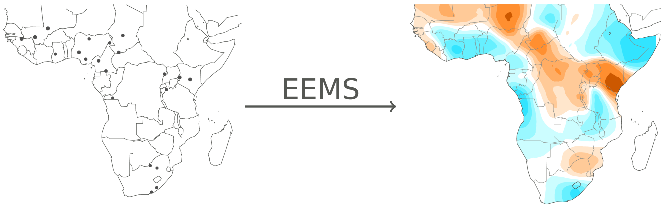
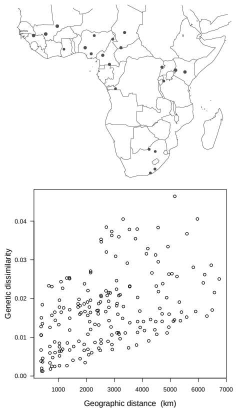


I presented this talk at SMBE 2015 in Vienna, in the session "PopGen in space! Theory and inference in spatial population genetics".

I will describe a method called EEMS, which analyzes genetic data with sampling locations to produce a visual representation of its spatial population structure like this contour plot.



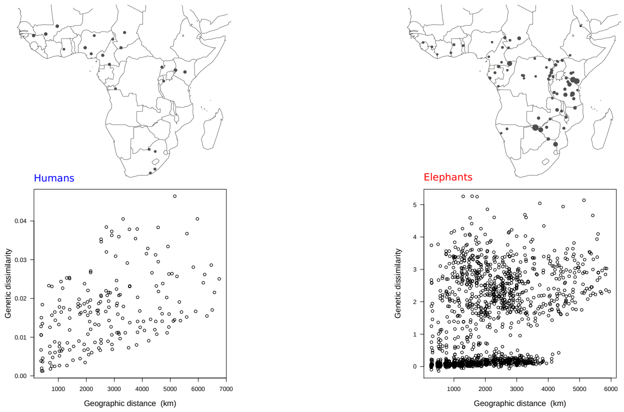
Spatially structured population

Isolation by distance

I've used a sample from Sub-Saharan Africa to plot pairwise genetic dissimilarities between sampling locations against their distance.

Here dissimilarity tends to increase with distance – a property called isolation by distance.

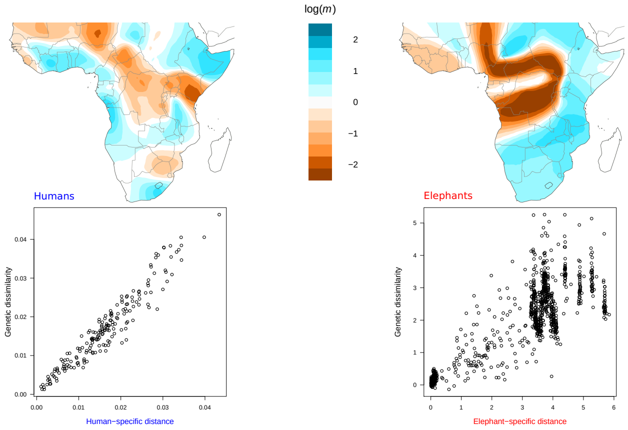
It arises when migration is (mostly) local so that more closely related individuals tend to live in greater proximity.



Under isolation by distance, the relationship between genetics and geography does not change across the habitat.

But in practice, this relationship can vary across space due to a combination of environmental and historical factors.

Here is a second Sub-Saharan African example, for which isolation by distance is not as good a description: many pairwise comparisons agree with isolation by distance but there are some elephants that are very dissimilar genetically even though they are close geographically.

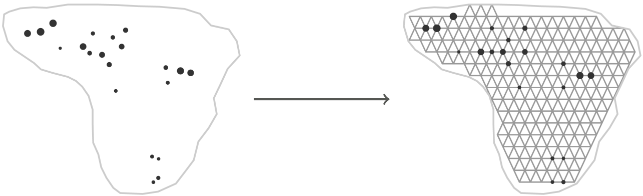


The goal of EEMS is to visualize the (possibly complex) relationship between genetics and geography. To achieve it, EEMS estimates local rates of effective migration to explain spatial patterns of population structure.

The colors in the contour plots represent the migration rates: shades of orange mean lower than average migration, shades of blue mean higher than average migration.

Isolation by distance would correspond to white across space. Regions in blue or orange indicate deviations from exact isolation by distance.

EEMS aims to model possible deviations from isolation by distance and it provides a better-fitting distance metric than geographic distance.



$D_{ij} = \frac{1}{p}\sum_{k=1}^{p}\left(z_{ik} - z_{jk}\right)^2$

i, j: samples, k: SNP

$\mathrm{E}\{D_{ij}\} = \Delta_{\delta(i)\delta(j)} = \Delta_{\alpha\beta}$

α, β: vertices

Given a sample with both genetic and geographic information, EEMS summarizes the genetic data as the matrix of average squared differences D across SNPs and captures the geographic information by assigning each sample to the closest vertex in a regular triangular grid.

EEMS also assumes that the expected genetic dissimilarity Δ between samples i, j is determined by their locations α, β.

Model Δ:

$$\Delta_{\alpha\beta} = \underbrace{\Delta_{\alpha\beta} - (\Delta_{\alpha\alpha} + \Delta_{\beta\beta})/2} + \underbrace{(\Delta_{\alpha\alpha} + \Delta_{\beta\beta})/2}$$

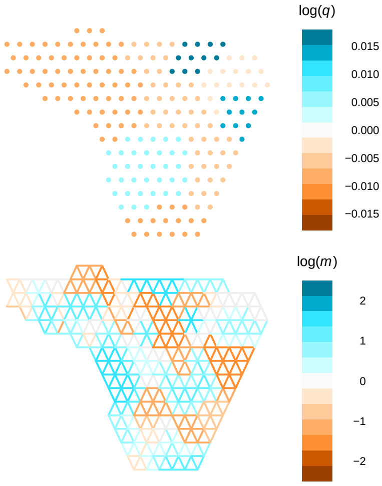
$$= B(m)_{\alpha\beta} + \left(W(q)_\alpha + W(q)_\beta\right)/2$$

Estimate m and q:

$$\Delta(m, q) \approx D$$

EEMS decomposes the expected genetic dissimilarities into "between" and "within" components of differentiation, B and W, which are functions of migration rates m and diversity rates q.

EEMS estimates the parameters m and q so that the expected dissimilarities Δ match the observed dissimilarities D.



$W(q) = \sigma^2 q$

$B(m) =$

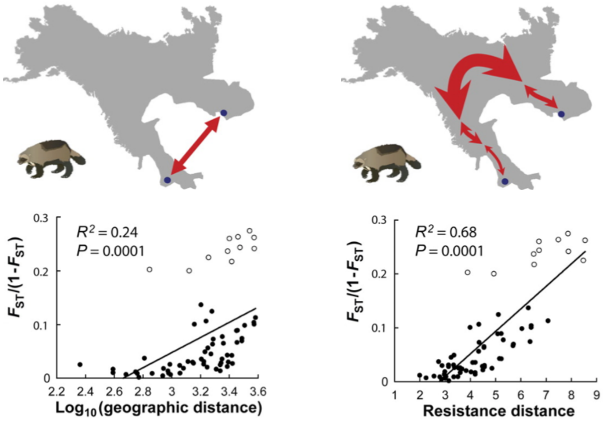
resistance distance

What are these parameters and how to compute the two components?

The population graph is fully parametrized: every vertex has a diversity rate q and every edge has a migration rate m. Parameters are plotted on the $\log_{10}$ scale and blue means "higher than average", orange means "lower than average".

The q parameters describe the genetic diversity within a location: the higher q is, the more dissimilar two individuals from that location are expected to be. The function W is simple – it is q scaled by a positive constant.

The m parameters describe the effective migration between neighboring locations. The function B is a distance metric for unconnected graphs (such as electrical circuits) – it is called resistance distance.



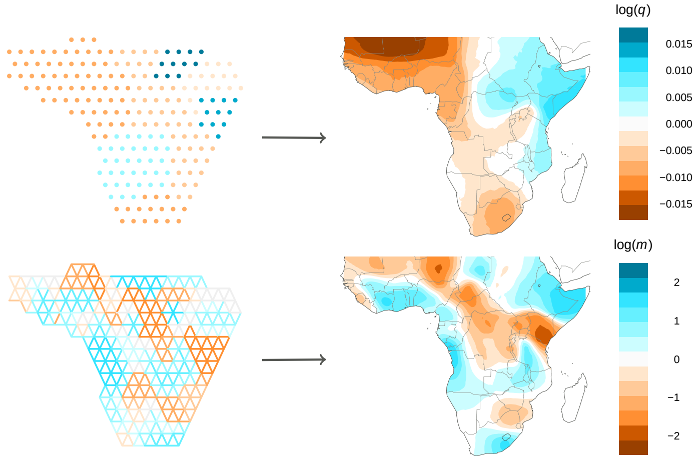
Isolation by resistance: McRae (2006), McRae & Beier (2007)

Resistance distance is useful because it approximates how genetic differences accumulate in a spatially distributed population.

It is more flexible than geographic distance because it can capture complexities such as the fact that some regions might be harder traverse or not accessible at all.

So resistance distance can lead to a better fit (closer to linear) to observed pairwise differentiation than geographic distance.

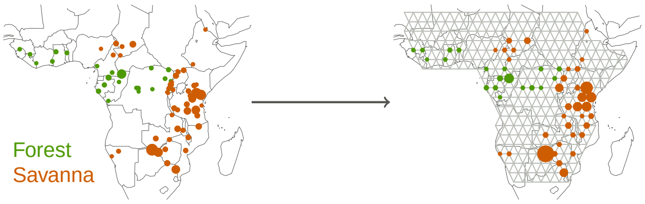
The resistance distance between vertices α and β approximates the expected coalescence time between an individual from α and another from β. And the longer the expected coalescence times, the more dissimilar the two individuals are expected to be.



EEMS uses Markov Chain Monte Carlo to estimate the parameters m and q. Ot produces smooth contour maps which represent the posterior mean of the migration and the diversity rates across space.

The top contour map says that human genetic diversity in Sub-Saharan Africa is highest in the Horn of Africa and decreases westwards and southwards.

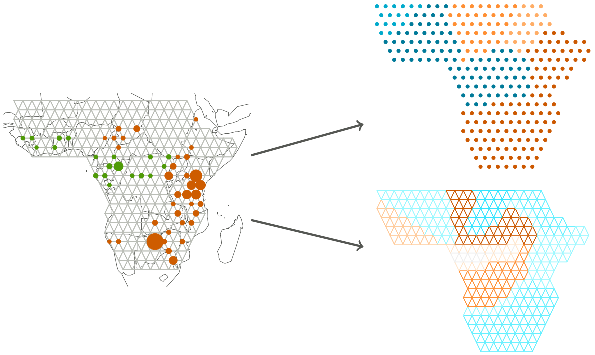
The bottom contour map says that human genetic similarity decreases more quickly inland than along the Atlantic coast where effective migration rates are higher.



I will summarize how EEMS works using the African elephant example.

1. EEMS uses the geographic information in a geo-referenced sample.

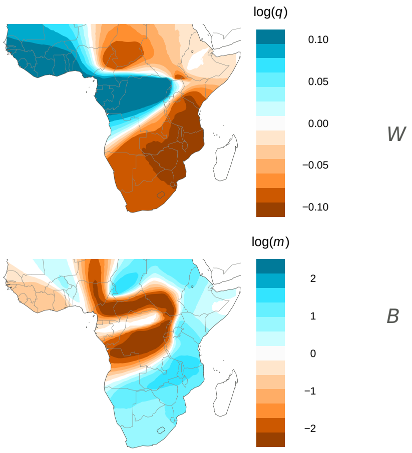
There are two subspecies – forest and savanna – and they occupy roughly the east and the west regions. First EEMS uses the available geographic information by assigning samples to vertices in a regular triangular grid.



2. EEMS estimates the posterior mean of diversity and migration rates, using Markov Chain Monte Carlo.

These parameters model two different aspects of population structure in space:

Effective diversity rates quantify the genetic similarity of two individuals from the same location. Effective migration rates quantify how genetic similarity decays across space.



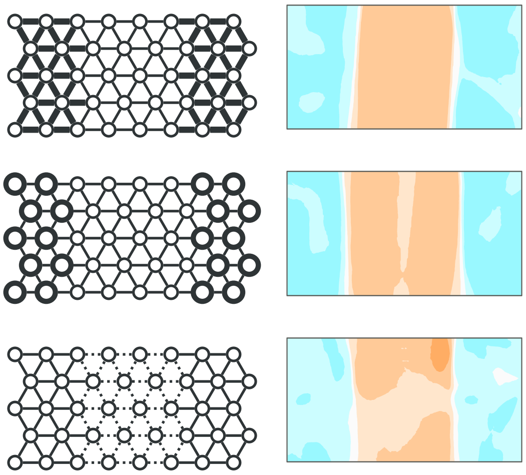
Effective migration is low in regions where genetic similarity decays quickly

3. EEMS provides insights into spatial patterns of genetic variation.

The top contour plot shows that two forest elephants tend to be more dissimilar than two savanna elephants.

The bottom contour plot shows that one forest and one savanna elephant are very strongly differentiated. The barrier means that elephants on either side are very different genetically even though they are relatively close geographically.

EEMS explains spatial patterns in genetic variation as deviations from isolation by distance, i.e., deviations from an exact linear relationship between genetic dissimilarity and geographic distance.



EEMS approximates a demographic model which evolves under equilibrium in time. This is an idealized situation not likely to hold in practice – it is just a device for building visualizations.

Here are three demographic scenarios that all produce a “barrier” of reduced effective migration through the central region but for different reasons

- Scenario with reduced migration rates
- Scenario with reduced effective population sizes
- Scenario with a split (no migration) after a point of time in the past.

These scenarios shows that EEMS produces meaningful representations of spatial structure even though the effective migration rates are not necessarily the actual migration rates.

EEMS: Petkova, Novembre & Stephens, bioRxiv 2014

The University of Chicago

Matthew Stephens
John Novembre
Hussein Al-Asadi
Benjamin Peter

University of Washington

Samuel Wasser

University of Copenhagen

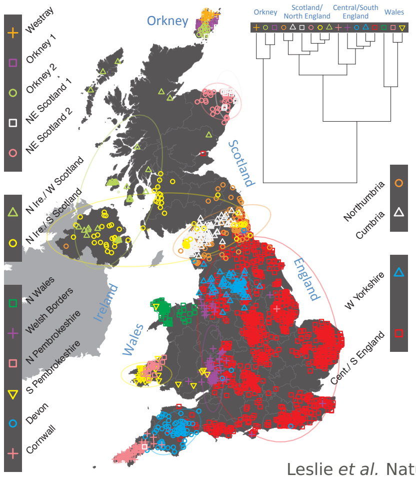
Ida Moltke

US National Institutes of Health

(NHGRI)

Summary: EEMS shows promise for visualizing patterns of genetic similarity in a spatial framework.

Thanks to everyone who contributed. . .



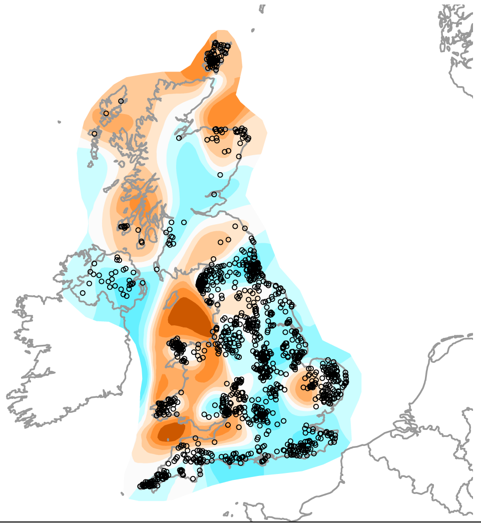
Leslie *et al.* Nature 2015

I didn’t have enough time to present the PoBI results.

PoBI, which stands for People of the British Isles, is a wonderful georeferenced dataset with very detailed geographic information.

These are the main fineStructure results for the PoBI dataset: the hierarchical tree and the clusters represent fine-scale population structure within the UK. Read the PoBI paper in Nature for details about the fineStructure method and its application to the PoBI data.

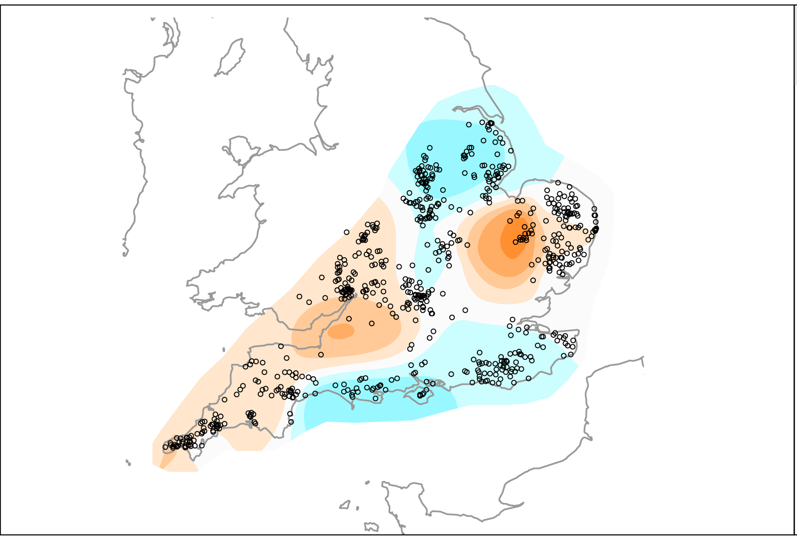
There is a large area (Central + South England) where no population structure is detected by fineStucture.



And here is the estimated effective migration surface for the PoBI data.

The areas that separate early in the fineStructure tree are generally inferred to have lower effective migration rates. This could explain why they break off first in the hierarchical tree and why they break down further into smaller clusters.

On the other hand, EEMS can’t represent the Northern Island population is a mixture of two spatially disjointed clusters. These are labeled N Ire./W Scotland (green triangles) and N Ire./S Scotland (yellow circles).



I am most interested in the England cluster (red squares) so I have analyzed a subset of the data to “zoom in” on that geographic area.

- Cornwall and Devon (pink pluses and blue circles): the westernmost counties in the south
- Welsh borders (purple pluses): a noticeable barrier to the south roughly corresponds to a geological feature – the Bristol channel
- Suffolk and Norfolk: the easternmost counties in England. These are separated by a “barrier” from the rest of England. No corresponding structure is detected by fineStructure, which is exciting, though the results should be carefully validated.