Landscape genetics - Ghana

## Landscape genetics on parasite mitochondrial data from the transition region of Ghana: Proof-of-concept of utility of landscape genetics to incorporate genetic data with the environmental data

## Background/Introduction

Different environmental features of the landscape affect the gene-flow and the dispersal of the parasites and the vectors and thus, the disease transmission dynamics. Landscape genetics combines population genetics, landscape ecology, and spatial analytical techniques to explicitly quantify the effect of landscape on evolutionary processes like gene-flow, drift, and selection (Manel et al 2003 (Balkenhol, 2016). These techniques have traditionally been used in the areas of species conservation, but have several potential implications in understanding the epidemiology of diseases and their control and elimination (Archie et al., 2009; Hemming-Schroeder et al., 2020; Lo et al., 2017; Real & Biek, 2007; Saarman et al., 2018; Schwabl et al., 2017)

Therefore, I investigated landscape genetic methods to incorporate genetic data to a geo-spatial framework. This was done with parasite mitochondrial data from a central transect across Ghana to identify environmental factors that might influence the population genetic structure of *Onchocerca volvulus*. In addition, I inferred connectivity between sampling locations of these parasites using landscape genetics.

*What is the concept of transmission zone?*

* <<definition of transmission zone – WHO guideline>>
* For practical implementation and prioritize limited resources/identify when and where the treatment is required and treatment interventions we need to “try” to delineate transmission zone
* *We might not be able to delineate it as the zone itself might not be static but dynamic/ change with respect to time.*
* For the transmission to persist, there should be movement of the parasite/pathogen which can be measured indirectly by genetic relatedness. Genetic relatedness gives us the idea about how common are the parasite samples, do they share similar traits which might be the result of the movement of the fraction of parasite population from one location to another. First, we need to measure the spatial pattern of genetic differentiation of the parasite population. Second, we can use those parameters of genetic differentiation to see which environmental features might govern the spatial pattern of genetic differentiation. Third, we can transform the most important environmental maps to a resistance surface maps and optimize the resistance surface maps based on genetic distances to be able to get an idea about the parasite movement and thus, the transmission zones.
* The way to estimate transmission is via prevalence/microfilarial estimates which alone is not sufficient for distinguishing if the locations belong to a different transmission zone. It is because the transmission zone is a spatial concept and it’s a connected complicated system without perfect sampling. To understand if the geographic locations are linked or not or to see if the samples are from different transmission zones, we are looking at genetic differences between the samples collected from different locations and looking at the environmental variables to extrapolate the connectivity. By doing so, we might be able to compensate for the fact that we have very patchy, incomplete data in the low resource setting where the disease is most prevalence and get the conclusions about transmission.

*Why transmission zones?*

Disease control programs have historically focused on government administrative units as the unit of intervention which has led to a situation where decisions are being made about the treatment and stopping treatments that do not really consider where the transmission is really occurring. There is a long-held view that onchocerciasis being a vector borne disease, the transmission is intensely focal but the evidence suggest otherwise (cross border transmission, transmission over long range, Crawford et al citations).

* The reason why we are focused on transmission zones is that we want to make sure that the focus of the intervention is at the correct scale and the evidence is suggesting that it is not.

*Why landscape genetics for transmission zone?*

* Transmission zone is a spatial concept, and we are trying to get an idea about it purely via a population genetics metrics/does not make sense without incorporating spatial information.
* Landscape genetics approach can be very good way to incorporate spatial information to the population genetic measures. We can include spatial information in the form of remote sensing images/satellite maps of different environmental and climate variables such as elevation, slope, distance to the water bodies, mean annual temperature, mean annual precipitation and so on. This allows us to test/understand how the physical environment affects/influences the population genetic structure of the parasite or the vector population.
* What is the likely probability of the parasite being transmitted to a particular location i.e. what’s the connectivity to the parasite populations between the locations of interest?
* We can go from DAPC to heatmaps which are the quantified resistance surface maps.
* The underlying hypothesis is that the genetic similarity between parasites in different geographical locations is a measure of gene flow and in order for gene flow to occur - parasites need to move between these locations - we can infer how the movement might occur in space with the help of resistance surface maps
* This will help us to define the infectious disease concept of a transmission zone - a geographical area within which the transmission is occurring

*Why transition region of Ghana?*

* Persistence of onchocerciasis transmission in Ghana might be either due to the movement of the vectors or the movement of the human hosts
* Onchocerciasis control started as a vector control in the transition region of Ghana as early as 1974 (citation, WHO report) and ivermectin is being distributed for more than two decades.

## Methods

### ### Sampling locations

From 97 individuals living with onchocerciasis

We had the data from 14 communities belonging to three river basins.

*The sampling locations of 163 Onchocerca volvulus from 14 communities from three river basins, viz. Black Volta/Tombe, Pru, and Daka, were used to assess environmental correlates with parasite presence.* Ethics approvals for sampling parasites from people is reported in Crawford et al., 2019. *A transect of 308 x 112 km was used for the landscape analysis. For landscape genetic analysis*, I used 189 SNPs from analysis of whole mitochondrial genome sequence data obtained from K. Crawford (Crawford et al., 2019).

FST estimates were calculated using Arlequin V 3.5 (Excoffier & Lischer, 2010) and were used as a measure of genetic distance.

*F-st and other genetic distance metrics were calculated based on the packages like Hierfstat, and graph4lg (citation, packages).*

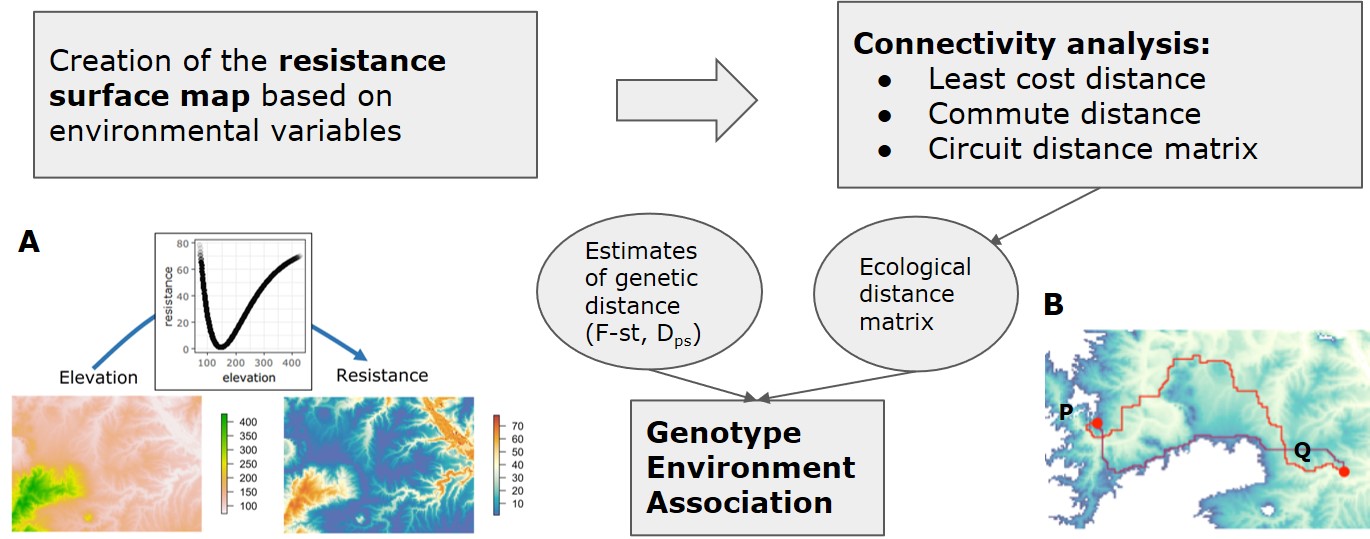
Map

Description automatically generated

Figure 1. Sample locations for the parasite data in transition region of Ghana.

### Analysis workflow

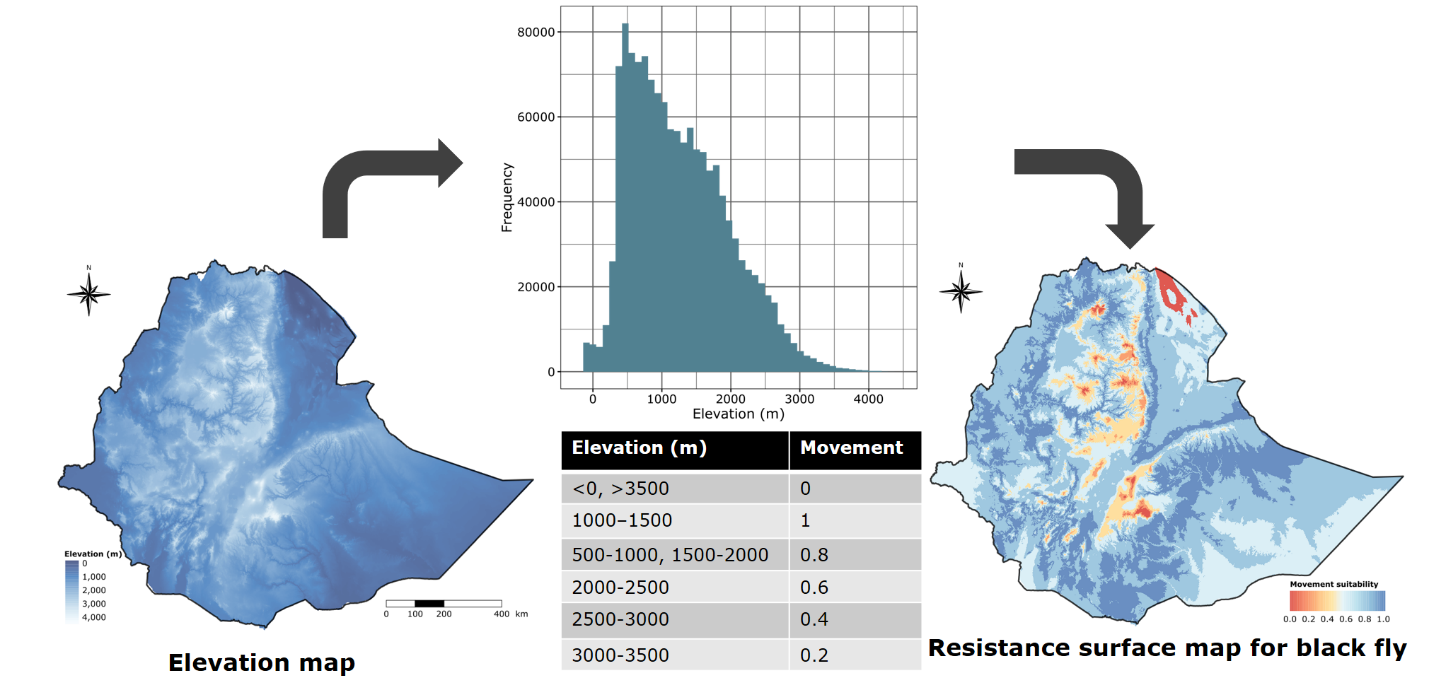
The landscape genetics analysis framework involved processing environmental rasters into a resistance surface map to infer the movement pattern of *O. volvulus* (Figure 2).



**Figure 2. Flowchart showing the analysis workflow for landscape genetics analysis**. Resistance surfaces (**A**) were obtained by applying transformation function to an environmental raster. Different distance matrix algorithms, including the least cost distance (**B**), were used to determine the ecological distance matrix between point P and Q (see text for descriptions of each distance algorithm).

#### Creation of the resistance surface map

Multiple environmental variables were downloaded for the study region. Pixels in the environmental raster were assigned different cost/resistance values for movement or gene flow. This approach was taken for the variables for which the suitability of the environment was known. Different ecological and spatial studies were referred to assign cost values to the environmental raster. Two different approaches were taken to encode a cost/resistance value to the range of values in the environmental variables. First, manually encoding cost values based on the published literature on the effect of environmental variables on onchocerciasis prevalence and vector distribution (Figure 3). Second, transforming environmental raster with different transformation function available in the *ResistanceGA* package (Figure 2A). Different transformation functions like as linear, exponential, monomolecular, ricker etc. are available in the package where cost values are assigned to the environmental raster accordingly.



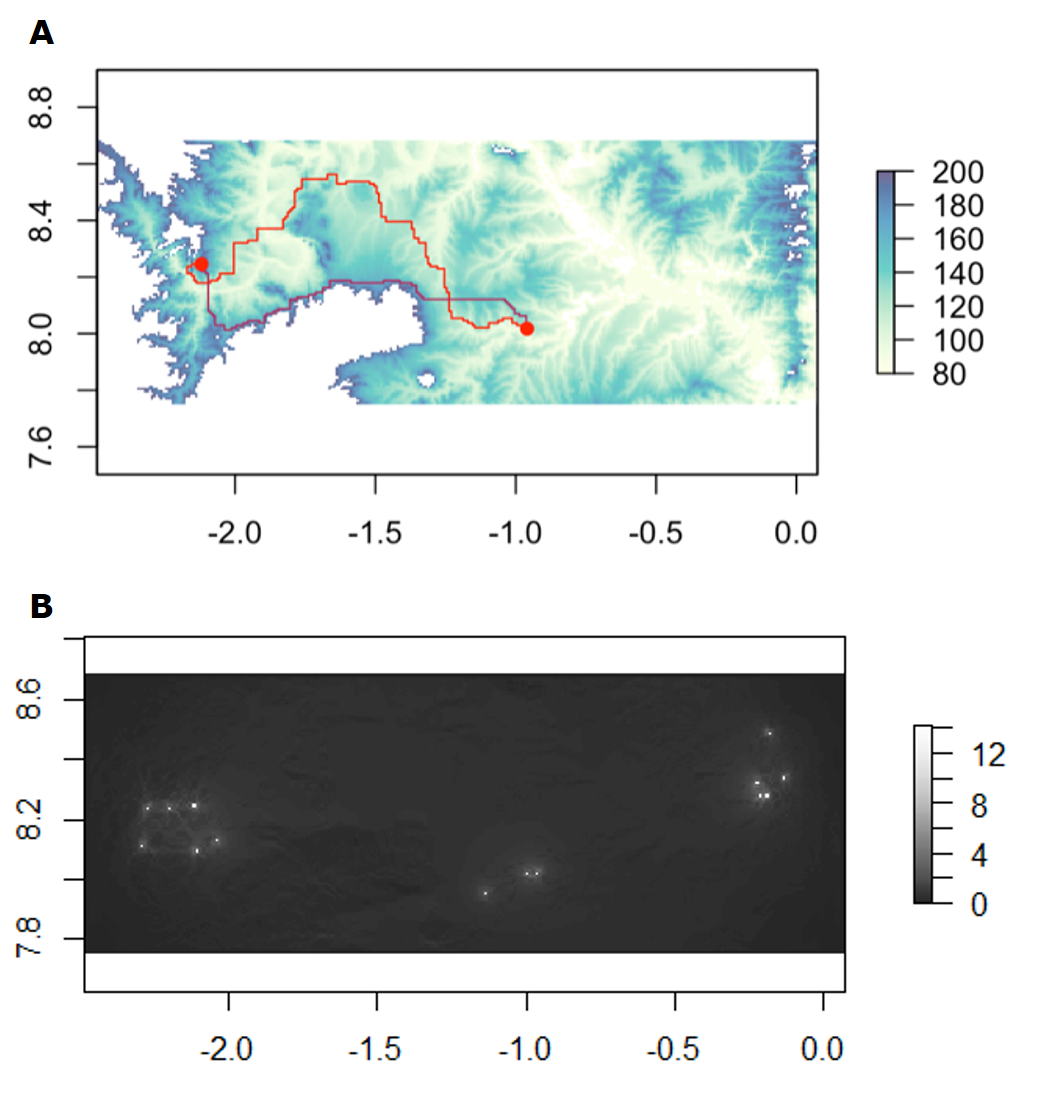
**Figure 3. An example of creation of the resistance surface map by manually encoding cost values for black fly movement based on the elevation in Ethiopia**. Range of values of elevation are assigned with a different cost for movement as shown on the table. The area with high suitability/low resistance for the movement of blackflies is shown with blue color whereas the area shaded red are the ones with high resistance for black fly movement.

* Resistance/friction surfaces are determined by biotic factors, abiotic factors, and the movement (*tse tse* review paper).

#### Connectivity analysis

The resistance surfaces derived from the individual environmental variables were in turn used to derive different connectivity metrics: the least-cost path, commute distance matrix, and the circuit distance matrix between the sample locations. The least-cost path for any location between point P and Q is based on the path of least ecological resistance an organism would follow while moving from one point to the other (Figure 4A). The least-cost path might not be the same when the path starts from point P and when it starts from point Q. However, biological organisms do not always follow the path of least resistance. To incorporate the multiple paths an organism might follow while traversing from one location to the other, I also considered circuit distance and commute distance.

Circuit distance is based on circuit theory and has been used in chemical, social, and neural networks (reviewed in? or other citation?), and more recently has been applied to model connectivity in heterogeneous landscapes (B. H. McRae, 2006; B. H. McRae et al., 2008). This is based on the fact that current, voltage and resistance in an electrical circuit demonstrate a good relationship with random walk. Therefore, the current density obtained using this algorithm can be used to measure connectivity or isolation between patches, and identify corridors for movement in a landscape. Circuitscape version 4.0.5 (Figure 4B) (B. McRae et al., 2016) was used to produce a circuit distance matrix for connectivity between sampling sites in this region of Ghana. Because this is a computationally intensive algorithm, the computation time increases as the number of pixels/cells in the environmental raster grid increases. Therefore, it is less efficient for optimising the resistance surfaces (Peterman, 2018) when compared to the more computationally efficient commute distance.



**Figure 4. The least-cost distance (A) between two locations and the circuit distance (B) between all the sample locations in Ghana.**

Commute distance is similar to circuit distance because it is also based on a random walk approach. It represents the random walk commute time between two locations, which is the number of edges traversed during movement from one location to the destination location and returning to the starting point on the resistance surface (van Etten, 2017). Thus, while circuit distance was used to generate connectivity maps, commute distance was calculated for the optimisation of resistance surfaces using the *gdistance* package.

#### Genotype-by-environment association

The commute distance matrix derived from each environmental variable were used for genotype-by-environment association tests. These distance values were used for mixed matrix regression with randomization (MMRR) analysis and Mantel tests. Those environmental variables that were significantly associated with the genetic distance matrix were used to further optimise the resistance surface. The surfaces were optimised based on the genetic distance using the *ResistanceGA* package. The individual resistance surfaces were combined to form a composite resistance surface map and were tested for model fit.

The environmental variables that were found to be significantly associated with the genetic distance were used for further analysis using the disease prevalence and species distribution map. Clustering analysis was performed on the prevalence maps to find any clusters of disease hotspots or clusters of isolated populations within a species.

## Results

### DAPC analysis

Whether you are looking at a finer geographical scale of individual communities or a larger geographical scale of river basins separated by couple of 100kms there is a weak spatial clustering. In other words, there is high degree of genetic relatedness across the whole of geographical space.

There appears to be a lot of movement of parasite populations between the sampling sites. However, the physical environment is not uniform, and the aim of landscape genetics is to understand the relation between the landscape features and the movement of the parasites measured indirectly via estimates of genetic differentiation.

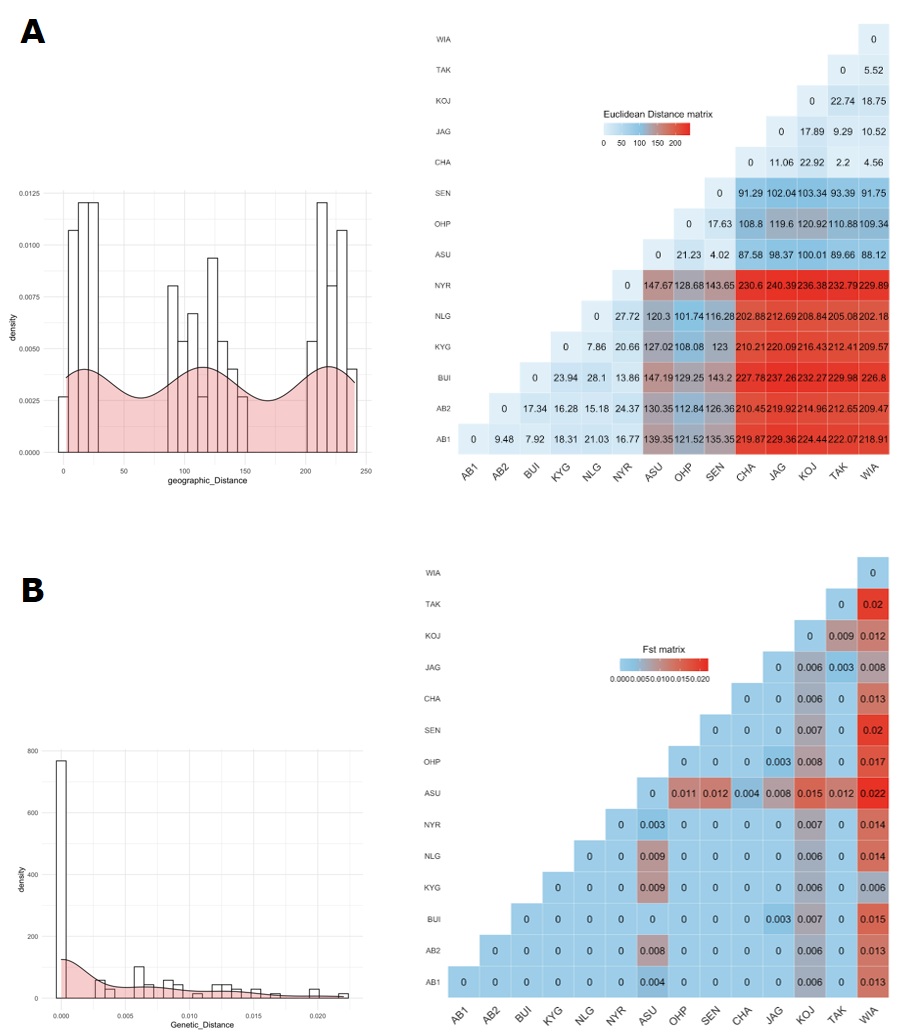
### sPCA

### tess3r

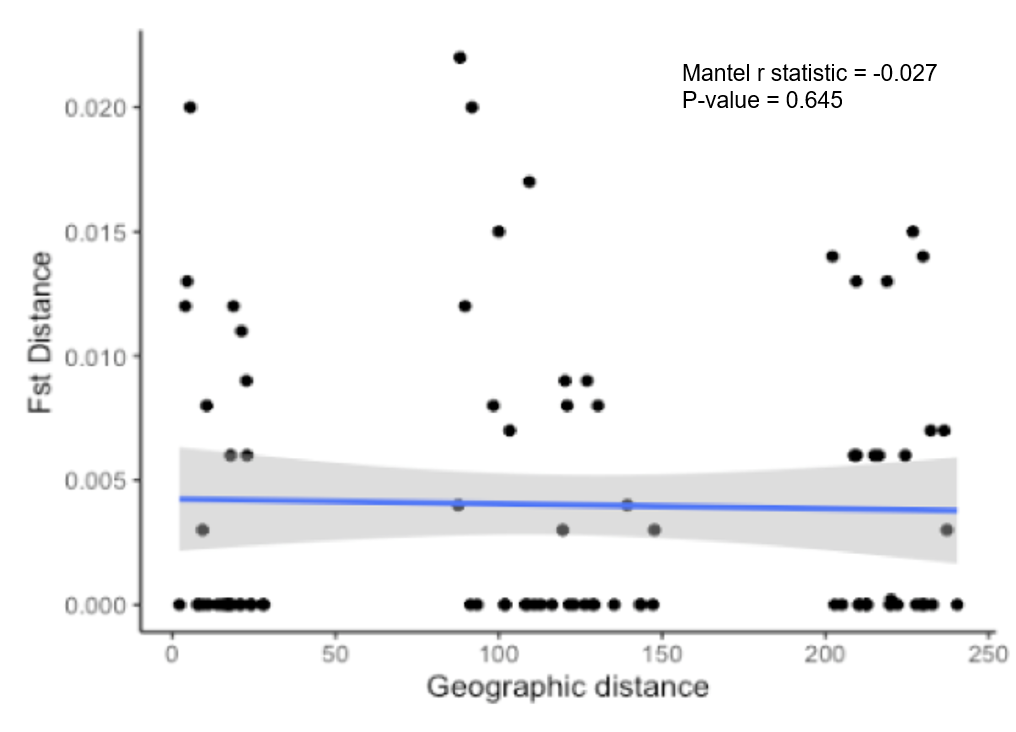
The relatedness map shows that samples are more related even if they are not geographically near to each other.

### Isolation-by-distance

To test for isolation-by-distance (IBD), a Euclidean distance matrix was calculated between the sample locations and pairwise FST was used as a genetic distance matrix. The Euclidean geographic distance between locations ranged from 2.2 km to 240.39 km. Three separate peaks in the histogram shows the presence of locations in three distinct river basins (Figure 5A). In contrast, genetic distance is 0 between most of the locations, indicating low genetic differentiation between the sampling sites in the transition region (Figure 5B). The highest FST value (0.022) was between Wiae (Daka river basin) and Asubende (Pru river basin), despite their geographical proximity. A Mantel test between FST and geographic distance matrix found no significant association (Figure 6). Therefore, there was absence of isolation-by-distance in the parasite populations.



**Figure 5. The Euclidean geographic distance matrix (A) and the genetic distance (B) matrix for the populations from the sample locations.** The histogram and the density graphs for the values in the pairwise matrix is shown on the left of the distance matrix.



**Figure 6. Relationship between the pairwise geographic distance and the pairwise FST distance between parasite populations in the transition region of Ghana.**

Geographic distance alone is not sufficient in explaining the spatial variation in the genetic differentiation spatially. We need to incorporate more complex variables like landscape features and need to be geographically explicit. The probability of transmission based on geographic distance alone can’t be claimed i.e., we can’t claim the transmission to be less likely between two sites simply because the sites are far apart.

Under normal circumstances the genetic similarity is going to be directly related to the distance between two points - but when we look at this - it's not the case.

The question is how we can explain the genetic dissimilarity matrix - we can look at a whole bunch of other environmental variables other than the straight-line distance

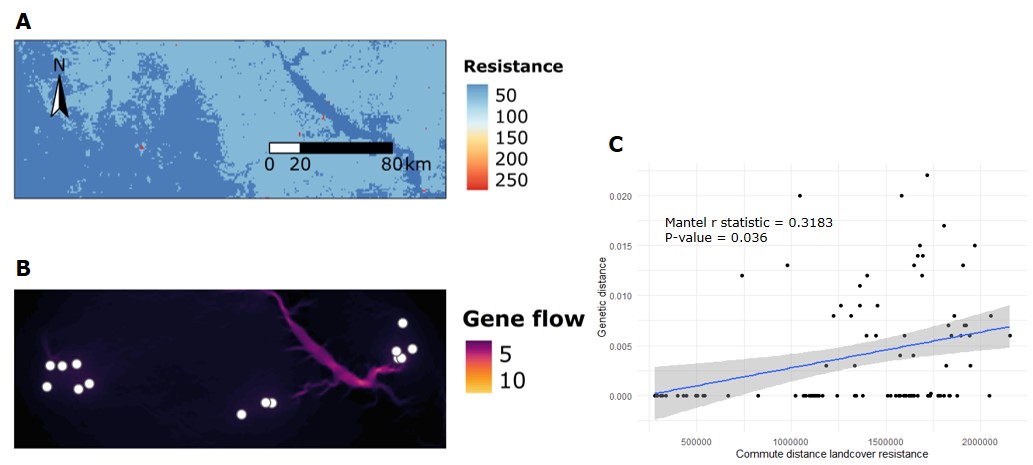
Surprisingly, elevation and distance to the water are strongly associated with the genetic dissimilarity - these two environmental variables explain most of the variation in the genetic distance matrix

the elevation and distance to the water must be correlated because the river does not flow uphills

Low resistance, high gene flow and connectivity

### Resistance surface maps and connectivity analysis

Resistance maps were created for different environmental variables using the monomolecular, reverse monomolecular, inverse monomolecular, and inverse reverse monomolecular functions available in the `ResistanceGA` package. These were done for six different environmental and socio-demographic variables, viz. altitude, land cover, potential evapo-transpiration, population density, distance to the nearest river, and Enhanced Vegetation Index (EVI). Both least-cost distance and the commute distance were tested for genotype-environment association using a Mantel test (Supplementary Table 1). Out of all of these, commute distance obtained from the resistance surface derived using inverse reverse monomolecular function from land cover was found to have a significant association (Mantel's r = 0.318, p-value = 0.036) with the genetic distance (Figure 7).

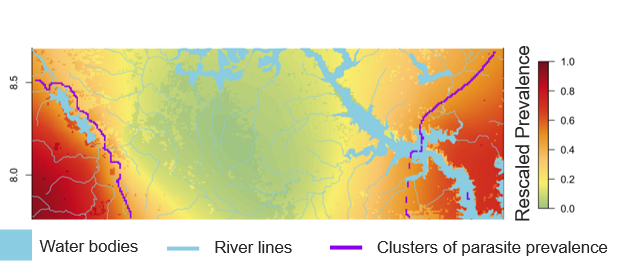


**Figure 7. The relationship between the commute distance matrix obtained from the resistance surface derived with land cover and the FST distance.** The resistance surface (A) was obtained using the inverse reverse monomolecular function based on land cover and the geneflow map was created based on the resistance surface. It showed significant positive correlation between the genetic distance and the circuit distance obtained from resistance surface map.

Based on the resistance surface map we can find the optimal path that the genes could be taking. Resistance maps gives us an idea about the resistance for the gene flow of the parasite populations based on the genetic similarities.

### Prevalence map and clustering analysis

Based on the positive correlation between commute distance based on land cover and FST, land cover was used for geostatistical modelling of the prevalence data of onchocerciasis in these locations. The onchocerciasis prevalence data was obtained from Osei-Atweneboana et al. (2007). Clustering analysis was done to locate any possible hotspots of onchocerciasis prevalence.



**Figure 8. Prevalence map obtained using land cover as an explanatory variable with clusters of high prevalence (red) in western and the eastern parts of Ghana.**

This demonstrates the incorporation of the parasite genetic data and environmental data into a geospatial framework. Here, I used the landscape genetics framework to observe the effect of environmental variables on the population structure of the parasites and infer connectivity among sampling locations.

## Discussion

* We are able to transform the genetic metrics of parasite relatedness spatially adjusting it for the specific landscape and ecological features of a particular area and infer about transmission zones.

*Limitations/Caveats*

* For the two parasites to be related they either need to have a shared ancestry *as a result of mating* or should have migrated from the other location.
* We are not sure about the temporal scale of the gene flow. The geneflow can happen not within generation but across several generations.

*Recommendations*

* We might be able to sample parasites in a spatially uniform way so that we can get the better estimates of the resistance surface map and thus good idea about the transmission zones.
* Applicable and transferable to other African settings

## Conclusion

We can use landscape genetics approach for genetic data to try to delineate where transmission zones might be based on genetic and landscape features and also identify the areas of high connectivity. We can get a clear picture about the natural transmission zones based on the biology of the parasite and the vector which can then be used for planning of interventions.

We can make quantitative predictions as we can create a heatmap of resistance

## Supplementary information

**Supplementary Table 1. Mantel correlation tests between the distance matrices obtained with different resistance surfaces from environmental variables and the F-st distance.**

## References

Archie, E. A., Luikart, G., & Ezenwa, V. O. (2009). Infecting epidemiology with genetics: A new frontier in disease ecology. *Trends in Ecology & Evolution*, *24*(1), 21–30. https://doi.org/10.1016/j.tree.2008.08.008

Balkenhol, N. (Ed.). (2016). *Landscape genetics: Concepts, methods, applications*. Wiley Blackwell.

Crawford, K. E., Hedtke, S. M., Doyle, S. R., Kuesel, A. C., Armoo, S., Osei-Atweneboana, M., & Grant, W. N. (2019). *Utility of the* Onchocerca volvulus *mitochondrial genome for delineation of parasite transmission zones* [Preprint]. Evolutionary Biology. https://doi.org/10.1101/732446

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*(3), 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Hemming-Schroeder, E., Zhong, D., Machani, M., Nguyen, H., Thong, S., Kahindi, S., Mbogo, C., Atieli, H., Githeko, A., Lehmann, T., Kazura, J. W., & Yan, G. (2020). Ecological drivers of genetic connectivity for African malaria vectors Anopheles gambiae and An. Arabiensis. *Scientific Reports*, *10*(1), 19946. https://doi.org/10.1038/s41598-020-76248-2

Lo, E., Hemming-Schroeder, E., Yewhalaw, D., Nguyen, J., Kebede, E., Zemene, E., Getachew, S., Tushune, K., Zhong, D., Zhou, G., Petros, B., & Yan, G. (2017). Transmission dynamics of co-endemic Plasmodium vivax and P. falciparum in Ethiopia and prevalence of antimalarial resistant genotypes. *PLOS Neglected Tropical Diseases*, *11*(7), e0005806. https://doi.org/10.1371/journal.pntd.0005806

McRae, B. H. (2006). ISOLATION BY RESISTANCE. *Evolution*, *60*(8), 1551–1561. https://doi.org/10.1111/j.0014-3820.2006.tb00500.x

McRae, B. H., Dickson, B. G., Keitt, T. H., & Shah, V. B. (2008). USING CIRCUIT THEORY TO MODEL CONNECTIVITY IN ECOLOGY, EVOLUTION, AND CONSERVATION. *Ecology*, *89*(10), 2712–2724. https://doi.org/10.1890/07-1861.1

McRae, B., Shah, V., & Edelman, A. (2016). Circuitscape: Modeling landscape connectivity to promote conservation and human health. *The Nature Conservancy*, *14*.

Osei-Atweneboana, M. Y., Eng, J. K., Boakye, D. A., Gyapong, J. O., & Prichard, R. K. (2007). Prevalence and intensity of Onchocerca volvulus infection and efficacy of ivermectin in endemic communities in Ghana: A two-phase epidemiological study. *The Lancet*, *369*(9578), 2021–2029. https://doi.org/10.1016/S0140-6736(07)60942-8

Peterman, W. E. (2018). ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms. *Methods in Ecology and Evolution*, *9*(6), 1638–1647. https://doi.org/10.1111/2041-210X.12984

Real, L. A., & Biek, R. (2007). Spatial dynamics and genetics of infectious diseases on heterogeneous landscapes. *Journal of The Royal Society Interface*, *4*(16), 935–948. https://doi.org/10.1098/rsif.2007.1041

Saarman, N., Burak, M., Opiro, R., Hyseni, C., Echodu, R., Dion, K., Opiyo, E. A., Dunn, A. W., Amatulli, G., Aksoy, S., & Caccone, A. (2018). A spatial genetics approach to inform vector control of tsetse flies ( *Glossina fuscipes fuscipes* ) in Northern Uganda. *Ecology and Evolution*, *8*(11), 5336–5354. https://doi.org/10.1002/ece3.4050

Schwabl, P., Llewellyn, M. S., Landguth, E. L., Andersson, B., Kitron, U., Costales, J. A., Ocaña, S., & Grijalva, M. J. (2017). Prediction and Prevention of Parasitic Diseases Using a Landscape Genomics Framework. *Trends in Parasitology*, *33*(4), 264–275. https://doi.org/10.1016/j.pt.2016.10.008

van Etten, J. (2017). R package gdistance: Distances and routes on geographical grids. *Journal of Statistical Software*, *76*(1), 1–21.