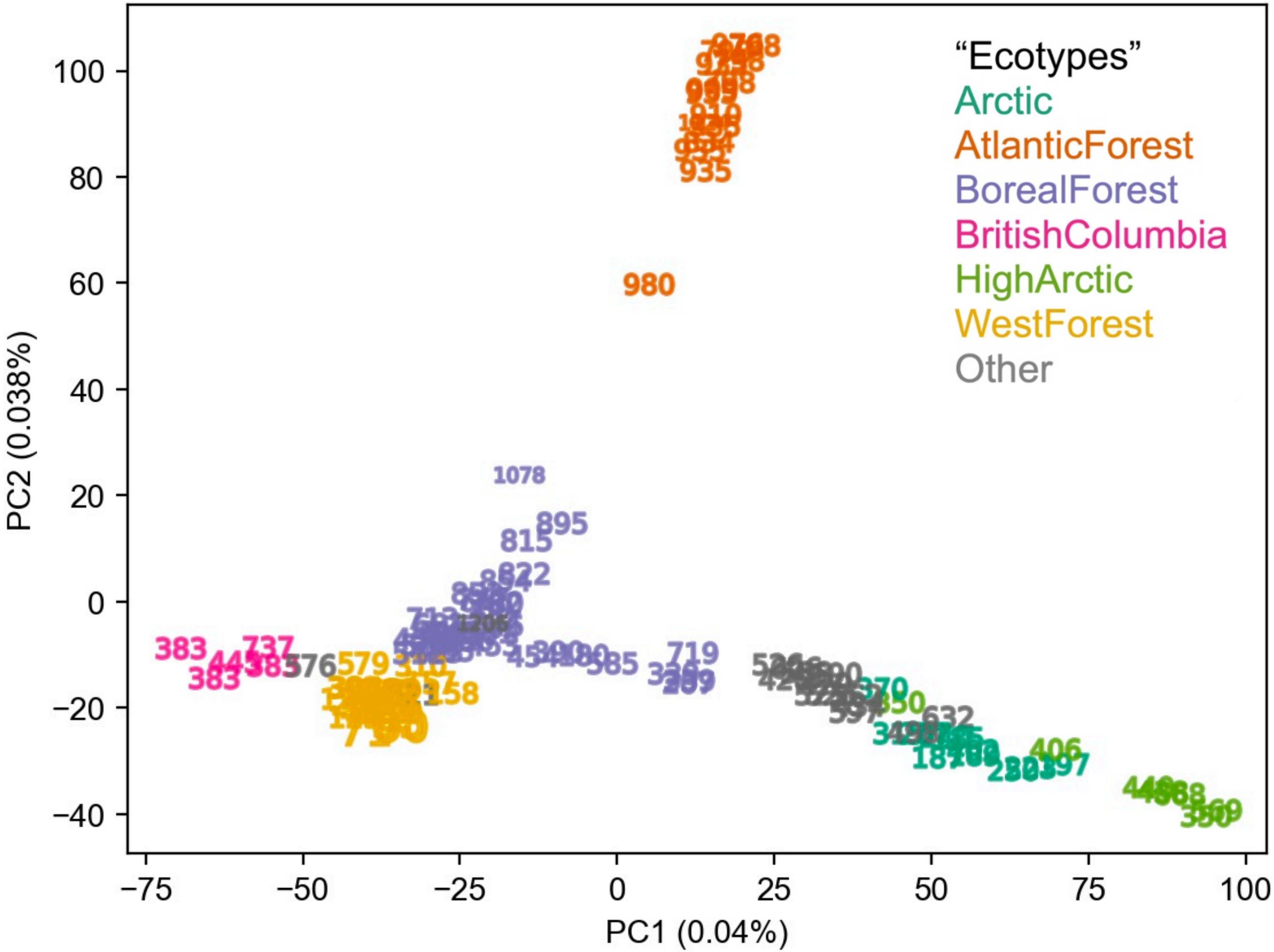
Conclusions

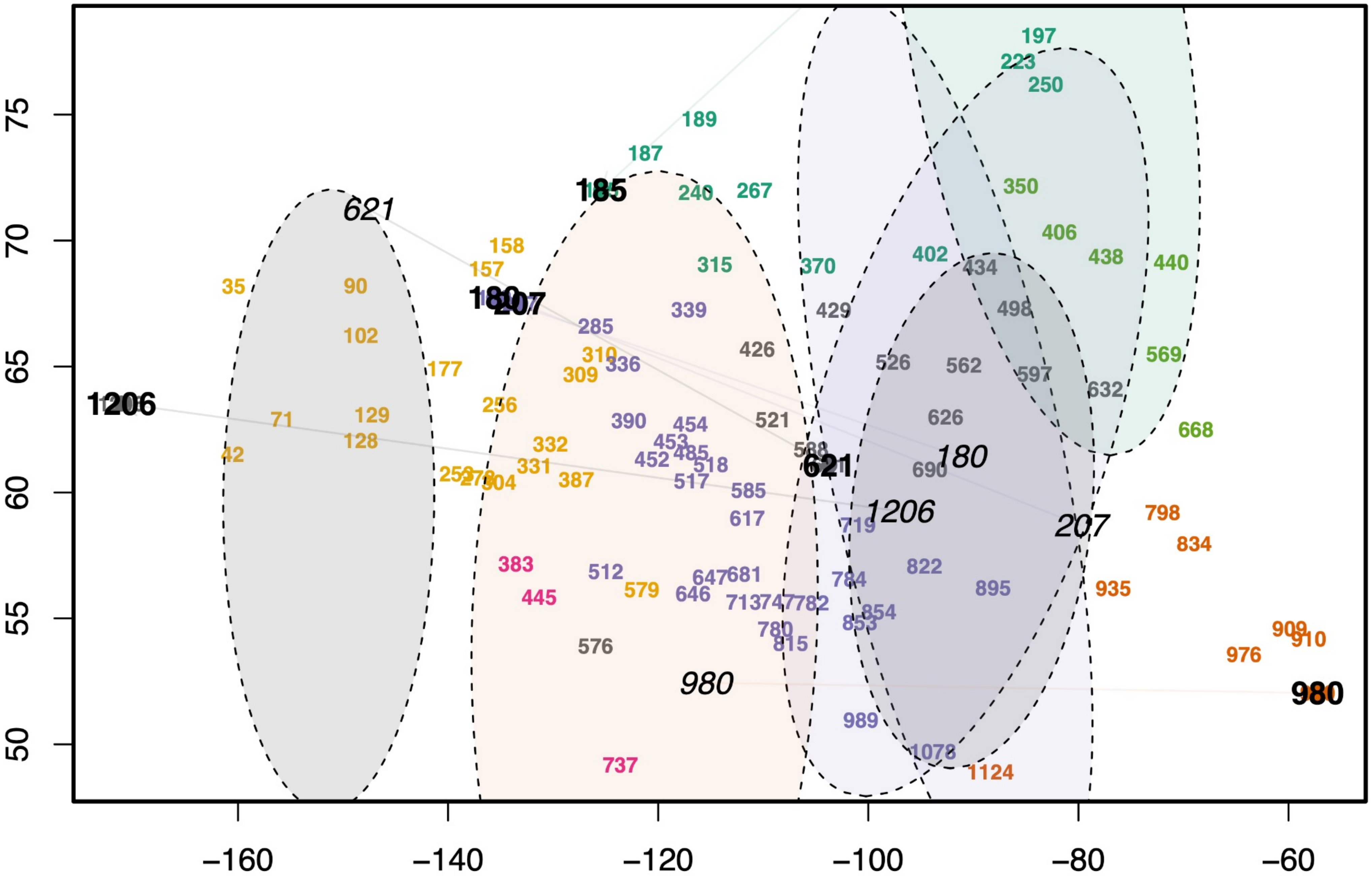
- `FEEMSmix`: a method to include long-range gene flow events in `FEEMS`
- Paints a fuller picture of the spatial patterns in genetic structure

⚠️ Caution ⚠️

- Interpret value of *c* as an informed suggestion, not as truth (e.g., if *c* > 0.5, it probably means high recent admixture)
- Reckon with uncertainty in source location (area-area vs point-point migration)


440
350
406
569
438
668
498
632
422
250
223
434
187
189
197
315
185
267
597
370
240
737
207
339
585
719
390
180
158
453
35
621
1206
895
512
713
815
1078
576
980
834
910
909
798
976
737
80
70
60
50
40
−140
−120
−100
−80
−60


"Ecotypes"
Arctic
AtlanticForest
BorealForest
BritishColumbia
HighArctic
WestForest
Other
PC1 (0.04%)
PC2 (0.038%)
−75
−50
−25
0
25
50
75
100
−40
−20
0
20
40
60
80
100
935
980
1078
895
815
822
1206
719
585
383
737
383
576
579
158
632
406
350



197
223
250
189
187
185
240
267
350
621
158
157
406
402
438
440
370
315
434
35
90
180
207
339
429
498
285
102
526
562
310
336
426
569
177
309
597
632
1206
256
71
129
390
454
521
626
668
128
332
453
485
42
452
518
588
621
180
331
690
253
278
304
387
517
585
1206
798
617
719
207
834
383
512
647
681
822
784
895
935
579
646
445
713
747
782
854
853
909
910
780
815
976
576
980
980
989
1078
737
1124
75
70
65
60
55
50
−160
−140
−120
−100
−80
−60